International Journal of Forecasting 39 (2023) 1272–1286

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Daily peak electrical load forecasting with a multi-resolution approach

Yvenn Amara-Ouali [a,*], Matteo Fasiolo [b], Yannig Goude [c], Hui Yan [c]

[a] University Paris-Saclay, France
[b] University of Bristol, United Kingdom
[c] EDF Lab Saclay, France

ARTICLE INFO

Keywords:
Generalised additive models
Neural networks
Peak load forecasting
Smart grids
Automated feature engineering
Multi-resolution

ABSTRACT

In the context of smart grids and load balancing, daily peak load forecasting has become a critical activity for stakeholders in the energy industry. An understanding of peak magnitude and timing is paramount for the implementation of smart grid strategies such as peak shaving. The modelling approach proposed in this paper leverages high-resolution and low-resolution information to forecast daily peak demand size and timing. The resulting multi-resolution modelling framework can be adapted to different model classes. The key contributions of this paper are (a) a general and formal introduction to the multi-resolution modelling approach, (b) a discussion of modelling approaches at different resolutions implemented via generalised additive models and neural networks, and (c) experimental results on real data from the UK electricity market. The results confirm that the predictive performance of the proposed modelling approach is competitive with that of low- and high-resolution alternatives.

1. Introduction

The electric daily peak load is the maximum of the electricity power demand curve over one day. Having an accurate forecast of the daily peak enables independent system operators (ISOs) and energy providers to better deliver electricity and optimise power plant schedules. The importance of such a forecast is increasing as the integration of intermittent renewable production sources progresses. In particular, renewable energy sources are at the bottom of the merit order curve, which makes them (currently) the most economical source of energy used to serve the market. However, they are intermittent and provide time-varying levels of power generation that are only partially under human control. If electricity demand is high and renewables cannot provide for it alone, ISOs must deliver electricity from sources with higher marginal costs (e.g. gas-fired plants) for stakeholders as well as for the environment in terms of carbon dioxide emissions. In such a context, accurately forecasting the peak demand magnitude and timing is essential for determining the generation capacity that must be held in reserve.

Electrical equipment is tailored to support a specific peak load. If the demand comes close to or exceeds the network capacity, it can lead to distribution inefficiencies and ultimately power system failures, such as blackouts. With the increasing number of electric vehicles (EVs) in circulation, a further source of stress is added to the electricity system. For instance, 46% of vehicles sold in Norway in 2019 were EVs (International Energy Agency, 2019). The challenge posed by the additional EV demand must be met by more tailored management systems and policies, if expensive infrastructural works are to be avoided. Dynamic electricity pricing schemes, for example, the Triads in the UK or the Global Adjustment in Ontario, Canada, have been developed to reduce the system peak load. Consumers who can correctly estimate and cut their use during peak events can unlock great

* Corresponding author.
E-mail address: yvenn.amara-ouali@universite-paris-saclay.fr (Y. Amara-Ouali).

https://doi.org/10.1016/j.ijforecast.2022.06.001
0169-2070/

savings. Peak demand forecasts will thus be key for the development of such policies.

To account for the increasing demand for electricity and to prevent system failures, smart grid technologies and policies are being implemented to foster communication between the various stakeholders in the electricity supply chain to achieve a more efficient use of energy. One major objective is to maximise the load factor. The load factor is the average load over a specific time period divided by the peak load over the same period. Maximising it leads to a more even use of energy over time, thus preventing system failures and surges in electricity prices. One of the most common ways to achieve load factor maximisation is peak shaving, which refers to the flattening of electrical load peaks. Three major strategies have been proposed for peak shaving, namely integration of the energy storage system (ESS), vehicle-to-grid (V2G) integration, and demand-side management (DSM) (Uddin, Romlie, Abdullah, Abd Halim, Abu Bakar, & Chia Kwang, 2018). ESS and V2G integration provide ancillary sources to balance the grid through batteries while DSM shifts consumer demand to flatten the peak. To be activated adequately, all these strategies require accurate forecasts of the demand peak magnitude (DP) and of the instant at which it occurs (IP).

This article proposes novel methods to forecast the DP and the IP by leveraging information at different time resolutions. In particular, the multi-resolution approach proposed here is illustrated in the context of two model classes: generalised additive models (GAMs) and neural networks (NNs). Both are state-of-the-art predictive models, widely used to forecast electrical load in industry and academia. The performance of the multi-resolution framework under both model classes is assessed using aggregate UK electricity demand data from the National Grid (National Grid, 2021).

The rest of the paper is structured as follows: Section 2 presents a literature review of daily peak forecasting methodologies. Section 3 introduces multi-resolution modelling using GAMs and NNs. Section 4 explains how the different models were setup in the high-resolution, low-resolution, and multi-resolution settings. Section 5 analyses the results of the models described in Section 4, using UK demand data.

2. Related work

This section provides an extensive literature review of peak forecasting methods. The review was conducted to identify gaps in the field. It includes methods ranging from probabilistic approaches to deep learning.

Probabilistic forecasts have been widely adopted in the context of load forecasting applications (e.g. Hong & Fan, 2016, for an overview), but little has been done on probabilistic peak demand forecasting. Two probabilistic setups commonly used for peak load forecasting were outlined by Jacob, Neves, and Vukadinović Greetham (2020). The first is block maxima (BM), where data are separated into time chunks of equal lengths and the maximum of each chunk is assumed to approximately follow a generalised extreme value (GEV) distribution. The second is peaks over threshold (POT), which approximates the distribution of the excess load over a threshold by a generalised Pareto distribution. While the POT and BM settings can be unified via point processes (Boano-Danquah, Sigauke, & Kyei, 2020), in this work we are mainly interested in the BM case.

In a long-term forecasting setting, McSharry, Bouwman, and Bloemhof (2005) used demand data at the daily resolution to forecast the magnitude and timing of the yearly peak (i.e. the day characterised by the largest total demand). They considered a forecasting lead-time of one full year and obtained a probabilistic forecast by simulating year-long trajectories for the weather variables and plugging them into a deterministic linear regression model. Similarly, Hyndman and Fan (2010) considered a long-term forecasting application, where the aim was to forecast the probability distribution of the annual and weekly peak electricity demand. They used semi-parametric additive models to capture the effect of covariates, such as temperature, on the demand and obtained a probabilistic forecast by adopting a simulation and scenario-based approach. Elamin and Fukushige (2018) used quantile regression methods to forecast the DP one day ahead. Even though they used quantile regression to obtain an upper bound on demand, quantile estimates at several probability levels could be used to estimate the full peak demand distribution. Also Gibbons and Faruqui (2014) modelled the DP via a quantile regression model, but their objective was post-processing daily estimates to forecast the annual demand peak, rather than modelling the DP probabilistically.

Multivariate regression models using multivariate adaptive regression splines (MARS) were proposed by Sigauke and Chikobvu (2010) to forecast the DP in South Africa. Explanatory variables including meteorological variables are aggregated at the daily resolution (e.g. average, minimum, and maximum temperature). The model outperforms piecewise polynomial regression models with an autoregressive error term. Sigauke and Chikobvu (2011) studied time series of the DP and illustrated its heteroscedastic structure. A SARIMA–GARCH error model and a regression-SARIMA–GARCH model are then proposed to forecast it at a short-term horizon. Results show that SARIMA-like models produce forecasts with an accuracy around 1.4 in mean absolute percentage error on a testing period.

Saxena, Aponte, and McConky (2019) proposed a hybrid model to forecast whether the following day will be a peak load day for the billing period for customers subject to a demand charge structure. They applied their model to optimise the electricity bill of an American University. Load data were provided every five minutes from January 2013 to April 2016. Here, the POT setup was used with a threshold depending on a monthly average and variance of the daily load. An original combination of four forecasts was proposed. First, a linear model is used to forecast the maximum daily load at a monthly horizon, which is then coupled to short-term load forecasting models (NN and ARIMA) to provide two forecasts. Two other forecasts were computed using binary classifiers (logistic regression and NN) and a synthetic minority over-sampling

technique (SMOTE) was used to balance the classes. The authors demonstrated that their methods led to better statistical accuracy and to reduced electricity bills.

NNs are among the most popular algorithms for peak load forecasting tasks because of their strong performance in non-linear modelling. Their flexibility is remarkable, but it is difficult to pick the right architecture and hyper-parameters for a specific problem. One of the first papers proposing an NN peak load forecasting method was produced by Dash, Liew, and Rahman (1995). According to the authors, NNs performed well on load forecasting problems, but they were much less performant on peak load forecasting tasks. A fuzzy NN was found to be more robust and accurate than a traditional NN structure. It involved an additional layer of fuzzification of the inputs before entering the only hidden layer of the network.

In a more traditional setup, Saini and Soni (2002) tested a fully connected neural network (FCNN) with different variants of back-propagation algorithms, where training was conducted separately in four periods of time during a year. Their work was further developed by Saini (2008), where numerous weather variables were included (e.g. temperature, rainfall, wind speed, evaporation per day, sunshine hours, and associated statistics). Similarly, different optimisation procedures were considered and it was found that an adaptive learning method based on the learning rate and momentum was the most performant. Amin-Naseri and Soroush (2008) combined a self-organising map with an NN to find better clusters of training data to improve forecasting performance. Some authors considered other form of networks. For instance, Abdel-Aal (2006) adopted abductive networks with the aim of obtaining a better intuition and a more automated way to address peak load forecasting. In particular, these networks split the overall problem into smaller and simpler ones along the network with abductive reasoning. It is based on an automated procedure which organises the data available into different chunks and deals with them separately.

