Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts

Bidong Liu, Jakub Nowotarski, Tao Hong, and Rafał Weron

***Abstract*—The majority of the load forecasting literature has been on point forecasting, which provides the expected value for each step throughout the forecast horizon. In the smart grid era, the electricity demand is more active and less predictable than ever before. As a result, probabilistic load forecasting, which provides additional information on the variability and uncertainty of future load values, is becoming of great importance to power systems planning and operations. This paper proposes a practical methodology to generate probabilistic load forecasts by performing quantile regression averaging on a set of sister point forecasts. There are two major benefits of the proposed approach. It can leverage the development in the point load forecasting literature over the past several decades and it does not rely so much on high-quality expert forecasts, which are rarely achievable in load forecasting practice. To demonstrate the effectiveness of the proposed approach and make the results reproducible to the load forecasting community, we construct a case study using the publicly available data from the Global Energy Forecasting Competition 2014. Compared with several benchmark methods, the proposed approach leads to dominantly better performance as measured by the pinball loss function and the Winkler score.**

***Index Terms*—Electric load forecasting, forecast combination, pinball loss function, prediction interval (PI), probabilistic forecasting, quantile regression, sister forecast, Winkler score.**

I. Introduction

PROBABILISTIC load forecasting refers to providing electric load forecasting output in the form of intervals, scenarios, density functions, or probabilities. Grid modernization has made the electricity demand more active and less predictable than ever before. In addition, the competitive and dynamic environment requires effective operations and planning of the power systems. Variability and uncertainty associated with the electricity demand is becoming a challenge to the utility industry. As a result, more and more decision making processes in the utility industry rely on probabilistic load forecasts. Typical applications of probabilistic load forecasting include stochastic unit commitment, probabilistic price forecasting, probabilistic transmission planning, and so forth [1], [2]. In the microgrid environment, probabilistic load forecasting is rather crucial, because the demand at individual household level or even distribution feeder level is quite volatile due to various demand response programs and feeder reconfiguration activities.

The load forecasting literature has focused on point forecasting, with researchers trying to forecast the expected value of future load using various techniques, primarily statistical techniques (such as regression models, exponential smoothing, and time series models), and artificial intelligence techniques (such as neural networks and support vector machines) [3]–[7]. Recent development on point load forecasting was mainly on forecasting with the data that has high temporal and/or spatial resolution. Examples include hierarchical load forecasting in the Global Energy Forecasting Competition (GEFCom) 2012 [8], household level load forecasting [9], weather station selection [10], and retail energy forecasting with customer attrition information [11]. There is also development on sharpening the accuracy of point load forecasts for microgrid applications [12].

The probabilistic load forecasting literature is quite limited. Hong and Fan [2] offered a tutorial review on probabilistic load forecasting. On short-term probabilistic load forecasting, Ranaweera *et al.* [13] proposed a two-stage neural network-based method to calculate the mean value and confidence intervals of daily peak load forecasts. Taylor and Buizza [14] estimated the variance of forecasting errors and prediction intervals (PIs) using load forecasts based on weather ensembles. Kou and Gao [15] proposed a sparse heteroskedastic model for forecasting the load of energy intensive enterprises. Another branch in short-term probabilistic load forecasting is on fuzzy interval load forecasting. Hong and Wang [16] applied fuzzy interaction regression to generate short-term load forecasts with possibilistic (fuzzy) intervals. Sáez *et al.* [17] developed fuzzy interval models for forecasting in microgrids.

On long-term probabilistic load forecasting, McSharry *et al.* [18] proposed an approach to one-year ahead annual peak demand forecasting using daily peak

Manuscript received February 2, 2015; revised March 26, 2015 and April 20, 2015; accepted May 11, 2015. Date of publication June 26, 2015; date of current version February 16, 2017. This work was supported in part by the National Science Centre (NCN), Poland, under Grant 2013/11/N/HS4/03649; in part by the Ministry of Science and Higher Education (MNiSW), Poland, Core Funding for Statutory Research and Development Activities; and in part by the Croatian Science Foundation under Grant IP-2013-11-2203. Paper no. TSG-00117-2015.

B. Liu and T. Hong are with the Energy Production and Infrastructure Center, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: bdliu0824@gmail.com; hongtao01@gmail.com).

J. Nowotarski is with the Department of Operations Research, Wrocław University of Technology, Wrocław 50-370, Poland, and also with the Energy Production and Infrastructure Center, University of North Carolina at Charlotte, Charlotte, NC, USA (e-mail: jakub.nowotarski@pwr.edu.pl).

R. Weron is with the Department of Operations Research, Wrocław University of Technology, Wrocław 50-370, Poland (e-mail: rafal.weron@pwr.edu.pl).

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

Digital Object Identifier 10.1109/TSG.2015.2437877

demand and daily weather data from The Netherlands. Hyndman and Fan [19] developed density peak load forecasts for the Australian energy market operator. Hong *et al.* [20] presented a methodology used for a large generation and transmission cooperative in the U.S. Hong and Shahidehpour [1] developed a case study with three U.S. power companies covering ten states. Most recent advancement in probabilistic load forecasting comes from the GEFCom2014 [21].

Within the limited literature on probabilistic load forecasting, none of the methodologies proposed so far are looking into generating probabilistic forecasts via forecast combination. This paper adds a new methodology to the probabilistic load forecasting literature. We apply the quantile regression averaging (QRA) technique [22] to a set of sister point forecasts and generate PIs of future electric loads. Sister forecasts are predictions generated from the same family of models, or sister models. While the sister models maintain a similar structure, each of them is built based on different variable selection processes, such as different lengths of calibration window and different group analysis settings. We will provide a more detailed coverage of sister forecasts in Section II-A.

