International Journal of Forecasting 32 (2016) 914–938

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Probabilistic electric load forecasting: A tutorial review

Tao Hong [a,*], Shu Fan [b]

[a] *University of North Carolina at Charlotte, USA*

[b] *Monash University, Australia*

ARTICLE INFO

Keywords:
Short term load forecasting
Long term load forecasting
Probabilistic load forecasting
Regression analysis
Artificial neural networks
Forecast evaluation

ABSTRACT

Load forecasting has been a fundamental business problem since the inception of the electric power industry. Over the past 100 plus years, both research efforts and industry practices in this area have focused primarily on point load forecasting. In the most recent decade, though, the increased market competition, aging infrastructure and renewable integration requirements mean that probabilistic load forecasting has become more and more important to energy systems planning and operations. This paper offers a tutorial review of probabilistic electric load forecasting, including notable techniques, methodologies and evaluation methods, and common misunderstandings. We also underline the need to invest in additional research, such as reproducible case studies, probabilistic load forecast evaluation and valuation, and a consideration of emerging technologies and energy policies in the probabilistic load forecasting process.

1. Introduction

Electric load forecasts have been playing a vital role in the electric power industry for over a century (Hong, 2014). The business needs of load forecasting include power systems planning and operations, revenue projection, rate design, energy trading, and so forth. Load forecasts are needed by many business entities other than electric utilities, such as regulatory commissions, industrial and big commercial companies, banks, trading firms, and insurance companies (Bunn & Farmer, 1985; Hong, 2010; Hong & Shahidehpour, 2015; Weron, 2006; Willis, 2002).

To avoid ambiguous and verbose presentation, we note that the rest of this paper uses the term "load forecasting" to refer to "electric load forecasting". We will use "PLF" as the abbreviation for both "probabilistic electric load forecasting" and "probabilistic electric load forecast". Nevertheless, we also recognize the similarities between electric load forecasting and the forecasting of other utilities, such as water and gas, in terms of forecasting principles, methodologies, techniques and even business requirements. We hope that this tutorial review is also beneficial to researchers and practitioners in other utility load forecasting areas.

There is not yet a gold standard for classifying the range of load forecasts. We can group the forecasting processes into four categories based on their horizons: very short term load forecasting (VSTLF), short term load forecasting (STLF), medium term load forecasting (MTLF), and long term load forecasting (LTLF). The cut-off horizons for these four categories are one day, two weeks, and three years respectively (Hong, 2010; Hong & Shahidehpour, 2015). A rough classification may lead to two categories, STLF and LTLF, with a cut-off horizon of two weeks (Hong & Shahidehpour, 2015; Xie, Hong, & Stroud, 2015). Fig. 1 depicts the load forecasting applications and classification. In this paper, we adopt the rough classification, though we occasionally use VSTLF and MTLF to refer to things that are specific to these categories.

Load forecasting traditionally refers to forecasting the *expected* electricity demand at *aggregated* levels. Long term

* Corresponding author.
E-mail addresses: hongtao01@gmail.com (T. Hong), Shu.Fan@monash.edu (S. Fan).

http://dx.doi.org/10.1016/j.ijforecast.2015.11.011

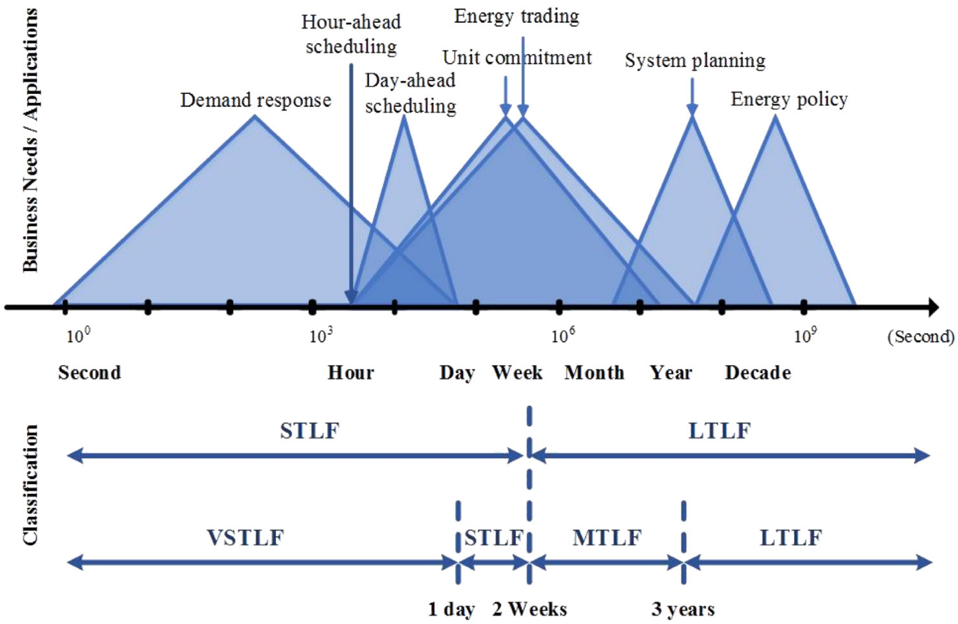

Fig. 1. Load forecasting applications and classification.

load forecasting at the small area level or the equipment (i.e., distribution transformer) level is called spatial load forecasting (SLF) (Hong, 2008; Willis, 2002; Willis & Northcote-Green, 1983). The massive smart meter deployment over the past decade has provided the industry with a huge amount of data that is highly granular, both temporally and spatially. The availability of this new data, together with the advancement of computing technologies and forecasting techniques, has converted spatial load forecasting into an emerging subject, hierarchical load forecasting (HLF). HLF covers forecasting at various levels, from the household level to the corporate level, across various horizons, from a few minutes ahead to many years ahead. The most significant development of HLF methodologies over the last decade was through the Global Energy Forecasting Competition 2012 (GEFCom2012), which was presented by Hong, Pinson, and Fan (2014).

Because the decision making process in the utility industry used to heavily rely on expected values, a load forecasting process typically results in point outputs, with one value at each step. Over the last decade, the increase in market competition, the aging infrastructure and renewable integration requirements have meant that PLF has become increasingly important for the planning and operation of energy systems. PLFs can be used for stochastic unit commitment, power supply planning, probabilistic price forecasting, the prediction of equipment failure, and the integration of renewable energy sources (Hong, 2014).

PLFs can be based on scenarios, though scenario-based forecasts are not probabilistic forecasts unless the scenarios are assigned probabilities. PLFs can be in the form of quantiles, intervals, or density functions. Note that there are two intervals that we often refer to in forecasting, namely prediction intervals and confidence intervals. A prediction interval is associated with a prediction, whereas a confidence interval is associated with a parameter. In PLF, almost all business applications require people to understand prediction intervals. However, many papers in the literature are misusing the term “confidence interval” to refer to prediction intervals. In this review, we follow the formal load forecasting terminology (Hong & Shahidehpour, 2015), regardless of the term used in the paper we are citing.

The literature on PLF is quite limited, particularly compared to that of either probabilistic forecasting in general (Gneiting & Katzfuss, 2014) or probabilistic wind power forecasting (PWPF) (Pinson, 2013; Zhang, Wang, & Wang, 2014). Nevertheless, PLF should be just as important as PWPF in the utility industry. For a medium sized US utility with an annual peak of 1 GW–10 GW, the typical day-ahead load forecasting error is around 3%, while the typical day-ahead wind power forecasting error is around 15%. If the wind penetration is around 20%, then, on average, the absolute errors of load forecasts are similar to those of wind power forecasts. As was discussed by Hong (2015), a load forecast error of 1% in terms of mean absolute percentage error (MAPE) can translate into several hundred thousand dollars per GW peak for a utility’s financial bottom line.

Table 1 summarizes the key features of the various load forecasting problems, namely their temporal and spatial resolutions, forecast horizons, and output formats. Fig. 2 shows the numbers of journal papers in the area of load forecasting since 1970s, with spatial load forecasting and hierarchical load forecasting being grouped together. From the late 1990s to the early 2000s, more effort was devoted to STLF than to LTLF, due mainly to the deregulation of the utility industry. Competition through electricity markets demanded improvements in STLF, while limitations in infrastructure investment reduced the need for LTLF. As the existing infrastructure has been approaching its design and age limits over the last decade, research in LTLF has been ramped up as well. The smart grid deployment has

Table 1
Key features of different load forecasting problems.

	Temporal resolution	Spatial resolution	Forecast horizon	Output format
LTLF	Monthly/annual	N/A	Years	Point
STLF	Hourly	N/A	Days	Point
SLF	Monthly/annual	Small area	Years	Point
HLF	Hourly	Premise	Hours to years	Point
PLF	Hourly	N/A	Hours to years	Density/interval

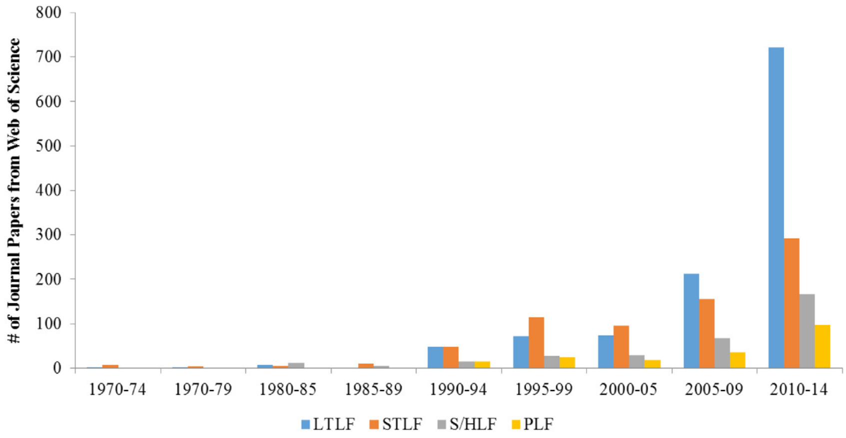

Fig. 2. Numbers of journal papers in the area of load forecasting since the 1970s.

also stimulated HLF development. PLF is the lowest bar for all time periods, but has a strong increasing trend over the past decade.

The literature contains thousands of papers on load forecasting. Researchers have published many different literature review articles on load forecasting techniques (Section 2), none of which has focused on PLF. This paper presents a tutorial review that is devoted to PLF across all forecasting horizons. Since most PLF studies to date have focused on the traditional point forecasting techniques and methodologies, we begin by reviewing several representative papers on point load forecasting (Section 3). The progress in PLF has been made by two groups, one from the application side, those who use load forecasts for specific business needs, and the other from the technical and methodological development side, those who develop load forecasting models. Section 4 of this paper reviews the major developments from each side.

Section 5 focuses on the production and evaluation of PLFs. We begin by dissecting the PLF problem into three elements, namely the input, model and output. The treatment of each may eventually lead to PLFs (Section 5.1). Although forecast evaluation is an important step in any forecasting process, the PLF evaluation methods have not yet been developed fully. Section 5.2 presents the properties of PLFs and the evaluation methods that have been used for PLF. As an emerging topic, PLF evaluation is still a long way from maturity. Section 5.3 discusses the integration aspect of PLF methods and techniques. Finally, Section 6 recommends several directions for future research that need joint efforts from a range of research communities.

Among the vast body of literature on load forecasting, and PLF in particular, there are many notable research outcomes that have generated significant value or are likely to be valuable for industry. There are also many errors and inconsistencies that need to be corrected or clarified. Instead of producing a comprehensive review that covers all papers in all relevant areas, we selected the references carefully so as to include the representative ones for people either to follow as excellent examples, or to avoid as counterexamples, so that the reference list of this tutorial review serves as a collection of useful papers.

2. Literature reviews

The literature on STLF is much more extensive than that on LTLF. This is also reflected by the literature reviews that have been published over the last thirty plus years. Of the 17 load forecasting review papers that we are going to discuss in this section, 13 are on STLF. Some of these STLF reviews are at the conceptual level, with qualitative analyses of the developments, results, and conclusions of the original papers (Abu-El-Magd & Sinha, 1982; Alfares & Nazeeruddin, 2002; Bunn, 2000; Gross & Galiana, 1987; Hippert, Pedreira, & Souza, 2001; Hong, 2010; Hong, Pinson et al., 2014; Metaxiotis, Kagiannas, Askounis, & Psarras, 2003; Tzafestas & Tzafestas, 2001). Some reviews perform empirical studies using quantitative analysis, with the aim of implementing, analyzing, evaluating, and comparing the different techniques reported in the literature using one or several new sets of data (Hong, 2010; Liu et al., 1996; Moghram & Rahman, 1989; Taylor & McSharry, 2007; Weron, 2006). In addition to examining the STLF reviews, we also discuss eight other review papers on load forecasting (Feinberg & Genethliou, 2005; Hong, 2014; Hong & Shahidehpour, 2015; Willis & Northcote-Green, 1983), electricity price forecasting (Weron, 2014), PWPF (Pinson, 2013; Zhang et al., 2014), and probabilistic forecasting (Gneiting & Katzfuss, 2014).

2.1. Conceptual reviews of STLF

STLF has been an active area of research for three decades. It would be difficult for researchers to follow even a fraction of the papers that are published each year. Conceptual reviews play a vital role in describing major developments, setting the stage for future research directions, and helping to point the researchers to notable references. However, a conceptual review does not add much value if it simply puts the papers into different categories (e.g., statistical techniques vs. artificial intelligence techniques) based on the techniques being used. The real value of conceptual reviews lies in the following aspects: (1) articulating the real-world applications of STLF; (2) presenting an authoritative point of view on the advantages and disadvantages of the methods and techniques; (3) discussing misconceptions and mistakes in the literature; (4) making recommendations as to future research needs; and (5) providing a high-quality list of references.

Abu-El-Magd and Sinha (1982) reviewed several statistical techniques for STLF, such as multiple linear regression, spectral decomposition, exponential smoothing, the Box–Jenkins approach, state space models, and some multivariate models. Their review focused on the system identification aspect of STLF, and discussed the merits and drawbacks of the different approaches. A significant portion of the discussion was on computational requirements and the applicability of these methods for online and offline applications. Advances in computing technologies over the past three decades mean that some of these discussions and recommendations concerning online and offline applications are no longer applicable in today's world. Nevertheless, in general, the paper offers a good summary of the major STLF techniques used prior to the early 1980s.

Gross and Galiana (1987) offered a tutorial review of STLF by organizing the contents based on the following five aspects: (1) applications of STLF; (2) factors that affect the load; (3) techniques for STLF; (4) practical considerations; and (5) some possible future directions. The review pointed out many practical issues that are well worth studying but still have not received much attention even today, such as error analysis, outlier detection, data cleansing, the human–machine interface, computational complexity, and so forth.

Bunn (2000) presented a review of short term load and price forecasting in the competitive power market. For load forecasting, the emphasis is on the segmentation of the forecast variables, forecast combination, and the use of neural networks for load forecasting.

Tzafestas and Tzafestas (2001) reviewed artificial intelligence (AI) techniques for STLF, such as artificial neural networks (ANN), fuzzy logic, genetic algorithms and chaos. In addition, hybrid AI methodologies, including the possible combinations with statistical models and knowledge-based methods, as well as among AI techniques, were also reviewed. The paper did not perform any quantitative experimentation, though it drew eight representative case studies from the literature to show the relative merits of the various forecasting methodologies under a range of geographic, weather and other peculiar conditions, together with the performances that each could achieve.

