ARTICLE SUBMITTED TO JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY

Probabilistic Load Forecasting Using Post-Processed Weather Ensemble Predictions

Nicole Ludwig[a,b], Siddharth Arora[c,d], and James W. Taylor[c]

[a]Machine Learning Cluster of Excellence, University of Tübingen, Tübingen, Germany; [b]Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany; [c]Saïd Business School, University of Oxford, Oxford, UK; [d]Mathematical Institute, University of Oxford, Oxford, UK

ARTICLE HISTORY
Compiled July 19, 2022

ABSTRACT
Probabilistic forecasting of electricity demand (*load*) facilitates the efficient management and operations of energy systems. Weather is a key determinant of load. However, modelling load using weather is challenging because the relationship cannot be assumed to be linear. Although numerous studies have focused on load forecasting, the literature on using the uncertainty in weather while estimating the load probability distribution is scarce. In this study, we model load for Great Britain using weather ensemble predictions, for lead times from one to six days ahead. A weather ensemble comprises a range of plausible future scenarios for a weather variable. It has been shown that the ensembles from weather models tend to be biased and underdispersed, which requires that the ensembles are post-processed. Surprisingly, the post-processing of weather ensembles has not yet been employed for probabilistic load forecasting. We post-process ensembles based on: (1) ensemble model output statistics: to correct for bias and dispersion errors by calibrating the ensembles, and (2) ensemble copula coupling: to ensure that ensembles remain physically consistent scenarios after calibration. The proposed approach compares favourably to the case when no weather information, raw weather ensembles or post-processed ensembles without ensemble copula coupling are used during the load modelling.

KEYWORDS
Probabilistic forecasting; electricity demand; weather predictions; ensemble model output statistics; ensemble copula coupling.

1. Introduction

Grid operators of electric utilities rely on accurate load forecasts to make informed decisions regarding electricity transmission and distribution. Over the past few years, modelling load has become increasingly challenging with the advancements in low carbon technologies, integration of renewable energy, and growth in unmetered and distributed small-scale renewable generation sources, all of which introduce more volatility into the energy system. Moreover, load and renewable energy supplies vary with

NL: nicole.ludwig@uni-tuebingen.de
SA: siddharth.arora@maths.ox.ac.uk
JWT: James.Taylor@sbs.ox.ac.uk

weather seasons (Taylor, 2003), and load also depends on human behaviour, which is often rather stochastic. To cope with these uncertainties, probabilistic forecasting of load at different hierarchies of the energy system has garnered attention (Arora & Taylor, 2016; Guo et al., 2018; Haben et al., 2019; Taieb et al., 2020; Taylor & Buizza, 2002; Taylor & Buizza, 2003; van der Meer et al., 2018). Probabilistic forecasts aim to quantify the uncertainty in the form of probability distributions over possible future events, which allows for informed decision-making compared to the case when only a point forecast is communicated. In the context of this study, probabilistic load forecasts are of particular interest for a range of energy applications, including reliability planning (Billinton & Huang, 2008), probabilistic load flow (Chen et al., 2008), stochastic unit commitment (Wang et al., 2011), and probabilistic energy price forecasting (Nowotarski & Weron, 2018).

Load at the national level exhibits prominent variability, due largely to periodic cyclicality and variations in weather patterns. While some studies do not use explicit weather information at all (Hu, 2017; Hu & Jiang, 2017), it is imperative, when modelling load in terms of weather as the basis for probabilistic load forecasting, to propagate the uncertainty from the weather variables through the load forecasting model (see, for example, Haupt et al., 2019). This can be achieved through the use of weather ensemble predictions generated from Numerical Weather Prediction (NWP) models.

In recent years, NWP models have become state-of-the-art in meteorology, with modern computing power allowing complex physical models to be run at a high resolution. These weather models describe the atmospheric processes using first principles, and, due to their nonlinearity and complexity, are solved with numerical approximations (Al-Yahyai et al., 2010). The outcome of a NWP model highly depends on the initial state of the atmosphere and the model's physical processes. To quantify these sources of uncertainty, the NWP model is run several times, with different initial conditions and/or a differently parameterized physical representation of the atmosphere. Each run of the NWP model provides a different scenario for the future of the weather variable, which is referred to as an ensemble member. Overall, the weather ensemble prediction encapsulates the degree of uncertainty in weather variables.

Tremendous advancements have been made in the area of NWP over the past few years. It has been shown that the forecast skill for lead times from three to 10-days ahead has been increasing by around one day per decade (Bauer et al., 2015). The improvements in weather predictions have primarily been attributed to progress in: (1) modelling the physical process - a detailed representation of the atmosphere, (2) ensemble forecasting - which encapsulates the uncertainty in initial conditions and model processes for a nonlinear complex system, and (3) model initialization - deriving the current state of the atmosphere and Earth's surface based on four-dimensional variational (4D-Var) data assimilation techniques that have been described as a major milestone in the field. For details, see Bauer et al. (2015) and Alley et al. (2019).

Unfortunately, raw ensemble predictions obtained from the NWP models are subject to underdispersion and bias. To deal with these shortcomings, statistical post-processing methods have been proposed for calibrating the weather ensembles (e. g. Baran & Lerch, 2018; Ben Bouallègue et al., 2016; Feldmann et al., 2015; Möller et al., 2015; Scheuerer & Buermann, 2014). Of these, two of the most commonly used methods for post-processing include Bayesian Model Averaging (see Raftery et al., 2005) and Ensemble Model Output Statistics (EMOS), originally also known as non-homogeneous Gaussian regression (NGR) (see Gneiting et al., 2007). Both of these post-processing methods have been shown to substantially improve the accuracy of

NWP ensemble predictions (Hagedorn et al., 2012; Wilks & Hamill, 2007).

It is worth noting that the raw ensemble outputs from the NWP model represent a multivariate dependence structure, as specified by the model equations. If the weather ensembles are treated as being independent during the post-processing, we end up with several univariate and independent distributions, which may be practically unrealistic. It is thus important to take into account the temporal dependencies, as well as dependencies between the weather variables during ensemble post-processing (Hu et al., 2016; Schefzik et al., 2013). This problem is well known in the meteorological literature, and studies have proposed methods to create dependent distributions via empirical copula approaches, most notably the Schaake shuffle (see Clark et al., 2004) and different forms of Ensemble Copula Coupling (ECC) (Schefzik et al., 2013).

Advancements in the NWP models over the years have, unfortunately, not adequately translated into more accurate modelling of load. In the energy forecasting literature, studies using ensemble weather information as model input are rare, although e. g. Al-Yahyai et al. (2010) show that NWPs are superior to station-based weather information. Even when weather ensemble predictions are used for forecasting in a range of diverse applications such as: e. g. wind power (Gensler, 2019; Heppelmann et al., 2015; Heppelmann et al., 2017; Nielsen et al., 2004), wind ramp events (Bossavy et al., 2013), solar power plant output (Thorey et al., 2018; Zamo et al., 2014) or load (Taylor & Buizza, 2002; Taylor & Buizza, 2003), the need for calibration and maintaining the dependency structures among the weather ensemble predictions is neglected, which is a major limitation in the modelling. Only Heppelmann et al. (2017) use post-processing and an approach to capture the dependencies, however, they investigate probabilistic wind power forecasts.