This paper has several features worth emphasizing.

1) The core technique, QRA, is new to the load forecasting literature. So far it has been only applied to electricity price forecasting [22]–[24]. It is also the core technique used in a top entry of the price forecasting track in GEFCom2014 [21].
2) The direct input to QRA can be generated based on point forecasts from individual models. This adds significant practical value. While most of the load forecasting literature is on point forecasting, the proposed method is able to leverage existing development in this research area.
3) Comparing with independent expert forecasts frequently discussed in the literature on forecast combination, the sister forecasts are much easier to generate. This is another aspect of practical value in our approach.
4) The data is publicly available. Most load forecasting papers are using proprietary data, which makes it impossible for others to reproduce the work. In this paper, we use publicly available data from the GEFCom2014 [21], so that future researchers can reproduce our work and easily compare their results with ours.

The remainder of this paper is organized as follows. In Section II, we discuss how to generate sister forecasts and review the QRA technique proposed in [22]. In Section III, we discuss two methods of evaluating probabilistic forecasts—the pinball function and the Winkler score. In Section IV, we present the case study based on GEFCom2014 data and compare the forecasting results of our approach with those of benchmark methods. In Section V, we discuss resolution, practicality, and possible extensions of our approach. Finally, in Section VI, we conclude this paper with discussion of the future work.

II. Methodology

The proposed methodology can be dissected into two steps, generating a set of sister load forecasts and applying QRA on the sister forecasts. In this section, we will introduce both steps conceptually. In Section IV, we will implement the proposed method on GEFCom2014 data.

A. Sister Models and Sister Forecasts

In this paper, we use a family of regression models to yield the sister point forecasts, mainly because regression analysis is a load forecasting technique that is transparent for reproduction and has been widely used in the industry [1], [6], [19], [25]. The fact that all of the four winning teams in the load forecasting track of GEFCom2012 used regression analysis further proves its usefulness for load forecasting [8]. Nevertheless, the proposed methodology does not limit itself to regression analysis. Other techniques, such as time series analysis and neural networks, can also be used to generate sister forecasts.

When developing a regression model for load forecasting, a key step is variable selection. In other words, given a large amount of candidate variables and their different functional forms, we have to select a subset of them to construct the load forecasting model. A variable selection process is the driver of such a model development process. The process may include several components, such as partition of the data, penalty functions, or error measures used to guide the selection process, and the threshold chosen by the forecaster to stop the selection process. Applying the same variable selection process to the same dataset, we should get the same subset of variables. On the other hand, different variable selection processes may lead to different subsets of variables being selected. We call the models constructed by different subsets of variables sister models if there are overlapping components in their variable selection processes. Consequently, the forecasts generated from these sister models are called sister forecasts. For instance, the so-called recency effect models in [26] are sister models, because they are all built based on the same model, Tao's Vanilla benchmark model [6], [8]

$$\hat{y} = \beta_0 + \beta_1 M_t + \beta_2 W_t + \beta_3 H_t + \beta_4 W_t H_t + f(T_t) \quad (1)$$

where $\hat{y}_t$ is the load forecast for time (hour) t; β_i are the parameters; M_t, W_t, and H_t are the month-of-the-year, day-of-the-week, and hour-of-the-day classification variables corresponding to time t, respectively; T_t is the temperature at time t; and

$$f(T_t) = \beta_5 T_t + \beta_6 T_t^2 + \beta_7 T_t^3 + \beta_8 T_t M_t + \beta_9 T_t^2 M_t + \beta_{10} T_t^3 M_t + \beta_{11} T_t H_t + \beta_{12} T_t^2 H_t + \beta_{13} T_t^3 H_t. \quad (2)$$

The recency effect refers to the fact that the current hour load is affected by the weather conditions in the preceding hours. In the load forecasting literature the term was coined by Hong *et al.* [26], who adopted it from psychology, where it means that when asked to recall a list of items in any order, people tend to begin recall with the end of the list, recalling those items best. The differences among these sister models are the amount of lagged temperature variables ($\sum_{\text{lag}} f(T_{t-\text{lag}})$, $\text{lag} = 1, 2, 3, \ldots$) and lagged daily moving average temperature variables ($\sum_d f(\tilde{T}_{t,d})$, $d = 1, 2, 3, \ldots$) added to (1), where

the daily moving average temperature of the dth day can be written as

$$\tilde{T}_{t,d} = \frac{1}{24} \sum_{\text{lag}=24d-23}^{24d} T_{t-\text{lag}}. \quad (3)$$

Then the family of sister recency effect models can be written as

$$\hat{y}_t = \beta_0 + \beta_1 M_t + \beta_2 W_t + \beta_3 H_t + \beta_4 W_t H_t + f(T_t) + \sum_d f\big(\tilde{T}_{t,d}\big) + \sum_{\text{lag}} f\big(\tilde{T}_{t-\text{lag}}\big). \quad (4)$$

By tuning the length of the training dataset (here: two or three years for parameter estimation) and the partition of the training and validation datasets for model selection (here: using the same four calibration schemes as in [26] that either treat all hourly values as one time series or as 24 independent series), we can obtain different "average-lag" (or d-lag) pairs, leading to different sister models.

B. Quantile Regression Averaging

Given a number of models to choose from, a forecaster faces the problem of selecting one of them or a combination of them as "the model." The forecasting literature suggests that as long as we are dealing with point forecasts, a simple average yields satisfactory results [27]. However, applying equal weights to interval forecasts will not ensure the nominal coverage rate, because the mixing of distributions is governed by different rules. In particular, the weights have to change with the quantile, and their estimation process is much more complex than that of point forecasting [22].