More recently, recurrent neural networks (RNNs) have been used by Yu, Niu, Tang, and Wu (2019) in the form of gated recurrent units (GRUs). In particular, a dynamic time warping (DTW) analysis was used to produce the GRU inputs. The DTW distance was used to find the most similar load curve to the one observed before the targeted load curve. Assuming that subsequent load curves are also very similar, they used the subsequent load curve from the training data to encode the inputs of the GRU network. A long short-term memory (LSTM) architecture has been used by Ibrahim and Hossain (2020) and was found to be more computationally efficient compared to FCNNs and other RNNs. Three statistical metrics were used to evaluate model performance: the mean absolute percentage error (MAPE), root mean squared error (RMSE), and mean bias error. In our work, statistical metrics, including the MAPE and RMSE, are also used to avoid introducing any bias towards a particular operational application.

The literature on deep learning peak load forecasting is sparse, but deep learning probabilistic load forecasting is much more common (e.g. Guo, Zhou, Zhang, & Yang, 2018, Yang, Hong, & Li, 2019, and Yang, Li, Gulliver, & Li, 2020). Such models do not explicitly focus on the DP or the IP, as the objective functions used to estimate their parameters are based on demand observed at a higher frequency (intra-day). The high-frequency forecasts thus obtained can be post-processed to produce a forecast for the DP.

Support vector regression (SVR) is another popular load forecasting method, based on structural risk minimisation instead of empirical risk minimisation as in NNs. El-Attar, Goulermas, and Wu (2009) used SVR in a local prediction framework. Recently, Kim, Lee, and Qureshi (2020) used an ensemble forecasting approach with other machine learning algorithms, such as boosting machines, tree-based methods, and bagging techniques. A compensation process based on an isolation forest was later added by analysing the predicted values of the ensemble models to detect outliers in the peak data. SVR was compared to NNs by Li, Zhao, Wang, Wang, Yang, and Yan (2018) for a control strategy of peak load and frequency regulation. LSTM NNs were used to forecast power load and improve the control strategy considered in this particular use case.

From this literature review, it can be concluded that a wide range of methodologies have been adopted in peak load forecasting applications. In most short-term applications, model inputs are manually chosen features that are defined at the same (daily) time resolution as the peak demand, which is the variable to be forecasted. Conversely, in long-term applications, weather variables are simulated at the original (high) resolution to produce demand forecasts at the same resolution, which are then post-processed to obtain low-resolution (e.g. yearly) peak forecasts. Hence, to the best of our knowledge, the existing literature on peak forecasting has not explored methods that are able to integrate both low- and high-resolution signals in a single model. However, in the field of functional data analysis, hybrid approaches have been used for clustering and forecasting functional data (e.g. Antoniadis, Paparoditis, & Sapatinas, 2006 and Cho, Goude, Brossat, & Yao, 2013). Therefore, this paper aims to exploit functional methods to tackle multi-resolution problems. From a feature engineering point of view, the goal is to automate feature extraction of high-resolution signals, that is, to let the model decide which hidden features to extract from the signal. This can be done with signal processing procedures such as tensor product decomposition, wavelets, or Fourier transforms (Amin et al., 2015).

The literature review also suggests that not much effort has been directed towards forecasting the IP, which is surprising because forecasting the IP is at least as important as forecasting the DP, for the purpose of short-term smart grid management and operational planning (Soman, Trivedi, Irwin, Kosanovic, McDaniel, & Shenoy, 2020). To fill this gap, the performance of multi-resolution methods is illustrated in this paper on both a DP and an IP forecasting problem.

3. Multi-resolution modelling

In this section, the multi-resolution modelling approach is introduced with its general principles. It is then developed formally and illustrated with GAMs and NNs.

3.1. General idea

The main idea behind multi-resolution modelling is to build a parsimonious model that is able to handle input and output variables that are available at different resolutions. In the context of DP load forecasting, low-resolution variables (e.g. day of the week or maximum daily temperatures) are observed daily, while high-resolution variables (e.g. temperatures or raw demand) are updated every hour or half-hour. Such problems are usually handled by manually placing all variables at the same resolution. In particular, one option is to take a high-resolution approach, which consists in doing the modelling at the highest available resolution, although this might require interpolating some of the low-resolution variables. Such an approach often lacks parsimony, as the low-resolution variables are brought to the higher resolution, thus increasing the size of the data that needs to be processed while adding no extra useful information. Another option is to take a low-resolution approach, that is, to transform the high-resolution variables into a set of manually chosen daily summaries or features. In this approach, the size of the data is reduced, but feature engineering is time consuming and some of the information contained in the high-resolution variables is lost in the process.

The multi-resolution approach proposed here aims at capturing all the information contained in the high-resolution variable, while avoiding explicit feature engineering and retaining the parsimony of the low-resolution approach. To describe the multi-resolution idea more formally, let us consider $\mathbf{y}_i = \{y_i(t)\}_{t\in\{1,\dots,T\}}$ the vector of electricity demand at each time step $t > 0$ of the day $i \in \mathbb{N}$, where T is the total number of daily steps (e.g. $T = 48$ for half-hourly steps). Then, the DP of day i is $\mathrm{DP}_i = max(\mathbf{y}_i)$, and IP_i is the time step corresponding to DP_i. Let $\mathbf{x}_i^{low}$ be the ith vector of covariates observed daily and let $\mathbf{x}_i^{high}$ be the corresponding vector of covariates containing information at the intra-day resolution. The multi-resolution approach exploits both sets of covariates as model inputs to obtain the forecasts of the $\hat{\mathrm{DP}}_i$ or the $\hat{\mathrm{IP}}_i$. That is,

$$\hat{\mathrm{DP}}_i = \psi_1(\mathbf{x}_i^{low}, \mathbf{x}_i^{high}) \tag{1}$$

$$\hat{\mathrm{IP}}_i = \psi_2(\mathbf{x}_i^{low}, \mathbf{x}_i^{high}) \tag{2}$$

where ψ_1 and ψ_2 represent the model for the DP and the IP, respectively. This general definition does not specify how the high-resolution inputs should be dealt with in practice. Several approaches could be considered, the aim being to process the information contained in a (possibly high-dimensional) signal vector, while avoiding information loss and retaining computational efficiency. In this paper, two options are considered. In particular, a description of how high-resolution covariates can be handled within GAMs and NNs is given below.

3.2. Particular instances of the multi-resolution approach

The multi-resolution approach is detailed firstly for GAMs, which, due to their performance and interpretability (Amato, Antoniadis, De Feis, Goude, & Lagache, 2021), are widely used in industry for load forecasting. Then, the multi-resolution approach is extended to NNs, which often perform well on load forecasting problems and enable the flexible handling of heterogeneous model inputs (Gao, Guo, & Wang, 2017).

3.2.1. Generalised additive models

First introduced by Hastie and Tibshirani (1999), GAMs are a semi-parametric extension of generalised linear models (GLMs) where the response variable, y_i, is assumed to follow a parametric probability distribution. That is, $y_i \sim \mathrm{Dist}(\mu_i, \boldsymbol{\theta})$, where μ_i and $\boldsymbol{\theta}$ are model parameters. While the elements of $\boldsymbol{\theta}$ do not depend on i, parameter μ_i is modelled as follows (Wood, 2017):

$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\gamma} + \sum_j f_j(\mathbf{x}_i), \tag{3}$$

where g is a monotonic transformation, which is simply the identity function in this paper. Two separate terms can be distinguished on the right-hand side of this equation: a parametric part $\mathbf{x}_i^T\boldsymbol{\gamma}$, where $\mathbf{x}_i$ is a vector of covariates while $\boldsymbol{\gamma}$ is a vector of regression coefficients; and a non-parametric part $\sum_j f_j(\boldsymbol{x}_i)$ which is a sum of smooth functions of covariates. The smooth effects are built via linear combinations of K_j basis functions, while the corresponding basis coefficients are penalised via generalised ridge penalties. The strength of the penalties is controlled via smoothing hyperparameters, which are selected using criteria such a generalised cross-validation.

In the context of forecasting DP_i, it is interesting to consider for $\mathrm{Dist}(\mu_i, \boldsymbol{\theta})$ a generalised extreme value (GEV) distribution. In fact, the GEV model is asymptotically justified for block maxima as $T \to \infty$ (Jacob et al., 2020). Thus, when enough steps are available throughout the day, the GEV distribution is particularly attractive for modelling the DP. The scaled-T (a scaled version of Student's t) distribution provides an alternative that is particularly suited for heavy tailed data such as peak load. The Gaussian distribution can be used as a baseline model. As for the IP, an ordered categorical (ocat) distribution based on a logistic regression latent variable is used. All of these distributions as well as GAM building and fitting methods are implemented in the *mgcv* R package (Wood, 2020).

Within the additive structure of GAMs, $\mathbf{x}_i^{low}$ and $\mathbf{x}_i^{high}$ can be treated as inputs for different smooth functions. The elements of $\mathbf{x}_i^{low}$ can be handled via separate standard smooth effects, which take scalars as inputs, while the joint effect of several elements of $\mathbf{x}_i^{low}$ can be captured via standard multivariate smooth effects. However, the $\mathbf{x}_i^{high}$ covariates have to be treated via functional smooth effects. The latter are smooth functions which take the vectors of high-resolution covariates as inputs and output a scalar. Therefore, functional GAMs permit the handling of each covariate at its original resolution, thus avoiding interpolation and guaranteeing parsimony.

In addition to the principle of parsimony, the goal is also to retain the time dependence of the covariates. In fact, it is important to ensure that the model is aware that each element of the high-resolution covariates has a different impact on the peak load distribution, as it belongs to a different time of day. However, modelling the

effect of each element separately would ignore temporal dependencies and might lead to overfitting or a lack of interpretability. A way to achieve a compromise when modelling high-resolution covariates is to make them interact with the time-of-day sequence via tensor product effects. Such effects can easily be integrated in GAMs, as explained in the following.