Hippert et al. (2001) focused on STLF with ANN. The specific aim of this review was to clarify the skepticism regarding the usage of ANN on STLF. Through a critical review and evaluation of around 40 representative journal papers published in the 1990s, the authors highlighted two facts that could have led to this skepticism. Firstly, the ANN models may be "overfitting" the data, possibly due to either overtraining or overparameterization. Secondly, although all of the proposed systems were tested on real data, most of the tests reported by the papers reviewed were not carried out systematically: some did not provide comparisons with standard benchmarks, while others did not follow standard statistical procedures in reporting the analysis of errors. Another contribution of Hippert et al. (2001) was their summary of the process of designing a STLF system. The design tasks were divided into four stages: data pre-processing, ANN design, implementation, and validation. Although the discussion was in the context of ANN, a significant portion was also applicable to other techniques.

Alfares and Nazeeruddin (2002) covered a wide range of techniques classified into nine categories: (1) multiple regression; (2) exponential smoothing; (3) iterative reweighted least-squares; (4) adaptive load forecasting; (5) stochastic time series; (6) autoregressive moving average models with exogenous inputs (ARMAX) based on genetic algorithms; (7) fuzzy logic; (8) ANN; and (9) expert systems. The paper described the methodologies briefly for each category, and discussed their advantages and disadvantages.

Metaxiotis et al. (2003) provided a chronological summary of the development of various AI techniques, such as expert systems (ES), ANNs, and genetic algorithms. The advantages of AI techniques in STLF were summarized both conceptually and qualitatively. However, there was no detailed discussion of disadvantages. Without any solid support, the paper concluded that AI techniques "*have matured to the point of offering real practical benefits*". Even now, it would be an exaggeration to consider AI to be mature enough to offer real practical benefits for STLF.

Hong (2010) reviewed 50 years of STLF literature from three points of view: the techniques, the variables being used, and the representative work being done by several major research groups. The review indicated that the recent advancements in statistical techniques and software packages had not been incorporated into the development of STLF methodologies as well as on the AI side. The review also pointed out the benchmarking issue in STLF.

All of the conceptual reviews discussed in this section refer to STLF at aggregated levels. Over the last decade, many countries around the globe have been modernizing their power grid. One major effort has been the deployment of smart meters and the related communication devices, which have introduced significant amounts of data, providing a challenge for traditional load forecasting practices. In response to this new challenge, the IEEE Working Group on Energy Forecasting organized the Global Energy Forecasting Competition 2012 (GEFCom2012), of which one track was on HLF. Hong, Pinson et al. (2014) reviewed the methodologies used by the top entries of the GEFCom2012. In the HLF track, all four of the winning entries applied regression analysis as part of the methodology, while two used gradient boosting machines. Some of these entries also performed additional tasks, such as modeling holidays, combining forecasts, and data cleansing.

2.2. Empirical reviews on STLF

Several researchers have also conducted quantitative case studies in order to compare and evaluate the various techniques for STLF, resulting in empirical reviews. However, some empirical reviews can be misleading, depending upon the expertise of the authors, as techniques may be put at a disadvantage if they are not applied properly. For instance, Liu et al. (1996) did not apply autoregressive models properly. A technique may also show a consistent superiority over others because the authors' expertise and/or the case study setup favors a particular technique. For instance, Taylor and McSharry (2007) showed the double seasonal Holt–Winters exponential smoothing method to be superior to its competing techniques, but this was mainly because the experiment was designed to favor this technique (as will be discussed later in the section). In general, there is not yet any single technique that is known to dominate all others for STLF; the important thing is the methodology used to apply the techniques. When reading empirical reviews, readers are encouraged to focus on the methodologies, rather than the conclusions as to the winning technique(s).

Moghram and Rahman (1989) evaluated five techniques: multiple linear regression, stochastic time series, exponential smoothing, state space methods, and knowledge-based expert systems. The authors began with a brief introduction of each technique, then used the five techniques to produce 24-hour-ahead forecasts. The case study was based on data from a southeastern utility in the US. The authors did not intend to build the finest model using each technique. Instead, they aimed to introduce the different techniques, so that interested readers could conduct further analyses in order to produce enhanced load forecasts.

Liu et al. (1996) compared three techniques: fuzzy logic, ANN, and autoregressive models. However, as presented in the paper, a mistake was made when applying AR to STLF. It is well known that the load series is not a stationary series, but the authors modeled the load series using AR directly, without performing any stationarity testing or differencing steps (Dickey & Fuller, 1979). Thus, the conclusion that "*the performances of fuzzy-logic-based and ANN-based forecaster are much superior to the one of AR-based forecaster*" was drawn based on an incorrect implementation. On the other hand, the design and implementation of the fuzzy-logic-based and ANN-based forecasters were not explained clearly either.

Weron (2006) reviewed a range of statistical techniques and concepts that could be used for modeling and forecasting the electricity demand, such as seasonal decomposition, mean reversion, heavy-tailed distributions, exponential smoothing, spike pre-processing, autoregressive time series, regime-switching models, interval forecasts, and so forth. A number of case studies and implementations of different techniques in MATLAB were provided, which could be useful for researchers and quantitative analysts in the load forecasting area.

Taylor and McSharry (2007) conducted an evaluation to compare models for 24-hour-ahead forecasting and select the best. Five methods were included in the discussion: autoregressive integrated moving average (ARIMA) modeling, periodic AR modeling, an extension of Holt–Winters exponential smoothing for double seasonality, an alternative exponential smoothing formulation, and a principal component analysis (PCA) based method. The case study was based on 30 weeks of intraday electricity demands from 10 European countries. However, a major issue with this paper is its experiment. All of the competing techniques are univariate models, and none rely on weather variables. Although regression analysis and ANN had been being used for STLF in the field for many years, the authors excluded them from the comparative analysis by citing a 1982 paper that indicated that multivariate modeling was impractical for online short term forecasting systems. The same is true of the study by Taylor (2008), which evaluated several methods, including ARIMA modeling, several exponential smoothing models and a similar day method, for VSTLF with forecast horizons of 10–30 min ahead.

Hong (2010) evaluated three representative techniques, namely multiple linear regression, ANN and fuzzy regression. The data used in this case study were from a medium-sized utility in the US. According to the evaluation results, the linear models outperformed the other two. However, the conclusion was limited again to the specific setup of the experiment; meaning that one should not generalize this conclusion to infer that linear models are superior in all cases. Nevertheless, the evaluation by Hong (2010) demonstrated that the variable selection processes of these techniques are inherently connected.

2.3. Other load forecasting reviews

Willis and Northcote-Green (1983) offered a tutorial review of spatial load forecasting. The review introduced the planning requirements for spatial load forecasting, described the load growth patterns, and reviewed several major spatial load forecasting methods, such as regression-based methods and land-use-based methods that rely on the simulation of urban growth.

Feinberg and Genethliou (2005) covered both STLF and LTLF. The authors discussed the factors that affect the accuracy of the forecasts, such as weather data, time factors, customer classes, and economic and end use factors. The review briefly examined various statistical and artificial intelligence techniques that have been tried for STLF and LTLF. In their discussion of future research directions, the authors pointed out that additional progress in load forecasting and its use in industrial applications could be achieved by providing short-term load forecasts in the form of probability distributions rather than point forecasts.

Hong (2014) reviewed the evolution of forecasting methodologies and applications in the energy industry. A significant portion of the review was devoted to load forecasting, though electricity price forecasting and renewable generation forecasting were also covered briefly. The primary audience of the review was forecasting practitioners. The pros and cons of various forecasting methods were discussed conceptually. The review emphasized the importance of conducting rigorous out-of-sample tests and respecting business needs. An interdisciplinary approach to energy forecasting, bringing together several disciplines, such as statistics, electrical engineering, meteorological science, and so forth, was favored.

Hong and Shahidehpour (2015) provided a comprehensive review of load forecasting topics, primarily for state governments and planning coordinators. In addition, the authors also presented case studies in three different jurisdictions, namely ISO New England, Exelon and North Carolina Electric Membership Corporation (NCEMC), to assist planning coordinators and their relevant state governments in applying innovative concepts, tools, and analysis to their forecasting regime. In these case studies, the authors followed the weather station selection methodology proposed by Hong, Wang, and White (2015), the variable selection methodology proposed by Hong (2010), and the long term probabilistic load forecasting methodology proposed by Hong, Wilson, and Xie (2014). The NCEMC case study by Hong and Shahidehpour (2015) was designed to increase the awareness of realistic load forecasting errors, as the forecast horizon stretches into the recession years, with the authors forecasting the load from 2009 to 2013 using historical data up to 2008.

2.4. Other notable reviews

PLF is the intersection between load forecasting and probabilistic forecasting. Although PLF is still an underdeveloped area, we do expect to take advantage of existing developments in both point load forecasting and probabilistic forecasting in general to advance the PLF research. While we have discussed over a dozen load forecasting reviews published over the past three decades, here we zoom out to the broad subject of probabilistic forecasting. We first discuss a few notable reviews that cover other areas of probabilistic energy forecasting, such as electricity price forecasting and wind power forecasting. We then discuss a recent review of probabilistic forecasting.

Weron (2014) offered a comprehensive review of electricity price forecasting, recognizing that there is a lot less in the literature on probabilistic price forecasting than on point price forecasting. The probabilistic price forecasting papers discussed are categorized as interval forecasts, density forecasts and threshold forecasts. In addition, the author acknowledged the lack of studies on the combination of probabilistic price forecasts prior to 2014, and discussed the most recent developments in this area.

Pinson (2013) provided a tutorial review on wind power forecasting, introducing the physical basics of wind power generation briefly and considering it as a stochastic process. By assessing the representative decision-making problems that require wind power forecasts as inputs, Pinson underlined the necessity of issuing the forecasts in a probabilistic framework. The review covered several major approaches to the forecasting of wind power in different forms, such as single-valued predictions, predictive marginal densities, and space–time trajectories. The challenges were discussed at the end, with a focus on new and better forecasts, forecast verification, and ways of bridging the gap between forecast quality and value.

Zhang et al. (2014) reviewed the state of the art of PWPF. They introduced three representations of wind power uncertainty, which were then used to split the forecasting methodologies into three categories: probabilistic forecasts (parametric and non-parametric), risk index forecasts, and space–time scenario forecasts. The authors also summarized the requirements and a framework for forecast evaluation. At the end, they discussed three challenges, namely the further improvement of wind power forecasts, the integration of wind power into energy markets, and forecasting with high-resolution data.

Gneiting and Katzfuss (2014) offered a selective overview of the state of the art in probabilistic forecasting. Their review covered theory, methodology, and a range of applications focusing on predictions of real-valued quantities, such as the inflation rate, temperature, and precipitation accumulation. A probabilistic wind speed forecasting case study was used to illustrate the concepts and methodologies.

3. Load forecasting techniques and methodologies

PLF is an emerging branch of the load forecasting problem, and therefore is not totally independent of point load forecasting. In this section, we provide an overview of representative load forecasting techniques and methodologies. Here, we use the word "technique" to refer to a group of models that fall in the same family, such as Multiple Linear Regression (MLR) models and Artificial Neural Networks (ANNs). On the other hand, we use "methodology" to represent a general solution framework that can be implemented with multiple techniques. For example, a variable selection methodology may be applicable to both MLR models and ANNs. While both techniques and methodologies are important for load forecasting practices, the literature has been dominated by papers that focus on trying out various techniques and their combinations, whereas the original research on load forecasting methodologies is quite limited.

3.1. Techniques

Load forecasting techniques are typically classified into two groups: statistical techniques and artificial intelligence techniques, though the boundary between the two is becoming more and more ambiguous, as a result of multidisciplinary collaborations in the scientific community. In this section, we will review four statistical techniques, namely MLR models, semi-parametric additive models, autoregressive and moving average (ARMA) models, and exponential smoothing models; and four AI techniques, namely ANN, fuzzy regression models, support vector machines (SVMs), and gradient boosting machines. We conclude this section with a high-level comparison of these load forecasting techniques.

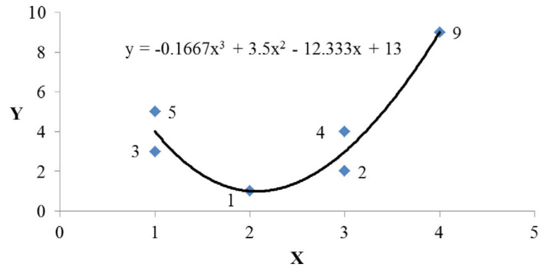

Fig. 3. Using a polynomial regression model to describe a nonlinear relationship.

3.1.1. Multiple linear regression models

Regression analysis is a statistical process for estimating the relationships among variables (Kutner, Nachtsheim, & Neter, 2004). MLR models have been used in the literature for both STLF and LTLF. The load or some transformation of the load is usually treated as the dependent variable, while weather and calendar variables are treated as independent variables. MLR requires the user or forecaster to specify a functional form among these variables. The parameters of the MLR models are often estimated using the ordinary least squares method.

When considering linear regression models for load forecasting, a typical misunderstanding is that they are not suitable for modeling the nonlinear relationships between the load and weather variables. Such a misunderstanding is used in many papers as the motivation for applying black-box techniques. In fact, the "linear" in linear regression refers to the linear equations that are used to solve the parameters, rather than the relationships between the dependent and independent variables. For instance, as Fig. 3 shows, polynomial regression models are in the family of MLR models, but can describe nonlinear relationships between the dependent and independent variables in the form of polynomials.

Papalexopoulos and Hesterberg (1990) proposed a regression-based approach to STLF. The proposed approach was tested using the Pacific Gas and Electric Company's (PG&E) data for the peak and hourly load forecasts of the next 24 h. This is one of the few papers that has focused fully on regression analysis for STLF. Several modeling concepts for using MLR for STLF were applied: the weighted least square technique, temperature modeling by using heating and cooling degree functions, holiday modeling by using binary variables, a robust parameter estimation method, etc. Through a thorough test, the proposed MLR model was concluded to be superior to the one PG&E used at the time. This paper provided a solid ground for applying regression analysis to STLF.

Ramanathan, Engle, Granger, Vahid-Araghi, and Brace (1997) developed 24 regression models, one for each hour of a day, with a dynamic error structure and adaptive adjustments to correct for the forecast errors of previous hours. The case study was conducted as part of a competition organized by the Electric Power Research Institute (EPRI) using data from a utility in the northwest of the US. The results showed that the regression models outperformed the other competitors' models.

Hong (2010) proposed an interaction regression based approach to STLF, emphasizing the interactions (or cross effects) among weather and calendar variables. The case study was based on a US utility that deployed the regression models in its production environment. Several special effects were modeled using regression analysis, such as the recency effect, weekend effect and holiday effect. Through comparisons with the models based on ANN and fuzzy regression, the linear models were shown to produce smaller errors than their competitors.

Hong, Wilson et al. (2014) developed a linear regression model for LTLF. The linear model started off as a STLF model, but was augmented with a macroeconomic indicator. It was then applied to various scenarios in order to generate the long term probabilistic load forecast. The authors showed that the models based on hourly data had smaller ex post forecasting errors than those based on monthly or daily data.

Charlton and Singleton (2014) presented a refined parametric model for STLF in the GEFCom2012. The model estimated the electricity demand as a function of the temperature and calendar variables. The authors set up a series of refinements of the model, explained the rationale for each, and used the competition scores to demonstrate that each successive refinement step increased the accuracy of the model's predictions. These refinements included combining models from multiple weather stations, removing outliers from the historical data, and treating public holidays specially.

