WILEY

CASE STUDY

Predicting the magnitude and timing of peak electricity demand: A competition case study

Daniel L. Donaldson[1] | **Jethro Browell[2]** | **Ciaran Gilbert[3]**

[1]School of Engineering, University of Birmingham, Birmingham, UK

[2]School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

[3]Independent Researcher, Glasgow, UK

Correspondence

Daniel L. Donaldson.
Email: d.l.donaldson@bham.ac.uk

Funding information

University of Birmingham

Abstract

As weather dependence of the electricity network grows, there is an increasing need to predict the time at which the network peak load will occur. Improving forecasts of peak hour can lead to more accurate scheduling of generation as well as the ability to use flexibility to improve system utilisation or defer capital investment. While there are extensive benchmark models for forecasting electricity demand, their efficacy at forecasting the time or shape of the peak remains to be seen. Global forecasting competitions provide a unique opportunity to compare multiple methodologies under common performance criteria and incentives. The methodology and results are detailed from the Big Data and Energy Analytics Laboratory Challenge 2022 used by the team 'peaky-finders' and investigates the suitability of using hourly methods to forecast daily peak magnitude, time, and shape. The resulting approach provides a reproducible ensemble benchmark against which to evaluate more complex methods. Results indicate that simple regression techniques can perform well and outperform more complicated methods during seasons with low hourly variability, however ensemble methods show higher accuracy overall. The results also highlight the significant impact of extreme weather on forecast accuracy, demonstrating the importance of forecasting processes that are resilient to extreme weather.

KEYWORDS

artificial intelligence and data analytics, load forecasting

1 | INTRODUCTION

Safe operation of the electric grid requires a continuous balance of supply and demand. In order to achieve this in a reliable, economic, and clean fashion, load forecasts must be produced on a variety of time horizons. Forecasts one to 10 years ahead (long-term) enable generation and transmission capacity planning whereas forecasts minutes to weeks ahead (short-term) are required for operation of the grid and generation scheduling [1]. A bibliometric search via Web of Science for prior work in the area of short-term load forecasting (STLF) using the query ALL = (short term AND ("load forecasting") AND (energy OR electric*)) returned nearly 3000 papers with the growth in published literature shown in Figure 1.

Despite the significant number of papers, inconsistencies exist in: (1) the robustness of data used for analysis and validation, (2) choice of evaluation metrics and (3) models used for comparison [2]. As a result, despite the numerous publications, it is difficult to discern the current 'state-of-the-art'. That being said there are some clear principles which are evident: (1) Ensemble models often outperform their individual components [3]; and (2) Careful selection of features is key, including the choice of weather data.

Forecasting competitions provide an effective means overcome this challenge by assessing the performance of a variety of methods and approaches on the same data, and aim to provide a fair comparison of their efficacy and benefit [4]. Even after the competition, future researchers can use the competition data to compare their new methods against the competition benchmarks, further supporting advancement of new methods and open science. Examples include the 2001 EUNITE competition [5], and the Global Energy Forecasting Competition

(GEFCOM) series held over 2012, 2014, and 2017 in which hundreds of competitors from industry and academia have tackled emerging energy industry forecasting tasks [6]. Competitions have also been held with industry including Western Power Distribution [7], the IEEE Computational Intelligence Society in partnership with E.ON [8], and the Day-Ahead Electricity Demand Forecasting Competition supported by BluWave-ai [3]. However, while many of these competitions required the development of hourly forecasts of which one of these hours was the peak, they were not expressly focused on peak demand magnitude, timing and shape.

For purposes of capacity or adequacy planning, the magnitude of the peak has been of primary importance [9]. This was sufficient as most generation was dispatchable, enabling operators to schedule in such a fashion to account for the uncertainty in the demand due to socio-economic and weather related factors. This has worked well and the annual peak demand often follows a consistent pattern each year, resulting in a fairly consistent peak hour. Seasonal peak demand has also been of interest to inform customer billing, such as the charges for industrial and commercial customers. One example of this is the use of 'triads'—'the three half-hour settlement periods of highest demand on the GB electricity transmission system between November and February (inclusive) each year, separated by at least 10 clear days'—by the GB transmission system operator to determine charges for customers based on the their peak usage [10]. However new technologies and use cases are requiring the production of forecasts that are more granular both spatially and temporally as seen in Figure 2 resulting in the
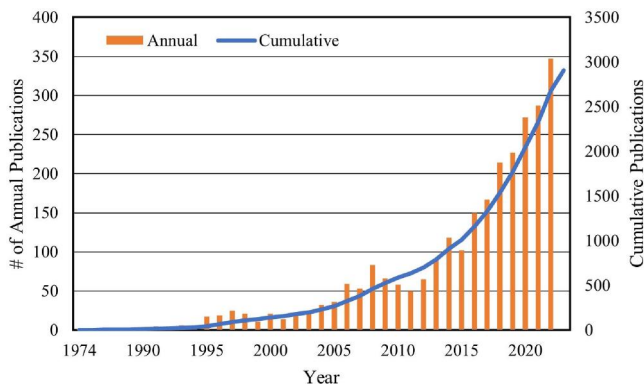

FIGURE 1 Growth in short-term load forecasting literature.
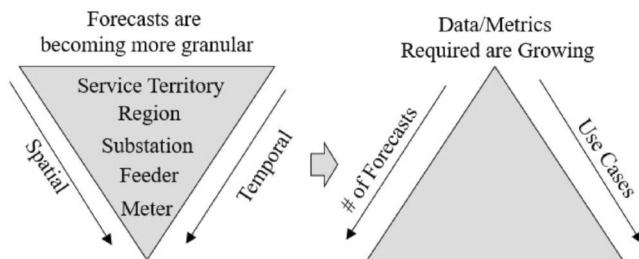

FIGURE 2 Forecasts are becoming more granular, requiring research and development of new metrics and methods.

need for evaluation of new methods and approaches. For example, knowing when the peak occurs can support more effective decision-making for energy storage dispatch [11, 12] as was the focus of the aforementioned WPD POD data challenge.

For utilities, growth in solar PV generation can shift the peak hour later in the day. This has required updates to tariffs, and re-education of consumers to encourage use of energy during times when it is more plentiful [13]. Examples of the significance of this shift can be seen at a regional level in California where historical data shows that the annual peak time for the California Independent System Operator has shifted from roughly 15:30 (1998–2000) to 17:00 (2020–2022) [14]. There have also been opportunities created for generation owners and distributed energy resource (DER) service providers to predict the day and time at which the peak will occur, enabling them to dispatch resources to maximise economic gains [11].