This study aims to bridge the gap between the field of meteorology (focusing on weather ensemble predictions from NWP models) and energy modelling by proposing and implementing the following set of best practices: (1) calibrating the raw weather ensemble predictions to correct for biases and dispersion errors; (2) maintaining the temporal and multivariate dependencies between the calibrated ensemble members; and (3) using the post-processed weather ensemble predictions to estimate the forecast distribution of load (as opposed to just producing a point estimate or a set of pre-specified discrete scenarios of load). Our post-processing comprises a two-stage approach, in the first stage we use EMOS for calibration, and in the second stage, we use ECC to ensure that the calibrated weather ensemble predictions remain physically consistent scenarios. Moreover, we are essentially revising the approach taken by Taylor and Buizza (2002) and Taylor and Buizza (2003), who employed raw weather ensemble predictions to generate multiple scenarios of load, which were then post-processed. We compare the out-of-sample load forecast accuracy obtained using the raw weather ensemble predictions (current practice) versus using post-processed ensembles (as proposed in this study). To the best of our knowledge, this is the first study that employs ensemble post-processing for probabilistic load forecasting.

The remainder of the paper is structured as follows. We introduce our data in Section 2 and describe the post-processing in Section 3. We present our forecasting methodology in Section 4. We evaluate the point and probabilistic forecast accuracy in Section 5 and finally, we conclude in Section 6.

2. National Load and Weather in Great Britain

In this section, we first describe the data for load followed by weather ensemble predictions.

2.1. Load Data

We employ load for Great Britain (GB), from 2006 to 2017, inclusive. The data is sampled every hour and obtained from National Grid (NG), the company responsible for the transmission of electricity in GB. As evident from Figure 1, load exhibits a prominent recurring within-year pattern (intrayear seasonality), whereby the demand is higher in winter than in summer. The national load is a summary of all flows on the transmission grid in GB, and therefore, it does not capture all distributed energy sources. With the recent growth in unmetered and small-scale renewable generation sources, this form of measuring national demand has resulted in an overall downward trend and increased variability in the load time series.

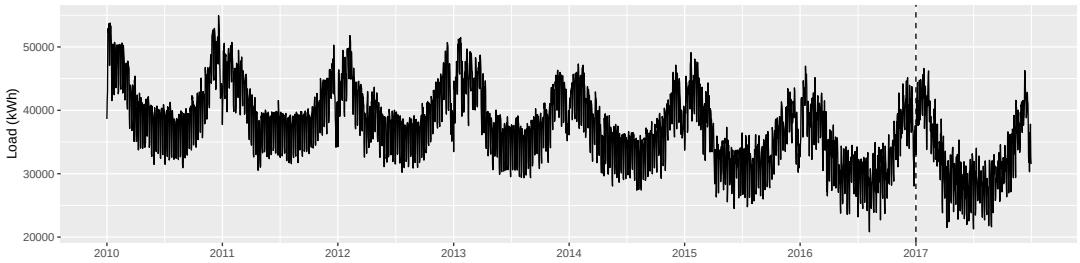

Figure 1. Load observed at midday in Great Britain for the period 1 January 2006 to 31 December 2017. The vertical dashed line denotes the non-overlapping split of the load time series into an in-sample data (2006-2016) and an out-of-sample data (2017).

Figure 2 presents the average daily load profiles. It can be seen that load is higher during typical working hours of the day and late evenings, and load on weekends is usually lower than on working days. Moreover, load exhibits a recurring within-week and within-day pattern. Load is overall lower on special days (such as public holidays) and proximity days (days adjacent or close to a public holiday that are not public holidays) compared to normal working days. Previous studies have typically focused on modelling the load for normal days while ignoring load on special days (Taylor & Buizza, 2002; Taylor & Buizza, 2003). Accommodating the special and proximity day effects during the modelling has been shown to result in improved load forecast accuracy across all days in the out-of-sample period (Arora & Taylor, 2018). We thus model load observed during both normal and anomalous periods.

In this study, we focus on modelling the load observed at midday. This is particularly relevant as the peak demand during summer months occurs around midday in GB. We use the first 10 years of data to train the model before testing it on the last available year 2017 [1]. The last year of the training data is used as the cross-validation hold-out sample. We generate forecasts by rolling the forecast origin through each midday in the one-year out-of-sample period. We identified 18 special and proximity days in the out-of-sample data. Although we focus on generating multi-step ahead load predictions for

[1] Although load data for 2018 is available, the corresponding weather data was not complete at the time, and we thus restricted the analysis to the complete data set we could obtain from both sources.

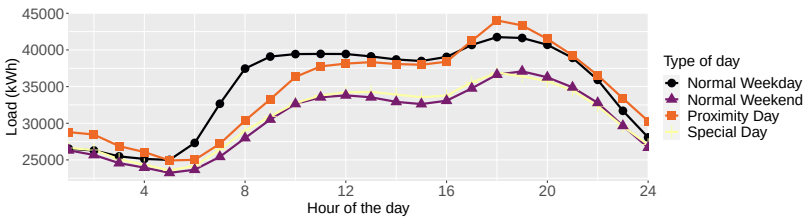

Figure 2. Average daily load profile for the following different day types: normal weekday, normal weekend, proximity day, and special day.

midday, the proposed methodology could easily be adapted for forecasting load across all periods of the day.

2.2. Weather Ensemble Predictions

We employ the actual weather data and corresponding ensemble predictions from the NWP model of the European Centre for Medium-Range Weather Forecasts (ECMWF). The actual weather is represented by the reanalysis data, while the uncertainty in the weather forecasts is represented by the ensemble forecasts. Weather predictions from the ECMWF comprise 51 ensemble members, whereby 50 ensemble members are constructed by perturbing the initial conditions and/or model processes, and one ensemble member is generated using the best estimate of the initial condition/process (CNT: ensemble control member). We additionally use the high resolution deterministic forecast (HRES) of the ECMWF in our ensemble. We employ ensemble predictions for the following weather variables: temperature (at 2m above ground), wind speed (perpendicular u- and v-components at 10m above ground) and total cloud cover. The wind speed is calculated from the u- and v-components with windspeed $= \sqrt{\mathrm{u}^2 + \mathrm{v}^2}$.

