International Journal of Forecasting 35 (2019) 1389–1399

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting

Tao Hong [a,*], Jingrui Xie [b], Jonathan Black [c]

[a] Department of Systems Engineering and Engineering Management, University of North Carolina at Charlotte, Charlotte, NC, USA
[b] Forecasting R&D, SAS Institute, Inc. Cary, NC, USA
[c] ISO New England, Holyoke, MA, USA

ARTICLE INFO

Keywords:
Load forecasting
Hierarchical forecasting
Forecasting competition
Energy forecasting
Probabilistic forecasting

ABSTRACT

The Global Energy Forecasting Competition 2017 (GEFCom2017) attracted more than 300 students and professionals from over 30 countries for solving hierarchical probabilistic load forecasting problems. Of the series of global energy forecasting competitions that have been held, GEFCom2017 is the most challenging one to date: the first one to have a qualifying match, the first one to use hierarchical data with more than two levels, the first one to allow the usage of external data sources, the first one to ask for real-time ex-ante forecasts, and the longest one. This paper introduces the qualifying and final matches of GEFCom2017, summarizes the top-ranked methods, publishes the data used in the competition, and presents several reflections on the competition series and a vision for future energy forecasting competitions.

1. Introduction

Forecasting is crucial to the energy sector. The previous two global energy forecasting competitions, namely GEFCom2012 and GEFCom2014, attracted hundreds of data scientists from industry and academia and tackled several forecasting problems in the energy industry, such as hierarchical forecasting and probabilistic forecasting (Hong et al., 2016; Hong, Pinson, & Fan, 2014). Considering the popularity and wide reach of GEFCom2012 and GEFCom2014 both during and after the competition periods, we decided to raise the challenge to a higher level in GEFCom2017, making it the most challenging energy forecasting competition in the series so far.

More than 300 students and professionals from over 30 countries formed 177 teams to compete in GEFCom2017. The theme was hierarchical probabilistic load forecasting, merging the challenges of the load forecasting tracks of the previous two global energy forecasting competitions. The ultimate problem of GEFCom2017 was to forecast the load for hundreds of delivery point meters, which is at a larger scale than the load forecasting problems of GEFCom2012 and GEFCom2014. To avoid the computational burden of grading too many submissions and potential issues with data transfer and communication, we decided to include a qualifying match that limited the number of teams entering the final match. This makes GEFCom2017 the first bi-level competition in the series of global energy forecasting competitions.

Table 1 summarizes the key features of the series of competitions. The complexity of the GEFCom2017 problem and the related rating and ranking methodologies require a flexible competition platform. To pursue that flexibility, we decided to use email for data transfers and LinkedIn as the discussion forum. The qualifying match was five months long, while the final was six weeks long after a one-month break. Thus, GEFCom2017 lasted more than seven months in total, making it the longest competition in the series.

The qualifying match includes two tracks. The defined-data track makes use of the data published by ISO New

* Corresponding author.
E-mail address: hongtao01@gmail.com (T. Hong).

https://doi.org/10.1016/j.ijforecast.2019.02.006
0169-2070/

Table 1
Summary of previous global energy forecasting competitions.

	GEFCom2012	GEFCom2014	GEFCom2017	
			Qualifying match	Final match
Platform	Kaggle	CrowdANALYTIX/LinkedIn	Email/LinkedIn	Email/LinkedIn
Duration	2 months	4 months	5 months	1.5 months
Subjects	Load; Wind	Load; Price; Wind; Solar	Load	Load
Load data	U.S. (unknown)	U.S. (unknown)	U.S. (ISONE)	U.S. (unknown)
Hierarchy level	2	1	3	4
Load series	20; 1	1	8; 5[a] +1; 1	183; 16; 1
Allow external data	No	No	Yes[b]	No
Real-time	No	No	Yes	No
Ex-ante	No	No	Yes	No
Rolling	No	Yes	Yes	No
Output	Point	99 quantiles	9 quantiles	9 quantiles

[a]There are six nodes at the second level (a.k.a. middle level or state level) of the load hierarchy. However, all states except for Massachusetts have the same loads as their individual zonal loads, so we did not ask the contestants to submit forecasts for these five states.
[b]The open-data track allows the usage of external data sources.

England (ISONE) plus U.S. federal holidays. While the majority of the load forecasting literature has used conventional data sources, such as temperatures and calendar information, we hoped to stimulate the usage of diverse data sources in order to further advance load forecasting methodologies and practice. Unlike the two previous competitions, the contestants in this qualifying match know the location of the load data. Thus, we created an open track that allows them to make use of external data sources, which is the only track in the series that has not put a constraint on external data.

The ISONE load data has three levels, as is shown in Fig. 1. The bottom level includes eight zones. The middle level contains the six states. While the state of Massachusetts is the sum of three zones, each of the other five states is the same as the corresponding zone at the bottom level. The top level is the sum of all eight zones. The final match is based on a pre-defined dataset from an anonymous U.S. utility company. The load dataset has four levels, as is shown in Fig. 2. The bottom level includes 183 delivery point meters. These meters are aggregated into 16 groups, which are then aggregated into two control zones (IOO2 and IOO3). The top level is the sum of all meters. GEFCom2017 is the only competition in the series to feature more than two levels of hierarchies in the load data. We wanted to challenge the contestants to take advantage of the hierarchies in their probabilistic load forecasting.

The qualifying match of GEFCom2017 is the first real-time ex-ante forecasting competition in the series. We asked the contestants to forecast the load of the next calendar month on a rolling basis. The setup of the final match was similar to GEFCom2012, where we kept a hold-out period in order to rank the forecasts from contestants. Instead of asking for 99 quantiles as in GEFCom2014, we asked the contestants to submit nine quantiles, from the 10th to 90th percentiles, with an increment of 10, so that each submission could fit into an email attachment.