In continuous time, the smooth effect for a high-resolution (functional) covariate, $x_i(u)$, can be written as follows:

$$f(x_i) = \int_0^T \phi(x_i(u), u)du \tag{4}$$

where ϕ is the time-dependent effect of the covariate, which needs to be estimated, while u is the time of day. In practice, on the ith day, $x_i(u)$ is observed at F discrete instants $0 \leq t_1 \leq \cdots \leq t_F \leq T$, and the corresponding values of $x_i(u)$ are stored in the vector $\boldsymbol{x}_i$. Hence, approximating the integral with a summation and constructing ϕ via a tensor product expansion leads to:

$$\begin{aligned}\hat{f}(\boldsymbol{x}_i) &= \sum_{r=1}^{F} \hat{\phi}(x_i(t_r), t_r) \\ &= \sum_{r=1}^{F}\sum_{k=1}^{K}\sum_{l=1}^{L} \beta_{kl} a_k(x_i(t_r)) b_l(t_r)\end{aligned} \tag{5}$$

where $\{a_k\}_{(k)\in\{1,\ldots,K\}}$ and $\{b_l\}_{(l)\in\{1,\ldots,L\}}$ are known spline basis functions, and $\{\beta_{kl}\}_{(k,l)\in\{1,\ldots,K\}\times\{1,\ldots,L\}}$ are parameters to be estimated. By using such effects, high-resolution information can be parsimoniously incorporated into the model, while retaining the temporal information contained in the covariates.

3.2.2. Neural networks

NNs are convenient machine learning algorithms for implementing a multi-resolution model. In fact, common architectures such as convolutional neural networks (CNNs) and RNNs already make use of inputs from different scales. Recent work was undertaken to make tensor inputs available for multi-layer perceptrons with MatNet (Gao et al., 2017), which further shows their versatility. From scalars to tensors, the flexibility of NNs is hard for other machine learning models to compete with.

An FCNN or CNN architecture, without its output layer, can be generally written as follows:

$$H_k(\mathbf{x}, \Theta) = h_k(\ldots h_3(h_2(h_1(\mathbf{x}, \theta_1), \theta_2), \theta_3)\ldots, \theta_k) \tag{6}$$

where k is the number of hidden layers of the NN, $h_{i,i\in\{1\ldots k\}}$ are the transformations made by the hidden layers (e.g. linear operation, activation, and dropout), and $\Theta = \{\theta_i\}_{,i\in\{1\ldots k\}}$ is the sequence of parameter vectors (weights and biases). In a multi-resolution approach, one part of the architecture will contain low-resolution information feeding an FCNN branch, and the other one will contain the reshaped high-resolution data feeding a CNN or RNN branch. In this paper, only CNNs were considered in depth for this latter branch, with the lags of the response provided as model inputs. The CNN enables a very close replication of the tensor product construction used for GAMs, thus creating a consistent setup for comparing both algorithms.

Even though the CNN and FCNN branches do not have similarly shaped inputs and outputs, the unit shapes can be transformed along the network to interact and be brought together without losing consistency. This process consists in flattening the tensor shapes in order to bounce back on vectorial inputs within some layer of the network. It is precisely this flexibility that can be leveraged to build a multi-resolution architecture (Fig. 1). More precisely, the CNN branch contains one convolutional block for each of the high-resolution time series. In this way, each tensor product of the GAM formula can find its equivalent in the CNN branch of the network. In fact, the multi-resolution NN architecture can be concisely written as follows:

$$\boldsymbol{\mu}_i = F_j(H_k(\mathbf{x}_{low}, \Theta), H_l'(\mathbf{x}_{high}, \Theta')) \tag{7}$$

where H_k is the FCNN which handles low-resolution terms while H_l' is the CNN which deals with the high-resolution information. Then, in the final part of the network, both outputs are concatenated (after flattening the CNN branch) and enter another FCNN F_j which can be reduced to the output layer when $j = 1$. Here, μ_i is the mean of the random output variable considered. This multi-resolution architecture is summarised in Fig. 1.

4. Experiments

On the DP and the IP forecasting tasks, the multi-resolution approach is compared to two alternative modelling approaches: a high-resolution approach and a low-resolution approach (Fig. 2). The low-resolution approach uses inputs aggregated at the daily level (e.g. maximum daily temperature or day of the week) to forecast the DP and the IP separately. The high-resolution approach uses inputs at the half-hourly level to forecast the half-hourly demand, and then it extracts the DP and the IP by taking the maximum of the half-hourly forecasted values and the corresponding time of day. Therefore, the high-resolution approach leverages all the information available by taking half-hourly inputs and outputs, while the low-resolution approach directly models the variables of interest (DP and IP) with fewer parameters to be estimated. The multi-resolution approach can be seen as a compromise, aimed at integrating the advantages of both approaches, and the following experiments are designed to assess whether it can outperform them.

The comparison includes baseline models: a naive persistence model, which simply consists of forecasting the DP and the IP based on the value taken by the target variable on the previous day; a low-resolution ARIMA (on daily peaks with horizon 1); and a high-resolution ARIMA aggregated forecast composed of 48 ARIMA models, each fitted on the half-hourly load of a specific time of day with horizon 1. That is, the high-resolution ARIMA produces 48 forecasts at horizon 1 instead of one forecast at horizon 48. All ARIMA models are fitted using the Hyndman and Khandakar (2007) algorithm without using exogenous information.

The performance metrics chosen for DP models are the mean absolute percentage error (MAPE) and the root

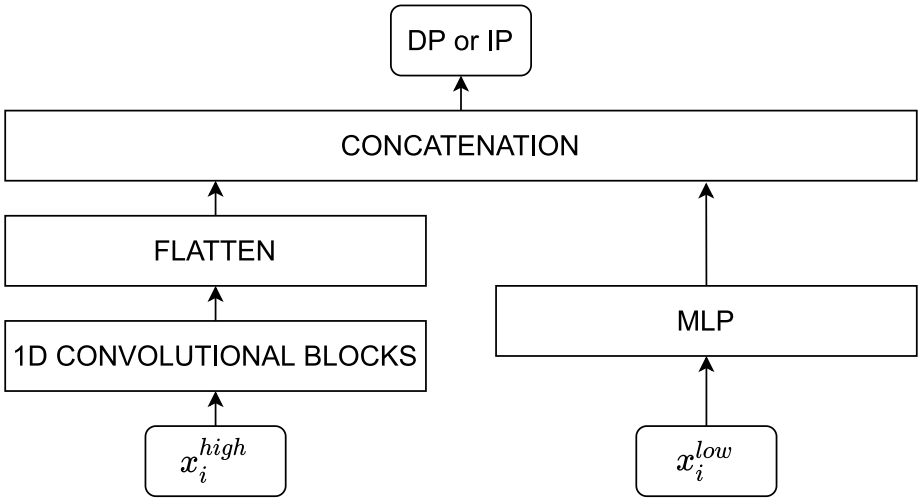

Fig. 1. Multi-resolution architecture for NNs with a multi-layer perceptron (MLP) taking low-resolution inputs, and a CNN with high-resolution inputs.

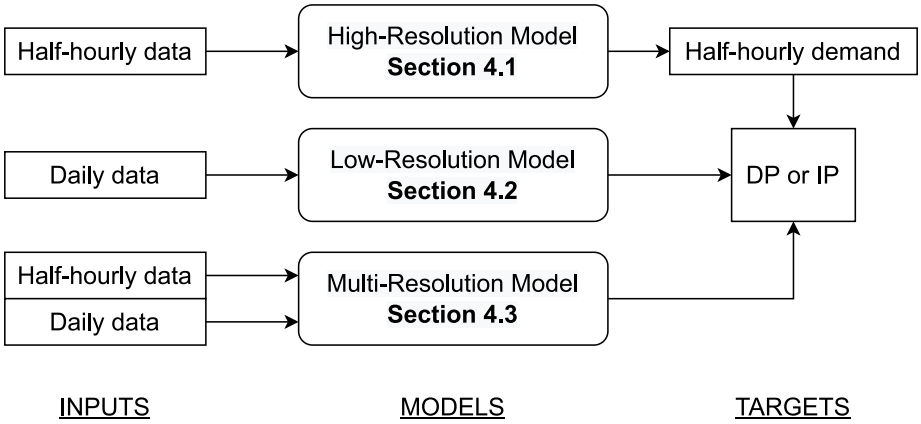

Fig. 2. The different modelling settings compared in this work.

mean squared error (RMSE). As for IP models, the RMSE is also used but the MAPE is substituted with a relaxed accuracy (R-accuracy) metric in the form of a binary loss function (equal to 1 if the IP forecasted is more than two instants away from the observed IP, and 0 if it is within two instants of the observed IP). While the R-accuracy metric is also relevant in operational settings where it is crucial to know the IP within a small time window, the RMSE penalises forecasts proportionally to their distance from the observed IP.

A rolling-origin forecasting procedure is used to replicate a realistic short-term load forecasting setup. Model parameters are updated on a monthly basis with consolidated data since, in an operational setting, threats to data validity and computational constraints can emerge when refitting a model too often using real-time data.

The data used in the experiments were the half-hourly load consumption (total national demand) between 2011-07-01 00:00:00 and 2016-06-30 23:30:00, available via the UK National Grid (2021) website. Temperature data at different locations (London, Sheffield, Manchester, Leeds, Cardiff, Bristol, Birmingham, Liverpool, Crosby, and Glasgow) were downloaded from the National Oceanic and Atmospheric Administration (2021) website. The temperature data were at an hourly resolution. They were interpolated (natural cubic spline interpolation) to obtain half-hourly data. Furthermore, if pop_s is the population of the nearest city to station s, a weighted mean temperature is calculated as follows:

$$\text{temp}(t) = \frac{1}{\sum_{a=1}^{10} pop_s} \sum_{s=1}^{10} pop_s T_{s,t} \tag{8}$$

where $T_{s,t}$ is the temperature recorded at time t by station s, and $\text{temp}(t)$ is the weighted mean temperature that is used in the modelling experiments. An exponentially smoothed version of the weighted mean temperature is also included in the model features: $\text{temp95}(t) = \alpha \times$

Table 1
High-resolution model inputs.