A plausible solution to this problem is to apply quantile regression to point forecasts of a number of individual forecasting models [22]. More precisely, the individual point forecasts and the corresponding observation (here: electric load, y_t) are put in a standard quantile regression setting, being treated as independent variables and the dependent variable, respectively [28]. This QRA method yields an interval forecast of the predicted process, but does not use the PI of the individual methods. The quantile regression problem can be written as follows:

$$Q_y(q|X_t) = X_t \beta_q \quad (5)$$

where $Q_y(q|\cdot)$ is the conditional qth quantile of the electric load distribution (y_t), X_t are the regressors (explanatory variables), and β_q is a vector of parameters for quantile q. The parameters are estimated by minimizing the loss function for a particular qth quantile

$$\min_{\beta_q} \left[\sum_{\{t:y_t \geq X_t\beta_q\}} q\big|y_t - X_t\beta_q\big| + \sum_{\{t:y_t < X_t\beta_q\}} (1-q)\big|y_t - X_t\beta_q\big| \right] = \min_{\beta_q} \left[\sum_t \big(q - 1_{y_t < X_t\beta_q}\big)\big(y_t - X_t\beta_q\big) \right] \quad (6)$$

where y_t is the actual load and $X_t = [1, \hat{y}_{1,t}, \ldots, \hat{y}_{m,t}]$ is a vector of point forecasts of m individual models. The choice of the number of individual models can be made arbitrarily (e.g., the best three models or all models, etc.).

III. Evaluating Probabilistic Load Forecasts

Prior to GEFCom2014, the state-of-the-art in probabilistic load forecasting evaluation was on ex-post forecasting error analysis [2]. GEFCom2014 formally brought probabilistic scoring to the load forecasting community, where the pinball loss function was used to evaluate the entries in the competition. To select models, evaluate the proposed method, and compare it to other benchmarks in this paper, we use two probabilistic scoring methods: 1) the pinball loss function, which penalizes for observations lying far from a given quantile and 2) the Winkler score, which additionally takes into account the width of the PI.

A. Pinball Loss Function

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

$$\text{Pinball}\big(\hat{y}_{t,q}, y_t, q\big) = \begin{cases} (1-q)\big(\hat{y}_{t,q} - y_t\big) & y_t < \hat{y}_{t,q} \\ q\big(y_t - \hat{y}_{t,q}\big) & y_t \geq \hat{y}_{t,q}. \end{cases} \quad (7)$$

Note that the pinball function is the function to be minimized in quantile regression, similar to (6). Summing up the pinball losses across all targeted quantiles (i.e., quantiles $q = 0.01, 0.02, \ldots, 0.99$) throughout the forecast horizon, we can obtain the pinball loss of the corresponding probabilistic forecasts. A lower score indicates a better PI.

B. Winkler Score

When dealing with multiple methods with similarly accurate levels of coverage, our preference is to choose the model that yields the narrowest intervals. The score function proposed by Winkler [29], now known as the Winkler (or interval) score, allows to jointly assess the (unconditional) coverage and interval width [25]. For a central $(1-\alpha) \times 100\%$ PI, it is defined as

$$\text{Winkler} = \begin{cases} \delta, & L \leq y_t \leq U \\ \delta + 2(L_t - y_t)/\alpha, & y_t < L_t \\ \delta + 2(y_t - U_t)/\alpha, & y_t > U_t \end{cases} \quad (8)$$

where L_t and U_t are, respectively, 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 a forecaster for a narrow PI. A lower score indicates a better PI.

IV. Case Study

A. GEFCom2014 Data and Sister Load Forecasts

The probabilistic load forecasting track of GEFCom2014 released seven years of hourly load history (2005–2011) and 11 years of hourly weather history from 25 weather stations (2001–2011) [21]. In this paper, the last six years (2006–2011) of load (Fig. 1) and weather data were used to conduct the case study. The simple average of all 25 weather stations was used as the virtual weather station for the territory. Note that the actual future

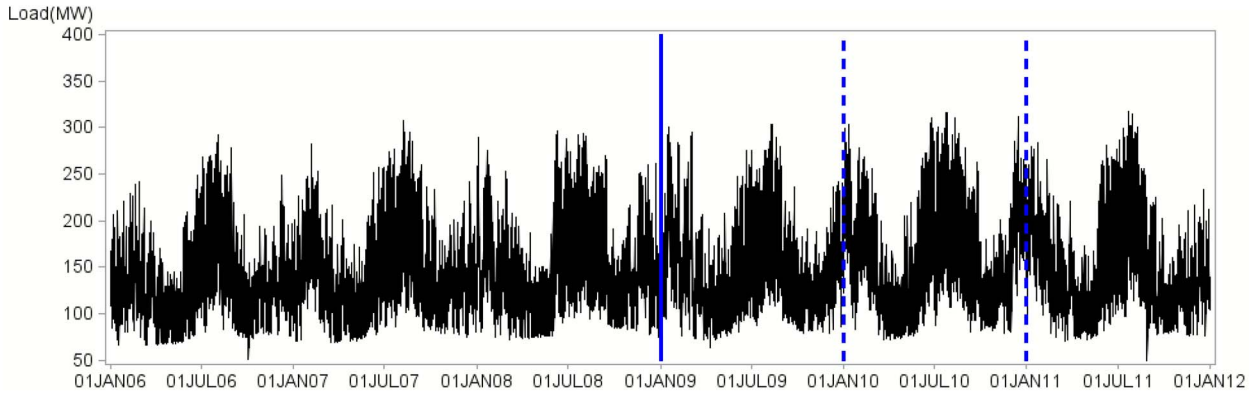

Fig. 1. Six years of hourly load data from GEFCom2014 (2006–2011). The first three years are used for the calibration of the sister (i.e., individual) models only and the latter three for validation and testing of the PI implied by the sister models and the QRA technique.