Wang, Liu, and Hong (2016) extended the recency effect modeling method proposed in Hong (2010) by including large number of lagged temperature and moving average temperature variables in the MLR models. The idea is to leverage the increased computing power to build large regression models to enhance the load forecast accuracy. Another finding from this paper is that developing 24 models with one for each hour may not result in better forecasts than one interaction regression model for all 24 h.

3.1.2. Semi-parametric additive models

The semi-parametric additive model falls within the regression framework, but is designed to accomodate some non-linear relationships and serially correlated errors. In particular, such models allow the use of nonlinear and non-parametric terms within the framework of additive models (Ruppert, Wand, & Carroll, 2003). In load forecasting, these generalized additive models are used to estimate the relationship between the load and explanatory variables such as temperature and calendar variables.

Hyndman and Fan (2010) developed two models for forecasting the long term peak demand for South Australia, namely a semi-parametric model for the half-hourly demand and a linear model for the annual median demand. The natural logarithms were used to transform the raw demand with major industry loads removed. The semi-parametric model captured calendar and temperature effects, as well as the effects from demographic and economic factors. In particular, the model was split into two separate models. One was a linear model (using linear regression), based on the seasonal demographic variables, economic variables, and degree days. The other one was a non-parametric model (using regression splines), based on the remaining variables, which are measured at half-hourly intervals. The models were then used to generate density forecasts with the simulated temperatures as inputs.

Fan and Hyndman (2012) applied a similar non-parametric additive model to STLF in the Australian national electricity market. In addition to the calendar and temperature effects, the models also incorporated the lagged demand, in order to capture the serial correlation within the demand series.

Goude, Nedellec, and Kong (2014) used generalized additive models to model the electricity demand over more than 2200 substations of the French distribution network, at both short- and middle-term horizons. These generalized additive models estimated the relationship between the load and explanatory variables such as temperatures, calendar variables, and so forth. This methodology showed good results on a case study of the French grid.

Nedellec, Cugliari, and Goude (2014) used semi-parametric additive models in the load forecasting track of GEFCom2012. They proposed a temporal multi-scale model that combined three components. The first component was a long term trend estimated by means of non-parametric smoothing. The second was a medium term component describing the sensitivity of the electricity demand to the temperature at each time step, and was fitted using a generalized additive model. Finally, local behaviors were modeled with a short term component. A random forest model was used for parameter estimation.

3.1.3. Exponential smoothing models

Exponential smoothing assigns weights to past observations that decrease exponentially over time (Hyndman & Athanasopoulos, 2013; Hyndman, Koehler, Ord, & Snyder, 2008). It does not rely on explanatory variables, meaning that it has lower data requirements than other widely used techniques such as MLR and ANN. Two notable papers in the literature are those by Taylor and McSharry (2007) and Taylor (2008), which were discussed in Section 2.2. In each review, some variations of exponential smoothing, such as double and triple seasonal exponential smoothing models, outperformed the other selected models that do not rely on weather variables.

Despite its success in some academic papers, exponential smoothing is rarely a top candidate in real-world STLF practice, as is reflected in the fact that none of the top entries to GEFCom2012 used exponential smoothing (Hong, Pinson et al., 2014). Since the electricity demand is driven strongly by the weather, changes in weather patterns can have a big effect on the load profiles. When weather conditions are volatile, techniques that do not use meteorological forecasts are often at a disadvantage.

3.1.4. Autoregressive moving average models

ARMA models provide a parsimonious description of a stationary stochastic process in terms of two polynomials, one an autoregression and the other a moving average (Box, Jenkins, & Reinsel, 2008; Brockwell & Davis, 2010; Hyndman & Athanasopoulos, 2013; Wei, 2005). Since the hourly electricity demand series is well-known to be non-stationary, ARIMA models, which are a generalization of ARMA models, are often used for load forecasting purposes. ARMA models can also be generalized to include exogenous variables, giving ARMAX models.

Weron (2006) provided a good coverage of various statistical techniques for load forecasting, such as exponential smoothing, regression models, autoregressive models, ARMA, ARIMA and ARMAX models. Two case studies based on data from California ISO were used to illustrate the modeling concepts.

3.1.5. Artificial neural networks

ANNs have been used extensively for load forecasting since the 1990s. The ANN is a soft computing technique that does not require the forecaster to model the underlying physical system explicitly (Hagan, Demuth, Beale, & De Jesús, 2014). In other words, the forecaster does not have to specify the functional form among the input and output variables, as must be done when building MLR models. By simply learning the patterns from the historical data, a mapping between the input variables and the electricity demand can be constructed, then adopted for the prediction. Many types of ANNs have been used for load forecasting, such as feedforward neural networks, radial basis function networks, and recurrent neural networks. The most popular estimation method is the back propagation algorithm. Researchers have been reporting fairly good results with ANN models, though many of the good results have been due to peeking into the future. Hippert et al. (2001) offered a critical review of the literature on ANN-based load forecasting, as was discussed in Section 2.1.

The best-known implementation of ANN models for STLF to date was from a project sponsored by EPRI. The solution was named ANNSTLF—artificial neural network short-term load forecaster (Khotanzad & Afkhami-Rohani, 1998). This load forecasting system included two ANN forecasters, one predicting the base load and the other forecasting the change in load. The final forecast was computed through an adaptive combination of these two forecasts. The ANNSTLF and its improved versions were later commercialized, and are used by a large number of utilities across the US and Canada.

3.1.6. Fuzzy regression models

Fuzzy regression is introduced in order to overcome some of the limitations of linear regression, such as the vague relationship between the dependent variable and the independent variables, insufficient numbers of observations, and hard-to-verify error distributions. The fundamental difference between the assumptions of the two techniques relates to the deviations between the observed and estimated values: linear regression assumes that these values are supposed to be errors in measurement or observations, while fuzzy regression assumes that they are due to the indefiniteness of the system structure.

Song, Baek, Hong, and Jang (2005) used fuzzy linear regression to forecast the loads during holidays, and the model showed a promising level of accuracy. The proposed approach forecasted the load based only on the previous load, without the input of weather information. A further improvement was achieved through the use of a hybrid model with fuzzy linear regression and general exponential smoothing (Song et al., 2005).

Hong and Wang (2014) proposed a fuzzy interaction regression approach to STLF. In a comparison with three models (two fuzzy regression models and one multiple linear regression model) without interaction effects, the proposed approach showed the best performance. The paper focused on the application of fuzzy regression to STLF, and provided several tips for fuzzy regression based forecasting. The paper indicated that, when improving the underlying linear model, one could observe a reduction in the fuzziness that was recognized originally by a deficient model.

3.1.7. Support vector machine

SVMs are supervised learning models with associated learning algorithms that analyze data and recognize patterns, often being used for classification and regression analysis. SVM has been shown to be very resistant to the problem of over-fitting, and eventually achieves good performances for solving time series forecasting problems.

Chen, Chang, and Lin (2004) provided the winning entry for the competition organized by the EUNITE network. In the competition, the task was to forecast the daily peak loads of the next 31 days. This winning entry was based on a SVM. More specifically, the model was based on winter data only, and did not use any temperature information. One of the conclusions from the paper was that temperatures (or other types of climate information) might not be useful in a MTLF problem. Although the competition focussed on MTLF, it led to SVM becoming notable in the field of STLF.

3.1.8. Gradient boosting

Gradient boosting is a machine learning technique for regression problems, and produces a prediction model in the form of an ensemble of weak prediction models. Unlike other boosting techniques, gradient boosting allows the optimization of an arbitrary differentiable loss function.

Ben Taieb and Hyndman (2014) used a gradient boosting method for the load forecasting track of GEFCom2012. Separate semi-parametric additive models were used for each hourly period, with component-wise gradient boosting being used to estimate each model, and univariate penalised regression splines as base learners. The models allowed the electricity demand to change with the time-of-year, day-of-week and time-of-day, and also on public holidays, with the main predictors being current and past temperatures, and past demand.

Lloyd (2014) used gradient boosting machines and Gaussian processes for the load forecasting track of GEFCom2012. The methods were generic machine learning and regression algorithms, with few domain-specific adjustments.

3.1.9. The myth of the best technique

Although all forecasts are wrong, researchers have long been pursuing the most accurate forecast. Very often people still put their hope in finding that best technique of all. We have reviewed a collection of papers that represent eight major techniques that have been applied to load forecasting. It is worth noting that there are many more techniques that have been tried for load forecasting. Over the past several decades, the majority of the load forecasting literature has been filled with attempts to determine the best technique for load forecasting. Although researchers have tried many different techniques for generating load forecasts, the number of original techniques is still countable, e.g., within 100. As original techniques are being exhausted, many researchers have started to combine them to come up with "new" hybrid techniques. Some of these hybrid techniques have been of some value in solving the load forecasting problem, e.g., fuzzy neural networks. However, most of them have made a minimal contribution to the literature. A typical way to create massive numbers of valueless papers is to use some soft computing techniques to estimate the parameters for a computationally intensive technique. For instance, a randomly generated idea could be an ANN-based STLF with wavelet transform and particle swarm optimization; or a hybrid ant colony and genetic algorithm for identifying the parameters of ARMAX load forecasting models.

To ensure publication, many authors manipulate their case studies so that the proposed technique beats its competitors, often as a result of magically peeking into the future. The reported accuracy of the proposed techniques is usually very impressive, sometimes too good to be true. Such research practices have several negative consequences:

(1) Virtually all papers show the superiority of various techniques on very specific datasets. This makes the conclusions hard to generalize, and is of little value for load forecasting practice.
(2) Due to an over-manipulation of the data and a lack of detailed information on the setup of experiments, the case studies presented by one research group can rarely be reproduced by another. This limits the progress of research and development.
(3) Many papers hide the weaknesses of the proposed techniques, usually resulting in misleading conclusions. Many other papers then cite these misleading conclusions without reproducing the results or even reading the original paper. This propagates the unverified findings, while burying any empirically validated work.

It is very important for researchers and practitioners to understand that *a universally best technique simply does not exist*. It is the data and jurisdictions that determine what technique we should use, rather than the other way around. We should always understand the business needs first, then analyze the data, and usually go through a trial-and-error process, to figure out which is the best technique for a specific dataset in a specific jurisdiction. Note that the forecasting error may also differ significantly for different utilities, different zones within a utility, and different time periods.

Here, we offer some general guidance about the strengths and weaknesses of different classes of techniques.

(1) Black-box models vs. non-black box models.

The most popular black-box technique in applications to load forecasting is ANN. ANNs do not offer any insights as to the form of the relationship between the load and its driving factors. As a result, ANNs are often avoided for regulatory purposes, due to their lack of interpretability. On the other hand, the application of ANNs does not require much by way of statistical background or skill in data analysis. With many software packages, such as MATLAB, offering comprehensive ANN model structures, the forecaster can simply use trial and error to investigate different ANN structures with various numbers of hidden neurons, hidden layers, elevation functions, etc. In the 1990s and early 2000s, the computational complexity of black-box models was often criticized by practitioners. However, advances in computing technologies over the last decade have gradually helped to alleviate the concerns about computing time.

Non-black box models, or interpretable models, offer insights into the relationship between the load and its driving factors. The most representative non-black box models in load forecasting are MLR models. The downside of these models is the requirement of statistical analysis skills, as forecasters have to designate the functional form of the relationship between the load and its driving factors. For instance, when modeling the relationship between load and temperature, the forecasters should select from among several candidate forms, such as 2nd order polynomial, 3rd order polynomial, and piece-wise linear functions.

(2) Univariate models vs. multivariate models.

Univariate models in load forecasting are those that do not rely on explanatory variables, which are primarily weather variables. The most common of these techniques are exponential smoothing and ARIMA. Their main advantage is that they do not rely on weather information. In other words, these univariate models can be used when weather data are unavailable or unreliable. Many system operators make historical load data freely available, but withhold the weather data. This means that it is quite convenient and sensible to conduct academic research on univariate techniques. On the other hand, accessing high quality weather data usually requires significant funding and domain knowledge, which raises the entry bar for the development of models that rely on weather information.

The most common multivariate models for load forecasting are MLR models, ANNs and support vector regression models. For STLF practice, the main advantage of these techniques over univariate ones is accuracy. This is because temperature is a major driving factor for the electricity demand. The temperature forecasts made using state-of-the-art weather forecasting techniques are quite reliable in the short term, i.e., within a few days. For long term load forecasting, the major advantage of multivariate models is their ability to perform what-if analyses, which are crucial for power systems planning and financial planning.

Since each technique has its own strengths and weaknesses, we can make use of the strengths of each by taking a multi-stage approach. For instance, we can use non-black box and multivariate models to capture the salient features of the electricity demand, then use black box and/or univariate models to forecast the residual series. Alternatively, we can also combine the forecasts from multiple techniques, which is considered to be best practice for load forecasting.

3.2. Methodologies

Most papers in the load forecasting literature simply present a single model and compare it with other models, then draw the unsound conclusion that one technique was better than the others. However, many papers, including some of those discussed in Section 3.1, also illustrate how a methodology can be used to solve the load forecasting problem or its sub-problems. These methodologies can usually be applied to multiple techniques. In this section, we will discuss a few of them, from classical ones such as the similar day method to recent ones such as weather station selection.

3.2.1. Similar day method

The idea of the similar day method is to find a day in the historical data that is similar to the day being forecasted. The similarity is usually based on day of the week, season of the year, and weather patterns. As was mentioned by Hong (2014), the similar day method was one of the first methods to be applied to load forecasting. Even now, many system operators still display the load and temperature profiles of the representative days on the wall of the operations room. Today, the similar day method is often implemented using clustering techniques. Instead of one similar day, the algorithms may identify several similar days or similar segments of a day, and then combine them to obtain the forecasted load profile.

3.2.2. Variable selection

For many techniques that rely on explanatory variables, an important step is determining which explanatory variables to use and their functional forms. Hong (2010) proposed a variable selection mechanism and applied it to three different techniques for STLF, namely linear regression, ANN and fuzzy regression. The results showed that, for each of the three techniques, the proposed mechanism was able to reduce the forecasting errors gradually. In a follow-up work, Wang et al. (2016) took a big-data approach to variable selection, where the algorithm allows selection of a large amount of lagged and moving average temperature variables to enhance the forecast accuracy. Several other papers have also showed a step-by-step refinement of the base models or captured the salient features one by one (e.g., Fan & Hyndman, 2012 and Nedellec et al., 2014), though they did not plug different techniques into the same modeling framework.

3.2.3. Hierarchical forecasting

The deployment of smart grid technologies has meant that the question of how hierarchies can be utilized to improve load forecasts has become an important topic

Table 2
Exemplary papers that reported valuable work currently being used by the industry.