While high-resolution spatial weather data is available in a grid format from the ECMWF, we use only weather data for the seven GB locations considered by NG. These are chosen to reflect the regions with highest load. The locations are shown in Figure 3. Using the weights shown in Figure 3, a weighted average of the weather data for these locations is used by NG as input to their load forecasting models. We follow the same approach, with a key difference being that, while NG consider only point forecasts for the weather variables, we use weather ensemble predictions. We obtain the ensemble predictions from ECMWF at each of the seven locations and calculate a weighted average over each ensemble member at all locations to give one set of ensemble predictions for each time point. This weighted average set of ensemble predictions is then used in the further post-processing and forecasting steps.

To capture the influence of weather on load adequately, we use the original weather variables from the NWP and, following NG, we additionally derive two new variables, namely *effective temperature* and *cooling power of wind*. We calculate the effective temperature as done in Taylor and Buizza (2003) using

$$\mathrm{TE}_t = \frac{1}{2}\mathrm{TO}_t + \frac{1}{2}\mathrm{TE}_{t-1}.$$

where for a given period t, TO_t denotes the average spot temperature of the previous four time steps, resulting in TE_t being an exponential smoothed form of TO_t. The rationale of using a smoothed form of temperature (TE_t) is to try and accommodate the slow and gradual change in human behaviour (and resulting demand response) to

Grid Point	Weight
Heathrow	0.28
Bristol	0.18
Birmingham	0.16
Hawarden	0.14
Glasgow	0.1
Leconfield	0.07
Leeming	0.07

Figure 3. Map of Great Britain with dots indicating the grid points that were used for the extraction of weather information. The points size indicated their corresponding weight, as also summarised in the table.

changes in the outside temperature. Additionally, the *cooling power of wind* variable, is based on the idea that wind speed changes electricity consumption behaviour only if the outside temperature is below a certain threshold. We again use the definition provided by NG, and used by Taylor and Buizza (2003), with this threshold at 18.3 °C

$$\mathrm{CP}_t = \begin{cases} W_t^{\frac{1}{2}} \left(18.3 - \mathrm{TO}_t\right) & \text{if } \mathrm{TO}_t < 18.3\,^{\circ}\mathrm{C} \\ 0 & \text{if } \mathrm{TO}_t \geq 18.3\,^{\circ}\mathrm{C}, \end{cases}$$

where for a given period t, CP_t denotes the *cooling power of wind*, W_t denotes the wind speed, and TO_t represents the average temperature. These two new variables (TE_t and CP_t) have a strong correlation with load, as evident from Figure 4. It can also be seen in Figure 4 that load and temperature have a nonlinear relationship, which could potentially be approximated using an asymmetric quadratic function. The correlation between the effective temperature and load shows that the rise in load is sharper during the winter months than in the summer months. This can be attributed to the higher use of electrical equipment for heating during winter (compared to the use of cooling equipment during summer) in GB. A rise in the *cooling power of wind* is associated with an overall increase in electricity demand. Additionally, although the load is generally higher during the weekdays as compared to the weekend, the correlation is strong in both cases.

3. Weather Ensemble Post-Processing

The raw weather ensemble predictions from the ECMWF are biased (Atger, 2003; Mass, 2003) and underdispersed (Eckel & Walters, 1998; Hamill & Colucci, 1997). The weather ensemble predictions thus need to be calibrated using post-processing methods. As pointed out by Gneiting and Katzfuss (2014), ensemble calibration aims to correct for the dispersion errors and biases in raw weather ensemble predictions, with the overall goal of maximizing the sharpness of post-processed ensemble prediction distributions subject to calibration. Calibration refers to the statistical consistency between the forecast distribution and actual observations, while sharpness refers to the concentration (or spread) of the forecast distribution. Rank histograms (also known as

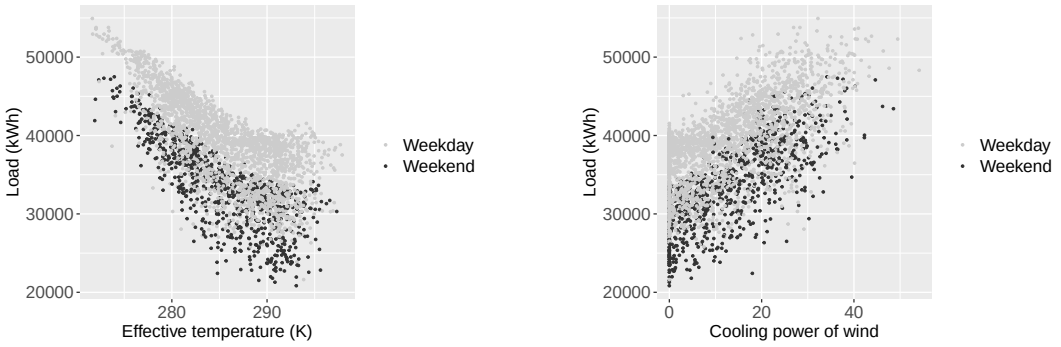

Figure 4. Scatter plot of load with effective temperature and cooling power of wind. Note: data for weekdays and weekends are denoted by grey and black dots, respectively.

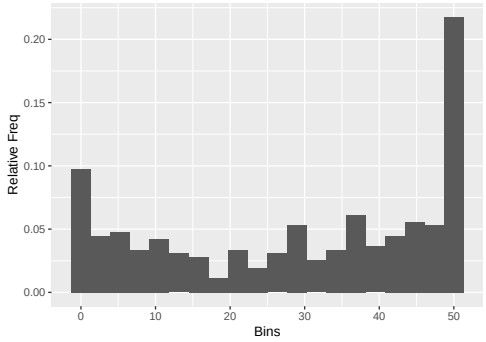

Figure 5. Rank histogram of the one-step-ahead raw temperature ensembles before any post-processing at midday.

Talagrand diagrams) can be used to assess the calibration of a probabilistic forecasting system. In an ideal forecasting system the verifying observations are equally likely to fall within any bin constructed from two ordered neighbouring ensemble members. The rank histogram distribution is thus ideally symmetric and flat with equal numbers of observations in each bin. However, while the uniform rank histogram is a necessary condition for calibration, it is not sufficient. The rank histogram of one-step-ahead raw temperature ensembles, as shown in Figure 5, is asymmetric, which indicates the presence of bias. The U-shape in Figure 5 indicates that the ensemble predictions are underdispersed. The presence of bias and dispersion errors in the weather ensemble predictions could result in a poor estimation of the load forecast distribution. To deal with this issue, we adopt a two-stage post-processing scheme. In the first stage, we calibrate the weather ensemble predictions using Ensemble Model Output Statistics (EMOS). In the second stage, we retain the multivariate dependency structures in calibrated weather ensemble predictions using Ensemble Copula Coupling (ECC). For ensemble post-processing, we use weather data for the period 1 January 2016 up to 31 December 2017, inclusive. We now describe the post-processing method.