Demand response and critical peak pricing programs introduced some value from knowing the time at which the peak was expected, as this information enabled more effective programs [15]. However, efforts to decarbonise the energy sector are elevating the importance of knowing the time at which the peak occurs to support decision-making for battery charging and discharging. While there are established forecasting methods and metrics for the prediction of electricity demand, the evaluation of these methods for peak hour prediction remains nascent. Research in this area includes using weather probability to predict the top k peak days [15], and ensemble machine learning to predict peak days and the peak hour on those days [16]. Authors in Ref. [17] have also introduced a toolkit designed to support comparison of such methods. Recently it was demonstrated in Ref. [18] that fusion of daily peak and half-hourly forecasts could improve the accuracy of demand forecasts, offering improvements of over 10% during peak hours for forecasts from the individual household to distribution substation level. However, these publications all used differing datasets, limiting the ability to directly compare the relative performance of methods. Furthermore, across existing works on peak load forecasting, there is a lack of consistency in the metric used for evaluation, and the horizon that is being evaluated. For example, error metrics for evaluation differ across works and include precision and/or recall [15–17], accuracy [16], mean squared error [11] or continuous ranked probability score [18]. The toolkit presented in Ref. [17] provides some first steps towards a means to unify evaluation and enable comparison, but is mostly focused on monthly and daily peak prediction rather than the magnitude, hour and daily shape. While it does include one model for peak hour forecasting (Long Short Term Memory—LSTM) the implementation uses demand from the prior 7 days to predict day ahead peak hour which may not always be available.

To address this gap in the established literature, a forecasting competition, the Big Data and Energy Analytics Laboratory (BigDEAL) Challenge 2022 (BDC22) was held from Nov-–Dec 2022 to provide a platform through which to evaluate methods for forecasting the magnitude, time, and shape of electricity demand. While forecasting competitions can provide benefit to the participants, the publication of methods from the

competitors provides a mechanism via which to enable the wider field to reproduce and evaluate future methods. To that end, this paper reflects the following contributions: (1) Presentation of the methodology used by our team 'peaky-finders' which finished fourth overall, and third in the track focused on the magnitude of peak demand. (2) Discussion of lessons learnt as a result of the competition which can further benefit practitioners and researchers in the field.

The economic value of improvements to the accuracy of peak timing may lead some end-users to develop bespoke forecasting methods for this application. However, as short term forecasting is often done already on an hourly basis, the time of the peak is already inherently predicted without requiring development of an additional method. Therefore, this paper proposes a methodology to extract forecasts of the peak hour and shape from traditional hourly forecasts that can provide a benchmark against which to measure the benefit of more complex methods. Lessons learnt from the forecasting competition that point to several areas for future research to further enhance forecasts of the peak hour.

The rest of the paper is structured as follows: Sections 2 and 3 present the competition's data and methodology. The modelling results and discussion are given in Sections 4 and 5. Finally Section 6 presents the conclusions and areas for future research.

2 | PEAK FORECASTING COMPETITION

BDC2022 had a theme of peak load forecasting and was organised by academics from UNC Charlotte. The competition began with a qualifying match from 31 October to 10 November 2022; 121 contestants from 27 countries formed 78 teams to take part [19]. The match was divided into three tracks whereby teams were tasked to provide ex-post forecasts of hourly loads, daily peak magnitude, and timing of daily peaks. Our team, 'peaky-finders' finished above 'Shreyashi's Recency Benchmark' in all three tracks, and therefore was invited to compete in the final match alongside 13 other teams and three individuals.

2.1 | Final match structure

For the Final match, teams were tasked to produce ex-ante daily peak load forecasts for six rounds for three U.S. local distribution companies (LDC). Historical data over the 3 year period from 2015 to 2017 was provided for training alongside historical actual temperature from six weather stations. In contrast to the qualifying match (and much of the forecasting literature) the competition used day-ahead temperature forecasts for each of the rounds making the results more relevant to real-world application. Actual data from the prior round and temperature forecasts for the next round were released every 3–5 days. The time-frame and data for each of the six rounds is shown in Table 1.

TABLE 1 Structure of BDC 2022 final match.

	Months covered	Date provided	Submission deadline
Round 1	Jan–Feb	16-Nov-22	20-Nov-22
Round 2	Mar–May	21-Nov-22	23-Nov-22
Round 3	Jun–Jul	24-Nov-22	27-Nov-22
Round 4	Aug	28-Nov-22	30-Nov-22
Round 5	Sep–Oct	01-Dec-22	04-Dec-22
Round 6	Nov–Dec	05-Dec-22	07-Dec-22

2.2 | Performance metrics

Past literature and competition discussions have demonstrated the importance of metrics on forecast evaluation, and that the method used can highly depend on the desired metric. As there is not an established set of performance metrics for peak time forecasting, the organisers of the competition selected the following metrics.

2.2.1 | Track 1—Magnitude

The error metric for this track (M) is calculated as the Mean Absolute Percentage Error (MAPE) as given in Equation (1).

$$M = \frac{1}{n}\sum_{d=1}^{n}|\frac{L_d^{pk} - \hat{L}_d^{pk}}{L_d^{pk}}| \tag{1}$$

where L_d^{pk} and $\hat{L}_d^{pk}$ are the actual and estimated peak load on a given day d; and n represents the number of days in the evaluation period.

2.2.2 | Track 2—Timing

The error metric for this track (T) is the weighted sum of absolute errors with cap calculated as

$$T = \sum_{d=1}^{n} w_d \Delta_t \tag{2}$$

where $\Delta_t = |t_d^{pk} - \hat{t}_d^{pk}|$ and the weights w_d are

$$w_d = \begin{cases} \Delta_t, & \text{if } \Delta_t \leq 1 \\ 2, & \text{if } 2 \leq \Delta_t \leq 4. \\ 10, & \text{otherwise} \end{cases} \tag{3}$$

2.2.3 | Track 3—Shape

First the 24 hourly load forecasts of each day will be normalised by the peak forecast of that day to obtain the shape of that day. The same was done for the actual load. The sum of

absolute errors is then calculated during the 5-h peak period (peak hour $+/- 2$ h) of every day

$$S = \sum_{d=1}^{n} \sum_{k=-2}^{2} |s_{t_d^{pk}+k,d} - \hat{s}_{t_d^{pk}+k,d}| \quad (4)$$

where s and $\hat{s}$ represent the normalised actual and estimated load at a point in time divided by the actual peak load of that day, and k represents the 5 h surrounding the peak.

Data cleansing was not performed, however one notable anomaly in the historical data is discussed in detail in Section 4.1. The historical data for the case studies is illustrated Figure 3a,b, showing the load time series and load-temperature relationship, respectively.

3 | METHODOLOGY

This section presents the approach used by the peaky-finders in BDC2022 along with the underlying theoretical foundation with an overview presented in Figure 4.