Papers	Forecasting systems or commercial solutions
Khotanzad and Afkhami-Rohani (1998)	ANNSTLF (a commercial STLF solution from EPRI)
Hong (2008); Willis (2002)	LoadSEER (a commercial spatial load forecasting solution from integral analytics)
Fan et al. (2009)	A STLF system used by Western Farmer Electric Cooperative
Hong (2010); Hong, Wilson et al. (2014)	SAS® Energy Forecasting (a commercial load forecasting solution from SAS)
Hyndman and Fan (2010); Hyndman and Fan (2014)	A LTLF system used by the Australian Energy Market Operator
Hong et al. (2015)	A weather station selection system used by NCEMC and many other US utilities
Xie et al. (2015)	A retail energy forecasting system used by Clearview Electric and several other US retail electricity providers

in the load forecasting community. The literature on hierarchical load forecasting is limited, but there are a few major milestones in the area. Hong (2008) implemented a hierarchical trending method for spatial load forecasting at a medium-sized US utility, which involved fitting S-curves for 3460 small areas and their aggregated levels through a constrained multi-objective optimization formulation. Fan, Methaprayoon, and Lee (2009) reported the results of a multi-region forecasting project at a Generation and Transmission (G&T) co-op. While the project was aimed at aggregate-level load forecasting, the methodology involved looking for the optimal combination of the regions in order to improve the forecasting accuracy. The authors used the average of all weather stations. Lai and Hong (2013) reported an empirical hierarchical load forecasting case study based on ISO New England data, which included several different ways of averaging weather stations and grouping loads. If we expand the concept of a hierarchy from geographic/spatial hierarchies to temporal hierarchies, there are many papers in the literature that use 24 different models to produce 24 forecasts for the 24 h of a day (e.g., Khotanzad & Afkhami-Rohani, 1998). Note that none of these hierarchical forecasting methods is limited to a specific technique. In fact, all of them can be implemented with regression models, semi-parametric models, ANNs, and so forth.

3.2.4. Weather station selection

Since the weather is a major factor driving the electricity demand, it is important to figure out the right weather stations to use for a territory of interest. Hong et al. (2015) provided the first original research paper devoted to weather station selection. Two case studies were provided, one based on a field implementation at NCEMC, and the other based on the data from the GEFCom2012. Although MLR models were used to illustrate the proposed methodology, models based on other techniques can also be plugged into this framework. The same weather station selection method was also adopted by Hong and Shahidehpour (2015) in their development of long-term load forecasts in several states of the US.

3.3. Novelty and significance

The ultimate goal of load forecasting research is to create knowledge that will be useful for load forecasting practice in the industry. Over the past three decades, very few scientific papers have actually presented research outcomes that are useful for the industry. One reason for this might be a misunderstanding of the idea of novelty. Table 2 highlights a few examples of papers that have reported valuable work that is currently being used in the industry. In this section, we will use some of these papers, together with other notable references, to illustrate what the novelty in the load forecasting content is. This section serves as a conclusion for the reviews of load forecasting techniques and methodologies. The analogy is also applicable to the discussions of PLF in the following sections.

Novelty is a basic requirement for scientific papers. If a paper presents nothing new or original, it has made no additional contribution to the state-of-the-art, and therefore would not be published by scholarly journals. Novelty in load forecasting includes the following aspects:

(1) New problems: identifying a new problem in the load forecasting arena. For instance, Fan et al. (2009) were solving a new short term load forecasting problem, where a utility's load can be broken down into several regions. Hong et al. (2015) were solving the weather station selection problem, which should be one of the first steps in a load forecasting process. Xie, Hong, Laing, and Kang (in press) were solving the load forecasting problem for retail electricity providers, whose customers may terminate their services at any time. New problems are hard to find, and are usually the result of working closely with the industry.

(2) New methodologies: proposing a new load forecasting methodology. New methodologies usually come with new problems. For instance, Fan et al. (2009) proposed a grouping method for multi-region forecasting, while Xie et al. (2015) proposed a two-step method in order to mitigate the risk of volatile customer counts due to marketing activities. Sometimes new methodologies can also be proposed for exisiting problems. For instance, Hong (2010) proposed a heuristic method for variable selection.

(3) New techniques: proposing or applying a technique that has not been tried previously for load forecasting. For instance, Hyndman and Fan (2010) used semi-parametric models to model the half-hourly demand for long term load forecasting. Xie et al. (2015) introduced the use of survival analysis to model customer attrition for retail energy forecasting. Sometimes researchers play the game of putting several techniques together and assuming the hybrid technique to be new. As was discussed in Section 3.1.9, most of these hybrid ones are of minimal value for load forecasting practice.

(4) New datasets: using new datasets to test new or existing methodologies and techniques. These case studies usually provide evidence as to whether the methodologies and techniques work well on another dataset or not. For instance, Khotanzad and Afkhami-Rohani (1998) applied ANN to a large set of data from many utilities.
(5) New findings: presenting a more in-depth analysis than has been done previously, resulting in some additional findings. For instance, Khotanzad and Afkhami-Rohani (1998) provided a new design of ANN structures that resulted in better forecasts than those in previous studies. Hong and Wang (2014) pointed out that a frequently cited paper misused the technique of fuzzy regression.

Nevertheless, novelty is not equivalent to significance. While reviewing the extensive literature, we have found that novel ideas inspired by real-world projects usually lead to findings that are of great significance. Therefore, we would like to encourage researchers to work closely with the industry in order to maximize the likelihood of making a significant contribution to the load forecasting field.

4. Probabilistic load forecasting: two perspectives

The PLF literature has been developed from two main angles. One is the application side, where researchers need PLFs as inputs for the decision making process. The other is the technical and methodological development side, where researchers are focusing on enhancing the forecast quality.

4.1. Applications

Load forecasts are used in virtually all segments of the power industry, and PLFs are no exception. The applications of PLF spread across power systems planning and operations. In this section, we review several important applications in which researchers have been moving from the traditional deterministic decision making framework to its probabilistic counterpart, with the PLFs as an input.

4.1.1. Probabilistic load flow

Load flow analysis, also known as power flow analysis, is an important part of power systems analysis. It involves the application of numerical analysis to a power system in its steady state, in order to obtain the magnitude and phase angle of the voltage at each bus, as well as the real and reactive power flowing in each line. In reality, the future state of a system is never 100% accurate. The uncertainties include generation outages, changes in network configuration, and load forecasting errors. Having recognized the necessity of incorporating these uncertainties into load flow analysis, researchers have been investigating probabilistic load flow analysis since the 1970s.

Borkowska (1974) proposed a methodology for the evaluation of power flow that involved a consideration of the node data uncertainty. Several load levels were given for each node, together with the associated probabilities, and the proposed methodology then found the corresponding set of branch flow values. Allan, Borkowska, and Grigg (1974) proposed a method for analyzing the power flow probabilistically. All of the nodal loads and the generation were defined as random variables. The outputs included the mean and standard deviation of each power flow and the probability density function of the overall balance of the power. The forecasted load was assumed to be a random variable following a normal distribution.

Another way to evaluate the probabilistic load flow problem is through the use of Monte Carlo simulation. This involves running many cases of deterministic load flows, which takes a significant computational effort. On the other hand, the results are quite accurate, since it utilizes the exact load flow equation directly. As was discussed by Allan, Silva, and Burchett (1981), these simulation results are often used as a benchmark for comparisons with other probabilistic load flow methods.

Chen, Chen, and Bak-Jensen (2008) provided a review of probabilistic load flow. In addition to covering basic techniques such as the two mentioned above, the authors also discussed other techniques that improved the accuracy and efficiency of the basic ones, as well as several applications, such as systems planning, voltage control, and the integration of distributed generation.

4.1.2. Unit commitment

Unit commitment determines when to run which generator and at what level, in order to satisfy the electricity demand. By its nature, this is an optimization problem that minimizes the costs subject to many constraints on the units and the system. Popular solution techniques include heuristic searches, dynamic programming, Lagrangian relaxation and mixed integer programming.

Zhai, Breipohl, Lee, and Adapa (1994) proposed a methodology for analyzing the effect of the load uncertainty on the probability of not having a sufficient committed capacity to compensate for unit failure and unexpected load variation. The point load forecast described the unconditional mean, while the unconditional variance described the unconditional uncertainty. The conditional mean and variance on the latest observed load were derived using a Gauss-Markov model. This was the first quantitative demonstration of the effect of load uncertainty on the unit commitment risk.

Douglas, Breipohl, Lee, and Adapa (1998) presented a study that analyzed the risk due to STLF uncertainty for the short term unit commitment. A Bayesian load forecaster was used to produce one- to five-day-ahead forecasts. The load was assumed to be a random variable that follows a normal distribution. The authors used a case study with utility-derived system data and temperature forecast data from the National Weather Service to find the expected cost of the uncertainty due to load forecast variation.

Valenzuela, Mazumdar, and Kapoor (2000) also analyzed the influence of the load forecast uncertainty on production cost estimates. Several increasingly comprehensive load models were considered, ranging from a Gauss model that assumed the load to follow a normal distribution, to a Gauss-Markov regression model that assumed the load to follow a Gauss-Markov process and was driven by temperature. Through a case study using two years of actual load and temperature data, the authors found out that "*a knowledge of the correlation that exists between the hourly loads and the temperature results in a reduction of the standard deviation associated with the conditional distribution of each hour's load*", which is similar to Hong and Wang's (2014) conclusion that the fuzziness could be reduced by improving the underlying model. This finding eventually led to the conclusion that "*for the particular day, including the temperature and the correlation between the hourly loads gives rise to a better estimation of the expected production costs*". This was a major step toward the usage of advanced predictive models for production cost estimation.

Hobbs (1999) analyzed the value of forecasting error reductions in terms of unit commitment costs. Instead of simulating the forecasting errors using some predefined probability distribution or stochastic models, this study was based on actual forecasting errors. The authors concluded that a 1% forecasting error reduction for a 10 GW utility could save up to $1.6 million annually, though it should be noted that these numbers were derived in the late 1990s, and therefore may not reflect today's costs. Nevertheless, the methodology used to reach this conclusion can still be used to evaluate the savings from forecasting improvements. A more recent study by Hong (2015) produced an estimated cost of a similar scale.

Wu, Shahidehpour, and Li (2007) proposed a stochastic model for long-term security-constrained unit commitment problems. In this paper, load forecasting uncertainties were modeled as a uniform random variable, represented by 5% of the weekly peak load. The authors divided the scheduling horizon into several time intervals, and created a few scenarios for each time interval, based on historical data, to reflect the representative days/hours chosen for each week/season.

Wang, Xia, and Kang (2011) proposed a full-scenario unit commitment formulation, which was then translated into an interval mixed integer linear programming problem. The proposed method was capable of acquiring the worst-case impact of a volatile node injection on the unit commitment. Load forecasting was outside the scope of this paper, although the authors assumed that the forecasting methods could forecast both the expected nodal load and the upper and lower limits of the prediction interval. In other words, the proposed methodology required a probabilistic load forecast as an input.

4.1.3. Reliability planning

Reliability is one of the most important aspects of generation and transmission operations and planning. A widely adopted reliability measure of the grid is the loss of load probability (LOLP), which refers to the probability that the generation supply will not be sufficient to support the electricity demand. Stremel (1981) presented a method that allowed the generation expansion criterion to be based upon a reliability target, where the reliability index was similar to LOLP. The load forecasts were assumed to fall within one of five scenarios (very low, low, median, high and very high) with different probabilities (0.09, 0.14, 0.54, 0.14 and 0.09, respectively). However, Stremel (1981) did not discuss how these probabilities were obtained.

Hoffer and Dörfner (1991) developed a model that could take into account the uncertainty of the peak load forecast and extreme load values when calculating the production cost. The load duration curve was assumed to be a piecewise linear function. While the traditional LOLP calculations relied on a load duration curve with a fixed peak load level, the peak load of Hoffer and Dörfner (1991) was assumed to be distributed exponentially. Later, Hoffer and Prill (1996) relaxed the assumptions by considering more advanced peak load distributions, such as gamma, beta and triangular distributions.

Hamoud (1998) proposed a probabilistic method for evaluating the interconnection assistance between power systems, defined as the amount of power that can be transferred from one system to another without violating the transmission limits or the system reliability level. The load forecasting accuracy was one of the key factors that affected the level of transfers. The uncertainty in the load forecast was not considered in the case study, though the author did mention that the proposed methodology can include the load forecast uncertainty.

Billinton and Huang (2008) examined the effects of load forecast uncertainty in a bulk system reliability assessment. Several important factors were considered, such as changes in the system composition, topology, load curtailment policies, and bus load correlation levels. The load forecast uncertainty was modeled as a normal distribution, with the forecasted peak load as the mean. Three uncertainty scenarios were discussed, with standard deviations of 0, 5% and 10% of the forecasted peak load.

As a key concept in power systems reliability, the operating reserve is the "backup", generating capacity to meet demand within a short time interval under abnormal conditions, e.g., a generator going down or some other disruption to the supply. Under normal conditions, the operating reserve is usually designed to be the capacity of the largest generator plus a fraction of the peak load. Chandrasekaran and Simon (2011) considered load forecast uncertainty for reserve management in a bilateral power market for the composite generation and transmission system. The load forecast uncertainty was modeled by a normal distribution, with the forecasted peak load as the mean and 2–5% of the forecasted load as the standard deviation. This normal distribution was then divided into seven intervals, of which the midpoints were used in the reliability calculation.

4.1.4. Other applications

Bo and Li (2009) proposed the concept and methodology of probabilistic locational marginal price (LMP) forecasting, which incorporated the load forecasting uncertainty into LMP simulation and price forecasting. Based on the assumption that the load forecasting error was a random variable that follows a normal distribution, the authors then derived the expected value of the probabilistic LMP and the upper and lower bounds of its sensitivity. In the case study, the standard deviations of the load forecasting error were assumed to be 1%, 3% and 5% of the forecasted load.

Matos and Ponce de Leao (1995) discussed distribution systems planning with fuzzy loads, where the load forecasting uncertainties were modeled using fuzzy numbers. To evaluate alternative distribution system designs, the authors also defined four attributes, using the fuzzy decision making framework where applicable: installation cost, operating cost, robustness and severity, and global indices. An example illustrating the proposed methodology was also provided. Ramirez-Rosado and Dominguez-Navarro (1996) proposed a similar approach to distribution systems planning, where the load and costs were modeled using fuzzy numbers.

In an electricity market with imperfect competition due to uncertainties in equipment outages, fuel prices, and other price drivers, the forecasted load has a direct effect on the solution of the optimal bidding strategy. Instead of using a normal distribution to model the load forecast uncertainty, Kabiri, Akbari, Amjady, and Taher (2009) proposed a fuzzy approach to modeling the uncertainty of the load forecast. Fuzzy game theory was utilized to develop the optimal bidding strategy for each generation company.

Load forecasts are an important input to the evaluation of power and energy loss for transmission planning. Traditionally, a normal distribution is assumed for modeling the load uncertainty. Nowadays, many utilities have started using the most probable load forecast, with unequal upper and lower bounds that do not follow a normal distribution. Li and Choudhury (2011) presented a method for combining fuzzy and probabilistic load models for the evaluation of transmission energy loss. The BC Hydro system was used to demonstrate the application of the method.

Volatile demand and intermittent renewable energy resources are challenging today's power systems operations. One possible solution may be to incorporate energy storage units. This idea introduces a new question for power systems planning: *how much storage does the power system need*? Dutta and Sharma (2012) aimed to identify the optimal storage size for a system consisting of a wind farm and a load, in order to meet certain specified reliability indices. The probability distribution of forecast errors was assumed to be Gaussian, with zero mean and a known standard deviation that might vary between intervals. The continuous probability distribution curve was discretized to quantize the forecasts into different levels for the stochastic linear program formulation.