This paper will introduce the qualifying match in Section 2 and the final match in Section 3, including the problem, rules, data, and techniques and methodologies used by the top teams. We will conclude the paper in Section 4 with our reflections on GEFCom2017 and a vision for future energy forecasting competitions.

Table 2
Important dates of the qualifying match.

Date	Activities
2016-10-14	Competition problems release
2016-12-01	Qualifying match starts
2016-12-15	Round 1 due date; forecast period: 2017-01-01 to 2017-01-31
2016-12-31	Round 2 due date; forecast period: 2017-02-01 to 2017-02-28
2017-01-15	Round 3 due date; forecast period: 2017-02-01 to 2017-02-28
2017-01-31	Round 4 due date; forecast period: 2017-03-01 to 2017-03-31
2017-02-14	Round 5 due date; forecast period: 2017-03-01 to 2017-03-31
2017-02-28	Round 6 due date; forecast period: 2017-04-01 to 2017-04-30
2017-03-10	Report and code due date

2. Qualifying match

The qualifying match was on a hierarchical probabilistic load forecasting problem at a small scale in terms of the size of the hierarchy. It aimed to attract and educate contestants with diverse backgrounds and to prepare them for the final match. The contestants were required to submit forecasts on a rolling basis for at least four of the six rounds. The report documenting the methodology and the code that allowed the organizers to validate the results were due within ten days after the last round. The qualifying match lasted about five months, from the problem release date to the due date of the report and the code. Table 2 lists the important dates. We used two benchmarks in the qualifying match, the *vanilla benchmark* for rating the forecasts from each team, and the *Rain benchmark* as the cutoff for determining the qualifying teams that entered the final match. This section covers the qualifying match in detail, including the data, two benchmark methods, the evaluation method, and a summary of the top-ranked methods.

2.1. Data description

The qualifying match consisted of two tracks: a defined-data track and an open-data track. For both

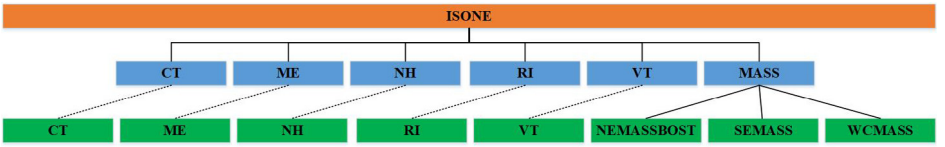

Fig. 1. Hierarchy of ISONE load data.

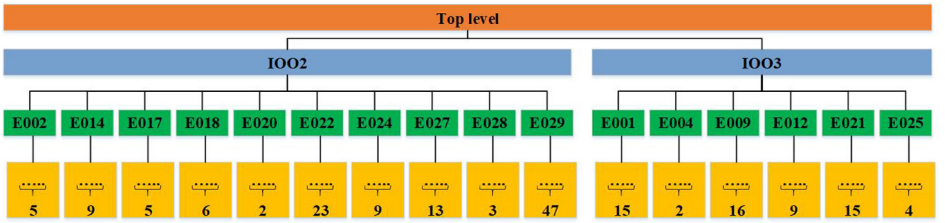

Fig. 2. Hierarchy of the final match data, where the numbers at the bottom level indicate the numbers of meters.

tracks, the contestants were required to provide probabilistic load forecasts for the eight load zones of ISONE at the zonal (eight series), state (one series that is the sum of three load zones under Massachusetts), and system (one series that is the sum of the eight zonal level series) levels, for a total of ten time series. For each of these ten series, the forecasts were generated for the next month in real-time in each round on a rolling basis for six rounds.

In the defined-data track, the contestants could only use the calendar, load (the "DEMAND" column), and temperature (the "Drybulb" and "DewPnt" columns) data provided by ISONE via the zonal information page of the energy, load and demand reports, which were available from 2003 to one or two months prior to the current date. Since ISONE updates the data periodically as part of their load settlement process, the dataset used in this qualifying match is published with this paper as part of the attachment, to allow the research community to benchmark their results against the winning methods in GEFCom2017. The contestants could also use the U.S. Federal Holidays published via the US Office of Personnel Management.

In the open-data track, the contestants were encouraged to explore various public and private data sources and to include the relevant data in the load forecasting process. The additional data might include but was not limited to the real-time data published by ISONE, weather forecast data from any weather service provider, information about the economy, the penetration of PV as published by US government websites, and information on local events.

2.2. Benchmarks

The *vanilla benchmark* aimed to provide some simple forecasts to offset the scale differences among different forecast series and rounds. The probabilistic load forecasts were generated using the shifted-date temperature scenario method that was reported by Xie and Hong (Xie & Hong, 2018a). The historical temperature series from 2005 to 2015 was shifted four days forward and four days backward to create 99 temperature scenarios. These 99 temperature scenarios were then fed into the *vanilla model* in Eq. (1) in order to create 99 point load forecasts for each of the forecast hours. The empirical distribution function was used to derive the required nine quantiles from these 99 point forecasts.