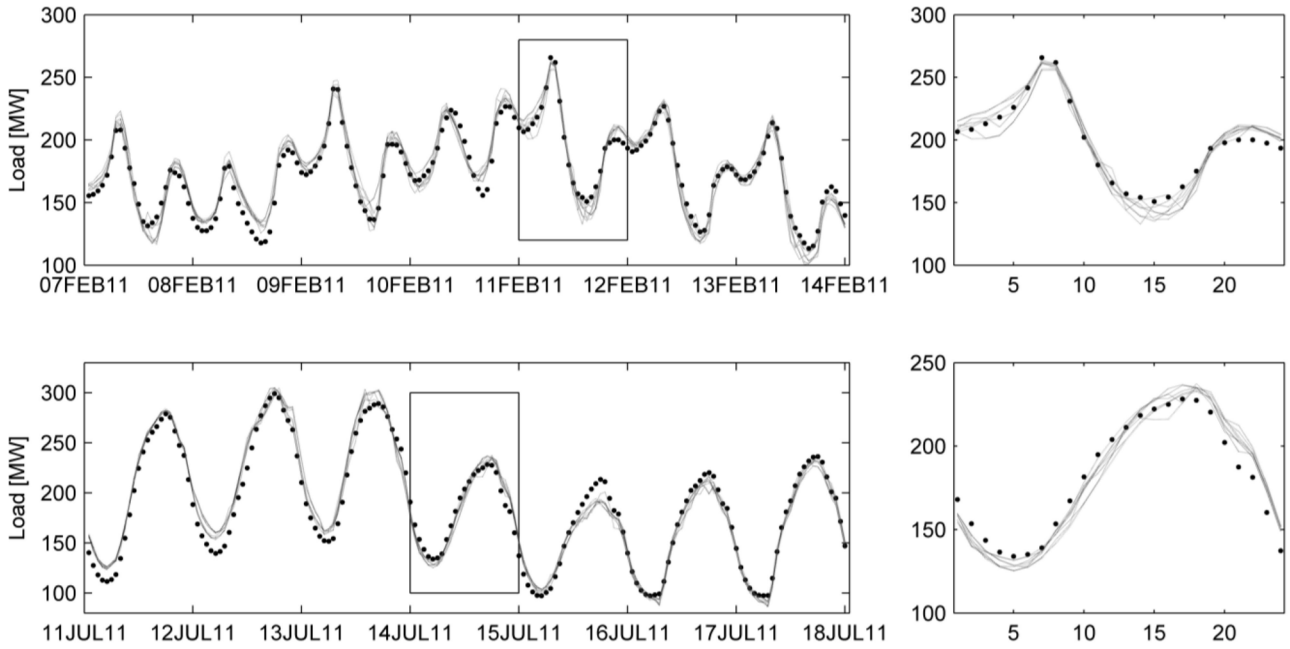

Fig. 2. Actual load (black dots) and eight sister forecasts (gray lines) in a winter week (upper left) and a summer week (bottom left). The right panels show the zoomed-in view on two days, which are indicated by rectangles in the left panels. Since the sister forecasts are similar by nature, the differences between them are hardly visible.

temperatures were used as a proxy for temperature forecasts. As discussed in [30], using actual temperatures to develop models is a common practice in load forecasting. Eight sister load forecasts in three years (2009–2011) were created as input to feed and compare QRA with benchmark methods that utilize the variability of a selected individual forecast (i.e., a sister forecast). The actual load (in black dots) and the eight sister forecasts (in gray lines) in a winter week and a summer week are shown in Fig. 2.

The first four sister models (denoted as Ind1–Ind4) were created based on two years (2007–2008) of training data with the four data selection schemes proposed in [26]. The next four sister models (denoted as Ind5–Ind8) were created based on three years (2006–2008) of training data with the same four data selection schemes. For all eight sister models, data for year 2009 was used as the validation dataset, which allowed selection of the average-lag (or d-lag) pair [26]. These eight sister models were then used to create eight sister 24-h ahead forecasts on a rolling basis for 2009 and 2010.

We then follow the same steps mentioned earlier to create eight sister models using 2010 as the validation year, with two (2008 and 2009) and three (2007–2009) years for training, respectively. These eight sister models are then used to create eight sister 24-h ahead forecasts on a rolling basis for 2011, i.e., they are re-estimated each day.

B. Two Naïve Benchmarks

In this paper, we first create two naïve benchmarks. The first one, named as *Vanilla*, is a scenario-based probabilistic load forecast generated using a similar methodology as discussed in [20]. Here, we are using the model specified in (1) as the underlying model, of which the parameters are estimated using

the most recent two years (2009 and 2010) of data, and the ten years (2001–2010) of weather history is used to generate ten weather scenarios. In total, we are getting ten load forecasts for each hour in year 2011. We then create 99 quantiles based on these ten forecasts. The percentiles 95–99 are set equal to the highest of the ten values, while the percentiles 1–5 are set equal to the lowest of the ten values. The percentiles from 6 to 94 are set to be linearly interpolated values for adjacent load forecasts. Note that this benchmark does not rely on sister forecasts or any forecast combination method at all. It does not rely on metrological forecast either. The variation among different quantiles is purely due to changes in historical weather scenarios.

The second naïve benchmark, named as *Direct*, is generated by directly creating 99 quantiles from the eight sister point forecasts using an analogous approach as for the Vanilla benchmark. In particular, the percentiles 94–99 (and 1–7) are set equal to the highest (and lowest) value of the eight forecasts. The remaining percentiles are linearly interpolated between the adjacent sister forecasts. Comparing with QRA models, which assign different weights to different sister forecasts, the benchmark *Direct* is treating each sister forecast the same. This also means that *Direct* does not require any calibration. This is in contrast to QRA whose weights are estimated based on the forecasting performance in the calibration period.