3.1. Ensemble Model Output Statistics

EMOS addresses the issue of both bias and underdispersion in raw weather ensemble predictions (Gneiting et al., 2005). Specifically, EMOS calibrates past ensembles using

the corresponding actual historical weather data, whereby the estimated parameters (from the training data) are used for calibrating the future ensemble members. This calibration is performed by estimating a distribution for the raw weather ensemble predictions, $x_1, \ldots x_M$ with M members, using a parametric distributional regression approach. In our post-processing, we use the distributions for temperature and wind speed as proposed by Gneiting (2014). The temperature ensembles are calibrated using a normal distribution $\mathcal{N}(\mu, \sigma^2)$ and the wind speed is modelled using a truncated normal distribution $\mathcal{N}_0(\mu, \sigma^2)$, where μ and σ^2 are calculated as

$$\mu = a_0 + a_{\text{HRES}} x_{\text{HRES}} + a_{\text{CNT}} x_{\text{CNT}} + a_{\text{ENS}} \frac{1}{50} \sum_{m=1}^{50} x_m$$

and

$$\sigma^2 = b_0 + b_1 \frac{1}{50} \sum_{m=1}^{50} \left(x_m - \frac{1}{50} \sum_{m=1}^{50} x_m \right)^2 .$$

The EMOS location parameters $a >= 0$ correct for the bias in the raw weather ensemble predictions, while the scale parameters $b >= 0$ adjust the spread and potentially tackle the issue of underdispersion. Cloud cover is post-processed with a multinomial logistic regression following the approach by Baran et al. (2021), Hemri et al. (2016) using the same intervals for quantization of the forecast values in order to correspond to oktas (see Table A4). We implement the same MLR as Baran et al. (2021) and refer to their paper for more detail.

To evaluate if EMOS helps improve the calibration of our weather variables, we investigate the Probability Integral Transform (PIT) of calibrated ensembles. If F denotes a fixed, non-random predictive CDF for an observation Y, the PIT is the random variable $Z_F = F(Y)$. It is known that if F is continuous and $Y \sim F$ then Z_F is standard uniform. The PIT for the temperature variable is shown in Figure 6. For a perfectly calibrated ensemble, all bins would have the same height (as denoted by the dashed red line). Compared to the relative frequency of raw ensemble as shown earlier (Figure 5), the ensemble distribution after calibration is more uniform (Figure 6).

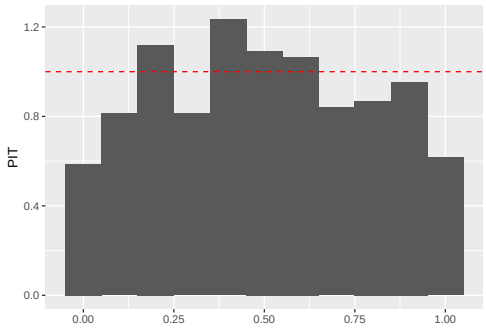

Figure 6. The probability integral transform for temperature after post-processing.

We train the EMOS parameters on the past 30 days and re-estimate every day. We tested different length of days for training (ranging from 10 to 100) and found that 30 days performed best on the test set in 2016. Figure 7 shows an example of a

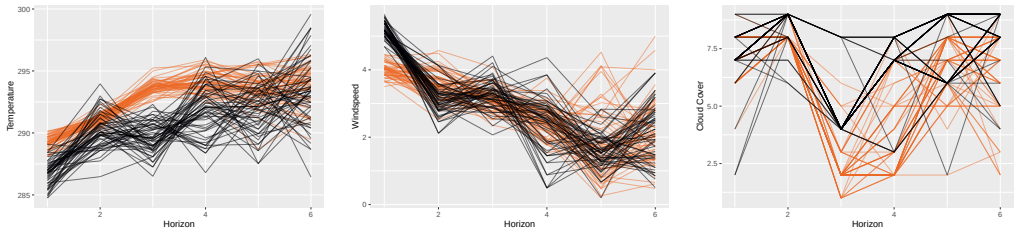

Figure 7. A trajectory of the original weather ensemble predictions (red) and the sampled realisations after post-processing (black) of a randomly selected six day period in the training data set.

post-processed ensemble trajectory over a randomly selected six day period.

3.2. Ensemble Copula Coupling

After calibrating the weather ensemble predictions based on univariate distributions, as specified using EMOS, we could potentially draw independent samples from the calibrated ensemble distributions, and use them as inputs in the load forecasting model. However, this could result in combinations of weather variables that are unlikely to happen in reality, because treating calibrated ensembles as being independent would result in a loss of the multivariate dependency structures of the raw ensembles. Thus, to ensure that the calibrated weather ensemble predictions maintain the original dependency structures, we use a reordering based ECC scheme similar to the Shaake shuffle (Schefzik et al., 2013) and the stratified sampling ECC (ECC-SS) approach by Hu et al. (2016).

ECC is based on the appropriate copula being defined in the form of a *reordering* process. The idea is that given a dependence structure "template" (Schefzik, 2017), the samples that are drawn from multiple univariate EMOS distributions can be reordered in such a way that their rank ordering resembles the rank ordering of the raw ensemble members for the same variables. The templates are based on the raw weather ensemble predictions, where we assume that the raw ensembles capture the correlations sufficiently. While several variants exist, we use a slightly adapted version of the stratified sampling ECC (ECC-SS) proposed by Hu et al. (2016).

The ECC-SS procedure essentially includes three steps. Figure 8 illustrates these steps in a scenario with six ensemble members, which are used to describe three distributions of weather variables at two different time steps. In the first step, we rank the ensemble member values. Thus, the raw ensemble members $x_1, \ldots, x_M$ with their order statistics $x_{(1)} \leq \ldots \leq x_{(M)}$ are used to generate a rank dependence structure at each time horizon via a permutation π, with $\pi(m) := \text{rank}(x_m)$ for $m \in \{1, \ldots, M\}$ (see Table (A) in Figure 8). In the second step, we impose this rank structure on the calibrated weather ensemble predictions. As we are left with a conditional distribution function following EMOS, we impose the rank order by first splitting the calibrated ensemble distribution into M equally spaced quantiles. Each quantile then represents a rank in the raw ensemble and thus M has the same size as the raw ensembles (see (B) and (C) in Figure 8). We chose the 0 and 100% quantiles to be equal to one standard deviation below the minimum and one standard deviation above the maximum, respectively. In the third step, we draw realisations $\tilde{x}$ from bins, i. e. the intervals between the quantiles to obtain more than M total samples and to efficiently sample from the tails of the distribution. In contrast to Hu et al. (2016), we do not fix the

(A)

	$t=1$			$t=2$		
	v	y	z	v	y	z
$\pi(1)$	1	1	3	4	6	1
$\pi(2)$	2	5	2	4	3	4
$\pi(3)$	5	6	6	1	4	3
$\pi(4)$	6	3	1	6	5	4
$\pi(5)$	3	4	5	5	3	5
$\pi(6)$	4	2	4	3	2	6