$$\hat{y}_t = \beta_0 + \beta_1 Trend_t + \beta_2 M_t + \beta_3 W_t + \beta_4 H_t + \beta_5 W_t H_t + f(T_t), \tag{1}$$

where $\hat{y}_t$ is the expected load; the βs are coefficients estimated using the ordinary least squares method; $Trend_t$ is a chronological trend; M_t, W_t and H_t are class variables representing the month of the year, day of the week, and hour of the day, respectively; and T_t is the temperature. The multiplications between the variables represent the cross-effects or interactions, and let

$$f(T_t) = \beta_6 T_t + \beta_7 T_t^2 + \beta_8 T_t^3 + \beta_9 T_t M_t + \beta_{10} T_t^2 M_t + \beta_{11} T_t^3 M_t + \beta_{12} T_t H_t + \beta_{13} T_t^2 H_t + \beta_{14} T_t^3 H_t. \tag{2}$$

The *Rain benchmark* aimed to set a higher bar than the *vanilla benchmark* by which to qualify teams to enter the final match. This benchmark was developed by and named after Jingrui (Rain) Xie, one of the GEFCom2017's organizers. It considered all of the factors in the *vanilla benchmark* plus the following:

(1) *Recency effect.* The *Rain benchmark* included the lagged dry-bulb temperature variables and their moving average, as reported by Wang, Liu, and Hong (2016). These augmented temperature variables were selected using the quantile score of the probabilistic load forecasts instead of the mean absolute percentage error (MAPE) of the point forecasts (Xie & Hong, 2018b).

Table 3
Rankings of the *Rain benchmark* in the GEFCom2017 qualifying match.

Round #	Defined-data track	Open-data track
1	6	4
2	9	3
3	9	3
4	22	7
5	18	5
6	11	4
Overall ranking	9	3

(2) *Relative humidity*. The difference between the dry-bulb and wet-bulb temperatures was used as a proxy for the relative humidity (RH) information. This derived RH variable and its polynomials were included in the forecasting model (Xie, Chen, Hong, & Laing, 2018).
(3) Other add-ons. The *weekend effect*, *holiday effect*, *outlier cleansing*, and *residual simulation* methods used by Jingrui Xie in GEFCom2014 and reported by Xie and Hong (2015) were also implemented in order to enhance the individual probabilistic forecasts.
(4) Forecast reconciliation. The proportional reconciliation technique and the optimal reconciliation technique discussed by Hyndman and Athanasopoulos (2014) were both tested. The results from these two reconciliation methods were fairly close. Thus, the proportional method was used for reconciliation due to its simplicity.

While the *Rain benchmark* was developed under the constraints set by the defined-data track, we also used the same benchmark for the open-data track. At the end of the qualifying match, the *Rain benchmark* ranked No. 9 in the defined-data track and No. 3 in the open-data track. Table 3 lists the rankings of the *Rain benchmark* in both tracks.

2.3. Evaluation

The submission was evaluated against the "DEMAND" column published by ISONE, and the quantile score was used to evaluate the skills of the probabilistic forecasts. Specifically, the quantile score is the average of the pinball loss, as defined in Eq. (3), over the nine required quantiles (i.e., 10th, 20th, …, 90th) and the forecast period, and was calculated for each of the ten time series. For the forecast submitted by team j for load series k in round i, the quantile score of the probabilistic forecast is denoted by S_{ijk}.

$$Pinball(\hat{y}_{t,q}, y_t, q) = \begin{cases} (1-p)(\hat{y}_{t,q} - y_t) & y_t < \hat{y}_{t,q} \\ p(y_t - \hat{y}_{t,q}) & y_t \geq \hat{y}_{t,q} \end{cases}, \quad (3)$$

where y_t is the actual load at time t, $\hat{y}_{t,q}$ is the prediction of the qth quantile at time t, and $p = q/100$.

For each round, the quantile score of each time series from each team was compared with the corresponding quantile score of the *vanilla benchmark*. For round i and load series k, the quantile score of the *vanilla benchmark* is denoted by B_{ik}. The relative improvement of each submission over the *vanilla benchmark* was used to rate and rank the teams. Specifically, the relative improvement for round i was calculated as $(1 - S_{ijk}/B_{ik})$ for each time series. The average improvement that team j accomplished over all ten time series was used as the rating for team j, denoted as R_{ij}. The rank of team j in round i was based on the rating R_{ij}, and is denoted by $RANK_{ij}$. The weighted average of the rankings from all six rounds was then used to rank the teams on the qualifying match leaderboard. Of these, the first five rounds were weighted equally, while the weight for the 6th round was doubled, based on the expectation that the contestants would have learned from practicing in the first five rounds and developed the most skillful solution ready to implement in the final round. The contestants with weighted average rankings that were higher than that of the *Rain benchmark* were qualified to enter the final match.

ISONE performs several rounds of load settlement in order to validate and correct the preliminary load values. A major update of the historical load data is released two to three weeks after the end of each month. For instance, the load history of April 2017 was published in mid-May 2017. To maintain the momentum, we did not plan to let the contestants sit idle for two or three months before joining the final match. Instead, we used an *interim score* to inform the contestants about their rankings during the qualifying match and to determine which teams qualified to enter the final match. The interim score of the first n rounds for team j is the simple average of its ratings for the first n rounds, which was defined as $\frac{1}{n}\sum_{i=1}^{n} R_{ij}$.

At the time when the initial invitations to participate in the final match were sent out on April 2, 2017, ISONE had just published the settled load values for February 2017. Therefore, the interim scores of the first three rounds were used to qualify the teams for the final match. Seven teams were ranked above the *Rain benchmark* in the defined-data track, along with two teams in the open-data track. Recognizing that the rankings may change for future rounds, we also included a few teams below but close to the *Rain benchmark*. In total, the first batch of invitations to enter the final match included ten teams. After ISONE released the settled load for March 2017, we invited two additional teams who had outperformed the *Rain benchmark* to participate in the final match on April 24, 2017.