C. Nine Advanced Benchmarks

In addition to the two naïve benchmarks introduced above, we also create nine advanced benchmarks by adding quantiles of the historical day-ahead prediction errors to nine point forecasts. The first eight point forecasts are the eight sister forecasts generated above (see Sections II-A and IV-A). The last point forecast is the best individual (BI) forecast in the moving calibration window (see Section IV-C) in terms of the mean absolute error (MAE). The BI forecast is essentially the result of a "pick the best one" model selection scheme, but we can also view it as a special case of forecast combination with degenerate weights equal to one for the sister model which performed best in the calibration window and zero otherwise. Naturally, as the calibration window is moved forward, a different sister model may be selected.

To obtain a desired PI for each point forecast, we first apply the logarithmic transformation to all data (historical loads and sister forecasts), then extract the residuals in the calibration window and take the appropriate quantiles of this empirical distribution. All 99 percentiles (i.e., $1\text{st}, 2\text{nd}, \ldots, 99\text{th}$) are added to the corresponding point forecast and are then used for evaluating the pinball function, with the 0.25 and 0.75 quantiles for computing the 50% PI and the 0.05 and 0.95 quantiles for computing the 90% PI. Note that this empirical approach resembles estimation of value-at-risk via historical simulation, a popular benchmark in the risk management literature [23]. Note also that the width of the PI constructed in this way is constant throughout each day in the test period for the log-transformed data. As a result, after applying the inverse transformation (i.e., the exponent), the PIs during off peak periods will be narrower than the PIs during peak periods.

D. Model Selection

The two naïve benchmarks do not require a model selection process, because their underlying models are predefined. Hence, we do not report their performance in the validation period (year 2010). On the other hand, we have to determine several control parameters for the nine advanced benchmarks and the QRA models. For the advanced benchmark methods, we need to determine: 1) which individual model to pick; and 2) the best calibration window length. For QRA, we need to determine: 1) the number of sister forecasts used for QRA; and 2) the best calibration window length. In this paper, we consider four calibration windows of different lengths (24 h times 91, 122, 183, or 365 days) within a rolling scheme, i.e., each day the calibration window is moved forward by one day (or 24 h) and the parameters of the models are re-estimated every day.

Overall, we consider 16 models (all calibrated over four different window lengths) that require model selection before proceeding to generating forecast in the hold-out sample:

1) seven QRA(m) models which take into account $m = 2, \ldots, 8$ BI (sister) models according to MAE;
2) eight sister models;
3) the BI model.

For all 16 models, we compute quantiles and PIs for all hours in the year 2010. This year is then used as the validation period to select the best "model Size—calibration window Length" or "(S,L)" pair. We consider three probabilistic scores: 1) the pinball loss function; 2) Winkler score for the 50% interval; and 3) Winkler score for the 90% interval. The validation results are shown in Table I. We can observe that for any of the three measures and any of the four calibration window lengths, the worst QRA model still outperforms the best Ind model and the BI model. In other words, the QRA models are dominantly better than the benchmark models on all three probabilistic scores in the validation (or post-sample fit or model selection) period.

The best (S,L) pairs in each of the nine subgroups are highlighted in bold in Table I. For each of the three measures, we select the best (S,L) pair from each of the three methods, i.e., QRA models, eight sister models and the BI model, respectively. For instance, on the pinball measure, we select QRA8 with the calibration window of 183 days, i.e., $(S, L) = (8, 183)$, Ind1 calibrated with 91 days, i.e., $(S, L) = (1, 91)$, and BI calibrated with 365 days, i.e., $(S, L) = (-, 365)$.

Note that for the nine advanced benchmark models, as well as the QRA models, we are working with log-transformed data. Likewise, after computing the 99 quantiles and the 50% and 90% PIs, we apply the inverse transformation (i.e., the exponent) and compare the probabilistic forecasts with the actual loads (not log-loads).

E. Forecasting Results

Applying the (S,L) pairs selected in Section IV-D to the test period, the year of 2011, we can obtain the probabilistic

TABLE I
MODEL SELECTION RESULTS FOR QRA AND THE NINE ADVANCED BENCHMARKS IN THE VALIDATION PERIOD (I.E., YEAR 2010)

	Pinball				Winkler (50%)				Winkler (90%)			
	Calibration window length (days)				Calibration window length (days)				Calibration window length (days)			
Model	91	122	183	365	91	122	183	365	91	122	183	365
QRA2	2.58	2.60	2.59	2.60	22.86	23.16	23.03	23.10	41.87	42.94	42.55	42.23
QRA3	2.57	2.58	2.57	2.59	22.76	22.93	22.84	22.97	42.13	42.69	42.14	41.79
QRA4	2.55	2.55	2.53	2.55	22.65	22.62	22.41	22.70	41.67	42.07	41.64	41.61
QRA5	2.52	2.52	2.51	2.52	22.42	22.41	22.28	22.36	41.56	41.69	41.43	41.28
QRA6	2.52	2.52	2.50	2.50	22.38	22.44	22.21	22.23	41.67	42.46	41.59	41.14
QRA7	2.52	2.51	2.50	2.50	22.37	22.34	**22.16**	22.20	41.79	42.22	41.62	**41.05**
QRA8	2.52	2.51	**2.50**	2.50	22.35	22.32	22.19	22.19	41.54	41.85	41.41	41.10
Ind1	**2.64**	2.64	2.64	2.65	**23.26**	23.34	23.35	23.48	43.43	43.45	43.84	**43.42**
Ind2	2.82	2.83	2.83	2.84	25.14	25.13	25.18	25.25	46.80	47.11	47.28	47.09
Ind3	2.84	2.85	2.84	2.84	25.19	25.24	25.20	25.26	46.60	46.43	46.50	46.21
Ind4	2.88	2.88	2.89	2.90	25.59	25.63	25.70	25.78	48.15	48.35	48.50	48.20
Ind5	2.91	2.89	2.90	2.90	25.89	25.77	25.78	25.82	46.08	45.48	45.93	45.29
Ind6	2.94	2.93	2.94	2.93	26.21	26.07	26.20	26.19	47.53	47.40	48.10	47.56
Ind7	3.00	2.99	2.99	2.97	26.71	26.67	26.62	26.50	47.91	47.44	47.65	46.75
Ind8	2.96	2.95	2.97	2.96	26.40	26.28	26.42	26.37	48.07	48.14	49.02	48.33
BI	2.69	2.67	2.69	**2.65**	23.76	23.67	23.81	**23.48**	44.57	44.24	44.71	**43.42**