Type	Name	Unit	Description
Weather	temp	[C]	Half-hourly temperature
	temp95	[C]	Half-hourly smoothed temperature
Calendar	dow	Categorical	Day of the week
	toy	None	Time of year (between 0 and 1)
	t	Categorical	Time of day (between 0 and 47)
Lag	load24	[10^1 GW]	Half-hourly load on the previous day
Output	load	[10^1 GW]	Half-hourly load

$\text{temp95}(t-1) + (1-\alpha) \times \text{temp}(t)$. It was computed using a smoothing parameter $\alpha = 0.95$, based on expert knowledge.

4.1. High-resolution approach

Forecasting the electricity hourly or half-hourly demand is a problem that has been extensively studied in the literature (Kuster, Rezgui, & Mourshed, 2017). It is well known that a common driver of electrical load is weather and, in particular, temperature. In addition, calendar information can be used to explain the seasonal variation of the demand. Finally, lagged demand values are highly informative for the subsequent values. These variables are summarised in Table 1.

The GAM chosen to implement this approach is $y_i(t) \sim N(\mu_i(t), \sigma^2)$, where the mean of the Gaussian distribution is modelled by:

$$\begin{aligned}\mu_i(t) = {} & \psi_1(\text{dow}_i) + \psi_2(\text{t}) + f_1^{20}(\text{toy}_i(t)) + f_2^{20}(\text{temp}_i(t)) \\ & + f_3^{24}(\text{temp95}_i(t)) \\ & + \text{ti}_1^{5,5}(\text{temp}_i, t) + \text{ti}_2^{5,5}(\text{temp95}_i(t), t) \\ & + \text{ti}_3^{5,5}(\text{load24}_i(t), t) \\ & + \text{ti}_4^{5,5}(\text{toy}_i(t), t)\end{aligned} \tag{9}$$

where the ψ functions are parametric effects, the f functions are univariate smooth effects, and the ti functions are bivariate tensor product smooth interactions. The number of basis functions used is indicated in the exponents. For instance, f_1^{20} uses 20 basis functions, and $\text{ti}_1^{5,5}$ uses five basis functions for each marginal. Thin-plate spline bases are used to build all smooth effects (Wood, 2003). The model structure (9) was decided on the basis of previous experience in the field and the statistical significance of each effect.

Many NN architectures could be considered for this problem. We want an architecture with the fewest layers possible and using the same model inputs as the GAM. Adding too many layers would lead to a drastic difference in degrees of freedom between the NN and the GAM, which is not realistic in a short-term load forecasting scenario. Furthermore, as we are not in the big data regime, adding too many layers may actually worsen the performance of the network.

Given that the universal approximation theorem (Cybenko, 1989 and Hornik, 1991) guarantees that a two-layer FCNN can approximate any measurable function on a compact support, an FCNN carefully built can approximate any non-linear function of the input variables with only one hidden layer. Therefore, an FCNN architecture was used to build an NN analogue of the high-resolution GAM baseline model.

In practice, there is no bound for the number of hidden units, which can lead to poor generalisation of the model when assessed on the test set. Therefore, a dropout layer was added after the hidden layer to foster the network generalisation. The outcome of the optimisation of hyperparameters led to an architecture which contains 50 neurons in the hidden layer and a dropout layer with a 10% dropout rate. The loss optimised is the mean squared error (MSE) with a Nesterov-accelerated adaptive moment (NADAM) optimiser (Dozat, 2016). In addition, the learning rate is 0.001, the number of epochs is 2000, and the batch size is 1024.

After obtaining the half-hourly demand forecast for the GAM and the NN, $\hat{\text{DP}}_i$ is estimated as the maximum daily value forecasted, and $\hat{\text{IP}}_i$ is estimated as the half-hour of the day during which $\hat{\text{DP}}_i$ occurred.

4.2. Low-resolution approach

In the low-resolution approach, all input variables are at the daily resolution (Table 2). Here, several distributions could be considered for GAMs. In particular, the scaled-T distribution, which is particularly suited for heavy tailed data, as well as the GEV family, which encompasses several extreme value distributions (Weibull, Gumbell, and Fréchet), are used to model the DP. For the IP forecasting task, the ordered-logit model implemented in the *mgcv* R package (Wood, 2020) is used. The low-resolution GAM can be written as follows:

$$\begin{aligned}\mu_i = {} & \psi_1(\text{dow}_i) + f_1^{10}(\text{IP24}_i) + f_2^{20}(\text{toy}_i(t)) + f_3^{20}(\text{DP24}_i) \\ & + f_4^{20}(\text{tempMax}_i(t)) + f_5^{20}(\text{temp95Max}_i(t)) \\ & + f_6^{20}(\text{tempMin}_i(t)) + f_7^{20}(\text{temp95Min}_i(t))\end{aligned} \tag{10}$$

For the DP, $\mu_i(t)$ is the location parameter of the distributions estimated, and the other parameters are assumed to be constants. For the IP, $\mu_i(t)$ is also the location parameter of a latent logistic distribution. Cut-off points are estimated in the course of model fitting and do not depend on the covariates. See Wood, Pya, and Säfken (2016) for details.

The same FCNN architecture as for the high-resolution approach was used (50 neurons in the hidden layer followed by a dropout layer). The optimal set of hyperparameters found is also equivalent. The difference between them lies in the inputs used (Table 2) and the response variable modelled, which here is directly the DP or the IP.

Table 2
Low-resolution model inputs.

Type	Name	Unit	Description
Weather	tempMax	[C]	Daily maximum temperature
	temp95Max	[C]	Daily maximum smoothed temperature
	tempMin	[C]	Daily minimum temperature
	temp95Min	[C]	Daily minimum smoothed temperature
Calendar	dow	Categorical	Day of the week
	toy	None	Time of year (between 0 and 1)
Lag	DP24	[10^1 GW]	Previous day peak demand
	IP24	Categorical	Previous day instant of peak
Output	DP or IP	[10^1 GW] or categorical	Daily demand peak or daily instant of peak

Table 3
Multi-resolution model inputs.

Type	Name	Unit	Description
Weather	matTem	[C]	Vector of half-hourly temperatures
	matTem95	[C]	Vector of half-hourly smoothed temperatures
Calendar	dow	Categorical	Day of the week
	toy	None	Time of year (between 0 and 1)
	matInt	Categorical	Vector of time steps (between 0 and 47)
Lag	matLag	[10^1 GW]	Vector of half-hourly load from previous day
Output	DP or IP	[10^1 GW] or categorical	Daily demand peak or daily instant of peak

The response structure for the DP is 1 neuron with a ReLU activation, while 48 neurons are used for the IP. Instead of the traditional softmax output used in classification problems, an ordinal output structure, more suited to model the IP, is implemented as formalised by Cheng, Wang, and Pollastri (2008). The observed response is structured as a vector of 1 and 0. If the peak was observed at $t \in \{1, \ldots, T\}$, all neurons before and including the tth one will be 1, and all neurons after will be 0. Therefore, sigmoidal activation functions are used.

4.3. Multi-resolution approach

The multi-resolution GAMs leverage the same level of information for model inputs as in the high-resolution GAMs. In addition, they directly target the DP response variable, as in the low-resolution approach. Tensor products defined in Section 3.2.1 are used to capture high-resolution information. The *mat* covariates presented in Table 3 are matrices of dimension $(N \times 48)$, where N denotes the number of observations of the response variable DP. The multi-resolution GAM model is:

$$\mu_i = \psi_1(\text{dow}_i) + f_1^{20}(\text{toy}_i) + ti_1^{15,10}(\text{matTem}_i, \text{matInt}_i) + ti_2^{5,5}(\text{matTem95}_i, \text{matInt}_i) + ti_3^{5,5}(\text{matLag}_i, \text{matInt}_i) \quad (11)$$

Unlike previous approaches, IP and DP lags are not directly included, as they can be captured by the model through the ti_3 tensor interaction. As for the low-resolution approach, Gaussian, scaled-T, and GEV distributions are considered for the DP, and the ordered categorical distribution is considered for the IP.

For the multi-resolution NN, the tensor product interactions will be replaced by convolution layers. The mechanism looked for through these convolution layers is essentially the same as for tensor products: extracting high-resolution information to directly model the DP or the IP. The high-resolution (half-hourly) data will be passed on to the convolution layers, while the low-resolution (daily) data will go through the same FCNN architecture used in the previous approaches. As shown in Fig. 3, these two sections of the architecture are then concatenated to produce the final forecast of the DP load. The output structure for the DP and the IP is the same as that detailed in Section 4.3, with one neuron for the DP and 48 neurons for the IP.

The convolutions used for the high-resolution information are 1D convolutions on two channels. Usually, only one convolution funnel is used to capture interactions between all inputs. Here, each tensor product interaction will be replicated as a unique convolutional block. Thus, three convolution blocks will independently extract the three high-resolution terms: matTem, matTem95, and matLag. The second channel of each block is the matrix containing the vectors of time steps matInt.

5. Results

The performance of the models for the DP and the IP forecasting tasks is evaluated using two statistical metrics. As a rolling-origin forecasting procedure was chosen, a transitional regime can be observed in the first few iterations, particularly for NNs, which usually perform better with a large amount of training data. Therefore, Table 4 (DP) and Table 5 (IP) present the models' performance on the last year of data, that is, from 2015-07-01 to 2016-06-30, inclusively.