3.1 | Explanatory variable creation

The explanatory variables we used throughout the competition can be grouped into four areas:
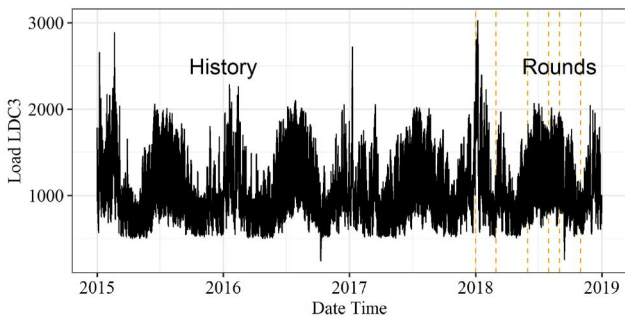
(a) Hourly load for one of the three Local Distribution Companies
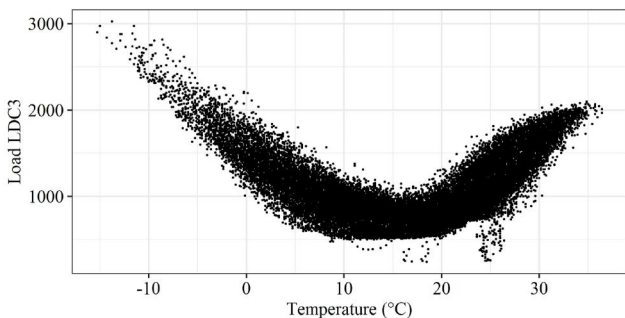
(b) Load vs. Temperature for one of the three Local Distribution Companies

FIGURE 3 Plots illustrating some of the main properties of load time series, such as annual seasonality and temperature response.

1. Linear Trend—to capture any longer term change in the demand
2. Calendar Effects—Hour, Day of the Week, and Month were all used as categorical variables to reflect seasonal differences in demand
3. Holiday Variables—Indication of whether a day was a holiday or not as holidays can influence the demand. All holidays were treated uniformly.
4. Temperature terms including lagged, rolling average and smoothed temperature

In practice, utilities can evaluate meteorological forecasts and build models to correct any systematic biases that may be present; however, only historical weather data was provided for the competition for model development. Therefore it was not possible to evaluate the accuracy of the weather forecasts being provided ahead of the first round. To account for the possible error in the forecast, three temperature scenarios were produced and three corresponding demand forecasts were generated using these scenarios and averaged to form a final demand forecast. When creating temperature scenarios for projecting future demand, one method to model the most likely weather conditions is a shifted date approach as described in Ref. [20] whereby the temperature values are shifted forward and backwards by a certain number of days to generate a range of temperature scenarios to use for forecasting. This approach was applied in Ref. [21] when creating forecasts for the sizing of non-wires alternatives and other temperature resampling approaches have also been proposed in Refs. [22, 23] to generate probabilistic forecasts.

While this competition did not require a probabilistic output, given the use of forecast temperature values, a shifted time approach based on Ref. [20] was used to account for uncertainty in the actual temperature. A new set of features was created by shifting the forecast temperature forward and backward by 1 h resulting in three distinct feature matrices for each hour: $F(T_{t-1})$, $F(T_t)$, and $F(T_{t+1})$. A forecast is produced using each set of features and then the resulting three forecasts are averaged to produce a single output. These time shifted versions will be referred to by appending '−T' to the corresponding forecasting model name.

3.2 | Forecasting models

Two main model families were used to produce the forecasts for our team. First, a Multiple Linear Regression (MLR) model
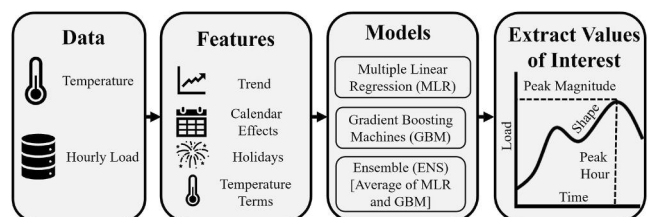

FIGURE 4 Overview of the approach.

based on the vanilla benchmark with recency used in GEF-COM 2017 [6, 24] and BFCom2018 [25]. The logarithm of the load (L) at time t is given by

$$
\begin{aligned}
\log(L_t) = \ &\beta_0 + \beta_1 D \times H + \beta_2 M + \beta_3 M \times T_t \\
&+ \beta_4 M \times T_t^2 + \beta_5 M \times T_t^3 + \beta_6 H \times T_t \\
&+ \beta_7 H \times T_t^2 + \beta_8 H \times T_t^3
\end{aligned} \quad (5)
$$

where D is day of the week; H is hour; M is month; and T is temperature. Additional variables from the list above were also incorporated as additional linear predictors, and the regression parameters are estimated via ordinary least squares.

Second, a Gradient Boosting Machine (GBM) was developed using a combination of the explanatory variables mentioned above and again the logarithm of the load was used as the target variable. Reference material is provided by Refs. [26, 27]. Prior research also demonstrates GBM to be effective for STLF [28, 29]. Here we use Gradient Boosting Trees, which comprise multiple decision trees that are combined to generate predictions [30]. For each decision tree, the input space is split into disjoint regions and each observation is assigned to the corresponding 'leaf' of the tree. Divisions between regions are selected to minimise some cost function measuring the fit between individual observations and a predictions computed based on the all observations assigned to a given leaf. This allows decision trees to capture non-linear relationships and interactions between input variables efficiently. New trees are added to improve the prediction of the model fit by minimising the negative gradient of the loss function from the model's previous iteration. The learning relationship is governed by several hyper-parameters, such as the number of trees (stopping criterion), the maximum number of splits in a single tree, and a learning rate that controls the contribution of each new tree to the model.

Finally, in addition to the individual MLR and GBM based models, a series of 'ensemble' models are produced by taking the average of the output of individual MLR and GBM models. These ensemble models are referred to using the abbreviation 'ENS'. Table 2 provides further detail of each of the ensemble models considered.

Other families of models may be considered for this task and have been widely reported in the literature. Classical time series methods were considered, for example, but the organisation of the competition into rounds prevented use of these methods due to lagged demand observations not being available. There is also large literature on deep learning for load forecasting, however, given the modest volume of training data available, weakness of these models in past forecasting competitions, and less experience from the team with these models, we did not consider applying them.

3.3 | Peak forecast production

For the competition, each of the above models was developed to produce an hourly forecast, before re-sampling the peak

TABLE 2 Description of individual and ensemble models evaluated.

Model	GBM included	MLR included	Temp. shift	Weight
GBM	Y	-	-	-
GBM-T	Y	-	Y	-
MLR	-	Y	-	-
MLR-T	-	Y	Y	-
MLR-L	-	Y	-	Load
MLR-W	-	Y	-	Pk
MLR-TL	-	Y	Y	Load
MLR-TW	-	Y	Y	Pk
ENS	Y	Y	-	-
ENS-L	Y	Y	-	Load
ENS-W	Y	Y	-	Pk
ENS-T	Y	Y	Y	-
ENS-TL	Y	Y	Y	Load
ENS-TW	Y	Y	Y	Pk

values from each day to generate the daily peak forecast. Given the focus on only the time around the peak hour, a matrix of weights was generated for the training data to penalise the model more significantly for errors made during peak time. Two weighting approaches were evaluated. The first was to weight the training data proportional to the overall daily load. The second was to use a Gaussian kernel $\mathcal{N}(x|\mu; \sigma^2)$ where x is the hour, μ is the peak hour, and σ^2 is set at 1.5 h. The value of σ^2 was selected heuristically for the competition but preliminary evaluations during the competition suggested larger values may not be as effective. These methods will be referred to as 'L' for the load based weighting and 'W' for Gaussian based weighting when describing the models. An example of each of these methods applied to a single day can be seen in Figure 5.