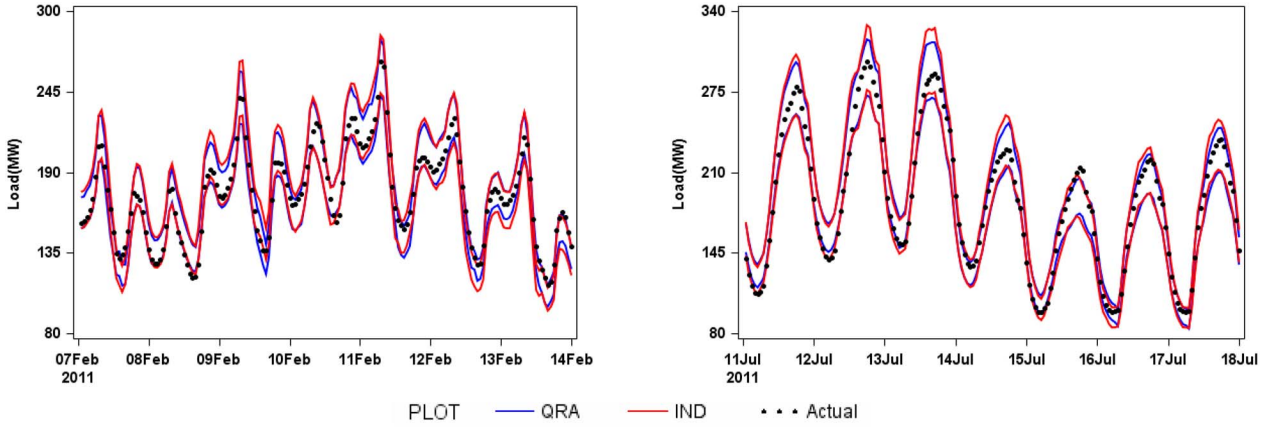

Fig. 3. Actual load (black dots) and two sets of 90% PIs (blue: QRA and red: Ind1) in the same winter week (left) and the same summer week (right) as in Fig. 2. While both PIs are able to cover most of actual values, the PIs of QRA are narrower than those of the individual model.

forecasts at desired quantiles. Fig. 3 shows the actual load and 90% PI from the QRA(8,183) model (combining all eight sister forecasts with a 183-day calibration window) and the selected individual Ind(1,91) model (the first sister model with a 91-day calibration window) for the same winter and summer weeks as presented in Fig. 2. While both PIs are able to cover most of actual values, the PIs of QRA are narrower than those of the individual model.

The probabilistic scores of the selected advanced benchmarks and the two naïve benchmarks are listed in Table II, which confirms that the QRA models are dominantly better than all the benchmark models on all three probabilistic scores. The proposed approach is a marriage between a forecasting combination method working on quantiles of the one-step ahead forecast errors (i.e., QRA) and an individual forecast generating method (i.e., sister forecasts). The advantage of the proposed approach has been demonstrated in three aspects through comparison against the benchmarks.

The outperformance over the advanced benchmarks confirms that combining forecasts results in better accuracy than each of the individual forecasts. This would have not been a surprising conclusion if the outcome of the combination were a point forecast. However, in this paper, the combination outcome is a probabilistic forecast, while the accuracy is measured using probabilistic scoring rules.

The outperformance over the benchmark *Direct* confirms the superiority of QRA over simple interpolation on a group of individual forecasts. Note that the outperformance was quite

TABLE II
FORECASTING RESULTS COMPARISON IN THE TEST PERIOD (I.E., YEAR 2010) FOR THE MODELS SELECTED IN THE VALIDATION PERIOD (SEE TABLE I)

Model	Pinball		Winkler (50%)		Winkler (90%)	
class	(S,L)	Score	(S,L)	Score	(S,L)	Score
QRA	(8,183)	2.85	(7,183)	25.04	(7,365)	55.85
Ind	(1,91)	3.22	(1,91)	26.35	(1,365)	56.38
BI	(-,365)	3.00	(-,365)	26.38	(-,365)	57.17
Direct	-	3.19	-	26.62	-	94.27
Vanilla	-	8.00	-	70.51	-	150.0

significant when measured by Winkler score on the 90% interval. In the utility industry, the 90% interval (or forecasts at high percentiles) is crucial to power systems planning and finance planning [1].

The significant outperformance of QRA models over the benchmark *Vanilla* confirms the advantage of combination-based probabilistic load forecasting over the traditional scenario-based method in the short term. The benchmark *Vanilla* appears to be much worse than the others. This is because it is a genuine ex-ante probabilistic forecast, while the others are ex-post forecasts relying on actual temperature observations in the holdout (or test) period. Since the temperature forecasts are fairly accurate in the short range, i.e., a few days ahead, we can select models based on ex-post forecasting accuracy, which is, again, a common practice in the industry as discussed in [10] and [30]. As the forecast horizon grows, the ex-ante forecasting accuracy difference between the benchmark *Vanilla* and the others is expected to decrease.