2.4. Black Analytics method

The Black Analytics method was developed by ISONE's forecasting manager Jonathan Black under the team name Black Analytics. Black submitted forecasts to the defined-data track for four of the six rounds, with rankings of 2, 2, 13, and 10, respectively. These rankings were better than some of those of the top five teams in the defined-data track. According to the GEFCom2017 rating and ranking rules, his worst ranking (that is, 13) was used to replace the rankings of the two missing rounds. As a result, Black Analytics ranked No. 6 on the qualifying match leaderboard.

This method begins by building a set of underlying point forecasting models for each zone and hour of a

day separately. These models share the following general form:

$$\hat{y}_t = \alpha_0 + \alpha_1 T + \alpha_2 H + \alpha_3 DM + \alpha_4 W_1 M + \alpha_5 W_2 M + \cdots + \alpha_{n+3} W_n M, \quad (4)$$

where T is a linear trend variable; H captures the holiday effects; D is a classification variable that representing the seven days of a week; M is a classification variable that represents the 12 months of a year; and W_n represents the weather variables. Following Liu, Nowotarski, Hong, and Weron (2017), we call this set of models sister models.

The holiday effects are modeled using the method documented by PJM Interconnection (PJM Interconnection, 2016). Dummy variables are used to indicate federal holidays, while fuzzy dummy variables that take on values between zero and one are used to capture the partial holiday effects of days that surround major holidays. In this competition, fuzzy dummy variables are applied to the surrounding days of the eight federal holidays other than Columbus Day and Veterans Day.

Table 4 lists the W_n variables in each sister model, such as coincident, lagged, and moving-average lagged functional forms of both the dry-bulb (DB) and dew-point (DP) temperatures. The lagged weather variable forms used here are an extension of the work of Wang et al. (2016) on forecasting using the *recency effect*, but with an increased granularity for the moving average values, which here include averages over the most recent 3, 12, 24, and 48 h rather than only daily average values. Model selection was performed on these sister point forecasting models based on their point load forecast performances, measured by the mean absolute percent error (MAPE).

Weather scenarios are developed using the shifted-date method reported by Xie and Hong (2018a). The historical temperature series was shifted seven days forward and seven days backward in order to generate temperature scenarios for each day in the forecasting period. Depending on the availability of historical data, the number of temperature scenarios for each forecast day was either 195 or 210. These weather scenarios are input to the point load forecasting models. Finally, the required quantile forecasts are derived from the remaining point forecasts.

2.5. Summary of methods and results

Out of 177 registered teams, including the co-organizer Jingrui Xie, 73 teams submitted entries to the defined-data track and 26 to the open-data track. By the end of the six rounds, 53 teams had completed the defined-data track with at least four submissions, while 20 had completed the open-data track.

Most of the teams that participated in the open-data track used either the same methods as in the defined-data track or ones that were revised only slightly. Therefore, we will discuss primarily the methodologies and results from the selected top teams of the defined-data track. Table 6 summarizes the methods from six aspects based on their reports, namely data cleansing, feature engineering for load, weather and calendar data, temperature scenario simulation, modeling techniques, forecast combination, and usage of the hierarchy. The ranking column of Table 6 lists the final rankings of each team on the defined-data (D) and open-data (O) tracks. If a team did not participate in one of the two tracks, that ranking will not be available, and thus will be shown by a dash. The *vanilla benchmark* is not ranked.

Data cleansing has been a focus in the competition series. As Table 5 shows, the peak load levels of the previous two competitions ranged from 1 MW to 3280 MW, while those of the ten series in this qualifying match range from 1036 MW to 27,622 MW. Typically, the load series at a higher level of aggregation presents fewer anomalies than the load series at a lower level. Moreover, ISONE carefully scrubs its data before publishing them. Therefore, we expect few anomalies in the data other than those caused by extreme events, such as heavy snowstorms. Of the 12 selected teams, five performed data cleansing, and only one of these ranked in the top five on the qualifying match leaderboard.

The contestants also made use of their domain knowledge in load forecasting in order to create features from the load, temperature, and calendar information that were defined by the track. Three teams derived some features from the load data. Team *Cassandra* implemented a neural network with three hidden layers for extracting the features from the load data. Teams *Orbuculum* and *UC3M* used different forms of the lagged load and the summary statistics of the load. In addition, team *It can be done* conducted logarithmic transformation on the load for several rounds. Temperature information was used extensively by most teams. The temperature features that were used most commonly in this competition were lags and summary statistics, which were similar to the temperature variables of Wang et al. (2016). All teams used calendar information to derive features such as the hour of the day, the day of the week, the month of the year, and the Fourier transformation, in order to capture the multiple seasonality in the load series.

The modeling techniques used in the qualifying match fell into three categories, namely regression analysis, time series analysis, and machine learning. Many teams derived quantile forecasts from point forecasts, which were generated by applying simulated temperature scenarios to point forecasting models. Of these, one team simulated the temperature scenarios from a normal distribution. The other five teams and the two benchmark models simulated the temperature scenarios using the shifted-date method (Xie & Hong, 2018a). The shift factor n is defined as the number of days by which the historical temperature series get shifted forward and backward, and works together with the number of years of historical temperature series k to decide the total number of simulated temperature scenarios. Four teams used all 11 years of temperature history, while the *Rain benchmark* removed some years with extreme temperature values for some forecast rounds. Four teams used a static shift factor across different load zones and forecast rounds. Team *4C* selected the optimal shift factors for different load zones and forecast rounds by evaluating the performances of the multiple linear regression models when using different numbers of simulated temperature scenarios. An alternative method used by several other teams

Table 4
Weather variables (W_n) used in regression models.