3.4 | Model selection across rounds

This section describes the models used for each round, the model selection process, and any rationale for changes. Given the reduction in training data in comparison to the qualifier, we used fewer explanatory variables to avoid over-fitting. Of the six overall weather stations, we used the MLR in Equation (5) to identify the weather stations that yielded the lowest MAPE for each LDC. Hyper-parameter tuning was performed for each model during the qualifying match; however, given the limited time for the final match optimal hyper-parameter selection was not repeated for each round and a fixed set of features was used. Further description of the model and rationale is given below for each round and Table 3 presents the models used for each round across all three tracks. Hyper-parameters are given in Section 4.
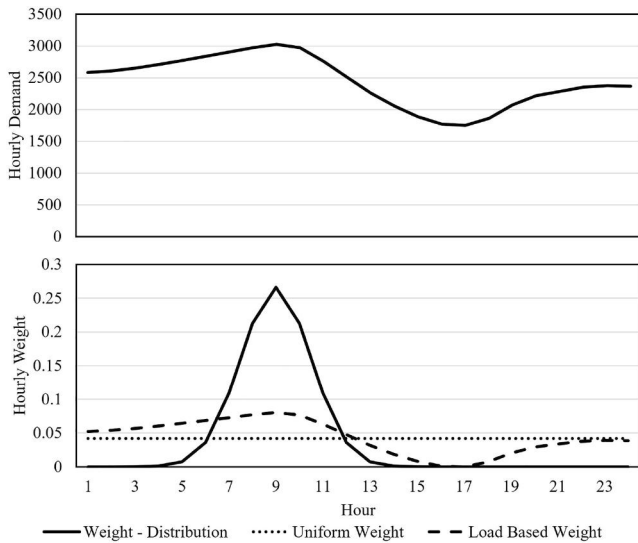
FIGURE 5 Comparison of the weighting approaches for an example day in January.

TABLE 3 Models used for each round of the competition.

	Models considered in ensemble		
	Track 1	Track 2	Track 3
Round 1	MLR-T, GBM-T	MLR-T, GBM-T	MLR-T, GBM-T
Round 2	MLR-T, GBM-T	MLR, GBM	MLR, GBM
Round 3	MLR-T, GBM-T	MLR-T, GBM-T	MLR-T, GBM-T
Round 4	MLR-T, GBM-T	MLR-T, GBM-T	MLR-T, GBM-T
Round 5	MLR-TW, GBM-T	MLR-TW, GBM-T	MLR-TW
Round 6	MLR-T, GBM-T	MLR-T, GBM-T	MLR-TL, GBM-T

3.4.1 | Rounds 1, 3, and 4

Hourly forecasts were generated using the GBM and MLR models then averaged to form a single forecast. The process was repeated two additional times with the temperature values shifted forward by 1 h and backward by 1 h. Then the average of these three forecasts (−1/0/+1) was taken to produce the final forecast for Track 3. The peak magnitude and hour of peak for Tracks 1 and 2, were selected from this hourly forecast.

3.4.2 | Round 2

For this round we hypothesised that shifting the temperature forward and backward might be contributing to increased error for Track 2 and 3 and excluded it for this round. However, we were not evaluating the forecast using the exact scoring metrics from Track 2 and 3, and unfortunately we would have performed better had we remained with the methodology from Round 1 (ENS-T) rather than ENS; 7% better in Track 2 and 0.7% better in Track 3 as shown in Tables 5 and 6.

3.4.3 | Round 5

In round 5 we introduced a kernel-based weighting scheme to give greater weight to more recent training samples. This was used only for the regression model, as the GBM implementation we used was not compatible with weights. We decided to exclude the GBM models for Track 3 as the performance of the MLR only model was better than the combined model with the GBM, and were being beat by the recency benchmark in Rounds 3 and 4. One other realisation was made after submitting our forecast for this round, that we had inadvertently excluded the August 2018 data from the training for the final model (as typically the prior round's data was used as a validation set for the model).

3.4.4 | Round 6

For this round the weather appeared more similar to Rounds 1 and 2. During these rounds, there were often a mix of peak hours in the morning and the evening, and therefore our weighting approach had to be altered to better reflect periods of high loading that may not be in proximity to the actual peak. We used this updated weight for the regression model for Track 3. However we only used it for Track 3 as when evaluating it using some of the prior rounds, we saw weighting yield much higher error in Track 2 and only marginal gains in Track 1.

4 | RESULTS

The approach was programed via R [31], making use of the following libraries/packages: dplyr [32], tibbletime [33], readxl [34], lubridate [35], Metrics [36], data.table [37], gbm [26], zoo [38], and tis [39]. Code is available at https://github.com/DLDonaldson/BigDEALChallenge2022_peakyfinders to reproduce this approach. This section explores and compares the models used throughout the competition on the full set of data.

For the MLR model results presented in this paper the following features were used as inputs into the MLR model given in Equation (5): Trend, Hour, Month, Weekday, Holiday, Weekday × Hour, third order polynomials of Temperature (T), lagged temperature (T_{lag1}, T_{lag2}, T_{lag3}, T_{lag6}), averaged temperature (T_{sma1}), smoothed temperature (T_{es995}, T_{es99}), and those polynomials crossed with Hour and Month. For the GBM model the following features were added: lagged temperature (T_{lag4}, T_{lag5}, T_{lag9}, T_{lag12}, T_{lag15}, T_{lag18}, T_{lag21}, T_{lag24}), averaged temperature (T_{sma2}, T_{sma3}) and the cross effects were removed, and only first order terms of each variable were included. The following hyperparameters were also set: distribution = 'Laplace', n.trees = 2000, n.minobsinnode = 300, interaction.depth = 3, bag.fraction = 0.8, shrinkage = 0.1, cv. folds = 5.

4.1 | Track 1

The aim of Track 1 was to produce forecasts of the daily peak load for each of the three LDCs. This was the highest performing track for the peaky-finders, placing 3rd overall and 1st in Round 3. The results for each round are shown in Table 4. The impact of ensembling, temperature shifting, and weighting each played a role in improving the overall forecast performance. Firstly, the individual GBM and MLR models were compared to the ensembled model. Results across each of the rounds indicate that ensembling approaches performed best overall, with the exception of the period from June to August. When considering the overall the shifted-time approach to handle the temperature forecast error, benefits were marginal prior to considering the weighting methods. However a similar benefit was not observed for the weighted approach, where a single temperature series yielded better performance overall.