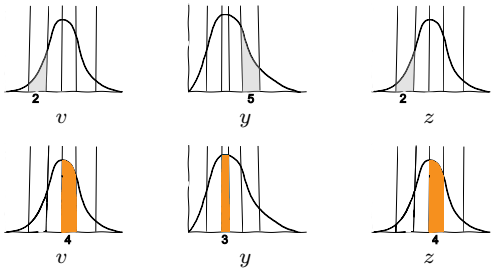

Figure 8. Symbolic explanation of ensemble copula coupling. The table (A) denotes the rank ordering of six ensemble members $x_1, \ldots, x_6$ for three weather variables (v, y, z), across two horizons $(t = 1, 2)$. The rank ordering in the table is derived from the raw ensembles (first step of ECC-SS). (B) and (C) show the EMOS output distributions at $t = 1$ and $t = 2$ respectively. To maintain the rank ordering of the six raw ensembles (as shown in A), we split the post-processed distributions into six equally spaced quantiles, whereby each quantile denotes a rank order (second step of ECC-SS). While drawing realisations from the multivariate weather distributions across different horizons $(t = 1, 2)$, we impose the original rank ordering (from A) to ensure that both the variable and temporal dependencies are maintained (third step of ECC-SS).

interval size to $\frac{1}{n}$, but use the quantiles such that the width of the bins adapt with the density. We then draw the realisations dependent on the rank structure from the raw ensembles, such that the calibrated and reordered ensemble $\hat{x}_1, \ldots, \hat{x}_M$ is given by

$$\hat{x}_1 := \tilde{x}_{(\pi(1))}, \ldots, \hat{x}_M := \tilde{x}_{(\pi(M))}.$$

Thus, we draw from the multivariate weather distributions at different time steps, while maintaining their dependency structures across both the weather variables and time.

4. Probabilistic Modelling of Load

To generate probabilistic load forecasts using post-processed weather ensemble predictions, we adopt a two-stage linear regression model. We adopt this approach as it is based on the model used in practice at the NG, and was also employed by Taylor and Buizza (2003). The first stage of the linear regression model can be described as

$$y_t = \beta_0 + \sum_{i=1}^{N} \alpha_i x_{t,i} + \sum_{j=1}^{M} \beta_j D_{t,j} + \sum_{k=1}^{K} \gamma_k C_{t,k} + \varepsilon_t,$$

where y_t is the dependent variable which in our case is the national electricity demand, $x_{t,i}$ are other variables describing the load and weather, $D_{t,j}$ are dummy variables. In our setting, the dummy variables include Friday, Saturday, Sunday, dummies for special days (e. g. public and bank holidays) and proximity days and dummy variables for the summer months (June, July, August) and winter months (December, January, February). The interaction terms $C_{t,k}$ are either between two variables or between a variable and a dummy variable, such as the interaction between temperature and wind, and the interaction between temperature and the weekend dummy.

The load variables $x_{t,i}$ can be modelled as a function of a time-specific component and a weather-specific component. For example, the time-specific component includes a counter of the day in a year, and an overall day counter for the whole data set, as well as quadratic and cubic terms of these counters. These time counters help accommodate

the seasonal patterns in load time series. The weather-specific component comprises weather variables to capture variations in load due to changing weather patterns. The model is trained on actual historical weather data for the three weather variables wind speed, temperature and cloud cover to describe the real relationship among the variables without introducing any bias through forecast values.

In the second stage, the error from the first stage is modelled with the help of autoregressive terms (using lags up to 4 weeks) in the form of

$$\varepsilon_t = a_0 + a_1\varepsilon_{t-1} + \ldots + a_{28}\varepsilon_{t-28} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2).$$

While a typical load forecasting strategy relies on using a single point forecast for each weather variable, we want to accommodate the uncertainty in weather predictions using the post-processed ensembles. Thus, the multivariate weather distributions are converted into load forecast distributions using Monte Carlo. We draw 1000 samples at each time step for each forecast horizon from the post-processed weather distributions while maintaining rank ordering (see Section 3.2) and use these as input into the linear regression model, resulting in our load forecast distribution.

Although we adopt a linear model, the transformation used for weather variables (such as, *cooling power of wind*) helps accommodate the nonlinear relationship between the weather variables and load. The objective of this study is not to compare different modelling approaches, instead, for a given well-established model (linear regression model in this case), we aim to assess the efficacy of post-processing weather ensemble predictions for probabilistic modelling of load. Thus, we compare the forecast accuracy of the linear regression model using a different set of input variables during the modelling, based on the following five alternative criteria:

Approach 1: ***No Weather*** - not incorporating any weather-related information in the modelling. A model with weather information would be expected to outperform this baseline model.

Approach 2: ***Actuals*** - using actual weather data as predictor variables. Although this information is not accessible at the time of forecasting, we use this model to provide an estimate of the upper limit on the load forecast accuracy that could theoretically be attained if perfect future weather information was available.

Approach 3: ***Raw Ensembles*** - using the raw weather ensemble predictions from the NWP models as predictor variables.

Approach 4: ***EMOS Ensembles*** - employing the post-processed (only EMOS) weather ensemble predictions.

Approach 5: ***ECC Ensembles*** - employing the post-processed (EMOS and ECC) weather ensemble predictions. This is our proposed approach.

The linear regression model (for the above five approaches) is trained using only the data from 2006 to 2016, and validated using the data from 2017. For the approaches with weather (Approach 2-5) and without weather (Approach 1) the parameters $(\beta_0, \alpha_i, \beta_j, \gamma_k)$ are estimated independently while using the same set of none weather-related variables, i. e., using the same set of dummy and load variables. In total, we consider 53 predictor variables including weather (Approach 2-5) and 47 pre-

dictor variables without weather (Approach 1) for modelling load. In ordinary least squares regression, this variety of features can lead to low predictive power and reduce model inference due to problems such as over-fitting, presence of noisy (or irrelevant) predictors, and multicollinearity. A common choice to overcome these problems is to use a regularisation technique, such as the *LASSO* (least absolute shrinkage and selection operator) (Hastie et al., 2013; Tibshirani, 1996). The LASSO regression forces the model coefficients of less salient features to go to zero. Although this shrinkage increases the bias, it improves the forecasting accuracy (Ludwig et al., 2015).

In contrast to other regularisation techniques, the LASSO technique uses an L_1 penalty term, which sets some coefficients to exactly zero (Hastie et al., 2013). The LASSO can, therefore, be used as a feature selection method. However, to efficiently use the LASSO method, the choice of the shrinkage parameter (λ) is essential. In our case, we use k-fold cross-validation on the training data set and choose λ as the largest value of λ such that the error is within one standard error of the minimum. Using LASSO, we select a total of 39 variables with non-zero coefficients in our model with weather variables (Approaches 2-5) and 34 in the model without weather variables (Approach 1). The full list of variables with non-zero coefficients (Table A2), as well as those with a coefficient of zero (Table A3), can be found in the supplementary materials.