Model family	Total no. of W_n variables	Dry-bulb (DB) temperature			Dew-point (DP) temperature		
		Coincident	Lags	Moving average	Coincident	Lags	Moving average
1	6	DB, DB^2	3, 24	3, 24	–	–	–
2	11	DB, DB^2	3, 24	3, 24	DP	3, 24	3, 24
3	10	DB, DB^2	3, 12, 24, 48	3, 12, 24, 48	–	–	–
4	19	DB, DB^2	3, 12, 24, 48	3, 12, 24, 48	DP	3, 12, 24, 48	3, 12, 24, 48

was to generate quantile forecasts directly using quantile forecasting models, such as quantile regression.

Based on the competition results, no one technique or methodology dominated the others. Nevertheless, there are some notable observations that are worth highlighting from the five highest-ranked teams in Table 6: all five teams used forecast combination; four teams used multiple techniques; four used regression; and three used gradient boosting machines. Most of the teams listed in Table 6 used more than one modeling technique and combined the forecasts.

None of the top six teams took advantage of the hierarchy information. Of the 12 teams selected, only four made use of the hierarchy when developing their forecasts. There may be multiple reasons for such a modest use of the hierarchy information. First, the research into hierarchical probabilistic load forecasting was at its infancy when GEFCom2017 was conducted. A hierarchical probabilistic forecasting paper using smart meter data as a case study was presented at a 2017 machine learning conference (Ben Taieb, Taylor, & Hyndman, 2017a), but not many people were aware that hierarchy information may help to improve probabilistic forecasts. While writing this paper, we also found two notable working papers on this subject (Ben Taieb, Taylor, & Hyndman, 2017b; Gamakumara, Panagiotelis, Athanasopoulos, & Hyndman, 2018). Secondly, the competition was intense, with only two months between the problem release and the due date of the first-round submission, and two weeks between two adjacent rounds. The contestants may not have had enough time to investigate hierarchical forecasting methods. Last but not least, the hierarchy in this qualifying match problem is small, with only eight zones at the bottom level, and as a result, the benefits of incorporating hierarchy information may not justify the additional complexity it introduces into the forecasting system. As was mentioned earlier, Jingrui Xie initially explored the hierarchy when developing the *Rain benchmark*, but eventually excluded it for this reason.

The investigations and innovations in the open-data track were rather limited. Some teams submitted the same forecasts to both tracks, while others made minor modifications to the methods used in the defined-data track when participating in the open-data track. Four of the teams listed in Table 6 participated in both tracks. Teams *Simple_but_good*, *4C*, and *Jingrui Xie* submitted the same forecasts for both tracks. Team *GeertScholma* included the solar data from the NOAA weather database when developing forecasts for the open-data track, which helped to improve their forecasts by about 5% in several rounds. Team *GeertScholma* also reported that some programming errors led to poor scores in earlier rounds. As was shown in Table 3, the *Rain benchmark* ranked No. 3 in the open-data track, but No. 9 in the defined-data track, indicating that the open-data track was not as competitive as the defined-data track.

Table 5
Peaks (MW) of the load series used in the competition series.

		Min (MW)	Max (MW)
GEFCom2012	Z1 – Z20	1	540
	Z21	3,280	3,280
GEFCom2014		316	316
GEFCom2017-Q	Top	27,622	27,622
	Middle	1,036	13,054
	Bottom	1,036	7,367
GEFCom2017-F	Top	2,185	2,185
	Aggregate	318	1,878
	Middle	1	623
	Bottom	0	207

Overall, other than a limited use of the hierarchy by some teams, most of the forecasting methodologies and techniques used in this qualifying match were more or less similar to those of the previous global energy forecasting competitions (Hong et al., 2016, 2014) and the recent literature on point and probabilistic load forecasting (Hong & Fan, 2016; Liu et al., 2017; Wang et al., 2016; Xie & Hong, 2018a). This was not surprising to us organizers, because we our aim with this qualifying match was to help the contestants to become familiar with the load forecasting literature and to get used to the pace of the competition.

3. Final match

The final match concerned the hierarchical probabilistic load forecasting problem at a larger scale than the qualifying match problem. We asked the contestants to provide probabilistic forecasts for 161 delivery point meters, a subset of the 183 meters that we provided. Twelve teams from the qualifying match were invited to the final match, and nine completed it. The final match data were released on April 18, 2017, and the submission was due on May 31, 2017. This section covers the data used in the final match and summarizes the results and the finalists' methods.

3.1. Data description

We provided the contestants with seven years of hourly load, temperature, and relative humidity data,

Table 6
Summary of the qualifying match methodologies.