With the exception of the high-resolution FCNN, the multi-resolution models perform better than the alternatives across all metrics (Table 4). The relatively strong performance of the high-resolution FCNN can be explained by the large amount of high-resolution data available, which suits the needs of NNs. Further, the FCNN contains more parameters to estimate and is thus more flexible than the

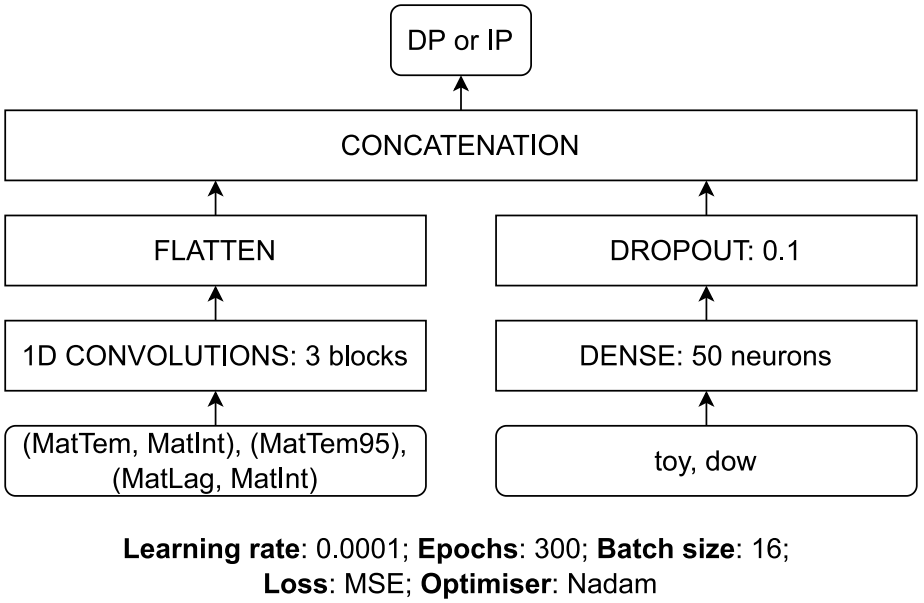

Fig. 3. Multi-resolution CNN architecture (input variable names are detailed in Table 3).

Table 4
Performance on the last year of data for the DP (best model and associated metrics are in **bold**).

Resolution	Model	Metrics	
		MAPE [%]	RMSE [MW]
NA	Persistence	4.38	34.3
High	ARIMA	4.08	27.8
	Gaussian GAM	2.43	15.5
	FCNN	1.47	10.3
Low	ARIMA	3.85	26.7
	Scat GAM	1.92	12.9
	GEV GAM	2.67	16.9
	Gaussian GAM	2.26	14.4
	FCNN	2.11	14.4
Multi	GEV GAM	1.52	10.3
	Scat GAM	**1.41**	**9.59**
	Gaussian GAM	1.42	9.63
	CNN	1.56	10.5

Table 5
Performance on last year of data for the IP (best model and associated metrics are in **bold**).

Resolution	Model	Metrics	
		R-Accuracy [%]	RMSE [half-hour]
NA	Persistence	79.4	5.36
High	Gaussian GAM	82.6	4.59
	FCNN	81.8	4.39
Low	Ocat GAM	79.1	4.22
	FCNN	83.2	4.40
Multi	Ocat GAM	79.4	4.08
	CNN	**83.5**	**3.85**

high-resolution GAMs, which require the user to manually specify how the effect of each input variable should be modelled. Nevertheless, the best model on all metrics is the scaled-T GAM, built using the multi-resolution approach. The GEV GAM performed worse than the other distributions, which is surprising given that the GEV distribution is asymptotically justified for BM. Interestingly, the shape parameter estimated was found to be close to 0, under which value the GEV model is simply a Gumbel distribution.

IP multi-resolution models have a similar or better performance than high- and low-resolution alternatives within the same model class on the RMSE metric (Table 5), and the multi-resolution CNN is the best model under all metrics. However, the metrics are affected by high sampling variability. The reasons for this are detailed below in this section, where we also argue that the mediocre performance of ocat GAMs for IP forecasting is not fundamental, but attributable to the insufficient flexibility of the specific ocat parametrisation adopted here.

To quantify the variability of the performance metrics considered so far, we used block-bootstrap resampling. As described by Messner, Pinson, Browell, Bjerregård, and Schicker (2020), for a test set of size N, we sample with replacement data blocks of fixed size $B = 7$ (i.e. one week) to obtain evaluation sets of size N. Repeating this procedure K times creates K metric samples, which can be used to estimate the metric's sampling variability. In particular, Fig. 4 shows block-bootstrapped boxplots for all metrics and models on the last year of data. Figs. 4(a–c) clearly demonstrate that the improvement obtained by adopting a multi-resolution approach is substantial and robust within the GAM model class. The HR-FCNN is competitive in terms of prediction but, as we discuss below, it is not easily interpretable and does not have the computational advantages of multi-resolution GAMs. For the IP problem, Figs. 4(e–d) make clear that the sampling variability is substantial (the reasons for which are discussed below).

As mentioned above, the rolling-origin forecasting setting may present a transitional regime during the first few

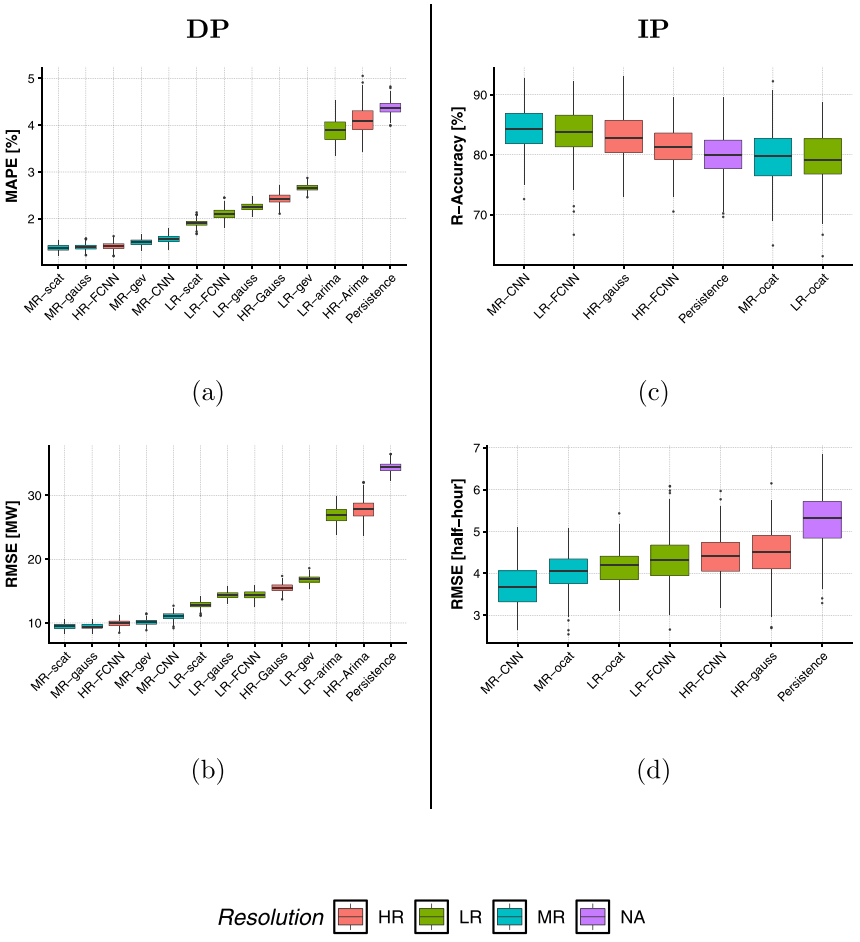

Fig. 4. Block-bootstrap boxplots of the two metrics considered for the DP models (a), (b) and IP models (c), (d) on the last year of data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

training iterations. Figs. 5 and 6 show the evolution of the different cumulative metrics calculated on the prediction signal updated on a monthly basis. Interestingly, the multi-resolution CNN for the DP (Fig. 5) starts off with a very bad prediction error on the first months. With more data, its performance rapidly improves across all metrics. The other models have a less dramatic performance trend, with the multi-resolution GAMs performing consistently better than the other models. The prediction error of these models oscillates during the first few months, which can be explained by the fact that the models did not have enough information to adequately estimate the yearly cycle, because they were fitted to only one year of data. After a year, the prediction errors stabilise.

For the IP forecasting task, the different metrics evolve with similar patterns (Fig. 6), but the seasonal oscillations in performance persist beyond the first year. Precisely, the performance slightly worsens after the first few months of each year. Fig. 7 explains why predicting the IP is harder in summer (April to August) than in winter (September to March). While winter daily demand profiles have a reliable evening peak, summer load profiles are flatter, and on some days the peak distribution becomes bimodal. That is, the daily peak might occur on the 25th half-hour (12:30 p.m.) or on the 35th half-hour (5:30 p.m.) with equal probability. This is also shown by the plot on the right in Fig. 8. Hence, it is clear that in the summer the IP point estimates might be unfairly penalised under the metrics considered here. This implies that a forecasting model might be better off providing an IP forecast that falls between the two peaks, as MR-CNN occasionally does (see Fig. 7). Such a forecast might improve the metrics but has little value in an operational setting. Note also that the ocat model struggles to capture an IP distribution that is unimodal or bimodal depending on the time of year. In particular, the ocat model used here is based on a standard ordered-logit parametrisation, which involves modelling the mean of a latent logistic random variable via an additive model. It is not possible to transform a unimodal distribution on the ordered categories (here, IP) into a bimodal one simply by controlling a location parameter. Hence, a more flexible model (e.g. Peterson & Harrell Jr, 1990) would be preferable.