In summary, PLFs can be used in most, if not all, places where single-valued load forecasts can be applied. For the past five decades, researchers working on the application side have been trying to create probabilistic load forecasts in order to meet various business needs. On the one hand, these attempts have confirmed the growing need for probabilistic forecasts. On the other hand, most of these forecasts have been based on immature methodologies, such as simulating load forecasts or load forecast errors using a normal distribution. This provides the load forecasting community with a great opportunity to contribute further to the power engineering field, with enhanced PLF methodologies.

4.2. Technical and methodological development

In this section, we will review the load forecasting community's formal PLF attempts. We begin by reviewing short term PLF, then consider long term PLF. At the end, we discuss interval forecasting without a probabilistic meaning.

4.2.1. Short term probabilistic load forecasting

Ranaweera, Karady, and Farmer (1996) proposed a two-stage method for calculating the mean value and prediction intervals of the 24-hour-ahead daily peak load forecasts. The first stage was to train a neural network with actual historical data, in order to generate forecasts without considering the uncertainties of the input variables. The second stage used the neural network parameters, outputs from the hidden and output neurons, and the mean and variance values of the input variables to calculate the mean and variance of the forecasted load through a new set of equations. The authors created 100 test cases using randomly generated temperature forecasts over a one-year period in order to compare the performances of the regular neural networks, which did not consider the uncertainties of the input variables, and the modified ones, which did consider the input variables' uncertainties. A MAPE value was calculated for each test case. Based on the average MAPE values, the modified neural networks outperformed the regular ones for point forecasting. The authors did not evaluate the probabilistic forecasts.

There was a notable error with the technical contents of Ranaweera et al. (1996), as the authors used future information when developing the ANN model. Two years of daily data were used in the case study, one year for training and the other for testing. Training aimed to decrease the error on the training set, and terminated when the test set error began to increase. In other words, the test set was used to determine the number of hidden neurons. The forecasts produced from such a process were not genuine forecasts, nor were they considered to be ex post forecasts.

Charytoniuk, Chen, Kotas, and Van Olinda (1999) proposed a nonparametric method for forecasting the customer demand, aggregated to the distribution level. The proposed method used information that is readily available at most utilities, such as load research data, the monthly energy consumption of a group of customers, customer class information, and hourly temperature forecasts. The load research data for each customer class was used to generate the probabilistic density function estimator of the temperature and the normalized demand distribution for each day type and season type. Then, the expected demand of a customer at a given time and temperature could be derived based on the distribution of the normalized demand at the same time and temperature, together with the monthly energy consumption of this customer. The authors also derived the limits of the aggregated customer demand, based on the demand distribution of individual customers. The paper used data from the Consolidated Edison Company of New York to construct the test cases. The relative root mean square error was used to evaluate the performances of point estimates of the expected demand. The performances of the interval forecasts were evaluated qualitatively by showing that the actual demand was contained by the estimated limits.

Taylor and Buizza (2002) investigated the use of weather ensemble forecasts for ANN-based STLF. An ensemble of weather forecasts consisted of several scenarios of a weather variable, each of which could be used to produce a load forecast. The case study results showed that an average of the load forecasts based on the weather forecast ensemble was more accurate than a point load forecast with a traditional point weather forecast as an input. The paper also used the rescaled variance of scenario-based load forecasts to estimate the variance of the load forecasting error and the load prediction intervals. The load forecast error variance was evaluated according to the R^2 value from the regression of the squared post-sample forecast errors on the variance estimates for the post-sample evaluation period. The prediction intervals were evaluated using Chi-square goodness-of-fit statistics.

Mori and Ohmi (2005) proposed an approach to STLF using a Gaussian process with hierarchical Bayesian estimation. A *Gaussian process* is a stochastic process for which any finite linear combination of samples has a joint normal distribution. The proposed approach was applied to one-step-ahead daily peak load forecasting. Based on a test case constructed using data from a Japanese power company, the Gaussian Process produced better point estimates than three other techniques, namely a multilayer perceptron ANN, a radio basis function network, and a support vector regression (SVR). The probabilistic forecasting performance was evaluated by counting the percentage of the predicted values that fell within the confidence limits. Mori and Kanaoka (2009) then applied a similar approach to temperature forecasting. Kurata and Mori (2009) used an information vector machine based method for short term load forecasting, and proposed a method for representing the predictive values and their uncertainty. Two years later, Mori and Takahashi (2011) proposed a hybrid intelligent method for probabilistic STLF. A regression tree was used to classify data into some clusters. Then a relevance vector machine was constructed in order to forecast the loads of each cluster using Bayesian inference. The proposed method was used to forecast both weather variables and one-step-ahead daily peak loads in a case study using data from a Japanese utility.

Fan and Hyndman (2012) proposed a modified bootstrap method for simulating the forecasting residuals and then generating prediction intervals for the electricity demand. The forecast distributions are evaluated by showing that all of the actual demands fall within the region from the forecasted distribution. The proposed methodology was validated through both out-of-sample tests and onsite implementation by the system operator.

Bracale, Caramia, Carpinelli, Di Fazio, and Varilone (2013) proposed a Bayesian-based solution for forecasting the probability density functions (PDF) of wind and solar power generation and consumer demand for a smart grid one hour ahead. The forecasting results were then used in a probabilistic steady-state analysis. The overall presentation of this paper was not quite clear, due to grammatical errors and the use of confusing mathematical notations and illustrations. At a high level, the probabilistic load forecasting portion of this work was handled in a simplified manner. The authors used the proposed Bayesian approach to forecast the PDF of the total active power of the consumers in a given class across all buses, which were assumed to be normally distributed. They then applied a simple point forecasting method to forecast the participation factors, which represent the probability that a consumer of a given class is connected to a given bus. The results from the two steps were then used to derive the forecasts of each consumer at each bus. The mean of the total active power was estimated based on a first-order Bayesian autoregressive time series model. The paper presented figures showing comparisons between the actual values and the forecasted mean and 5th and 95th percentiles. There were no comparisons with alternative approaches.

Migon and Alves (2013) proposed a class of dynamic regression models for STLF. In addition to a comprehensive discussion on modeling the salient features such as trend, seasonality, patterns in special days, and dependency on weather variables, the authors also explored the facilities of dynamic regression models, including the use of discount factors, subjective intervention, variance learning, and smoothing/filtering. The data used in the paper were from a Brazilian southeastern submarket. While the majority of the paper was on point forecasting, the forecasts were presented with prediction intervals.

Kou and Gao (2014) proposed a sparse heteroscedastic model for day-ahead PLF in energy intensive enterprises (EIE). They argued that the EIE load series was a heteroscedastic time series, due to the start-up and shutdown of some high power consuming production units. Such time series could be modeled using a heteroscedastic Gaussian process (HGP), which is an extension of the standard Gaussian process (GP) with a second GP governing the noise-free output. To reduce the computational complexity of HGP, the authors sparsified the base model using the L1/2 regularizer. The case study data were from a steel plant in China. The proposed sparse heteroscedastic model was compared to GP, splines quantile regression (SQR), SVR, and backpropagation neural networks, and showed a superior performance in terms of point forecasts. It was also compared with GP and SQR for the probabilistic forecasting outputs. The proposed approach also outperformed its competitors for the negative log predictive density, reliability and sharpness.

Quan, Srinivasan, and Khosravi (2014) applied and extended a method called LUBE (lower upper bound estimation) to develop prediction intervals using neural network models. This paper incorporated some of the comments made by Pinson and Tastu (2014) in order to revise the core LUBE method published in several early papers. However, the results were still questionable. For instance, in three testing periods of one week each, 503 of the 504 actual observations fell in the 90% prediction interval.

Liu, Nowotarski, Hong, and Weron (in press) used quantile regression to combine a group of point load forecasts in order to generate probabilistic load forecasts. The core methodology, quantile regression averaging (QRA), was originated from probabilistic electricity price forecasting (Maciejowska, Nowotarski, & Weron, 2015). Another novel aspect of the study by Liu et al. (in press) related to the point forecasts being fed to QRA, as these point forecasts were generated from sister models, which were selected via similar variable selection processes proposed by Wang et al. (2016).

4.2.2. Long term probabilistic load forecasting

Morita, Kase, Tamura, and Iwamoto (1996) applied a grey dynamic model to the production of forecast intervals for the long term electricity demand. The case study was based on 14 years of annual peak data, with the first eight years being used for parameter estimation and the last six years for error analysis. The authors compared the simplest form of a grey dynamic model with a simple linear regression model, where *Year* was the only independent variable. The results showed that the actual annual peak demand exceeded the upper bound of the 90% prediction interval predicted by the simple linear regression model, but stayed under that of the grey dynamic model.

This paper, however, had a major technical flaw that made the case study unable to support the conclusion that the proposed model is "*a new efficient tool for load forecasting*". The data used in this paper showed an obvious level shift starting from the 12th year, where the simple linear regression model exceeded the upper bound. In reality, such upward shifts are usually the result of booms in the economy, policy changes, or other external factors, none of which were included in the modeling process in this case study. In this particular situation, models that were sensitive to interventions and could put more weight on the recent data would win the contest. However, such models usually lose the game when the intervention is a temporary change instead of a permanent one. In other words, the case study in this paper in fact favored the proposed model, which was unfair to the competing model.

McSharry, Bouwman, and Bloemhof (2005) proposed an approach to one-year-ahead annual peak demand forecasting. The case study was based on 10 years (1991–2000) of daily peak demand and daily weather data from a province in the Netherlands. The weather variables included temperature, wind speed and luminosity. The one-year-ahead weather time series were simulated using the method of surrogates, in an attempt to preserve the distribution, autocorrelation, and cross-correlations of the original weather data. Four years of history were used to forecast the next year, while the forecasts were generated on a rolling basis from 1995 to 2000. The uncertainty due to model inadequacy was incorporated by sampling the in-sample fit errors. The probabilistic forecasts included both the magnitude (the probability density function of the peak demand) and timing (the probability that the peak demand will occur on various dates).

Grenier (2006) reviewed the evolution of the load forecasting models and applications at Hydro Quebec from 1981 to 2007. The authors briefly discussed their probabilistic forecasting and normalization methodology. They produced hourly probabilistic forecasts for the 18- to 24-month planning horizon using simulated weather history. The simulation spread 38 years of weather history across seven days, so that each day of the planning period had 266 hourly load profiles based on historical weather conditions. Then, these hourly profiles were used to extract probabilistic daily, weekly and seasonal forecasts. The same model that generated these hourly forecasts was used in their normalization process, the main purpose of which was to normalize the monthly energy. In the mid-1990s, the normalized energy was defined as the energy under normal climatic conditions. After 2000, they introduced an upward adjustment to the historical temperatures from 1971 to 2000, to reflect a warming effect. The normalized monthly energy was redefined as the average of 30 monthly energy values under these 30 adjusted scenarios.

Hor, Watson, and Majithia (2006) proposed a two-stage methodology for forecasting 90 years of future load demand patterns, with a consideration of four future climate change scenarios and socioeconomic scenarios. As the first step, ARIMA was used to predict daily load patterns. The authors found out that the residuals could be fitted better using the Student-t distribution than the normal distribution. The residual series was then modeled using generalized autoregressive conditional heteroscedasticity (GARCH). GARCH was also used to provide the unconditional variance relating to the long term behavior of the time series, assuming no explicit knowledge of the past. Since the lead time was significant in the proposed work, the authors applied unconditional volatility forecasts.

Magnano and Boland (2007) proposed a methodology for simulating the half-hourly electricity demand and demonstrated its application to PLF. The case study was based on data from South Australia. They began by removing the annual and daily seasonality, modeled using Fourier series, and the weekly seasonality, modeled using dummy variables. Then, they modeled the relationship between loads and temperatures using polynomial functions of the temperature, and removed the effects of temperatures. Finally, they modeled the residual series using ARMA processes. The probabilistic forecasts were generated by adding these components in the reverse order. This modeling process was validated by comparing the generated distribution with the actual via Kolmogorov–Smirnov (K–S) statistics. The probabilistic annual peak demand forecasts were validated qualitatively by comparing the forecasted 90th percentile with the past annual peaks. This proposed approach did not incorporate any macroeconomic or demographic information as trend variables, though the authors did mention that this could be one potential future enhancement.

Hyndman and Fan (2010) proposed a systematic methodology for forecasting the density of the long term peak electricity demand. Of the extensive body of literature on load forecasting, this was one of the few papers to report a methodology that was actually being used by power companies in practice. The case study considered in this paper was again from South Australia. The proposed methodology included three stages: modeling, simulating and forecasting, and evaluating. In the modeling stage, the authors developed two models, a semi-parametric model for the half-hourly demand and a linear model for the annual median demand. In the second stage, the authors designed a seasonal bootstrap method for generating temperature scenarios for half-hourly demand simulation. They also used annual economic and demographic scenarios to forecast the annual median demand. The two results were then combined in order to generate the density forecast for annual peaks up to 10 years ahead. The authors evaluated both ex-post and ex-ante forecasts in the evaluation stage. The ex-ante probability density functions for the half-hourly weekly and annual peak demands were compared to the actual peaks and ex-post forecasts. A visual inspection indicated that all of the actual peaks fell within the region predicted by the ex-ante forecast distribution.

The methodology proposed by Hyndman and Fan (2010) was first applied in practice for forecasting the demand for South Australia. It was then used by the Australian Energy Market Operator to forecast its entire Australian territory. The methodology has also undergone many improvements and developments, which were documented by Hyndman and Fan (2014). Two major additions were the identification of model changes over time and the incorporation of PV generation into the long-term forecasting process. In addition, the authors also allowed for climate change by introducing a positive mean and a larger standard deviation to the noise term that was added to the temperature bootstrap samples.

Kankal, Akpınar, Kömürcü, and Özşahin (2011) forecasted Turkey's energy consumption using socioeconomic and demographic variables, applying artificial neural networks and regression analysis. The forecasts were five different scenarios based on different assumptions about the population, GDP, imports and exports. Although the paper reported scenario-based forecasts, there were no probabilities assigned to the various scenarios. Eventually, the forecasts were used to provide upper and lower bounds for describing the possible range of the future energy consumption.

Hong, Wilson et al. (2014) proposed a practical solution to long term load forecasting, by dissecting the problem into three elements, namely predictive modeling, scenario analysis, and load normalization. In the predictive modeling stage, the authors began by developing a regression-based STLF model using Hong's methodology (Hong, 2010), which was then augmented with the inclusion of the gross state product as a macroeconomic indicator. In the scenario analysis stage, 30 weather scenarios were generated using 30 years of weather history. A total of 90 load forecasts were created using 30 weather scenarios interacted with three economy scenarios. Load normalization was performed in a similar way to the creation of scenario-based load forecasts, with the only difference being in the year used as the validation data for model selection. When forecasting the load in year y, the year $y - 1$ was used for model selection. When normalizing the load of year y, the year y itself was used for model selection.

The methodology proposed by Hong, Wilson et al. (2014) was commercialized by SAS, a large analytics software vendor, and deployed to many utilities worldwide. Their paper made three major contributions. Firstly, it demonstrated that models developed based on hourly data have much smaller ex post forecasting errors than those based on monthly data. Secondly, it reported a field implementation of the integrated forecasting methodology proposed by Hong (2010), which helped utilities enhance the efficiency and consistency of their load forecasting processes. Lastly, it proposed the concept of *load normalization against weather*, and provided practical guidelines for the load normalization procedure. The same methodology was also used by Hong and Shahidehpour (2015) for conducting the load forecasting case study.