During the competition, the weighting methods were not developed until later rounds, and therefore were not implemented until Round 5. Unfortunately, this was the only round in which using the weighted methods resulted in worse performance. As a result, the approach for round 6 reverted back to the standard model, where continued use would have instead reduced the MAPE from an average of 4.19%–4%. Overall across all rounds, the use of ENS-TW would have improved the team performance by 0.07%. Fundamentally, the use of the weighting factor gives a higher priority to the performance of the model in the hours surrounding the daily peak. Therefore, the performance of the model in other hours is sacrificed in order to achieve this objective and model

selection should be based on the bespoke error metric considered rather than the overall hourly error.

One significant anomaly in the results was observed in the performance during round 5. This round in particular included a significant disturbance in the load data caused by a hurricane that struck the area causing power outages, and therefore a reduction in load. Data for the load during this period and the corresponding disruption to the load-temperature relationship can be seen in Figures 6 and 7.

4.2 | Track 2

The aim of Track 2 was to predict the hour in which the peak demand occurs. The number of days are different across each
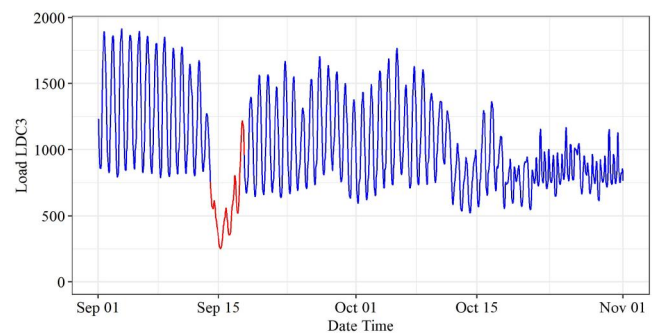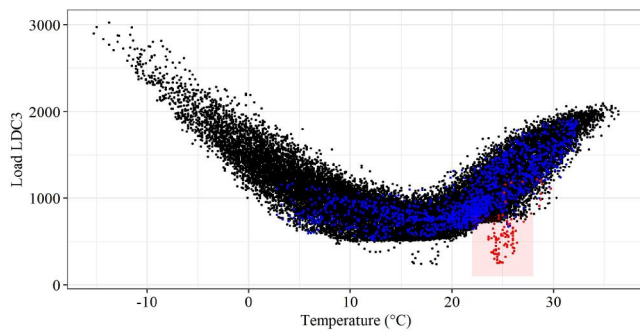

FIGURE 6 Unexpected event shown in the load for Round 5 (highlighted in red).

T A B L E 4 Results from Track 1 with the best performing model indicated in bold and underlined and the model used in the competition shaded in yellow.

| Model | Time period/Round | | | | | | | | Model information | | | |
	Train 2015–2017	Jan–Feb	Mar–May	Jun–Jul	Aug	Sep–Oct	Nov–Dec	Average	GBM included	MLR included	Temp. shift	Weight
GBM	2.61%	6.18%	4.96%	5.00%	3.34%	7.68%	4.40%	5.26%	Y	-	-	-
GBM-T	2.54%	6.10%	4.93%	4.89%	**3.30%**	7.66%	4.34%	5.21%	Y	-	Y	-
MLR	3.35%	5.41%	5.31%	4.91%	3.69%	8.20%	4.34%	5.31%	-	Y	-	-
MLR-T	3.33%	5.44%	5.34%	4.89%	3.71%	8.21%	4.32%	5.32%	-	Y	Y	-
MLR-L	3.11%	5.25%	5.22%	4.74%	3.53%	8.31%	4.14%	5.20%	-	Y	-	Load
MLR-W	3.10%	5.43%	5.24%	4.60%	3.51%	8.47%	4.49%	5.29%	-	Y	-	Pk
MLR-TL	3.08%	5.28%	5.25%	4.75%	3.57%	8.30%	4.12%	5.21%	-	Y	Y	Load
MLR-TW	3.16%	5.35%	5.19%	**4.59%**	3.49%	8.48%	4.42%	5.25%	-	Y	Y	Pk
ENS	2.77%	5.23%	5.01%	4.74%	3.48%	**7.65%**	4.20%	5.05%	Y	Y	-	-
ENS-L	2.66%	5.15%	5.00%	4.69%	3.42%	7.74%	4.13%	5.02%	Y	Y	-	Load
ENS-W	2.50%	**4.97%**	**4.92%**	4.68%	3.38%	7.79%	4.03%	**4.96%**	Y	Y	-	Pk
ENS-T	2.74%	5.18%	5.03%	4.72%	3.47%	7.66%	4.19%	5.04%	Y	Y	Y	-
ENS-TL	2.63%	5.10%	5.01%	4.67%	3.41%	7.74%	4.10%	5.01%	Y	Y	Y	Load
ENS-TW	**2.47%**	5.03%	4.93%	4.67%	3.36%	7.83%	**4.00%**	4.97%	Y	Y	Y	Pk

round, and therefore normalisation is performed by dividing the total error metric for the round by the number of days in that round. This enables comparison of the performance from one round to another. The results are provided in Table 5. Performance reveals the ability to predict the time within 1–2 h of the true time on average. Improvement was also observed from the shifting of temperature. Even after weighting, models including shifts in temperature were optimal for three of the six rounds.
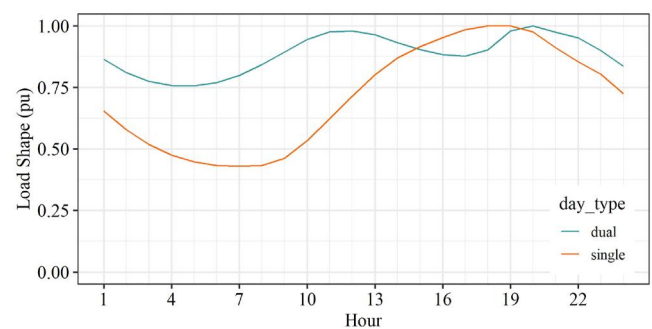
In this round, the decision-making regarding the model to use yielded improvements to forecast performance in one round (Round 5) and worse performance in another (Round 2). Overall this variation netted out and the overall average of the

ensembled methods demonstrates close performance (variation from 1.37 to 1.41). However, the error varies significantly across rounds due to some time of year having days with multiple periods (making it difficult to detect which will be the peak) and others with a clear single peak. This can be seen in Figure 8 where one day has a clear peak at hour 18, whereas another day has a peak at hour 20, and another hour with almost the same magnitude 8 h earlier.

Seasonal challenges result when trying to anticipate peak hour. This is a result of the underlying variability in peak hour. Similar effects may be observed with growth in solar photovoltaic generation. Understanding the link between this variability and the accuracy of forecasting approaches can provide insight into the reliability of peak related scheduling or energy optimisation efforts.