While the lower values of the pinball function can be interpreted as accuracy gains compared to individual models, what is particularly worth emphasizing is the QRA dominance in Winkler scores. Recall from Section III-B that this measure is a combination of the (unconditional) coverage and interval width. When evaluating these two attributes together, the consistent outperformance indicates that QRA is still more accurate than all the benchmarks.

V. Discussion

A. Resolution

The QRA-based PIs are constructed jointly for all hours of the test period. Unlike the empirical PI of each individual model, the interval widths generated from QRA are not constant over time. This is due to the fact that a prediction in QRA is the weight (β_q) times a time varying matrix of m individual models, where the weight is constant over all 24 h from a day-ahead. Consequently, the regression can yield different PI widths for low and high levels of load. In other words, the resolution of QRA-based PIs is expected to be more realistic than the empirical PIs of each individual model, which has been measured and confirmed through pinball loss function and Winkler scores, as shown in Tables I and II.

B. Practicality

One of the reasons for using point forecasts (but not probabilistic forecasts) as the input to QRA is their availability. For decades, load forecasters have focused on obtaining accurate point forecasts. Computing probabilistic load forecasts, on the other hand, is generally a much more complex task and has not been discussed in the literature nor developed by practitioners so extensively. Therefore, we find QRA particularly attractive from a practical point of view and expect its widespread use in probabilistic forecasting. As discussed in [22], QRA can also be viewed as an extension of combining point forecasts, in particular the least absolute deviation regression [27].

The proposed approach relies on sister forecasts, which have two methodological advantages over combining independent expert forecasts. First, it is difficult to gather many independent experts to develop forecasts for one business entity, i.e., a utility, retailer, or trading firm. The compromised solution of pursuing independent expert forecasts is usually to have one forecaster develop different forecasts using different techniques. In reality, the forecaster may not have the equal expertise in all the techniques she/he is using. Moreover, it is difficult to manage the personal bias on the favorable techniques. Second, because the sister forecasts are being developed consistently following the variations of several combinations of model selection strategies, the models are naturally transparent and easy to manage. On the other hand, independent experts often serve as a blackbox. Structural changes in the models can hardly be passed to the forecaster in a timely fashion. When the forecaster is combining the forecasts based on previous models, the structural changes in one or several individual models may jeopardize the performance of the forecaster.

C. Extensions

In this paper, we use regression models to generate the sister forecasts for the reasons mentioned in Section II-A. Again, regression analysis is not the only technique that fits the proposed approach. Most, if not all, forecasting techniques require some sort of model selection process. We can generate sister forecasts from different techniques, such as artificial neural networks, support vector machines, and fuzzy regression [6], [7], [16]. Although recency effect models are used as the underlying models to generate sister forecasts, the proposed method can also incorporate other special effects, such as holiday effect.

While the case study is based on 24-h ahead sister load forecasts, the proposed approach does not constrain itself to short-term load forecasting. Since many utilities are still generating point forecasts as the output of the long-term load forecasting process, they can apply QRA to a set of long-term point forecasts to generate interval forecasts in the same way as described in Section II-B.

Another extension is on peak load forecasting accuracy. While all the error measures presented in this paper were calculated on all the hours over the test (or validation) period, the proposed approach is also applicable to optimizing peak performance. For instance, the sister forecasts can be generated based on error measures on daily peaks. And the calibration window and number of sister forecasts in the combination can also be determined based on probabilistic scores on daily peaks.

VI. Conclusion

Among the limited literature on probabilistic load forecasting, there is no formal study on the forecast combination-based approach to generating PIs. In this paper, we propose a novel methodology that applies QRA to a set of load forecasts from sister models. This technique allows us to use the individual

sister point forecasts as independent variables and the corresponding observed load as the dependent variable in a standard quantile regression setting. The practical value of the proposed methodology is significant in the sense that it can leverage existing development of point load forecasting in the literature and does not rely on high quality expert forecasts. In the case study, using the data from the GEFCom2014 probabilistic load forecasting track, we show that the proposed methodology can generate better PIs than the benchmark methods do, according to the pinball loss function and Winkler scores. There are a few directions for the future work, such as exploring the applications of QRA on independent expert forecasts, and testing QRA on other areas of energy forecasting, e.g., renewable generation forecasting.

REFERENCES

[1] T. Hong and M. Shahidehpour, *Load Forecasting Case Study*, Nat. Assoc. Regul. Comm., Washington, DC, USA, 2015.

[2] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecast.*, 2016. [Online]. Available: http://www.drhongtao.com/articles

[3] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 44–55, Feb. 2001.

[4] B.-J. Chen, M.-W. Chang, and C.-J. Lin, "Load forecasting using support vector machines: A study on EUNITE competition 2001," *IEEE Trans. Power Syst.*, vol. 19, no. 4, pp. 1821–1830, Nov. 2004.

[5] R. Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. Hoboken, NJ: Wiley, 2006.

[6] T. Hong, "Short term electric load forecasting," Ph.D. dissertation, Dept. Elect. Eng., North Carolina State Univ., Raleigh, NC, USA, 2010.

[7] S. Fan and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 134–141, Feb. 2012.

[8] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *Int. J. Forecast.*, vol. 30, no. 2, pp. 357–363, 2014.

[9] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 411–419, Jan. 2014.

[10] T. Hong, P. Wang, and L. White, "Weather station selection for electric load forecasting," *Int. J. Forecast.*, vol. 31, no. 2, pp. 286–295, 2015.