4.2.3. Interval forecasting without a probabilistic meaning

Sometimes researchers use the term interval forecasting to refer to any forecasting process that provides intervals as the output, though these interval outputs may not have any probabilistic meaning. Such non-probabilistic interval forecasts are usually the result of converting high-resolution load data into multiple low-resolution interval time series.

Garcia-Ascanio and Mate (2010) converted the hourly load into 24 monthly interval-valued electricity demand series, one for each hour of the day, and then applied a vector autoregressive model to the interval time series. The proposed method showed a superior performance in the comparison with an interval multi-layer perceptron method based on a case study using data from a Spanish power system operator. The root mean square scaled error was used to evaluate the forecasts.

Xiong, Bao, and Hu (2014) also converted hourly load series into 24 monthly interval time series. They proposed a hybrid approach for forecasting the interval of the electricity demand, which employs a decomposition-ensemble strategy using two techniques, the bivariate empirical mode decomposition (BEMD) and support vector regression (SVR). The case study was based on PJM data. The authors developed a BEMD-SVR model for generating one-step-ahead forecasts. The interval U Theil statistics were used to evaluate the interval forecasts. In a comparison with four other models, namely the classical EMD-SVR model, the interval Holt's exponential smoothing model, a vector error correction model, and an interval multi-layer perceptron model, the proposed BEMD-SVR model showed a superior performance.

The two papers mentioned above, together with many other similar ones, shared two major issues, due to the initial treatment of the hourly load series. Firstly, rather than making comparisons with other interval forecasting methods, the authors should have compared the proposed approach first with methods that could forecast the original time series directly. Neither paper provided any evidence to show that the conversion was an effective approach. Secondly, although these papers applied interval forecasting methods, the results did not have any probabilistic meaning, because the proposed approach was equivalent to obtaining a point forecast of the hourly load series and then extracting the monthly maximum and minimum loads by hour of the day.

5. Producing and evaluating probabilistic load forecasts

In this section, we discuss the production and evaluation of probabilistic forecasts. Production and evaluation should always come together, because we have to make the model selection decisions based on an evaluation of the interim forecasting results. While there are many parameters to be fine-tuned as part of the forecasting process, we conclude this section with a discussion of integration, a systems engineering approach to PLF.

5.1. Production

The probabilistic load forecasting process can be dissected naturally into three elements: (1) input scenario

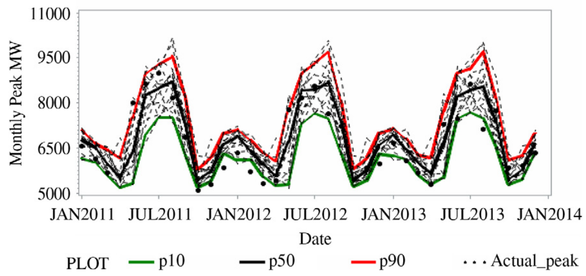

Fig. 4. A three-year probabilistic forecast for a US utility.

generation with simulated predictors; (2) model-dependent interval construction and probabilistic forecasting models; and (3) post-processing through residual simulation or forecast combination. Each of these may provide a probabilistic output.

5.1.1. Input

When feeding multiple scenarios (e.g., temperature scenarios) to a point load forecasting model, multiple point forecasts can be generated to form a probabilistic forecast. If each input scenario is assigned a probability, each point forecast will have the same probability as the corresponding input scenario. Load forecasters often simulate a few dozen to a few thousand different weather scenarios and assign them equal probabilities. For instance, Fig. 4 shows a three-year probabilistic monthly peak load forecast for a US utility (Hong & Shahidehpour, 2015). The dashed lines are 10 point forecasts based on 10 years of weather history. After assigning them equal probabilities, we can derive the important quantiles, such as the 10th, 50th and 90th percentiles.

The simulation methods used in the literature can be categorized into three groups based on the level of complexity of the implementation.

(1) The simplest method is to use the original weather history (or forecasts) directly (Hong, Wilson et al., 2014) used 30 years of weather history to create 30 scenarios for one-year-ahead PLF (Taylor & Buizza, 2002) used 51 weather ensemble members to produce 51 forecasts for one-to ten-day-ahead PLF.
(2) The modest method involves reorganizing the original weather history through rotation or bootstrap methods. Grenier (2006) spread 38 years of historical weather data over seven days, creating 566 weather scenarios. PJM (2014) generated 507 scenarios based on 39 years of data by shifting the weather data forward and backward by one to six days (Hyndman & Fan, 2010) applied a bootstrap method in order to create 2000 years of temperature profiles for each year to be forecasted.
(3) The relatively complicated method is to simulate the historical weather data based on mathematical models McSharry et al. (2005) first took the Fourier transform of the original weather series, then kept the amplitudes and altered the phases so as to generate additional weather scenarios.

While we can easily differentiate between the complexities of these three simulation methods, we have not found any formal study evaluating the quality of the forecasts derived from these methods. This is due largely to the underutilization of formal probabilistic forecast evaluation measures in the past.

5.1.2. Model

The modeling techniques for producing PLFs can be categorized into two groups, based on their original purpose, whether point or probabilistic forecasting:

(1) Some techniques were originally designed for point forecasting, and researchers later extended the methods to enable the construction of prediction intervals, so that the techniques can be used to quantify the future uncertainties around point forecasts. For instance, exponential smoothing has long been a popular point forecasting technique in the industry. Hyndman et al. (2008) later developed the modeling framework so that it could incorporate prediction intervals.
(2) Some techniques were developed with the probabilistic forecasting feature, such as nonparametric probability density estimation (Charytoniuk et al., 1999), Bayesian models (Bracale et al., 2013), sparse heteroscedastic models (Kou & Gao, 2014), and quantile regression (Liu et al., in press).

Some of these interval construction methods have already been implemented in commercial statistical software packages on top of existing point forecasting techniques. Many forecasting software packages today offer the option of plotting prediction intervals for linear regression models. For instance, Levi (1994) built a point forecasting model for medium term weekly peak load forecasting, and at the end showed the prediction intervals generated using a regression-based forecasting model.

On the other hand, the original probabilistic forecasting techniques (e.g., quantile regression) have not received much attention from the load forecasting community over the past thirty years. Instead, some of these techniques have been used for generating point load forecasts, which are essentially the expected values derived from the probabilistic forecasts. One possible explanation for the underdevelopment of these probabilistic forecasting techniques for load forecasting is the fact that their point forecast accuracy may not be as good as those from point forecasting techniques. Before the establishment of formal evaluation methods for PLF, people may have underestimated the power of these probabilistic forecasting techniques based on their underperformance in point forecasting accuracy.

There has not been any formal study comparing these two groups of techniques or the techniques within each group. Again, this is due largely to the underutilization of proper evaluation measures. With respect to computational complexity, when comparing multiple linear regression (a representative point forecasting technique with interval construction capability) with quantile regression (a representative probabilistic forecasting technique), the latter requires much more computing resources than the former with the same number of variables.

5.1.3. Output

Another method of producing probabilistic forecasts is to post-process the point forecasts. This method also takes advantage of the existing point load forecasting literature. The post-processing can be done in two different ways.

(1) One can produce a density forecast by applying the probability density function of residuals to the point forecast. The most straightforward way of doing this is to use a normal distribution to model the forecasting errors, as was done in most of the papers discussed in Section 4. Xie et al. (in press) developed a comprehensive case study showing several myths and tips about the usage of normal distributions for simulating load forecast residuals in the PLF context. A useful conclusion was that adding normal distributions to point load forecasts helps to improve the probabilistic forecasts from deficient underlying models. However, the improvement diminishes as the underlying model is improved. More sophisticated residual simulation methods have also been reported by the load forecasting community. For instance, Hor et al. (2006) modeled the residuals series using GARCH.

(2) Another approach to post-processing is through the combination of several point forecasts. Liu et al. (in press) used quantile regression to average a group of point load forecasts in order to produce PLFs. The QRA-based PLFs were compared with those produced from individual models, which were essentially based on adding simulated residuals to the original point forecasts, showing superior levels of accuracy based on several probabilistic scores.

Although several recent papers have presented some formal studies on the post-processing of point load forecasts in order to produce PLFs, there are still many technical issues to be investigated, such as how to improve the PLFs from advanced underlying point forecasting models; whether QRA works well for combining independent forecasts; and what other techniques can be used to combine point forecasts in order to produce PLFs.

5.2. Evaluation

Since expected values can be derived from probabilistic forecasts, it seems that we can apply the criteria and metrics used for evaluating point load forecasts to the evaluation of probabilistic load forecasts. For short term PLF, we can use the conventional point forecast error metrics, such as MAPE and the mean absolute error (MAE), to evaluate the expected value of the PLF at each step. However, evaluating the expected value may not be particularly realistic for long term PLF, because we rarely have access to a sufficiently long and stable load history to enable us to conduct out-of-sample tests for a horizon of several years. Nevertheless, we can still use point forecast error metrics (e.g., MAPE) to evaluate the ex post forecasting accuracy for long term PLF (Hong, Wilson et al., 2014). For additional details about point forecast evaluation, see Hyndman and Koehler (2006) and Tashman (2000).

Researchers have developed many different evaluation metrics specifically for probabilistic forecasting (Gneiting & Katzfuss, 2014). While some of them have been applied to probabilistic wind power forecasting (Zhang et al., 2014), few have been applied to probabilistic load forecasting. In other words, research into the evaluation of probabilistic load forecasts is almost nonexistent. A few papers have reported rigorous evaluations of probabilistic forecasts, such as those of Arora and Taylor (2016), Liu et al. (in press) and Xie et al. (in press). This lack of well-established evaluation methods is a primary reason for the underdevelopment of the PLF literature.

Here, we begin by discussing the three main attributes of probabilistic forecasts. We then review the metrics that have been used for PLF and have the potential for wide adoption in the industry. We also briefly cover some metrics that have not yet been tested fully for PLF, but that show promising academic value. Because probabilistic forecast evaluation is a fairly well-studied area, we would refer the audience to some valuable papers that offer comprehensive coverages of probabilistic scores.

5.2.1. Criteria

The most commonly used attributes for probabilistic forecast evaluation are reliability, sharpness and resolution, which were covered by Pinson, Nielsen, Møller, Madsen, and Kariniotakis (2007) in their discussion of the quality of probabilistic wind power forecasts. Since this was the first time that these terms had been raised formally in the load forecasting community, we introduce the concepts here instead of leaving it to the references.

Reliability refers to how *close* the predicted distribution is to the actual one. In econometrics, reliability is related to the *unconditional coverage* of a prediction interval. For instance, if a predicted 50% interval covers 50% of the observed load values, then this interval forecast is said to be reliable. Reliability for probabilistic forecasting is similar to bias for point forecasting, where a high reliability corresponds to a low bias. Because in the electric power industry the term reliability is more commonly used to describe the ability of power systems to perform the required functions under stated conditions, the remainder of this paper uses the unconditional coverage as a substitute for reliability as a probabilistic forecast attribute.

Sharpness refers to how *tightly* the predicted distribution covers the actual one. If the maximum and minimum values of the observed loads are very close to the upper and lower bounds of the predicted 99% interval, then this interval forecast is said to be sharp. Sharpness for probabilistic forecasting is similar to the range of errors for point forecasting, where a high sharpness corresponds to a low range of errors.

Resolution refers to how much the predicted interval *varies* over time. Resolution is similar to the *independence* (or clustering) of exceedances of a prediction interval in econometrics. Note that the conditional coverage is the sum of the unconditional coverage and independence (Weron, 2014). Returning to the previous example, if the width of the predicted 50% interval stays the same throughout the forecast horizon, then this interval forecast is said to have no resolution. Resolution for probabilistic forecasting is similar to the variance of forecasts for point forecasting, where a high resolution corresponds to a high variance of forecasts, and no resolution corresponds to a constant forecast.

The sharpness and resolution together capture the electricity demand's dependence on the driving factors, such as weather and calendar variables. For instance, the uncertainty of the electricity demand is supposed to be smaller at midnight, when the weather conditions are smooth and people do not have many activities but are asleep in bed, than in the late afternoon and early evening. In order for the probabilistic forecast to reflect this feature, the 50% interval should be wider in the late afternoon than at midnight.

5.2.2. Measures

Forecast evaluation is an important part of load forecasting, not only to allow the forecasters to select models at the modeling stage, but also to allow stakeholders (i.e., managers, regulators, rate payers and investors) to understand the performance of the forecasting system. Over the past several decades, researchers in the forecasting community have developed many measures for evaluating the point forecast accuracy (Hyndman & Koehler, 2006), though the one used most widely by the electric power industry is still MAPE, due mainly to its simplicity and transparency. In addition, the weaknesses of MAPE, such as difficulties in handling small and zero denominators, are not very relevant for traditional load forecasting problems, because the load at the aggregated level is rarely zero or approaching a very small number.

We believe that the electric power industry will follow a similar path in its adoption of metrics for probabilistic forecast evaluation. Regardless of the number of comprehensive measures proposed in the academic world, including those reviewed by Gneiting and Katzfuss (2014), the industry will use the most simple and transparent ones that can satisfy their business needs. Meanwhile, we would also like to recognize that the probabilistic forecast evaluation methods in the literature may not be received well by the business consumers of load forecasts due to their complexity and lack of transparency.

In the context of load forecasting, most, if not all, business needs in the electric power industry require forecasts in discrete form, as opposed to a continuous probability distribution. Nevertheless, the density forecasts are still quite useful, in the sense that we can derive quantiles from them. For measures for the evaluation of density forecasts, such as the continuous ranked probability score (CRPS), we refer readers to notable papers such as those by Gneiting and Katzfuss (2014) and Zhang et al. (2014). Here, we discuss two groups of measures for evaluating PLFs in discrete form. One group involves evaluating the unconditional coverage only, while the other includes comprehensive measures.

(1) Measures for unconditional coverage only

The simplest measure of the unconditional coverage is the Kolmogorov–Smirnov (KS) statistic. The two-sample KS test can be used to test whether two samples belong to the same distribution. The KS statistic is the maximum vertical distance between the cumulative distribution functions (CDFs) of the two samples. The smallest KS statistic indicates the best forecasted distribution. Magnano and Boland (2007) used the KS statistic to compare the real and simulated half-hourly load profiles.

Since the KS statistic is based only on the maximum vertical distance between two CDFs, it is not very sensitive for establishing the distance between two distributions. For instance, the two distributions may cross each other multiple times, but the maximum deviation between the distributions may still be small. Even if we modified the KS statistic so that it measured the sum of square differences between two distributions, giving the Cramer–von Mises statistic, it would still be insensitive when the difference between the curves is most prominent near the beginning or end of the distributions. The Anderson–Darling (AD) statistic was developed to help overcome such problems, though it has not yet been applied to PLF. Weron (2006) discussed the application of the AD statistic to electricity price forecasting.

Although the aforementioned statistics can be used to assess the unconditional coverage of a probabilistic forecast, they do not evaluate its sharpness or resolution. In other words, one can use all of the historical data to obtain the probability distribution of the load, then apply this same distribution to each step of the forecast horizon. If the calibration is done properly, the KS statistic will be minimized. However, this approach does not capture any salient features of the electricity demand or the associated uncertainties.

(2) Comprehensive measures

In order to consider the sharpness and resolution in the evaluation, we have to assess the forecast at each quantile. The GEFCom2014 was the first time that the load forecasting community formally applied a comprehensive measure, the pinball loss function, to the evaluation of probabilistic load forecasts.