FIGURE 7 Same unexpected event. The data overall is shown in black, round 5 in blue, and the data for September 14–17 (The dates of a hurricane which hit the United States) shown in red.
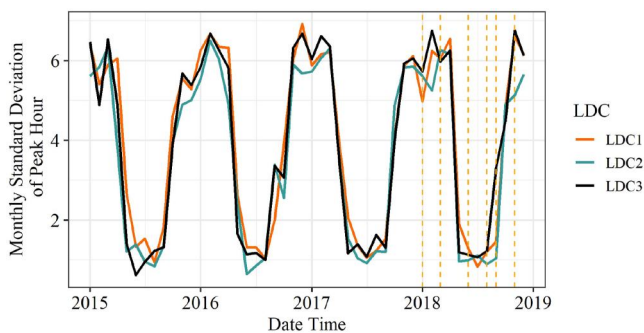

FIGURE 8 Differences in daily load shape across 2 days.

TABLE 5 Results from Track 2 with the best performing model indicated in bold and underlined and the model used in the competition shaded in yellow. The Error metric has been normalised by the number of days in each round to enable ease of comparison of model performance across rounds.

| Model | Time period/round | | | | | | | | Model information | | | |
	Train 2015–2017	Jan–Feb	Mar–May	Jun–Jul	Aug	Sep–Oct	Nov–Dec	Average	GBM included	MLR included	Temp. shift	Weight
GBM	1.08	1.45	1.25	1.14	1.52	1.83	2.17	1.56	Y	-	-	-
GBM-T	1.04	1.46	1.20	1.15	1.52	1.68	1.95	1.49	Y	-	Y	-
MLR	1.13	1.65	1.17	1.19	1.18	1.86	2.20	1.54	-	Y	-	-
MLR-T	1.11	1.51	1.09	1.15	__1.10__	1.85	2.19	1.48	-	Y	Y	-
MLR-L	1.09	1.73	1.12	1.09	1.15	1.84	2.07	1.50	-	Y	-	Load
MLR-W	1.72	2.45	1.43	__0.99__	1.22	2.01	2.75	1.81	-	Y	-	Pk
MLR-TL	1.05	1.69	1.14	1.09	1.24	1.87	2.13	1.53	-	Y	Y	Load
MLR-TW	1.71	2.11	1.41	1.01	1.27	1.83	2.75	1.73	-	Y	Y	Pk
ENS	0.99	1.18	1.16	1.04	1.27	1.84	2.00	1.41	Y	Y	-	-
ENS-L	__0.95__	__1.10__	1.11	1.03	1.24	1.82	1.91	__1.37__	Y	Y	-	Load
ENS-W	1.09	1.44	1.11	1.05	1.31	1.67	1.93	1.42	Y	Y	-	Pk
ENS-T	0.99	1.17	1.07	1.04	1.30	1.77	1.96	1.39	Y	Y	Y	-
ENS-TL	__0.95__	1.20	__0.98__	1.07	1.28	1.80	1.96	1.38	Y	Y	Y	Load
ENS-TW	1.07	1.43	1.08	1.03	1.28	__1.63__	__1.86__	1.38	Y	Y	Y	Pk

To track how much difference there is in the peak hour from one day to the next across the region of interest, the monthly standard deviation in peak hour is used. For the region of interest, there is a significant difference across seasons with the monthly standard deviation in peak hour ranging from roughly 1 hour in the summer months to six in winter months as shown in Figure 9. The performance across rounds also varies with the period from March to July resulting in the least error overall. In some rounds with little variation Rounds 3 and 4, simple MLR based methods yielded the highest accuracy. This was evidenced in the competition where the benchmark model outperformed many of the competitors in these rounds. In this Track, similar benefits of ensembling with

rounds 1, 2, and 4–6 showing improvements from the shifts in temperature.

4.3 | Track 3

In this round similar findings as prior Tracks were seen as far as the model performance. The overall results for each of the methods for Track 3 can be found in Table 6. First, overall, the ensemble methods outperformed the methods reliant on a single approach. Second, there was limited difference between the average performance of the ensemble models over the 6 rounds. However, what is not apparent is the significance of a shift in the daily shape metric. Finally, larger differences occur in the natural variation over the year than in between the different models. This track was the lowest performing round for the 'peaky-finders' and the only track in which the team scored below the recency benchmark.

5 | DISCUSSION

5.1 | Forecast evaluation metrics

With new areas of forecasting come new metrics. Where over time standards and methods of performance have been developed around acceptable 'MAPE' or other demand forecasting metrics, work remains to be done to link the forecast

FIGURE 9 Data for all three local distribution companies shows the peak hour varies much more significantly in winter than summer.

TABLE 6 Results from Track 3 with the best performing model indicated in bold and underlined and the model used in the competition shaded in yellow. The Error metric has been normalised by the number of days in each round to enable ease of comparison of model performance across rounds.

Model	Time period/round								Model information			
	Train 2015–2017	Jan–Feb	Mar–May	Jun–Jul	Aug	Sep–Oct	Nov–Dec	Average	GBM included	MLR included	Temp. shift	Weight
GBM	0.089	0.119	0.108	0.102	0.098	0.121	0.172	0.120	Y	-	-	-
GBM-T	0.084	0.115	0.103	0.102	0.095	0.112	0.168	0.116	Y	-	Y	-
MLR	0.085	0.127	0.103	__0.085__	__0.075__	0.102	0.173	0.111	-	Y	-	-
MLR-T	0.081	0.122	0.101	0.087	0.077	0.101	0.173	0.110	-	Y	Y	-
MLR-L	0.081	0.119	0.101	__0.085__	0.076	0.099	0.164	0.108	-	Y	-	Load
MLR-W	0.090	0.156	0.099	__0.085__	0.082	0.102	0.162	0.114	-	Y	-	Pk
MLR-TL	0.077	0.116	0.098	0.087	0.078	0.099	0.163	0.107	-	Y	Y	Load
MLR-TW	0.089	0.144	0.094	0.086	0.080	0.097	0.158	0.110	-	Y	Y	Pk
ENS	0.077	0.109	0.096	0.087	0.082	0.101	0.165	0.107	Y	Y	-	-
ENS-L	0.075	0.106	0.096	0.087	0.083	0.101	0.161	0.105	Y	Y	-	Load
ENS-W	0.074	0.110	0.092	0.086	0.083	0.100	0.154	0.104	Y	Y	-	Pk
ENS-T	0.075	0.107	0.095	0.090	0.082	0.099	0.165	0.106	Y	Y	Y	-
ENS-TL	0.074	__0.104__	0.095	0.090	0.082	0.099	0.160	0.105	Y	Y	Y	Load
ENS-TW	__0.072__	0.106	__0.091__	0.089	0.082	__0.096__	__0.153__	__0.103__	Y	Y	Y	Pk

metric to the business decisions that would be affected by metrics for peak hour forecasting. This is true not only for this competition, but for the larger field as the separation of performance metrics from the business decisions being informed can result in wasted model development efforts and suboptimal use of resources.