We generate probabilistic load forecasts by rolling the forecast origin through each day in the out-of-sample period (2017). After each week, we re-estimate the regression coefficients for the linear regression model. As we stated earlier, the parameters for EMOS are estimated once using the cross-validation hold-out sample (2016).

5. Forecast Evaluation

In this section, we evaluate the out-of-sample point and probabilistic load forecast accuracy using the following three performance scores: the mean absolute percentage error (MAPE), the root mean squared error (RMSE), and the continuous ranked probability score (CRPS). While the first two error measures quantify the point forecast accuracy, with the MAPE being scale independent and the RMSE putting a heavier penalty on large deviations, the CRPS quantifies probabilistic forecast accuracy summarising calibration and sharpness. The MAPE and RMSE are defined as

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2},$$

with y_t being the actual load at time point t and $\hat{y}_t$ being the forecast load value at this time point, while N denotes the number of observations. It has been shown by Gneiting (2011) that for model evaluation based on a quadratic loss function, the mean of the density forecast is the optimal forecast. Similarly, if the evaluation is based on a symmetric piecewise linear loss function, then the optimal forecast is the median of the density forecast. Thus, for evaluation using the RMSE and MAPE, we use the mean

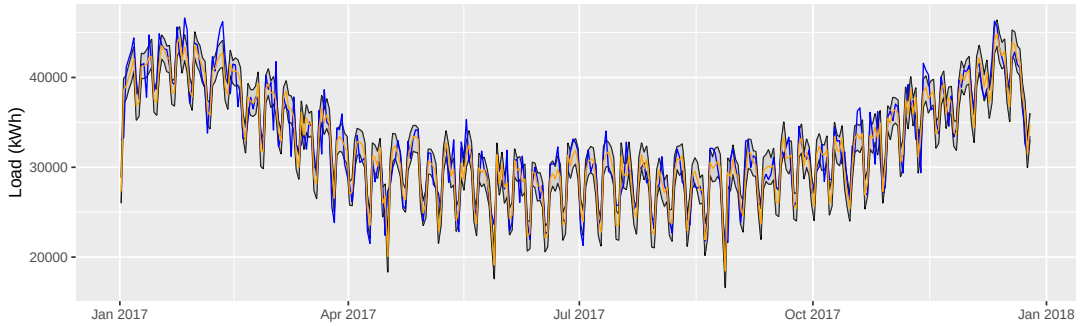

Figure 9. Probabilistic load forecasts for one-day-ahead with 95% prediction interval (grey area) and the median forecast (in orange) along with corresponding true load (in blue). Note: the forecasts were generated using the linear model with post-processed weather ensemble predictions.

and median of the distributional forecast, respectively. The CRPS is then defined as

$$\text{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\text{CRPS}(F_i, y_i),$$

where

$$\text{CRPS}(F, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 dz,$$

with F being the predictive cumulative distribution function of load, y the verifying observation and $\mathbb{1}$ denoting an indicator function. A lower value for the CRPS indicates greater probabilistic forecast accuracy. We report the average CRPS computed over all observations in the out-of-sample period. Additionally, as we want to assess whether we can improve the forecasting accuracy through post-processed ensembles, we calculate a CRPS skill score. A skill score is the percentage by which a model is more accurate than a baseline model.

Using the approach with post-processed weather ensemble predictions, we summarize the one-step-ahead probabilistic load forecasts for the out-of-sample data in Figure 9, where we plot the 95% prediction interval (grey area) and the median forecast (in orange) along with corresponding actual load observations (in blue). It is encouraging to see that the prediction intervals encapsulate the majority of actual observations. The reliability diagram across all horizons for the best performing model is shown in Figure 11. For each probability level, the reliability diagram presents the proportion of out-of-sample observations that fell below the corresponding quantile forecasts. In our case, we can see that less than 95% of the observations fall into the 95% prediction interval.

In Table 1, we present the MAPE, RMSE, CRPS and CRPS skill score for lead times ranging from one day to six days ahead. For the skill scores, we use the model with no weather information as the baseline (Approach 1: ***No Weather***). Crucially, the incorporation of weather information resulted in a substantial improvement in both the point and probabilistic load forecast accuracy across all lead times considered in this study. This result highlights the importance of using weather information in load forecasting models. Encouragingly, the model with ECC post-processed weather ensemble

Table 1. Comparison of results for different scores, models and forecast horizons for the linear regression models. The best performing model using ensemble weather information is highlighted in bold.

Model	Horizon	MAPE	RMSE	CRPS	CRPS Skill
Actuals	1	4.53	1806.16	1045.96	49.82
ECC Ensembles	1	**5.01**	**1993.27**	**1171.71**	**33.74**
EMOS Ensembles	1	5.40	2172.49	1283.95	22.05
Raw Ensembles	1	5.49	2203.80	1315.65	19.11
No Weather	1	6.71	2699.47	1567.06	0.00
Actuals	2	4.54	1798.55	1047.64	49.86
ECC Ensembles	2	**5.17**	**2065.18**	**1199.69**	**30.86**
EMOS Ensembles	2	5.51	2202.39	1285.11	22.16
Raw Ensembles	2	5.46	2200.43	1291.19	21.59
No Weather	2	6.75	2703.62	1569.95	0.00
Actuals	3	4.44	1770.47	1034.63	55.59
ECC Ensembles	3	**5.05**	**1984.02**	**1161.49**	**38.60**
EMOS Ensembles	3	5.46	2178.88	1279.54	25.81
Raw Ensembles	3	5.51	2185.80	1300.82	23.75
No Weather	3	6.85	2768.88	1609.83	0.00
Actuals	4	4.56	1792.64	1054.74	53.56
ECC Ensembles	4	**5.17**	**2020.67**	**1174.88**	**37.86**
EMOS Ensembles	4	5.40	2144.26	1248.97	29.68
Raw Ensembles	4	5.44	2163.35	1257.70	28.78
No Weather	4	6.91	2780.69	1619.69	0.00
Actuals	5	4.50	1778.50	1036.36	56.87
ECC Ensembles	5	**5.13**	**2027.18**	**1161.87**	**39.92**
EMOS Ensembles	5	5.39	2160.75	1241.04	31.00
Raw Ensembles	5	5.30	2145.39	1226.87	32.51
No Weather	5	6.89	2777.34	1625.71	0.00
Actuals	6	4.42	1753.50	1019.43	57.04
ECC Ensembles	6	**5.30**	**2098.22**	**1233.68**	**29.76**
EMOS Ensembles	6	5.42	2178.61	1277.93	25.27
Raw Ensembles	6	5.32	2149.30	1262.16	26.84
No Weather	6	6.87	2736.51	1600.88	0.00

predictions outperformed the model with raw ensembles and EMOS post-processed ensembles overall, which points towards the need to post-process the weather ensemble predictions accounting for dependency structures for load forecasting applications.

To summarize the performance of different models across multiple lead times, we present the MAPE, RMSE and CRPS skill scores in Figure 10.