The pinball loss function is an error measure for quantile forecasts. If $\hat{y}_{t,q}$ is the load forecast at the qth quantile and y_t is the load value actually observed, then the pinball loss function can be written as:

$$Pinball(\hat{y}_{t,q}, y_t, q) = \begin{cases} (1-q)(\hat{y}_{t,q} - y_t) & y_t < \hat{y}_{t,q} \\ q(y_t - \hat{y}_{t,q}) & y_t \geq \hat{y}_{t,q}. \end{cases}$$

Note that the pinball function is the function to be minimized in a quantile regression. By summing the pinball losses across all targeted quantiles (i.e., quantiles $q = 0.01, 0.02, \ldots, 0.99$) over the forecast horizon, we can obtain the pinball loss of the corresponding probabilistic forecasts. A lower score indicates a better prediction interval.

Another comprehensive measure is the Winkler score, which was recently used for evaluating PLFs in the load forecasting literature by Liu et al. (in press). The Winkler score allows a joint assessment of the unconditional coverage and interval width. For a central $(1-\alpha) \times 100\%$ prediction interval, it is defined as:

$$Winkler = \begin{cases} \delta, & L_t \leq y_t \leq U_t \\ \delta + 2(L_t - y_t)/\alpha, & y_t < L_t \\ \delta + 2(y_t - U_t)/\alpha, & y_t > U_t, \end{cases}$$

where L_t and U_t respectively are the lower and upper bounds of the PI computed on the previous day, $\delta_t = U_t - L_t$ is the interval width, and y_t is the actual load at time t. The Winkler score gives a penalty if an observation (the actual load) lies outside the constructed interval, and rewards forecasters for a narrow prediction interval. A lower score indicates a better prediction interval.

5.3. Integration: a systems engineering approach

The methods discussed in Section 5.1 can also be used together to create comprehensive simulations. For instance, McSharry et al. (2005) conducted a simulation with 10,000 runs, using different weather scenarios and different error realizations. The quantile regression averaging method proposed by Liu et al. (in press) was applied to individual point forecasts, with the aim of post-processing the outputs of different load forecasting systems.

Building a PLF system requires the integration of input scenario generation, modeling techniques, and post-processing methods. The optimal setup may vary depending upon the evaluation metrics used, such as those discussed in Section 5.2. Furthermore, the winning modeling technique for one input scenario generation method may not provide the best fit for another input scenario generation method. Ultimately, we should take a systems engineering approach to PLF, or load forecasting in general. This requires researchers to consider all of the components simultaneously, instead of sequentially.

GEFCom2012 recognized "integration" as one of the six challenges for the load forecasting community, though none of the winning teams tacked this challenge explicitly (Hong, Pinson et al., 2014). In point load forecasting, Hong et al. (2015) were the first to tackle the integration challenge, and proposed a weather station selection method that answers two questions, how many and which to select, simultaneously instead of sequentially. In probabilistic load forecasting, the first attempt that addressed the integration challenge was that of Xie et al. (in press), who investigated how the performances of residual simulation methods based on different grouping strategies varied as the quality of the underlying point forecasting model was enhanced.

6. Future research needs

Since PLF is a branch of load forecasting, many traditional load forecasting research problems can be considered in the content of PLF. On the other hand, the utility industry has been evolving rapidly over the past 10 years, while the new environment has also led to the emergence of research problems that are quite specific to PLF. In this section, we highlight a few areas of PLF that are underdeveloped.

6.1. Traditional problems

We first discuss the problems that have existed since the inception of PLF, but have not yet been studied fully:

(1) Reproducible and comprehensive empirical studies

Load forecasting is an area that relies heavily on empirical studies. Thousands of empirical studies have been presented as part of the development of the point load forecasting literature, although many of them are not sound technically, due to a lack of reproducibility and comprehensiveness. We hope that the research community has learned the lessons of the past, so that researchers will be able to develop empirical studies based on public data with sufficient detail being provided to enable others to reproduce the work. Moreover, the empirical studies should be comprehensive, so that their conclusions can be generalized to other situations.

(2) Leveraging the point load forecasting literature

While the majority of the load forecasting literature relates to point load forecasting, the way in which the point load forecasting techniques and methodologies should be applied to probabilistic load forecasting is an important research question. In other words, is a better point forecasting model able to produce better probabilistic forecasts?

(3) Scenario generation for PLF

In Section 5.1, we discussed three scenario generation methods for PLF. It is still unclear whether any one of those methods is better than the others. One future research direction could be to test the scenario generation methods on various datasets, with the aim of discovering the advantages and disadvantages of each method.

(4) Error measures for PLF

Thus far, the error measures proposed in the probabilistic forecasting literature have not yet been studied fully in the PLF context. Therefore, gaining an understanding of their usage in the evaluation of PLFs would be a useful first step. Furthermore, we need to identify or propose error measures that are practical for utility applications. Such measures may not be perfect in theory, but should be simple to compute and communicate. For instance, at least one of the measures should be scale-free.

(5) Forecast combination methods

Combination is known to be able to enhance the accuracy of point forecasts. In PLF, Liu et al. (in press) combined point forecasts in order to generate probabilistic forecasts, which is a research direction that overlaps with point (2) above. Another research direction could be the combination of probabilistic forecasts. In other words, can a combination of PLFs outperform individual probabilistic forecasts?

(6) Hierarchical PLF

Although the load forecasting community recently made some progress in hierarchical load forecasting through GEFCom2012 (Hong, Pinson et al., 2014), hierarchical PLF is yet to be considered. A major challenge of hierarchical PLF lies in the reconciliation of forecasts; that is, how can the probabilistic load forecasts be reconciled while maintaining the integrity of each forecast and enhancing their quality?

(7) High-performance computing

Many PLF techniques and methodologies have heavy computational requirements. The problem of designing forecasting algorithms to take advantage of today's computing technologies, such as parallel, in-memory and in-database processing, so as to speed up the forecasting procedures, is a promising direction for future PLF research.

(8) Valuation of PLF improvement

Only a few studies have discussed the benefits of point load forecast improvement, such as those by Hobbs (1999) and Hong (2015). The valuation of PLF improvement is another unstudied but important field. This would require the joint efforts of the load forecasting community and the power engineering community.

6.2. New problems

The changes in the industry over the past century have introduced many research problems for the research community. These newly identified problems often motivate and stimulate the development of new methodologies, and even new techniques.

At the inception of the electric power industry, the field of load forecasting started with LTLF for planning purposes. Gradually, operational needs required people to study STLF. In the late 1980s, the deregulation of the utility industry meant that a lot of attention was directed to the research on STLF. At the same time, transmission and distribution planning needs required spatial load forecasts, which initiated the development of hierarchical load forecasting. In the late 2000s, grid modernization efforts led to hierarchical load forecasting with high resolution data. That made people realize that evaluating load forecasts at the household level is challenging, so Haben, Ward, Vukadinovic Greetham, Singleton, and Grindrod (2014) invented a new error measure for evaluating household-level forecasts. The deployment of smart meters also provided more granular information about the customers' electricity usage, which gave retail electricity providers great opportunities to enhance their load forecasts (Xie et al., in press) revealed the retail energy forecasting problem, along with a plausible solution that is being used by a US retailer.

The popularity of PLF is due mainly to the need to understand the uncertainties associated with the decision making processes in today's industry. Below we discuss five new problems that have introduced the greatest-ever uncertainties for the electric power industry.

(1) Climate variability

A popular topic over the last decade has been climate change or global warming. With the wide debates and disagreements among scientists, politicians and the general public, climate change is gradually becoming a religion. Nevertheless, many utilities want to incorporate the effects of climate change in their long term load forecasting models. The current PLF practices do not fully address how to do this, in terms of either the level shift as a result of climate change, or, in particular, the change in the demand densities. There are at least three challenges in this research direction, such as how to model the input weather variables to reflect climate change; how to model changes in human behavior as a result of rising temperatures; and how to manage the statistical model, given that future temperature values will fall outside the range of the training data.

(2) Electric vehicles

The growing market penetration of electric vehicles (EV) has introduced another challenge for PLF. Currently, heating, ventilating, and air conditioning (HVAC) loads represent the majority of the residential electricity consumption in the US and many other countries. These loads are driven largely by the weather, and primarily temperature. The EV loads, on the other hand, are driven mainly by human activities. In the short term, the diurnal load profile is affected by the daily work schedules of EV drivers. In the long term, the adoption rate of EVs will be vital for the long term energy and peak demand forecasts. The question of how to model the EV load and incorporate EV load forecasts into PLF is another emerging research problem.

(3) Wind and solar power generation

Unlike the EV load, which is driven mainly by human activities, wind and solar power generation are driven mainly by the weather. The utility-scale wind and solar farms are metered individually, allowing us to develop wind and solar power forecasts based on data from each farm. Then, the net demand forecast is the load forecast less the renewable generation forecast. The greater challenge is introduced by rooftop solar panels, many of which are unmetered, making it difficult to account for the generation of solar power. Moreover, the penetration of rooftop solar panels in the long term depends upon energy policies and tax incentives, which are highly unpredictable.

(4) Energy efficiency

Energy efficiency means using less energy to provide the same service, and is one of the methods that people use to try to tackle the challenge of limited energy resources. Examples include LED light bulbs and energy-efficient HVAC systems and other appliances, such as refrigerators, washers, dryers, and TVs. Some utilities offer energy efficiency programs to encourage customers to replace their existing appliances or light bulbs with their energy efficient counterparts.

The influence of these energy efficiency programs on the long term load may be mixed. When an end user simply replaces their big old CRT TV with a new, thin, energy-efficient LED TV, the total energy consumption from watching TV may drop, assuming that the end user does not spend more time watching TV. However, if the end user buys a new, energy-efficient refrigerator and puts the old one in the garage to cool beers, the total energy consumption for refrigeration is actually increased. Therefore, we cannot simply use the marketing dollars spent on the energy efficiency programs to predict the long term load. Instead, utilities have to either conduct surveys or perform comprehensive data analyses to incorporate energy efficiency in long term PLF.

(5) Demand response

Demand response refers to changes in end users' electricity usages in response to changes in the electricity price or incentive payments. This is one method that people use to try to tackle the challenge of meeting the peak demand. Instead of expanding the infrastructure, utilities can try to slow down the peak demand growth. Typical demand response programs include a time-of-use rate (the end users pay a very low rate during off-peak hours and a very high rate during peak hours) and a critical peak rebate (the end users are getting case rebates by the utilities to curtail their load during peak hours).

Such demand response programs may have a significant effect on the short term load forecasting process (Hong & Wang, 2012). The traditional day-ahead load forecasting process only requires a one-off forecast based on the historical load and weather data and the weather forecast. Once demand response programs are involved, utilities have to do one backcast and two forecasts.

The *backcast* is based on the historical load and weather data during regular hours that were not affected by the demand response programs. By comparing this backcast with the actual metered loads, the utility can obtain an indication of how much these programs have reduced the load.

The *first forecast* is again based on the historical load and weather data during regular hours that are not affected by the demand response programs or the weather forecast. This forecast indicates what the load will be if no demand response programs are triggered.

The *second forecast* is based on what we have learned from the backcasting step, where we estimated the relationship between the curtailed load and the weather. In combination with the weather forecast, we can predict how much of the load will be curtailed by a demand response program. This forecast can help utilities to determine which demand response program(s) to trigger at what time.

Because the demand response programs are not triggered frequently, estimating the curtailed load in the backcasting step and forecasting the curtailed load in the second forecasting step are both very challenging, meaning that a probabilistic output is more desirable than point estimates and forecasts. For this reason, probabilistic demand response forecasting is the last emerging research problem.

6.3. Final remarks

In this paper, we provide a tutorial review of probabilistic electric load forecasting, which is a new branch of the load forecasting problem. We are not aiming for a comprehensive review. Instead, we try to limit the list of references to representative papers that can serve as either role models or counter-examples. We hope that this paper will offer some insights for researchers and practitioners in the area of load forecasting, to assist in the further development of useful models and methodologies. Moreover, we recognize that probabilistic load forecasting is an area that can take advantage of developments from multiple fields, such as statistics, electrical engineering, computer science, and meteorological science. We hope that this paper can provide the broader scientific community with enough background knowledge and good sources of references, so that more researchers can contribute to this new, challenging and important area of probabilistic load forecasting.

Acknowledgments

We would like to acknowledge valuable comments from Rob J. Hyndman, Pierre Pinson, Rafal Weron, and two anonymous reviewers, which led to significant improvements in the presentation quality relative to our original submission.

References

Abu-El-Magd, M. A., & Sinha, N. K. (1982). Short-term load demand modeling and forecasting: a review. *IEEE Transactions on Systems, Man and Cybernetics*, *12*, 370–382.

Alfares, H. K., & Nazeeruddin, M. (2002). Electric load forecasting: Literature survey and classification of methods. *International Journal of Systems Science*, *33*, 23–34.

Allan, R., Borkowska, B., & Grigg, C. (1974). Probabilistic analysis of power flows. *Proceedings of the IEE*, *121*, 1551–1556.

Allan, R. N., Silva, A. M. L. da, & Burchett, R. C. (1981). Evaluation methods and accuracy in probabilistic load flow solutions. *IEEE Transactions on Power Apparatus and Systems*, *PAS-100*(5), 2539–2546.

Arora, S., & Taylor, J. W. (2016). Forecasting electricity smart meter data using conditional kernel density estimation. *Omega*, *59*(A), 47–59.

Ben Taieb, S., & Hyndman, R. J. (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting*, *30*(2), 382–394.

Billinton, R., & Huang, D. (2008). Effects of load forecast uncertainty on bulk electric system reliability evaluation. *IEEE Transactions on Power Systems*, *23*(2), 418–425.

Bo, R., & Li, F. (2009). Probabilistic LMP forecasting considering load uncertainty. *IEEE Transactions on Power Systems*, *24*(3), 1279–1289.

Borkowska, B. (1974). Probabilistic load flow. *IEEE Transactions on Power Apparatus and Systems*, *PAS-93*, 752–759.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: forecasting and control* (4th ed.). Wiley.

Bracale, A., Caramia, P., Carpinelli, G., Di Fazio, A. R., & Varilone, P. (2013). A Bayesian-based approach for a short-term steady-state forecast of a smart grid. *IEEE Transactions on Smart Grid*, *4*(4), 1760–1771.

Brockwell, P. J., & Davis, R. A. (2010). *Introduction to time series and forecasting* (2nd ed.). Springer.

Bunn, D. W. (2000). Forecasting loads and prices in competitive power markets. *Proceedings of the IEEE*, *88*(2), 163–169.

Bunn, D. W., & Farmer, E. D. (1985). *Comparative models for electrical load forecasting*. Wiley.

Chandrasekaran, K., & Simon, S. P. (2011). Reserve management in bilateral power market for composite system with load forecast uncertainty. In *2011 International conference on recent advancements in electrical, electronics and control engineering* (pp. 5–12).

Charlton, N., & Singleton, C. (2014). A refined parametric model for short term load forecasting. *International Journal of Forecasting*, *30*(2), 364–368.

Charytoniuk, W., Chen, M. S., Kotas, P., & Van Olinda, P. (1999). Demand forecasting in power distribution systems using nonparametric probability density estimation. *IEEE Transactions on Power Systems*, *14*, 1200–1206.