5.2 | Temperature error

One of the issues that our team sought to deal with was the impact of errors in temperature forecasts on load forecast accuracy. The results demonstrate that the approach provided incremental benefits across all three rounds. However, the computational cost was tripled due to the need to calculate three times the number of forecasts. With advance knowledge of the types and magnitude of the forecast temperature error, other approaches to make the forecasting process more robust to temperature errors would be of benefit. Due to the time limitations in the competition the shifts in hour were limited to ±1 however similar to the evaluation of the number of lags to use in Ref. [24] empirical determination of the most suitable shifts in temperature could be done. Another alternative approach could be to create a single synthetic temperature series with which to forecast rather than combining the forecast results. Furthermore, in practice the availability of ensemble forecasts may provide a better means to account for uncertainty in weather parameters.

5.3 | Competition decision making

One of the most significant challenges that arose during the competition was selecting the forecast model to use for each round. Upon the completion of the round and the receipt of the actual data, there was a relatively short time period in which the decision had to be made to continue with the existing approach, or try to further improve the forecasting performance. Such adjustments to the models used did not always improve forecast performance. For example, decisions on the forecasting methodology based judgmentally on prior rounds led to selection of sub-optimal models for Rounds 2 (lack of temperature correction), and 6 (lack of weighting in Tracks 1 and 2). Another example is that the choice to use the weighting approach in Round 5 for Track 1 resulted in reduced performance and therefore the weighting was not used for Track 1 in Round 6 (thought it would have provided benefit). Ultimately, forecaster judgement of when to implement a new model (evaluated based on historical data) will always carry some risk of reduced performance when exposed to new data. This highlights the benefit of a causal understanding of how changes in forecast model lead to improved forecast performance. If better understanding of the underlying reason why features or models better perform (as opposed to 'the metric improved by xx%') can be gained then forecasters can have more confidence that they will continue to provide benefit in the future.

5.4 | Relevance of competition to practice

Large-scale forecasting competitions can provide valuable insight into the performance of forecasting methods and serve to compare newly proposed methods with established state-of-the-art, and to show that these methods work in practice. They can also promote emerging challenges and stimulate research activity, as in the case of daily peak load forecasting, which has received relatively little attention in the academic literature to date. Predicting the timing of daily peak load is certainly distinct from classical time series prediction, though we did not treat is as such here. For example, others have addressed the property that there is, by definition, exactly one peak each day via ordinal regression [18] and cardinal points [40, 41]. However, forecast users may be interested in other definitions of peaks. We saw examples of days with two distinct peaks in this competition; both may be relevant to energy system operation. It may even be preferable to integrate the forecasting and decision-making process, as was demonstrated for peak shaving using battery energy storage in Ref. [12]. The competition set-up prevented the use of recent observations and therefore classical time series methods, so the potential benefits of these methods could not be studied.

Unusual or previously unseen conditions are of particular interest to forecast practitioners as they can introduce substantial uncertainty and may have a large impact on energy systems. The BDC22 included a period affected by a hurricane, which disrupted power supplies and reduced electricity load substantially for several days. This event was unpredictable following a strategy of 'learning from data' (competition data at least) as nothing like it was present in the provided historic data. In practice, the operational forecasters would have been well aware of the inclement weather and able to make adjustments to their forecasts, perhaps even based on past experience of similar events. Competitors in BDC22 may also have been aware of this particular hurricane, which hit the organisers' home state, or discovered it after noticing the unusual temperature data, and been able to make adjustments. We are therefore sceptical as to the relevance and benefit of including such events in ex-post competitions with limited explanatory data and suggest that future forecasting competitions aim to recreate the reality of operational forecasting where such events are concerned.

6 | CONCLUSION

This paper presents the approach used by the team 'peaky-finders' to predict the hourly magnitude, time, and shape of the peak load using an ensemble of MLR and GBM models. The results indicate that hourly models for forecasting electricity demand provide reasonable performance when identifying the time of the peak. As the volatility grows, the ability of a single approach to predict the time and magnitude of the demand remains feasible. However, more bespoke additions such as the proposed weighting scheme can be used to increase accuracy for the hour of peak. However, this necessitates the production and maintenance of several disparate forecasts which increases

complexity. Therefore, utilities must identify the sensitivity of their scheduling and dispatch decisions to these factors in order to determine the point at which bespoke modelling is required, for example, should days be considered as containing just one or multiple peaks? Calculating and tracking the variability of peak hour over time is one useful metric by which utilities can evaluate the need for bespoke peak hour forecasting model development. As this work presents the outcomes of the BDC2022 forecasting competition, which included data from one region of the United States, future work should evaluate the performance of similar methods on diverse data from multiple regions of the world to investigate the differences that seasonality and load composition have on the performance of peak load forecasting methods.

AUTHOR CONTRIBUTIONS

Daniel L. Donaldson: Conceptualization; formal analysis; methodology; validation; writing – original draft. **Jethro Browell**: Conceptualization; methodology; validation; writing – review & editing. **Ciaran Gilbert**: Conceptualization; methodology; validation; writing – review & editing.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data from the BigDEAL Challenge 2022 was used as the basis for this manuscript. The code that support the findings of this study are openly available on Github at https://github.com/DLDonaldson/BigDEALChallenge2022_peakyfinders.

ORCID

Daniel L. Donaldson https://orcid.org/0000-0003-3419-3624
Jethro Browell https://orcid.org/0000-0002-5960-666X
Ciaran Gilbert https://orcid.org/0000-0001-6114-7880