[11] J. Xie, T. Hong, and J. Stroud, "Long term retail energy forecasting with consideration of residential customer attrition," *IEEE Trans. Smart Grid*. [Online]. Available: http://dx.doi.org/10.1109/TSG.2014.2388078

[12] N. Amjady, F. Keynia, and H. Zareipour, "Short-term load forecast of microgrids by a new bilevel prediction strategy," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 286–294, Dec. 2010.

[13] D. K. Ranaweera, G. G. Karady, and R. G. Farmer, "Effect of probabilistic inputs on neural network-based electric load forecasting," *IEEE Trans. Neural Netw.*, vol. 7, no. 6, pp. 1528–1532, Nov. 1996.

[14] J. W. Taylor and R. Buizza, "Neural network load forecasting with weather ensemble predictions," *IEEE Trans. Power Syst.*, vol. 17, no. 3, pp. 626–632, Aug. 2002.

[15] P. Kou and F. Gao, "A sparse heteroscedastic model for the probabilistic load forecasting in energy-intensive enterprises," *Int. J. Elect. Power Energy Syst.*, vol. 55, pp. 144–154, Feb. 2014.

[16] T. Hong and P. Wang, "Fuzzy interaction regression for short term load forecasting," *Fuzzy Optim. Decis. Making*, vol. 13, no. 1, pp. 91–103, 2014.

[17] D. Sáez, F. Ávila, D. Olivares, C. Cañizares, and L. Marín, "Fuzzy prediction interval models for forecasting renewable resources and loads in microgrids," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 548–556, Mar. 2015.

[18] P. E. McSharry, S. Bouwman, and G. Bloemhof, "Probabilistic forecasts of the magnitude and timing of peak electricity demand," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 1166–1172, May 2005.

[19] R. J. Hyndman and S. Fan, "Density forecasting for long-term peak electricity demand," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 1142–1153, May 2010.

[20] T. Hong, J. Wilson, and J. Xie, "Long term probabilistic load forecasting and normalization with hourly information," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 456–462, Jan. 2014.

[21] T. Hong, *et al*., "Probabilistic energy forecasting: State-of-the-art 2015," *Int. J. Forecast.*, to be published.

[22] J. Nowotarski and R. Weron, "Computing electricity spot price prediction intervals using quantile regression and forecast averaging," *Comput. Stat.*, 2015. [Online]. Available: http://ideas.repec.org/s/wuu/hscode.html

[23] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *Int. J. Forecast.*, vol. 30, no. 4, pp. 1030–1081, 2014.

[24] K. Maciejowska, J. Nowotarski, and R. Weron, "Probabilistic forecasting of electricity spot prices using factor quantile regression averaging," *Int. J. Forecast*, 2015. DOI: 10.1016/j.ijforecast.2014.12.004.

[25] Y. Goude, R. Nedellec, and N. Kong, "Local short and middle term electricity load forecasting with semi-parametric additive models," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 440–446, Jan. 2014.

[26] P. Wang, B. Liu, and T. Hong. *Electrical Load Forecasting With Recency Effect: A Big Data Approach*. [Online]. Available: http://www.drhongtao.com/articles, accessed Jun. 21, 2015.

[27] J. Nowotarski, E. Raviv, S. Trück, and R. Weron, "An empirical comparison of alternate schemes for combining electricity spot price forecasts," *Energy Econ.*, vol. 46, pp. 395–412, Nov. 2014.

[28] R. W. Koenker, *Quantile Regression*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[29] R. L. Winkler, "A decision-theoretic approach to interval estimation," *J. Amer. Statist. Assoc.*, vol. 67, no. 337, pp. 187–191, 1972.

[30] R. Nedellec, J. Cugliari, and Y. Goude, "GEFCom2012: Electric load forecasting and backcasting with semi-parametric models," *Int. J. Forecast.*, vol. 30, no. 2, pp. 375–381, 2014.

Bidong Liu received the B.S. degree in resource and environmental science from Zhejiang University, Hangzhou, China, in 2007, and the Master of Economics degree from North Carolina State University, Raleigh, NC, USA, in 2008. He is currently pursuing the Ph.D. degree in infrastructure and environmental systems from the University of North Carolina at Charlotte, Charlotte, NC.

His current research interest include electric load forecasting.

Jakub Nowotarski received the M.Sc. degree in financial mathematics from the Wrocław University of Technology, Wrocław, Poland, in 2013, where he is currently pursing the Ph.D. degree in management.

His current research interest include probabilistic electricity price and load forecasting.

Mr. Nowotarski was a recipient of the Best Ph.D. Student Presentation from the Conference on Energy Finance (EF 2014), Erice, Italy.

Tao Hong received the B.Eng. degree in automation from Tsinghua University, Beijing, China, in 2005, and the Ph.D. degree in operations research and electrical engineering from North Carolina State University, Raleigh, NC, USA, in 2010.

He is the Director of the Big Data Energy Analytics Laboratory, a NCEMC Faculty Fellow of Energy Analytics, a Graduate Program Director, and an Assistant Professor with Systems Engineering and Engineering Management, University of North Carolina at Charlotte, Charlotte, NC.

Dr. Hong is the Founding Chair of the IEEE Working Group on Energy Forecasting and the General Chair of the Global Energy Forecasting Competition.

Rafał Weron received the Ph.D. degree in financial mathematics and the Habilitation (higher doctorate) degree in economics.

He is a Professor of Economics with the Wrocław University of Technology, Wrocław, Poland. He is a Consultant to financial, energy, and software engineering companies. His current research interests include developing risk management and forecasting tools for the energy industry, and computational statistics as applied to finance and insurance. He has authored the book entitled *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach* (Wiley, 2006) and co-authored four other books, and has published over 80 peer-reviewed book chapters and journal articles.