Team	Ranking (D/O)	Data cleansing	Feature engineering			Temperature scenarios	Modeling techniques	Forecast combine	Hierarchy information
			Load	Temperature	Calendar				
It Can Be Done	2 / -	No/No discussion	No (transformation)	Yes	Yes	No	Quantile regression, quantile regression forest, and gradient boosting machine	Yes	No
Orbuculum	3 / -	Yes	Yes (transformation)	Yes	Yes	No	Gradient boosting machine, quantile random forest, and neural network	Yes	No
dmlab	4 / -	No/No discussion	No	No	Yes	Normal distribution	XGBoost, quantile gradient boosted regression trees, and generalized additive models	Yes	No
Simple but good	5 / 2	Yes	No	Yes	Yes	No	Multiple linear regression and quantile regression	Yes	No
Black Analytics	6 / -	No/No discussion	No	Yes	Yes	Shifted-date ($k = 13$ or 14, $n = 7$)	Multiple linear regression	Yes	No
Cassandra	8 / -	Yes	Yes	No	Yes	No	One-shot quantile forecasts, average monthly spread, automated time series forecasting in R, temporary hierarchical forecasting in R (thief), and hierarchical time series forecasting in R (hts)	Yes	Yes
Rain benchmark	9 / 3	Yes	No	Yes	Yes	Shifted-date (k varies, $n = 6$)	Multiple linear regression	No	No
QUINKAN	11 / -	Yes	No	Yes	Yes	No	Quantile regression and generalized additive models	Yes	Yes
GeertScholma	12 / 1	No/No discussion	No	Yes	Yes	Shifted-date ($k = 11$, $n = 13$)	Multiple linear regression and autoregression	Yes	No
The boosts are made forecasting	13 / -	Yes	No	No	Yes	Shifted-date ($k = 11$, $n = 4$)	Gradient boosting, hierarchical forecasting, residual re-sampling	No	Yes
4C	17 / 8	No/No discussion	No	Yes	Yes	Shifted-date ($k = 11$, n varies)	Neural network	No	No
UC3M	22 / -	Yes	Yes	Yes	Yes	No	Univariate seasonal ARIMA, seasonal factor models, and regression techniques	Yes	Yes
Vanilla benchmark	- / -	No	No	Yes	Yes	Shifted-date ($k = 11$, $n = 4$)	Multiple linear regression	No	No

together with hierarchy information that maps the meter IDs to the three higher levels. The weather data included temperature and relative humidity values from 28 anonymous weather stations. The locations of the delivery point and the weather stations were not provided. These real-world datasets may contain various data quality issues that load forecasters encounter on a daily basis.

Of the 183 delivery point meters provided in the historical period, 161 are active in the forecasting period, and we asked the contestants to provide probabilistic forecasts for these 161 meters. Given that the use of hierarchical information in the qualifying match had been very limited, we did not ask the contestants to provide probabilistic forecasts at higher levels. The submission was one-year-ahead probabilistic forecasts for the 161 delivery points in five quantiles, namely the 10th, 25th, 50th, 75th, and 90th percentiles.

The average of the pinball loss across the five required percentiles throughout the forecast period was calculated to serve as the quantile score of the probabilistic forecast for each meter. A team receives a ranking for each meter, and the average of the 161 rankings was used as the final score for each team. Teams with lower scores were ranked higher in the final leaderboard.

3.2. Summary of the methods and results

One of the nine teams in the final match did not provide enough details regarding their methodology. Table 7 summarizes the methodologies of the other eight teams from seven aspects, namely data cleansing, weather station selection, feature engineering, temperature scenario generation, modeling techniques, forecast combination, and the use of hierarchy information or some other form of meter grouping. Overall, most teams used very similar methods to the qualifying match.

Five of the eight teams selected in the final match conducted data cleansing prior to the modeling process. Specifically, the data cleansing efforts included removing spikes or zeros from the data, imputing missing values, correcting recording errors, trimming early history due to structure changes, etc. Of these five teams, team *QUINKAN* used pre-defined rules and inspected the data manually; team *GeertScholma* implemented some pre-defined rules while relying on the model fit residuals; team *Cassandra* plotted a heat map to help visualize the data; and teams *Simple_but_good* and *UC3M* used pre-defined rules.

Unlike the qualifying match, the data for the final match came from an unknown U.S. utility. Thus, there was no information on the association between the 161 meters and the 28 weather stations. Two teams conducted weather station selection: team *QUINKAN* selected the weather station(s) for different meters by evaluating the performance of a GBM, and team *GeertScholma* used the goodness-of-fit of a polynomial temperature model to select one weather station for each meter. The other teams did not assign separate weather stations to individual meters.

Many teams utilized the same feature engineering and modeling techniques for the final match as for the qualifying match. This may be due in part to the desirable performance of the model in the qualifying match. Additional models were developed for the load series with characteristics that differed from those in the qualifying match, such as retired meters and load series that did not respond to weather conditions.

Similarly to the qualifying match, the hierarchy information was not considered by most of the teams. Only three of the nine teams utilized the hierarchy information or grouped the meters by some other data-driven technique.

4. Discussion

This section will discuss our reflections from organizing GEFCom2017, along with our vision for future competitions.

4.1. Winner prediction

Due to the lag between the time when the contestants submitted their forecasts and the time when ISONE published the settled load data, the organizers had to predict the possible winners and invite them to the final match prior to all of the actual load values being available. In GEFCom2017, this prediction was based on the interim score, with a few buffer seats to include a few teams that were close to the qualifying bar but did not pass it. Team *QUINKAN*, which ranked No. 1 in the final match, was one of the two teams that was invited in the second batch due to their improved ranking. In fact, they did not pass the qualifying bar once ISONE had published all of the load values for the full six rounds.

The prediction of competition outcomes is nontrivial. While we must accept that teams may have varying performances between qualifying and final matches, we hope to include teams that can deliver outstanding performances in the final match. Running the first bi-level competition in the series, we struck lucky this time by inviting a few teams that did not pass the qualifying bar to the final competition. This could be a good lesson for future competition organizers.

4.2. Contestant retention

Fig. 3 presents information on contestant retention from one round to the next for both the defined-data and open-data tracks. The left vertical axis shows the number of teams. The blue bar shows the number of teams participating in both the current round and the preceding one. The green bar shows the number of teams who participated in the current round but not the preceding one, though they may or may not have participated in the earlier round(s). The red bar shows the number of teams from the preceding round that did not join this round. The right vertical axis shows the drop rate from round to round.