REFERENCES

1. Shahidehpour, M., Yamin, H., Li, Z.: Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management. John Wiley and Sons (2003)
2. Hong, T., et al.: Energy forecasting: a review and outlook. IEEE Open Access J. Power Energy 7, 376–388 (2020). https://doi.org/10.1109/oajpe.2020.3029979
3. Farrokhabadi, M., et al.: Day-ahead electricity demand forecasting competition: post-covid paradigm. IEEE Open Access J. Power Energy 9, 185–191 (2022). https://doi.org/10.1109/oajpe.2022.3161101
4. Hyndman, R.J.: A brief history of forecasting competitions. Int. J. Forecast. 36(1), 7–14 (2020). m4 Competition. https://doi.org/10.1016/j.ijforecast.2019.03.015
5. Chen, B.-J., Chang, M.-W., Lin, C.-J.: Load forecasting using support vector machines: a study on eunite competition 2001. IEEE Trans. Power Syst. 19(4), 1821–1830 (2004). https://doi.org/10.1109/tpwrs.2004.835679
6. Hong, T., Xie, J., Black, J.: Global energy forecasting competition 2017: hierarchical probabilistic load forecasting. Int. J. Forecast. 35(4), 1389–1399 (2019). https://doi.org/10.1016/j.ijforecast.2019.02.006
7. Distribution, W.P.: Presumed Open Data: Data Science Challenge (2021). [Online]. https://www.nationalgrid.co.uk/pod-data-science-challenge
8. Triguero, I.: IEEE-CIS Technical Challenge on Energy Prediction from Smart Meter Data. IEEE Dataport (2020). https://dx.doi.org/10.21227/2npg-c280.s
9. Gross, G., Galiana, F.: Short-term load forecasting. Proc. IEEE 75(12), 1558–1573 (1987). https://doi.org/10.1109/proc.1987.13927
10. National Grid ESO: What Are Electricity Triads? [Online]. https://www.nationalgrideso.com/document/130641/download
11. Borghini, E., et al.: Data-driven energy storage scheduling to minimise peak demand on distribution systems with PV generation. Energies 14(12), 3453 (2021). https://doi.org/10.3390/en14123453
12. Singleton, C., Grindrod, P.: Forecasting for battery storage: choosing the error metric. Energies 14(19), 6274 (2021). https://doi.org/10.3390/en14196274
13. Pacific Gas and Electric (PG&E): Time-of-Use Residential Transition Frequently Asked Questions [Online]. https://www.pge.com/pge_global/common/pdfs/rate-plans/rate-plan-options/time-of-use-base-plan/TOU-Transition-FAQs.pdf
14. California Independent System Operator: California ISO Peak Load History 1998 through 2022 [Online]. https://www.caiso.com/documents/californiaisopeakloadhistory.pdf
15. Jiang, Y.H., et al.: Predicting peak-demand days in the Ontario peak reduction program for large consumers. In: Proceedings of the 5th International Conference on Future Energy Systems, Ser. E-Energy '14, pp. 221–222. Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2602044.2602076
16. Fu, T., et al.: Predicting peak day and peak hour of electricity demand with ensemble machine learning. Front. Energy Res. 10 (2022). https://doi.org/10.3389/fenrg.2022.944804
17. Bovornkeeratiroj, P., et al.: Peaktk: an open source toolkit for peak forecasting in energy systems. In: ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS), Ser. COMPASS '22, pp. 324–339. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3530190.3534791
18. Gilbert, C., Browell, J., Stephen, B.: Probabilistic load forecasting for the low voltage network: forecast fusion and daily peaks. Sustain. Energy Grids Networks 34, 100998 (2023). https://doi.org/10.1016/j.segan.2023.100998
19. Hong, T.: Bigdeal Challenge 2022 [Online]. http://blog.drhongtao.com/
20. PJM Interconnection: Load Forecasting Model Whitepaper [Online]. https://www.pjm.com/~/media/library/reports-notices/load-forecast/2016-load-forecast-whitepaper.ashx (2016)
21. Donaldson, D.L., Jayaweera, D.: Weather informed need attributes for non-wires alternatives in distribution networks. In: 2020 IEEE Power and Energy Society General Meeting (PESGM), pp. 1–5. IEEE, Montreal (2020). https://doi.org/10.1109/PESGM41954.2020.9281917
22. Hong, T., Wilson, J., Xie, J.: Long term probabilistic load forecasting and normalization with hourly information. IEEE Trans. Smart Grid 5(1), 456–462 (2014). https://doi.org/10.1109/tsg.2013.2274373
23. Hyndman, R.J., Fan, S.: Density forecasting for long-term peak electricity demand. IEEE Trans. Power Syst. 25(2), 1142–1153 (2010). https://doi.org/10.1109/tpwrs.2009.2036017
24. Wang, P., Liu, B., Hong, T.: Electric load forecasting with recency effect: a big data approach. Int. J. Forecast. 32(3), 585–597 (2016). https://doi.org/10.1016/j.ijforecast.2015.09.006
25. Donaldson, D.L., Khan, Z.A., Jayaweera, D.: Improving Forecast Accuracy Using a Synthetic Weather Station: An Incremental Approach and Bfcom2018 Lessons Learned. In: IEEE Milan PowerTech, pp. 1–6 (2019). https://doi.org/10.1109/PTC.2019.8810869
26. Greenwell, B., et al.: GBM: Generalized Boosted Regression Models. R package version 2.1.8.1 (2022). [Online]. https://CRAN.R-project.org/package=gbm
27. Boehmke, B.: UC Business Analytics R Programming Guide (2018). [Online]. http://uc-r.github.io/gbm_regression

28. Zhang, N., et al.: Comparison of three short-term load forecast models in southern California. Energy 189, 116358 (2019). https://doi.org/10.1016/j.energy.2019.116358

29. Hu, M., et al.: Impacts of building load dispersion level on its load forecasting accuracy: data or algorithms? Importance of reliability and interpretability in machine learning. Energy Build. 285, 112896 (2023). https://doi.org/10.1016/j.enbuild.2023.112896

30. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 29(5), 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451

31. R Core Team: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2022). [Online]. https://www.R-project.org/

32. Wickham, H., et al.: dplyr: A Grammar of Data Manipulation. R package version 1.0.10 (2022). [Online]. https://CRAN.R-project.org/package=dplyr

33. Vaughan, D, Dancho, M.: tibbletime: Time Aware Tibbles. R package version 0.1.7 (2022). [Online]. https://CRAN.R-project.org/package=tibbletime

34. Wickham, H., Bryan, J.: readxl: Read Excel Files. R package version 1.4.1 (2022). [Online]. https://CRAN.R-project.org/package=readxl

35. Grolemund, G., Wickham, H.: Dates and times made easy with lubridate. J. Stat. Software 40(3), 1–25 (2011). [Online]. https://doi.org/10.18637/jss.v040.i03

36. Hamner, B., Frasco, M.: Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.4 (2018). [Online]. https://CRAN.R-project.org/package=Metrics

37. Dowle, M., Srinivasan, A.: Data.Table: Extension of 'data.frame'. R package version 1.14.4 (2022). [Online]. https://CRAN.R-project.org/package=data.table

38. Zeileis, A., Grothendieck, G.: zoo: S3 infrastructure for regular and irregular time series. J. Stat. Software 14(6), 1–27 (2005). https://doi.org/10.18637/jss.v014.i06

39. Hallman, J.: tis: Time Indexes and Time Indexed Series. R package version 1.39 (2021). [Online]. https://CRAN.R-project.org/package=tis

40. Taylor, J.W., Majithia, S.: Using combined forecasts with changing weights for electricity demand profiling. J. Oper. Res. Soc. 51(1), 72–82 (2000). https://doi.org/10.1057/palgrave.jors.2600856

41. Bunn, D., Farmer, E.D.: Comparative models for electrical load forecasting. 1 (1985). [Online]. https://www.osti.gov/biblio/6256333

How to cite this article: Donaldson, D.L., Browell, J., Gilbert, C.: Predicting the magnitude and timing of peak electricity demand: a competition case study. IET Smart Grid. 1–12 (2023). https://doi.org/10.1049/stg2.12152