Chen, B.-J., Chang, M.-W., & Lin, C.-J. (2004). Load forecasting using support vector machines: a study on EUNITE competition 2001. *IEEE Transactions on Power Systems*, *19*, 1821–1830.

Chen, P., Chen, Z., & Bak-Jensen, B. (2008). Probabilistic load flow: A review. In *3rd International conference on deregulation and restructuring and power technologies* (pp. 1586–1591).

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*, 427–431.

Douglas, A. P., Breipohl, A. M., Lee, F. N., & Adapa, R. (1998). Risk due to load forecast uncertainty in short term power system planning. *IEEE Transactions on Power Systems*, *13*(4), 1493–1499.

Dutta, S., & Sharma, R. (2012). Optimal storage sizing for integrating wind and load forecast uncertainties. In *2012 IEEE PES innovative smart grid technologies* (pp. 1–7).

Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, *27*, 134–141.

Fan, S., Methaprayoon, K., & Lee, W.-J. (2009). Multiregion load forecasting for system with large geographical area. *IEEE Transactions on Industry Applications*, *45*(4), 1452–1459.

Feinberg, E. A., & Genethliou, D. (2005). Load forecasting. In *Applied mathematics for restructured electric power systems: optimization, control, and computational intelligence* (pp. 269–285).

Garcia-Ascanio, C., & Mate, C. (2010). Electric power demand forecasting using interval time series: A comparison between VAR and iMLP. *Energy Policy*, *38*(2), 715–725.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and its Application*, *1*, 125–151.

Goude, Y., Nedellec, R., & Kong, N. (2014). Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE Transactions on Smart Grid*, *5*(1), 440–446.

Grenier, M. (2006). Short-term load forecasting at Hydro-Quebec TransEnergie. In *2006 IEEE power engineering society general meeting*.

Gross, G., & Galiana, F. D. (1987). Short-term load forecasting. *Proceedings of the IEEE*, *75*, 1558–1573.

Haben, S., Ward, J., Vukadinovic Greetham, D., Singleton, C., & Grindrod, P. (2014). A new error measure for forecasts of household-level, high resolution electrical energy consumption. *International Journal of Forecasting*, *30*(2), 246–256.

Hagan, M. T., Demuth, H. B., Beale, M. H., & De Jesús, O. (2014). *Neural network design* (2nd ed.). Martin Hagan.

Hamoud, G. (1998). Probabilistic assessment of interconnection assistance between power systems. *IEEE Transactions on Power Systems*, *13*, 535–542.

Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems*, *16*, 44–55.

Hobbs, B. F. (1999). Analysis of the value for unit commitment of improved load forecasts. *IEEE Transactions on Power Systems*, *14*, 1342–1348.

Hoffer, J., & Dörfner, P. (1991). Reliability and production cost calculation with peak load forecast uncertainty. *International Journal of Electrical Power and Energy Systems*, *13*, 223–229.

Hoffer, J., & Prill, M. (1996). On the models of peak load forecast uncertainty in probabilistic production costing algorithms. *International Journal of Electrical Power and Energy Systems*, *18*, 153–160.

Hong, T. (2008). *Spatial load forecasting using human machine co-construct intelligence framework*. North Carolina State University.

Hong, T. (2010). *Short term electric load forecasting*. North Carolina State University.

Hong, T. (2014). Energy forecasting: past, present, and future. *Foresight: The International Journal of Forecasting*, *2014*(32), 43–49.

Hong, T. (2015). Crystal ball lessons in predictive analytics. *EnergyBiz*, *2015*, 35–37. Spring.

Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, *30*(2), 357–363.

Hong, T., & Shahidehpour, M. (2015). Load forecasting case study. National association of regulatory utility commissioners. Retrieved from http://naruc.org/Grants/Documents/Load%20Forecasting%20Case%20Study%20_%20UNCC.pdf.

Hong, T., & Wang, P. (2012). On the impact of demand response: load shedding, energy conservation, and further implications to load forecasting. In *2012 IEEE PES general meeting*.

Hong, T., & Wang, P. (2014). Fuzzy interaction regression for short term load forecasting. *Fuzzy Optimization and Decision Making*, *13*(1), 91–103.

Hong, T., Wang, P., & White, L. (2015). Weather station selection for electric load forecasting. *International Journal of Forecasting*, *31*, 286–295.

Hong, T., Wilson, J., & Xie, J. (2014). Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid*, *5*(1), 456–462.

Hor, C. -L., Watson, S. J., & Majithia, S. (2006). Daily load forecasting and maximum demand estimation using ARIMA and GARCH. In *2006 International conference on probabilistic methods applied to power systems*.

Hyndman, R. J., & Athanasopoulos, G. (2013). *Forecasting: principles and practice*. Retrieved from http://otexts.org/fpp/.

Hyndman, R. J., & Fan, S. (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, *25*, 1142–1153.

Hyndman, R. J., & Fan, S. (2014). *Monash electricity forecasting model*. Retrieved from http://robjhyndman.com/working-papers/mefm/.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Springer Series in Statistics*, *Forecasting with exponential smoothing: the state space approach*.

Kabiri, M., Akbari, S., Amjady, N., & Taher, S. A. (2009). Consideration of load forecast uncertainty in calculating the optimal bidding strategy of generating companies. In *IEEE Eurocon 2009* (pp. 622–630).

Kankal, M., Akpınar, A., Kömürcü, M. İ., & Özşahin, T. Ş. (2011). Modeling and forecasting of Turkey's energy consumption using socio-economic and demographic variables. *Applied Energy*, *88*(5), 1927–1939.

Khotanzad, A., & Afkhami-Rohani, R. (1998). ANNSTLF—Artificial neural network short-term load forecaster generation three. *IEEE Transactions on Power Systems*, *13*(4), 1413–1422.

Kou, P., & Gao, F. (2014). A sparse heteroscedastic model for the probabilistic load forecasting in energy-intensive enterprises. *International Journal of Electrical Power and Energy Systems*, *55*, 144–154.

Kurata, E., & Mori, H. (2009). Short-term load forecasting using informative vector machine. *Electrical Engineering in Japan*, *166*(2), 23–31.

Kutner, M. H., Nachtsheim, C., & Neter, J. (2004). *Applied linear statistical models*. McGraw-Hill/Irwin.

Lai, S., & Hong, T. (2013). *When one size no longer fits all: electric load forecasting with a geographic hierarchy*. SAS White Paper.

Levi, V. A. (1994). A medium-term probabilistic model for forecasting loads with a small proportion of air conditioners. *Electric Power Systems Research*, *29*(1), 57–67.

Li, W., & Choudhury, P. (2011). Including a combined fuzzy and probabilistic load model in transmission energy loss evaluation: Experience at BC hydro. In *IEEE power and energy society general meeting*.

Liu, B., Nowotarski, J., Hong, T., & Weron, R. (2015). Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid* (in press) http://dx.doi.org/10.1109/TSG.2015.2437877.

Liu, K., Subbarayan, S., Shoults, R. R., Manry, M. T., Kwan, C., Lewis, F. L., & Naccarino, J. (1996). Comparison of very short-term load forecasting techniques. *IEEE Transactions on Power Systems*, *11*, 877–882.

Lloyd, J. R. (2014). GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. *International Journal of Forecasting*, *30*(2), 369–374.

Maciejowska, K., Nowotarski, J., & Weron, R. (2015). Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *International Journal of Forecasting*.

Magnano, L., & Boland, J. W. (2007). Generation of synthetic sequences of electricity demand: application in South Australia. *Energy*, *32*, 2230–2243.

Matos, M. A., & Ponce de Leao, M. T. (1995). Electric distribution systems planning with fuzzy loads. *International Transactions in Operational Research*, *2*(3), 287–296.

McSharry, P. E., Bouwman, S., & Bloemhof, G. (2005). Probabilistic forecasts of the magnitude and timing of peak electricity demand. *IEEE Transactions on Power Systems*.

Metaxiotis, K., Kagiannas, A., Askounis, D., & Psarras, J. (2003). Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. *Energy Conversion and Management*, *44*, 1525–1534.

Migon, H. S., & Alves, L. C. (2013). Multivariate dynamic regression: modeling and forecasting for intraday electricity load. *Applied Stochastic Models in Business and Industry*, *29*(6), 579–598.

Moghram, I., & Rahman, S. (1989). Analysis and evaluation of five short-term load forecasting techniques. *IEEE Transactions on Power Systems*, *4*, 1484–1491.

Mori, H., & Kanaoka, D. (2009). GP-based temperature forecasting for electric load forecasting. In *TENCON 2009–2009 IEEE region 10 conference*.

Mori, H., & Ohmi, M. (2005). Probabilistic short-term load forecasting with Gaussian processes. In *Proceedings of the 13th international conference on intelligent systems application to power systems*.

Mori, H., & Takahashi, A. (2011). Hybrid intelligent method of relevant vector machine and regression tree for probabilistic load forecasting. In *2nd IEEE PES international conference and exhibition on innovative smart grid technologies* (pp. 1–8).

Morita, H., Kase, T., Tamura, Y., & Iwamoto, S. (1996). Interval prediction of annual maximum demand using grey dynamic model. *International Journal of Electrical Power and Energy Systems*, *18*, 409–412.

Nedellec, R., Cugliari, J., & Goude, Y. (2014). GEFCom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting*, *30*(2), 440–446.

Papalexopoulos, A. D., & Hesterberg, T. C. (1990). A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, *5*, 1535–1547.

Pinson, P. (2013). Wind energy: forecasting challenges for its operational management. *Statistical Science*, *28*, 564–585.

Pinson, P., Nielsen, H. A., Møller, J. K., Madsen, H., & Kariniotakis, G. N. (2007). Non-parametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy*, *10*(6), 497–516.

Pinson, P., & Tastu, J. (2014). Discussion of "Prediction intervals for short-term wind farm generation forecasts" and "Combined nonparametric prediction intervals for wind power generation". *IEEE Transactions on Sustainable Energy*, *5*(3), 1019–1020.

PJM (2014). *PJM Load Forecast Report January 2014*. PJM Resource Adequacy Planning Department, Retrieved from: http://www.pjm.com/~/media/documents/reports/2014-load-forecast-report.ashx.

Quan, H., Srinivasan, D., & Khosravi, A. (2014). Uncertainty handling using neural network-based prediction intervals for electrical load forecasting. *Energy*, *73*, 916–925.

Ramanathan, R., Engle, R., Granger, C. W. J., Vahid-Araghi, F., & Brace, C. (1997). Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting*, *13*(2), 161–174.
Ramirez-Rosado, I. J., & Dominguez-Navarro, J. A. (1996). Distribution planning of electric energy using fuzzy models. *International Journal of Power and Energy Systems*, *16*, 49–55.
Ranaweera, D. K., Karady, G. G., & Farmer, R. G. (1996). Effect of probabilistic inputs on neural network-based electric load forecasting. *IEEE Transactions on Neural Networks*, *7*, 1528–1532.
Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.
Song, K.-B., Baek, Y.-S., Hong, D. H., & Jang, G. (2005). Short-term load forecasting for the holidays using fuzzy linear regression method. *IEEE Transactions on Power Systems*, *20*(1), 96–101.
Stremel, J. P. (1981). Generation system planning under load forecast uncertainty. *IEEE Transactions on Power Apparatus and Systems*, *PAS-100*(1).
Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*, 437–450.
Taylor, J. W. (2008). An evaluation of methods for very short-term load forecasting using minute-by-minute British data. *International Journal of Forecasting*, *24*, 645–658.
Taylor, J. W., & Buizza, R. (2002). Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems*, *17*, 626–632.
Taylor, J. W., & McSharry, P. E. (2007). Short-term load forecasting methods: An evaluation based on European data. *IEEE Transactions on Power Systems*, *22*, 2213–2219.
Tzafestas, S., & Tzafestas, E. (2001). Computational intelligence techniques for short-term electric load forecasting. *Journal of Intelligent and Robotic Systems*, *31*, 7–68.
Valenzuela, J., Mazumdar, M., & Kapoor, A. (2000). Influence of temperature and load forecast uncertainty on estimates of power generation production costs. *IEEE Transactions on Power Systems*, *15*, 668–674.
Wang, P., Liu, B., & Hong, T. (2016). Electric load forecasting with recency effect: a big data approach. *International Journal of Forecasting*, *32*, 585–597.
Wang, Y., Xia, Q., & Kang, C. (2011). Unit commitment with volatile node injections by using interval optimization. *IEEE Transactions on Power Systems*, *26*, 1705–1713.
Wei, W. W. S. (2005). *Time series analysis: univariate and multivariate methods* (2nd ed.). Pearson.
Weron, R. (2006). *Modeling and forecasting electricity loads and prices a statistical approach*. Wiley.
Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, *30*(4), 1030–1081.
Willis, H. L. (2002). *Spatial electric load forecasting* (2nd ed.). CRC Press.
Willis, H. L., & Northcote-Green, J. E. D. (1983) Spatial electric load forecasting: A tutorial review. *Proceedings of the IEEE*, http://dx.doi.org/10.1109/PROC.1983.12562.
Wu, L., Shahidehpour, M., & Li, T. (2007). Stochastic security-constrained unit commitment. *IEEE Transactions on Power Systems*, *22*, 800–811.
Xie, J., Hong, T., Laing, T. D., & Kang, C. (2015). The myths of residual simulation for probabilistic load forecasting. *IEEE Transactions on Smart Grid* (in press) http://dx.doi.org/10.1109/TSG.2015.2447007.
Xie, J., Hong, T., & Stroud, J. (2015). Long term retail energy forecasting with consideration of residential customer attrition. *IEEE Transactions on Smart Grid*, *6*(5), 2245–2252.
Xiong, T., Bao, Y., & Hu, Z. (2014). Interval forecasting of electricity demand: A novel bivariate EMD-based support vector regression modeling framework. *International Journal of Electrical Power & Energy Systems*, *63*, 353–362.
Zhai, D., Breipohl, A. M., Lee, F. N., & Adapa, R. (1994). Effect of load uncertainty on unit commitment risk. *IEEE Transactions on Power Systems*, *9*, 510–517.
Zhang, Y., Wang, J., & Wang, X. (2014). Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, *32*, 255–270.

Tao Hong is the Director of BigDEAL (Big Data Energy Analytics Laboratory), NCEMC Faculty Fellow of Energy Analytics, Graduate Program Director, Assistant Professor of Systems Engineering and Engineering Management, and Associate of the Energy Production and Infrastructure Center at the University of North Carolina at Charlotte. He is the Founding Chair of the IEEE Working Group on Energy Forecasting, General Chair of the Global Energy Forecasting Competition (http://gefcom.org), lead author of the online book *Electric Load Forecasting: Fundamentals and Best Practices* (http://otexts.org/elf), and author of the blog Energy Forecasting (http://blog.drhongtao.com). Dr. Hong received his B.Eng. in automation from Tsinghua University in Beijing and his Ph.D. with co-majors in operations research and electrical engineering from North Carolina State University.

Shu Fan received his B.S., M.S., and Ph.D. degrees in the Department of Electrical Engineering of China's Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995, 2000, and 2004, respectively. He conducted postdoctoral research sponsored by the Japanese Government at Osaka Sangyo University from 2004 to 2006, and was a Visiting Assistant Professor at the Energy Systems Research Center at the University of Texas at Arlington from 2006 to 2007. At present, he is a Senior Research Fellow at Monash University, Clayton, Australia. His research interests include energy system forecasting, power system control, and high-power power electronics.