Several factors may have contributed to the high drop rate during Round 2. First, the due date of Round 2 was right before New Year's Day (see Table 2), which is a holiday season in many countries. Second, the teams were only required to join four out of the six rounds. Considering that the next round (i.e., Round 3) required the same forecasting period as Round 2, some teams may have skipped Round 2 and participated in Round 3 instead. This may have been the reason for the high drop rate in Round 4 as well, as it required the same forecasting period as Round 5.

4.3. Model evolution

As was shown earlier in Table 2, the 2nd and 3rd rounds of the qualifying match asked for forecasts for the same forecasting period, as did the 4th and 5th rounds. We designed the competition in this way so as to offer the contestants the opportunity to modify their forecasts as they improved their models. We were expecting more changes in the open-data track than in the defined-data track, mainly because some additional data becomes

Table 7
Summary of final match methodologies.

Team	Ranking	Data cleansing	Weather station selection	Feature engineering			Temperature scenarios	Modeling techniques	Forecast combine	Meter grouping / hierarchy information
				Load	Weather	Calendar				
QUINKAN	1	Yes	Yes	No	Yes	Yes	No	Quantile regression and generalized additive model	No	Yes
dmlab	3	No / No discussion	No	No	Yes	Yes	Shifted-date (no discussions on the value of k and n)	Quantile gradient boosted regression trees		No
Orbuculum	4	Yes	No	Yes (transformation)	No	Yes	No	Gradient boosting machine, quantile random forest, and naive forecast	Yes	No
GeertScholma	5	Yes	Yes	No	Yes	Yes	Shifted-date ($k = 7$, $n = 2$)	Multiple linear regression and autoregression	Yes	No
Cassandra	6	Yes	No	No	No	Yes	No	Neural network-based quantile forecasting model, and time series models	No	Yes
It Can Be Done	7	No / No discussion	No	No (transformation)	No	Yes		Gradient boosting machine, quantile random forest	Yes	No
Simple_but_good	8	Yes	No	No	No	No	No	Time series models	No	No
UC3M	9	Yes	No	No	Yes	Yes	No	Linear regression model, factorial model, profiling	Yes	Yes

available in real time, such as preliminary load data and weather forecasts.

Table 8 shows whether the teams that participated in both rounds of the same forecasting period modified their forecasts, and the outcomes of such modifications. For instance, 52 teams participated in both Round 2 and Round 3 of the defined-data track, of which 27 submitted exactly the same forecasts. Of the 25 teams that modified their forecasts, 17 had better forecasts in Round 3 than in Round 2. Overall, the percentage of teams that changed their forecasts was lower in the open-data track than in the defined-data track. This may be due in part to the limited time available for the teams to build significantly better models. On the other hand, it could indicate the need to use real-time preliminary data in short- and medium-term load forecasting.

4.4. Research opportunities

When designing the competition, we had a few challenges in mind for testing the contestants. While the contestants have touched most of them with varying degrees of emphasis, we think that the competition and the associated datasets can help to stimulate the research in these areas further.

Load patterns are typically smoother and easier to capture at high voltage levels than at medium and low voltage levels. Several issues have to be addressed in order to forecast accurately at lower levels, such as the proper selection of weather stations for the individual zones, the detection and fixing of anomalies, and accounting for load transfers among distribution feeders. Making use of the hierarchies is also crucial to improving the forecast accuracy. Furthermore, probabilistic load forecasting in the presence of hierarchical information is an emerging topic.

In addition to the challenges that we put into the competition problem intentionally, Team *QUINKAN* brought up another important issue at the 2017 International Symposium on Energy Analytics when they were presenting their winning methodology, which included data cleansing for weather data. Although the weather data that we released to GEFCom2017 was of high quality, Team *QUINKAN* still identified numerous data issues that could potentially harm the load forecasts. Weather data cleansing, a subject rarely touched upon in the load forecasting literature, would be another useful future research direction.

From a competition organization perspective, we find it important to have a robust rating and ranking method, or a pool of robust methods. Langville and Meyer (2013) discussed a wide range of rating and ranking methods for different kinds of competitions. Since forecasting competitions are great vehicles for driving innovation and advancement in forecasting research, an investigation of rating and ranking methods would be a valuable contribution to the forecasting literature. Furthermore, predicting the outcomes of forecasting competitions is another topic that deserves investigation.