We expect all methods that include weather information to have a positive skill score. It can be seen from Figure 10 that the ECC (Approach 5) and EMOS (Approach 4) post-processed weather ensembles, as well as the raw weather ensemble predictions (Approach 3) perform significantly better than the baseline that uses no weather data. The ECC post-processed weather ensembles outperform the raw ensembles, as well as EMOS post-processed ensembles based on all three skill scores. Only for a forecasting horizon of 6 days ahead do the raw ensembles achieve a slightly higher MAPE accuracy. However this is not a significant difference to the other post-processed ensemble MAPE scores. Overall, compared to the baseline model, using post-processed weather ensemble predictions can improve the point and probabilistic load forecast accuracy of the model by up to 40% with the ECC post-processed ensembles performing best. To ensure that the differences between the forecast models are statistically significant, we compare the models pairwise using the Diebold-Mariano test (Diebold et al., 1995).

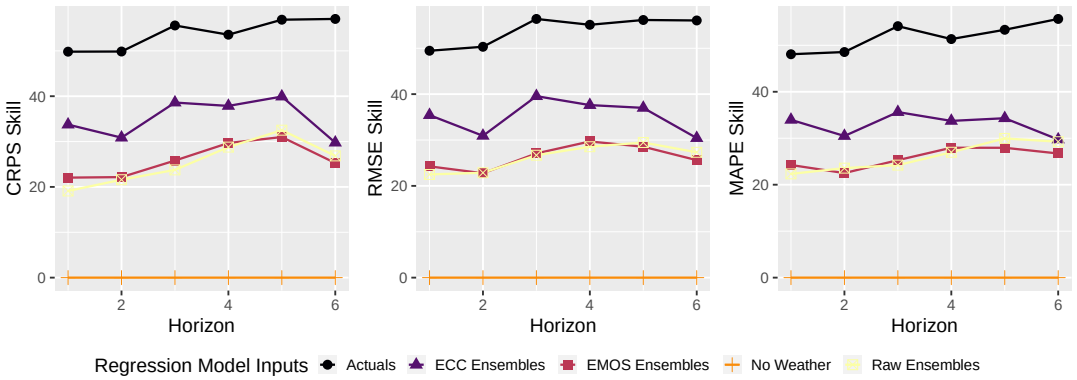

Figure 10. Skill scores for CRPS, RMSE and MAPE for the different weather inputs and all forecast horizons.

For our approach, the ECC ensembles, we can reject the null hypothesis that this model performs worse or equal to another model, for all models except the one using actual weather data. The test statistics and their corresponding p-values for horizon one can be found in Table A1. Finally, we also take a look at the quantile decomposition of the CRPS score for horizon one across all models using $\overline{\text{CRPS}}_n^f = \int_0^1 \overline{\text{QS}}(\alpha) dz$ with

$$\overline{\text{QS}}_N^f(\alpha) = \frac{1}{N} \sum_{t=1}^{N} \text{QS}_\alpha(\hat{F}_t(\alpha)^{-1}, y_t), \text{ and}$$
$$\text{QS}_\alpha(F^{-1}(\alpha), y) = 2(\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha)(F^{-1}(\alpha) - y).$$

Following (Gneiting & Ranjan, 2011) we plot the mean quantile score against α showing the quantile decompositions of the mean CRPS score (see Figure 11).

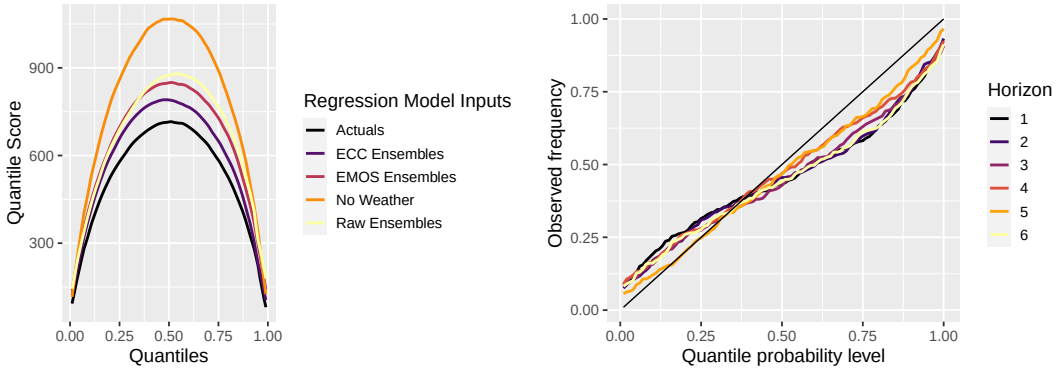

Figure 11. Quantile decomposition of the mean continuous probability score for the different models (left) and coverage rate for the best performing ensemble forecast model (ECC Ensembles) over all horizons (right).

6. Conclusion

In this study, we proposed and implemented a best practice to generate probabilistic forecasts of electricity demand using weather ensemble predictions. We used data for three weather variables (temperature, wind speed, and cloud cover), obtained from a 51-member ensemble system and a high resolution deterministic forecast. For load forecasting, we investigated the efficacy of using ensemble post-processing, as opposed to using raw weather ensemble predictions from NWP systems. This paper has shown how to post-process the weather ensemble predictions by accounting for temporal correlations and correlations between the weather variables. We showed that calibrating the weather ensemble predictions while accounting for their multivariate dependencies using a copula-based coupling approach improves the probabilistic load forecast accuracy, resulting in a CRPS that is noticeably better than a model that does not include any weather information. The post-processed ensembles outperform the raw ensembles, which highlights the advantage of careful post-processing for improved load forecast accuracy.

The proposed modelling framework could potentially be adapted to other energy applications, such as wind and solar power generation. A useful line of future work would be to investigate this post-processing approach for modelling electricity demand at different layers of the energy hierarchy, including the low voltage level or at various locations also accounting for spatial dependencies. It would also be worth investigating the use of machine learning for accommodating the nonlinear relationship between post-processed weather ensemble predictions and load in a nonparametric modelling framework.