It is interesting to verify the performance of each model for IP forecasting via a bespoke metric. In particular, let t_i^{m} be the observed IP on day i, and let $\hat{t}_i^{\mathrm{m}}$ be the

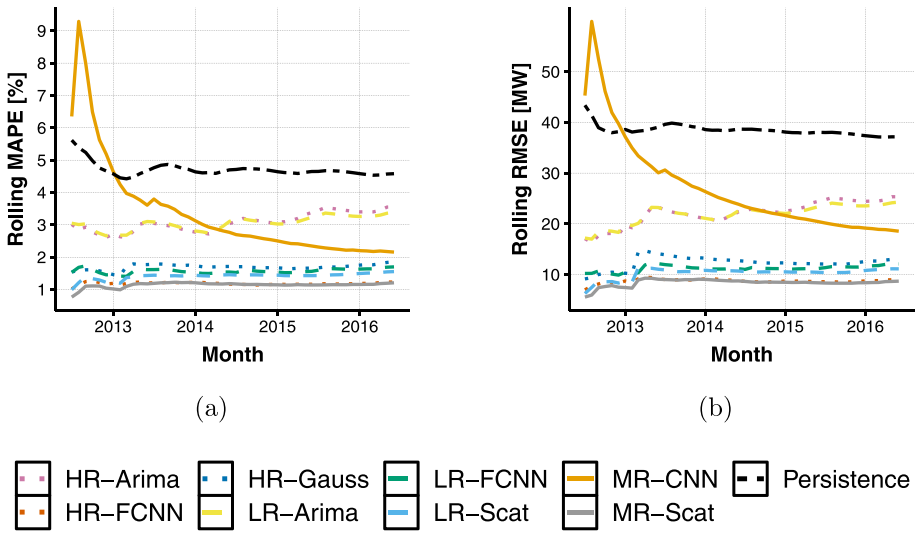

Fig. 5. Cumulative evolution of forecasting metrics for each of the monthly updated DP models: (a) MAPE, (b) RMSE. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

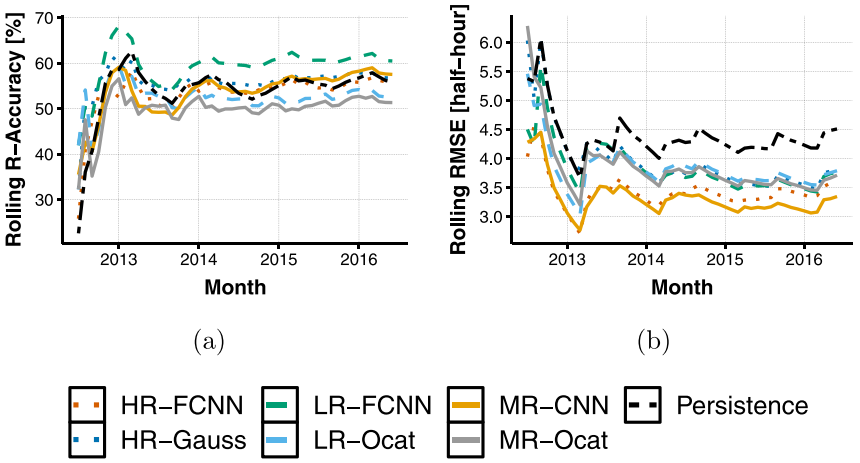

Fig. 6. Cumulative evolution of forecasting metrics for each of the monthly updated IP models: (a) R-accuracy, (b) RMSE. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

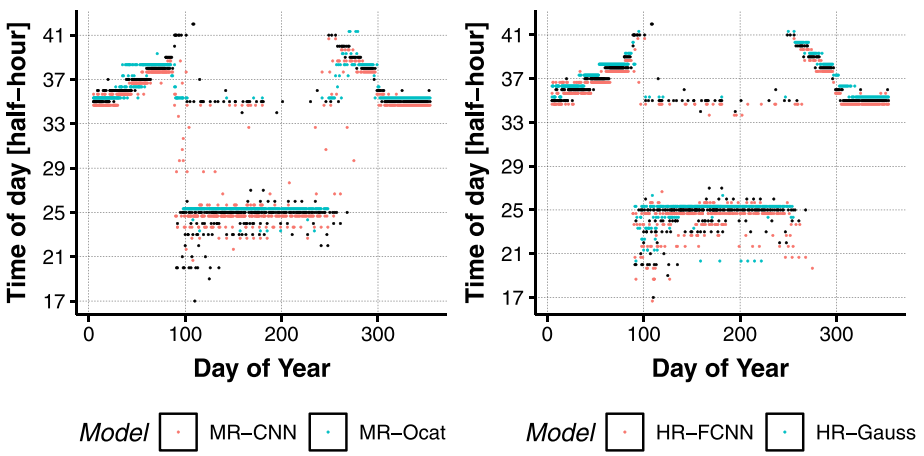

Fig. 7. Left: observed IP as a function of the day of year (black) and corresponding predictions from MR-CNN (red, shifted downward for visibility) and MR-ocat (blue, shifted upward). Right: same plot for HR-FCNN (red, downward) and HR-Gauss (blue, upward). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

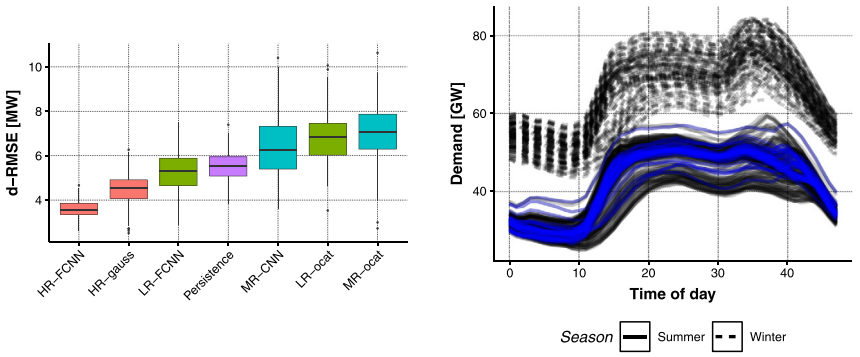

Fig. 8. Left: Block-bootstrap boxplots of the d-RMSE metric for the IP problem. Right: daily demand profile curves during winter (shifted upward by 15 GW) and summer. The blue curves are profiles with a small absolute difference between the morning and evening peaks (<50 MW). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

corresponding forecast. We propose the following metric:

$$\text{d-RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} (y_{t_i^{\text{m}}} - y_{\hat{t}_i^{\text{m}}}) \right)^{1/2} \tag{12}$$

which is based on the difference between the daily peak demand and the demand at the predicted IP (where 'd' stands for demand). This metric is more relevant to operations than the MSE. For instance, in peak shaving applications, providing a forecast $\hat{t}_i^{\text{m}}$ very different from t_i^{m} might not be a problem if $y_{t_i^{\text{m}}}$ and $y_{\hat{t}_i^{\text{m}}}$ are similar, which is what d-RMSE quantifies. Fig. 8 shows a bootstrapped boxplot of the d-RMSE for each model. Interestingly, high-resolution methods are best here, by a substantial margin in the case of HR-FCNN.

The results obtained so far do not provide reliable evidence in favour or against the adoption of a multi-resolution approach for IP forecasting. In fact, the poor forecasting performance of MR-ocat is arguably attributable to the particular ordered-logit parametrisation used here. MR-CNN does well using standard, statistically motivated losses, but it is inferior to high-resolution approaches on an operationally relevant one (d-RMSE). It would be interesting to verify whether fitting the MR-CNN model by minimising the d-RMSE directly (rather than the MSE, as done here) would lead to better results. We leave this, and the search for a more flexible distribution for ordered categorical responses, for future work.

Implementing the multi-resolution approach on the DP forecasting problem is more straightforward. Hence, the results discussed so far are positive and reliable. We further verify their significance by performing Diebold and Mariano (1995) (DM) tests on the absolute and squared error losses. The null hypothesis of the tests is: 'both forecasts have the same expected loss'. The results of the DM tests are available in Fig. 9, which confirms that, within the GAM class, the multi-resolution forecasts are significantly different to the low-resolution and high-resolution approaches under both metrics.

It is interesting to quantify the complexity or parsimony of the models considered so far. The Akaike information criterion (AIC) can be interpreted as a measure of parsimony, but it requires computing the effective number of model parameters, and we are not aware of any method that would allow estimating them across all the model classes considered here. Fig. 10 shows the AICs of low- and multi-resolution GAMs. The multi-resolution approaches have a consistently smaller AIC than the low-resolution approaches. Furthermore, the slopes indicate that with more data the gap continues to increase.

For NNs, parsimony is highly dependent on the chosen architecture. In our case, the low-resolution and high-resolution NNs have a very similar architecture, with only one hidden layer and a dropout layer. Only the inputs, outputs, and number of observations vary. On the other hand, the multi-resolution NN (Fig. 3) requires the use of convolutional layers which are leveraged to extract the high-resolution information. The extraction process requires multiple layers, which forces the multi-resolution CNN to have a larger number of parameters than the low-resolution and high-resolution NNs.

The results discussed in this section show that multi-resolution approaches are superior to low- and high-resolution alternatives for the DP forecasting problem. The forecasting performance of the high-resolution FCNN and the multi-resolution GAMs is not significantly different but, in an operational peak demand forecasting context, the multi-resolution GAM would be preferred because it can be decomposed into additive components, which can be more easily interpreted (and manually adjusted) by operational staff. In addition, note that adopting a multi-resolution approach can bring substantial computational advantages, which are easy to quantify within the GAM model class. In particular, the GAM model matrix $\mathbf{X}$ in the multi-resolution case has T times fewer rows than in the high-resolution case, where T is the number of daily observations (i.e. $T = 48$ for half-hourly data). Therefore, T times less memory is used, and many computations frequently required during GAM model fitting (such as $\mathbf{X}^T\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is a diagonal matrix) will take less time.