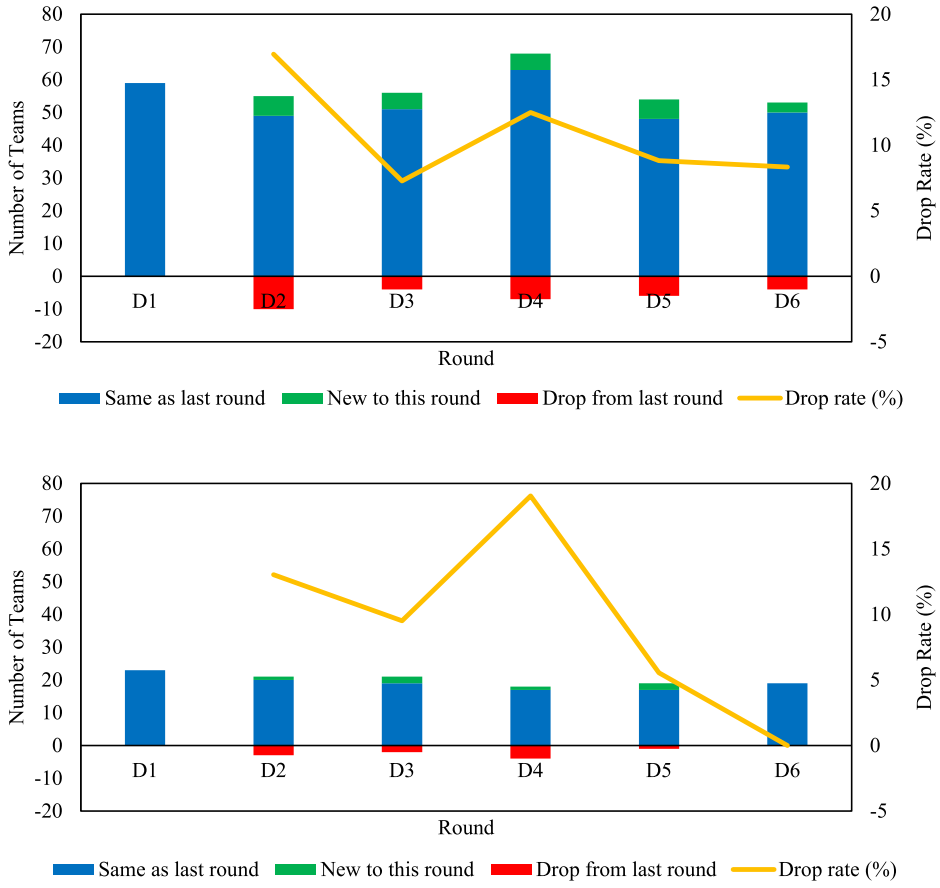

Fig. 3. Contestant retention (Upper panel: Defined-data track; Lower panel: Open-data track).

Table 8
How the forecast method evolved.

		Total # of teams responding to both rounds	Same	Better	Worse
Defined-data track	R2–R3	52	27	17	8
	R4–R5	49	36	5	8
Open-data track	R2–R3	19	9	3	7
	R4–R5	17	10	3	4

4.5. Future competitions

Competition is an effective way of generating interest from the broad community, recognizing useful forecasting models, and inspiring new research ideas. GEFCom2017 will not be the last one in the series. We have been actively seeking novel ideas and the associated datasets for future competitions. Currently, we are considering a few problems, such as ex ante point load forecasting, load forecasting at the household level, net load forecasting with a consideration of rooftop solar generation, and power distribution outage forecasting.

In addition to attracting thousands of data scientists across the globe to tackle their energy forecasting problems, the three global energy forecasting competitions have also inspired many others to organize their own public competitions. For instance, a U.K. gentailer RWE npower has been organizing gas and electricity load forecasting competitions every year as a way of recruiting student interns; RTE, the French transmission system operator, has organized two forecasting challenges in 2017 and 2018 respectively; Tokyo Electric Power Company organized a competition in 2017 in order to identify the best forecasting technologies that they may be able to use; and Tao Hong from the University of North Carolina at Charlotte has opened his in-class competitions (known as BigDEAL Forecasting Competitions) to the public multiple times so that his students can compete with and learn from the other invited contestants. We sincerely hope that more and more researchers and practitioners can contribute to these energy forecasting competitions.

Acknowledgments

We would like to acknowledge the reviews provided by Geert Scholma, Juan Quintana, Slawek Smyl, Grace Hua, Ilias Dimoulkas, and Julian De Hoog, whose comments helped to improve the accuracy and clarity of this paper further.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2019.02.006.

References

Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2017a). Coherent probabilistic forecasts for hierarchical time series. In *Proceedings of the 34th international conference on machine learning, Vol. 70* (pp. 3348–3357).

Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2017b). Hierarchical probabilistic forecasting of electricity demand with smart meter data. Retrieved from http://souhaib-bentaieb.com/pdf/jasa_probhts.pdf.

Gamakumara, P., Panagiotelis, A., Athanasopoulos, G., & Hyndman, R. J. (2018). Probabilistic forecasts in hierarchical time series. Retrieved from https://www.monash.edu/business/ebs/research/publications/ebs/wp11-2018.pdf.

Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: a tutorial review. *International Journal of Forecasting*, *32*(3), 1–32.

Hong, T., Pierre, Pinson, Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913.

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

Hyndman, R. J., & Athanasopoulos, G. (2014). Optimally reconciling forecasts in a hierarchy. *International Journal of Applied Forecasting*, *35*, 42–48.

Langville, A. N., & Meyer, C. D. (2013). *Who's #1?: The science of rating and ranking*. Princeton, New Jersey: Princeton University Press.

Liu, B., Nowotarski, J., Hong, T., & Weron, R. (2017). Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid*, *8*(2), 730–737.

PJM Interconnection (2016). Load forecasting model whitepaper. Retrieved October 8, 2018, from https://www.pjm.com/~/media/library/reports-notices/load-forecast/2016-load-forecast-whitepaper.ashx.

Wang, P., Liu, B., & Hong, T. (2016). Electric load forecasting with recency effect: A big data approach. *International Journal of Forecasting*, *32*(3), 585–597.

Xie, J., Chen, Y., Hong, T., & Laing, T. D. (2018). Relative humidity for load forecasting models. *IEEE Transactions on Smart Grid*, *9*(1), 191–198.

Xie, J., & Hong, T. (2015). GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation. *International Journal of Forecasting*, *32*(3), 1012–1016.

Xie, J., & Hong, T. (2018a). Temperature scenario generation for probabilistic load forecasting. *IEEE Transactions on Smart Grid*, *9*(3), 1680–1687.

Xie, J., & Hong, T. (2018b). Variable selection methods for probabilistic load forecasting: empirical evidence from seven states of the United States. *IEEE Transactions on Smart Grid*, *9*(6), 6039–6046.