6. Conclusion

This paper proposed a novel modelling approach that uses both high-resolution and low-resolution information to forecast the daily electrical load peak magnitude and timing. The results demonstrate that this multi-resolution

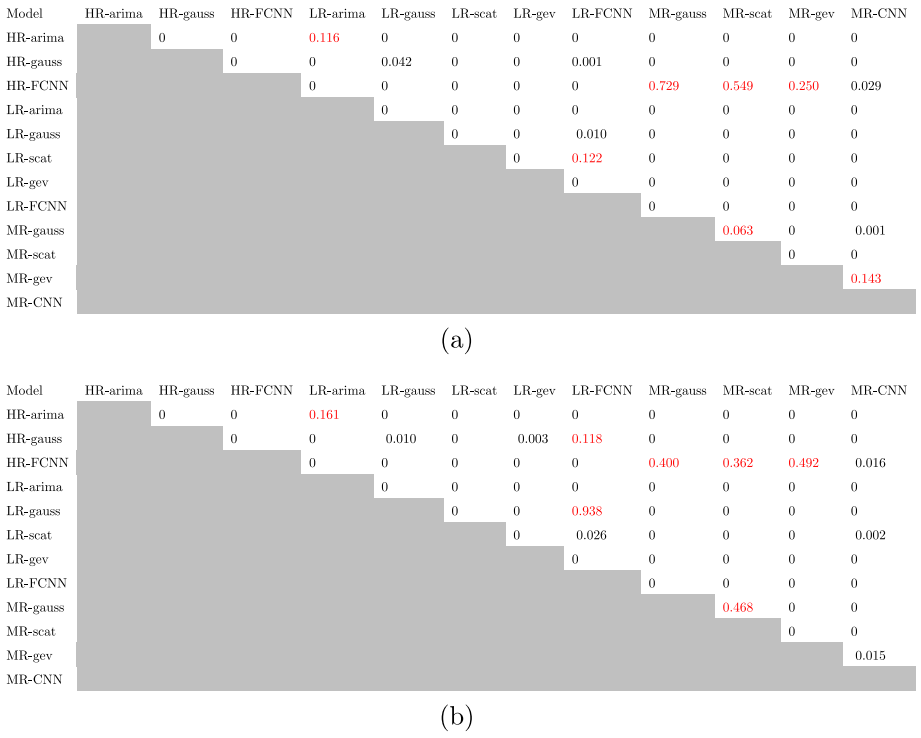

Fig. 9. *P*-values from the Diebold–Mariano test for DP forecasts. The test used is from the *multDM* package in R (Drachal, 2020). In black, the null hypothesis is rejected at the 5% threshold, and both forecasts are significantly different. In red, the null hypothesis is not rejected at the 5% threshold, and both forecasts cannot be significantly differentiated; (a) absolute errors (b) squared errors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

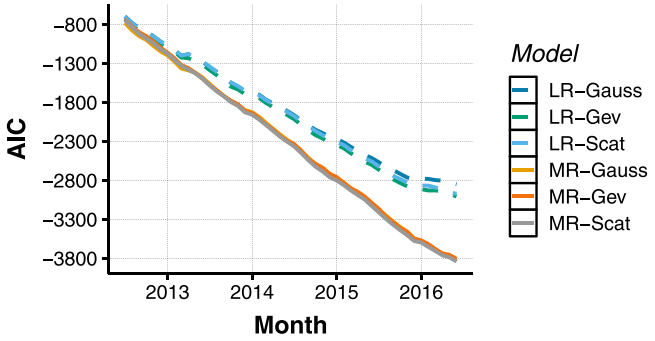

Fig. 10. AIC for the low-resolution and multi-resolution DP GAMs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

approach is flexible enough to be applied to different model classes and that it provides competitive predictive performance. In particular, GAMs and NNs with similar input structures were used to implement the multi-resolution approach and to compare its performance to that of low-resolution, high-resolution, and persistence alternatives. On UK aggregate demand data, the multi-resolution models performed significantly better across all metrics when forecasting peak magnitude. In addition to improved predictions, adopting a multi-resolution approach enables faster computation via data compression and leads to more parsimonious models, as demonstrated by the consistently lower AIC scores achieved by multi-resolution models within the GAM model class.

The results on the peak timing forecasting problem are mixed, but interesting. A multi-resolution neural network does marginally better than the alternatives when performance is assessed via standard statistical metrics. However, the corresponding forecast is occasionally inappropriate (falling between the morning and evening peaks) and inferior to high-resolution alternatives when assessed via an operationally motivated metric. The results suggest that the multi-resolution neural network should be fitted to data by minimising a problem-specific performance metric directly. For instance, one could consider financial metrics on billing periods, as done by Saxena et al. (2019). The multi-resolution GAM does poorly on the peak timing problem, but this is attributable to

the insufficient flexibility of the ordered logit parametrisation used here. Obtaining stronger evidence in favour or against the use of multi-resolution methods for the peak timing problem would require solving the issues just mentioned, which could be the subject of further work.

The forecasting methods presented here could be extended in several ways. Firstly, the set of models described in this paper could be used within an aggregation of experts or ensemble methods, which might lead to more accurate forecasts. Secondly, the benefits of multi-resolution methods were demonstrated in a context where the covariates were available at different temporal resolutions, but the underlying idea could be generalised to other settings, such as spatiotemporal data or individual customer data (see Fasiolo, Wood, Zaffran, Nedellec, & Goude, 2020 for an example application of functional quantile GAMs to residential electricity demand data). Thirdly, the multi-resolution approach could be extended to RNNs and in particular LSTMs, which are known to be powerful NN architectures for time series modelling. Finally, this paper focused on day-ahead daily peak magnitude and time forecasting, but multi-resolution methods could be applied to other short-term windows (e.g. weekly). However, estimating monthly or yearly peaks would require a different approach, because the number of observed demand peaks would be too low.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

One of the co-authors of the paper, Matteo Fasiolo, was partially funded by EPSRC grant EP/N509619/1.

Appendix A. Supplementary data

The raw datasets used in this paper are available on the National Grid (2021) and National Oceanic and Atmospheric Administration (2021) websites. The R code as well as the data prepared for the experiments in this paper are available at the following link: https://github.com/yvenn-amara.

References

Abdel-Aal, R. (2006). Modeling and forecasting electric daily peak loads using abductive networks. *International Journal of Electrical Power & Energy Systems*, *28*(2), 133–141. http://dx.doi.org/10.1016/j.ijepes.2005.11.006, URL https://linkinghub.elsevier.com/retrieve/pii/S0142061505001390.

Amato, U., Antoniadis, A., De Feis, I., Goude, Y., & Lagache, A. (2021). Forecasting high resolution electricity demand data with additive models including smooth and jagged components. *International Journal of Forecasting*, *37*(1), 171–185. http://dx.doi.org/10.1016/j.ijforecast.2020.04.001, https://www.sciencedirect.com/science/article/pii/S0169207020300583.

Amin, H. U., Malik, A. S., Ahmad, R. F., Badruddin, N., Kamel, N., Hussain, M., et al. (2015). Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques. *Australasian Physical & Engineering Sciences in Medicine*, *38*(1), 139–149. http://dx.doi.org/10.1007/s13246-015-0333-x, URL http://link.springer.com/10.1007/s13246-015-0333-x.

Amin-Naseri, M., & Soroush, A. (2008). Combined use of unsupervised and supervised learning for daily peak load forecasting. *Energy Conversion and Management*, *49*(6), 1302–1308. http://dx.doi.org/10.1016/j.enconman.2008.01.016, URL https://linkinghub.elsevier.com/retrieve/pii/S0196890408000174.

Antoniadis, A., Paparoditis, E., & Sapatinas, T. (2006). A functional wavelet-kernel approach for time series prediction. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *68*(5), 837–857. http://dx.doi.org/10.1111/j.1467-9868.2006.00569.x, URL http://doi.wiley.com/10.1111/j.1467-9868.2006.00569.x.

Boano-Danquah, J., Sigauke, C., & Kyei, K. A. (2020). Analysis of extreme peak loads using point processes: An application using South African data. *IEEE Access*, *8*, 146105–146115. http://dx.doi.org/10.1109/ACCESS.2020.3015259, URL https://ieeexplore.ieee.org/document/9163096/.

Cheng, J., Wang, Z., & Pollastri, G. (2008). A neural network approach to ordinal regression. In *2008 IEEE international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1279–1284). Hong Kong, China: IEEE, http://dx.doi.org/10.1109/IJCNN.2008.4633963, URL http://ieeexplore.ieee.org/document/4633963/.

Cho, H., Goude, Y., Brossat, X., & Yao, Q. (2013). Modeling and forecasting daily electricity load curves: a hybrid approach. *Journal of the American Statistical Association*, *108*(501), 7–21. http://dx.doi.org/10.1080/01621459.2012.722900, http://www.tandfonline.com/doi/abs/10.1080/01621459.2012.722900.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*, 303–314.

Dash, P., Liew, A., & Rahman, S. (1995). Peak load forecasting using a fuzzy neural network. *Electric Power Systems Research*, *32*(1), 19–23. http://dx.doi.org/10.1016/0378-7796(94)00889-C, URL https://linkinghub.elsevier.com/retrieve/pii/037877969400889C.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263, URL http://www.jstor.org/stable/1392185.

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Drachal, K. (2020). multDM: Multivariate version of the Diebold-Mariano test. URL https://CRAN.R-project.org/package=multDM.

El-Attar, E. E., Goulermas, J. Y., & Wu, Q. H. (2009). Forecasting electric daily peak load based on local prediction. In *2009 IEEE power & energy society general meeting* (pp. 1–6). Calgary, Canada: IEEE, http://dx.doi.org/10.1109/PES.2009.5275587, URL http://ieeexplore.ieee.org/document/5275587/.

Elamin, N., & Fukushige, M. (2018). Quantile regression model for peak load demand forecasting with approximation by triangular distribution to avoid blackouts. *International Journal of Energy Economics and Policy*, *8*(5), 119–124.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., & Goude, Y. (2020). Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 1–11. http://dx.doi.org/10.1080/01621459.2020.1725521, https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1725521.

Gao, J., Guo, Y., & Wang, Z. (2017). Matrix neural networks. In F. Cong, A. Leung, & Q. Wei (Eds.), *Advances in neural networks – ISNN 2017* (pp. 313–320). Cham: Springer International Publishing.

Gibbons, C., & Faruqui, A. (2014). Quantile regression for peak demand forecasting. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.2485657, URL http://www.ssrn.com/abstract=2485657.

Guo, Z., Zhou, K., Zhang, X., & Yang, S. (2018). A deep learning model for short-term power load and probability density forecasting. *Energy*, *160*, 1186–1200. http://dx.doi.org/10.1016/j.energy.2018.07.090, URL https://linkinghub.elsevier.com/retrieve/pii/S0360544218313872.

Hastie, T., & Tibshirani, R. (1999). *Generalized additive models*. Boca Raton, Fla: Chapman & Hall/CRC.

Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, *32*(3), 914–938. http://dx.doi.org/10.1016/j.ijforecast.2015.11.011, URL https://linkinghub.elsevier.com/retrieve/pii/S0169207015001508.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257. http://dx.doi.org/10.1016/0893-6080(91)90009-T, URL https://linkinghub.elsevier.com/retrieve/pii/089360809190009T.

Hyndman, R. J., & Fan, S. (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, *25*(2), 1142–1153. http://dx.doi.org/10.1109/TPWRS.2009.2036017, URL http://ieeexplore.ieee.org/document/5345698/.

Hyndman, R. J., & Khandakar, Y. (2007). Automatic time series forecasting: the forecast package for R. *Technical report*, (p. 31). Clayton VIC, Australia: Monash University, Department of Econometrics and Business Statistics.

Ibrahim, I. A., & Hossain, M. J. (2020). LSTM neural network model for ultra-short-term distribution zone substation peak demand prediction. In *2020 IEEE power & energy society general meeting (PESGM)* (pp. 1–5). Montreal, QC, Canada: IEEE, http://dx.doi.org/10.1109/PESGM41954.2020.9281973, URL https://ieeexplore.ieee.org/document/9281973/.

International Energy Agency (2019). *Global EV outlook 2019: scaling-up the transition to electric mobility*. OECD, http://dx.doi.org/10.1787/35fb60bd-en, URL https://www.oecd-ilibrary.org/energy/global-ev-outlook-2019_35fb60bd-en.

Jacob, M., Neves, C., & Vukadinović Greetham, D. (2020). *Forecasting and assessing risk of individual electricity peaks*. Cham: Springer International Publishing, Springer, Imprint, URL https://link.springer.com/book/10.1007/978-3-030-28669-9, OCLC: 1141446837.

Kim, D.-H., Lee, E.-K., & Qureshi, N. B. S. (2020). Peak-load forecasting for small industries: A machine learning approach. *Sustainability*, *12*(16), 6539. http://dx.doi.org/10.3390/su12166539, URL https://www.mdpi.com/2071-1050/12/16/6539.

Kuster, C., Rezgui, Y., & Mourshed, M. (2017). Electrical load forecasting models: A critical systematic review. *Sustainable Cities and Society*, *35*, 257–270. http://dx.doi.org/10.1016/j.scs.2017.08.009, URL https://linkinghub.elsevier.com/retrieve/pii/S2210670717305899.

Li, J., Zhao, Q., Wang, H., Wang, W., Yang, Y., & Yan, C. (2018). Analysis of deep learning control strategy about peak load regulation and frequency regulation with distribution thermal storage electric boiler. In *2018 5th IEEE international conference on cloud computing and intelligence systems (CCIS)* (pp. 461–464). Nanjing, China: IEEE, http://dx.doi.org/10.1109/CCIS.2018.8691145, URL https://ieeexplore.ieee.org/document/8691145/.

McSharry, P., Bouwman, S., & Bloemhof, G. (2005). Probabilistic forecasts of the magnitude and timing of peak electricity demand. *IEEE Transactions on Power Systems*, *20*(2), 1166–1172. http://dx.doi.org/10.1109/TPWRS.2005.846071, URL http://ieeexplore.ieee.org/document/1425617/.

Messner, J. W., Pinson, P., Browell, J., Bjerregård, M. B., & Schicker, I. (2020). Evaluation of wind power forecasts—An up-to-date view. *Wind Energy*, *23*(6), 1461–1481. http://dx.doi.org/10.1002/we.2497, http://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1002/we.2497, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/we.2497.

National Grid (2021). ESO data portal: Home | national grid electricity system operator. URL https://data.nationalgrideso.com/.

National Oceanic and Atmospheric Administration (2021). National oceanic and atmospheric administration | data discovery portal. URL https://data.noaa.gov/datasetsearch/.

Peterson, B., & Harrell Jr, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *39*(2), 205–217.

Saini, L. M. (2008). Peak load forecasting using Bayesian regularization, resilient and adaptive backpropagation learning based artificial neural networks. *Electric Power Systems Research*, *78*(7), 1302–1310. http://dx.doi.org/10.1016/j.epsr.2007.11.003, URL https://linkinghub.elsevier.com/retrieve/pii/S0378779607002258.

Saini, L., & Soni, M. (2002). Artificial neural network based peak load forecasting using Levenberg–Marquardt and quasi-Newton methods. *IEE Proceedings-Generation, Transmission and Distribution*, *149*(5), 578. http://dx.doi.org/10.1049/ip-gtd:20020462, URL https://digital-library.theiet.org/content/journals/10.1049/ip-gtd_20020462.

Saxena, H., Aponte, O., & McConky, K. T. (2019). A hybrid machine learning model for forecasting a billing period's peak electric load days. *International Journal of Forecasting*, *35*(4), 1288–1303. http://dx.doi.org/10.1016/j.ijforecast.2019.03.025, URL https://linkinghub.elsevier.com/retrieve/pii/S016920701930144X.

Sigauke, C., & Chikobvu, D. (2010). Daily peak electricity load forecasting in South Africa using a multivariate non-parametric regression approach. *ORiON*, *26*(2), http://dx.doi.org/10.5784/26-2-89, URL http://orion.journals.ac.za/pub/article/view/89.

Sigauke, C., & Chikobvu, D. (2011). Prediction of daily peak electricity demand in South Africa using volatility forecasting models. *Energy Economics*, *33*(5), 882–888. http://dx.doi.org/10.1016/j.eneco.2011.02.013, URL https://linkinghub.elsevier.com/retrieve/pii/S0140988311000491.

Soman, A., Trivedi, A., Irwin, D., Kosanovic, B., McDaniel, B., & Shenoy, P. (2020). Peak forecasting for battery-based energy optimizations in campus microgrids. In *Proceedings of the eleventh ACM international conference on future energy systems* (pp. 237–241). Virtual Event Australia: ACM, http://dx.doi.org/10.1145/3396851.3397751, URL https://dl.acm.org/doi/10.1145/3396851.3397751.

Uddin, M., Romlie, M. F., Abdullah, M. F., Abd Halim, S., Abu Bakar, A. H., & Chia Kwang, T. (2018). A review on peak load shaving strategies. *Renewable and Sustainable Energy Reviews*, *82*, 3323–3332. http://dx.doi.org/10.1016/j.rser.2017.10.056, URL https://linkinghub.elsevier.com/retrieve/pii/S1364032117314272.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *65*(1), 95–114. http://dx.doi.org/10.1111/1467-9868.00374, URL http://doi.wiley.com/10.1111/1467-9868.00374.

Wood, S. (2017). *Generalized additive models: an introduction with R* (2nd ed.). Chapman and Hall/CRC, http://dx.doi.org/10.1201/9781315370279, URL https://www.taylorfrancis.com/books/9781498728348.

Wood, S. (2020). Mgcv: Mixed GAM computation vehicle with automatic smoothness estimation. URL https://CRAN.R-project.org/package=mgcv.

Wood, S., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. http://dx.doi.org/10.1080/01621459.2016.1180986, URL https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1180986.

Yang, Y., Hong, W., & Li, S. (2019). Deep ensemble learning based probabilistic load forecasting in smart grids. *Energy*, *189*, Article 116324. http://dx.doi.org/10.1016/j.energy.2019.116324, URL https://linkinghub.elsevier.com/retrieve/pii/S0360544219320195.

Yang, Y., Li, W., Gulliver, T. A., & Li, S. (2020). Bayesian deep learning-based probabilistic load forecasting in smart grids. *IEEE Transactions on Industrial Informatics*, *16*(7), 4703–4713. http://dx.doi.org/10.1109/TII.2019.2942353, URL https://ieeexplore.ieee.org/document/8844831/.

Yu, Z., Niu, Z., Tang, W., & Wu, Q. (2019). Deep learning for daily peak load forecasting–a novel gated recurrent neural network combining dynamic time warping. *IEEE Access*, *7*, 17184–17194. http://dx.doi.org/10.1109/ACCESS.2019.2895604, URL https://ieeexplore.ieee.org/document/8629072/.