Acknowledgements

The authors would like to thank two anonymous reviewers for their useful comments. The research detailed in the current paper was based on data from the ECMWF obtained through an academic licence for research purposes. This work was supported by the German Research Foundation (DFG) as part of the Research Training Group GRK 2153: Energy Status Data – Informatics Methods for its Collection, Analysis and Exploitation and under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

References

Alley, R. B., Emanuel, K. A., & Zhang, F. (2019). Advances in weather prediction. *Science (New York, N.Y.)*, *363*(6425), 342–344. https://doi.org/10.1126/science.aav7274

Al-Yahyai, S., Charabi, Y., & Gastli, A. (2010). Review of the use of numerical weather prediction (nwp) models for wind energy assessment. *Renewable and Sustainable Energy Reviews*, *14*(9), 3192–3198. https://doi.org/10.1016/j.rser.2010.07.001

Arora, S., & Taylor, J. W. (2016). Forecasting electricity smart meter data using conditional kernel density estimation. *Omega*, *59*, 47–59. https://doi.org/10.1016/j.omega.2014.08.008

Arora, S., & Taylor, J. W. (2018). Rule-based autoregressive moving average models for forecasting load on special days: A case study for france. *European Journal of Operational Research*, *266*(1), 259–268. https://doi.org/10.1016/j.ejor.2017.08.056

Atger, F. (2003). Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review*, *131*(8), 1509–1523. https://doi.org/10.1175//1520-0493(2003)131⟨1509:SAIVOT⟩2.0.CO;2

Baran, Á., Lerch, S., El Ayari, M., & Baran, S. (2021). Machine learning for total cloud cover prediction. *Neural Computing and Applications*, *33*(7), 2605–2620. https://doi.org/10.1007/s00521-020-05139-4

Baran, S., & Lerch, S. (2018). Combining predictive distributions for the statistical postprocessing of ensemble forecasts. *International Journal of Forecasting*, *34*(3), 477–496. https://doi.org/10.1016/j.ijforecast.2018.01.005

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Ben Bouallègue, Z., Heppelmann, T., Theis, S. E., & Pinson, P. (2016). Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Monthly Weather Review*, *144*(12), 4737–4750. https://doi.org/10.1175/MWR-D-15-0403.1

Billinton, R., & Huang, D. (2008). Effects of load forecast uncertainty on bulk electric system reliability evaluation. *IEEE Transactions on Power Systems*, *23*(2), 418–425. https://doi.org/10.1109/TPWRS.2008.920078

Bossavy, A., Girard, R., & Kariniotakis, G. (2013). Forecasting ramps of wind power production with numerical weather prediction ensembles. *Wind Energy*, *16*(1), 51–63. https://doi.org/10.1002/we.526

Chen, P., Chen, Z., & Bak-Jensen, B. (2008). Probabilistic load flow: A review, In *Third international conference on electric utility deregulation and restructuring and power technologies, 2008. drpt 2008*, IEEE / Institute of Electrical and Electronics Engineers Incorporated. https://doi.org/10.1109/DRPT.2008.4523658

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., & Wilby, R. (2004). The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, *5*(1), 243–262. https://doi.org/10.1175/1525-7541(2004)005⟨0243:TSSAMF⟩2.0.CO;2

Diebold, F. X., Mariano, R. S., Diebold, F., & Mariano, R. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–63. https://econpapers.repec.org/RePEc:bes:jnlbes:v:13:y:1995:i:3:p:253-63

Eckel, F. A., & Walters, M. K. (1998). Calibrated probabilistic quantitative precipitation forecasts based on themrf ensemble. *Weather and Forecasting*, *13*(4), 1132–1147. https://doi.org/10.1175/1520-0434(1998)013⟨1132:CPQPFB⟩2.0.CO;2

Feldmann, K., Scheuerer, M., & Thorarinsdottir, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous gaussian regression. *Monthly Weather Review*, *143*(3), 955–971. https://doi.org/10.1175/MWR-D-14-00210.1

Gensler, A. (2019). Wind power ensemble forecasting.

Gneiting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting*, *27*(2), 197–207. https://doi.org/10.1016/j.ijforecast.2009.12.015

Gneiting, T. (2014). Calibration of medium-range weather forecasts, In *Ecmwf technical memoranda*.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, *133*(5), 1098–1118. https://doi.org/10.1175/MWR2904.1

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using thresholdand quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, *29*(3), 411–422. https://doi.org/10.1198/jbes.2010.08110

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. https://doi.org/10.1016/j.energy.2018.07.090

Haben, S., Giasemidis, G., Ziel, F., & Arora, S. (2019). Short term load forecasting and the effect of temperature at the low voltage level. *International Journal of Forecasting*, *35*(4), 1469–1484. https://doi.org/10.1016/j.ijforecast.2018.10.007

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., & Palmer, T. N. (2012). Comparing tigge multimodel forecasts with reforecast-calibrated ecmwf ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *138*(668), 1814–1827. https://doi.org/10.1002/qj.1895

Hamill, T. M., & Colucci, S. J. (1997). Verification of eta–rsm short-range ensemble forecasts. *Monthly Weather Review*, *125*(6), 1312–1327. https://doi.org/10.1175/1520-0493(1997)125⟨1312:VOERSR⟩2.0.CO;2

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2. ed., corr. at 7th printing.). New York, NY, Springer.

Haupt, S. E., Garcia Casado, M., Davidson, M., Dobschinski, J., Du, P., Lange, M., Miller, T., Mohrlen, C., Motley, A., Pestana, R., & Zack, J. (2019). The use of probabilistic forecasts: Applying them in theory and practice. *IEEE Power and Energy Magazine*, *17*(6), 46–57. https://doi.org/10.1109/MPE.2019.2932639

Hemri, S., Haiden, T., & Pappenberger, F. (2016). Discrete Postprocessing of Total Cloud Cover Ensemble Forecasts. *Monthly Weather Review*, *144*(7), 2565–2577. https://doi.org/10.1175/MWR-D-15-0426.1

Heppelmann, T., Ben Bouallegue, Z., & Theis, S. (2015). Exploring the calibration of a wind forecast ensemble for energy applications. *EGU General Assembly Conference Abstracts*, 11897.

Heppelmann, T., Steiner, A., & Vogt, S. (2017). Application of numerical weather prediction in wind power forecasting: Assessment of the diurnal cycle. *Meteorologische Zeitschrift*, *26*(3), 319–331. https://doi.org/10.1127/metz/2017/0820

Hu, Y.-C. (2017). Electricity consumption prediction using a neural-network-based grey forecasting approach. *Journal of the Operational Research Society*, *68*(10), 1259–1264. https://doi.org/10.1057/s41274-016-0150-y

Hu, Y.-C., & Jiang, P. (2017). Forecasting energy demand using neural-network-based grey residual modification models. *Journal of the Operational Research Society*, *68*(5), 556–565. https://doi.org/10.1057/s41274-016-0130-2

Hu, Y., Schmeits, M. J., van Andel, S. J., Verkade, J. S., Xu, M., Solomatine, D. P., & Liang, Z. (2016). A stratified sampling approach for improved sampling from a calibrated ensemble forecast distribution. *Journal of Hydrometeorology*, *17*(9), 2405–2417. https://doi.org/10.1175/JHM-D-15-0205.1

Ludwig, N., Feuerriegel, S., & Neumann, D. (2015). Putting big data analytics to work: Feature selection for forecasting electricity prices using the lasso and random forests. *Journal of Decision Systems*, *24*(1), 19–36. https://doi.org/10.1080/12460125.2015.994290

Mass, C. F. (2003). Ifps and the future of the national weather service. *Weather and Forecasting*, *18*(1), 75–79. https://doi.org/10.1175/1520-0434(2003)018⟨0075:IATFOT⟩2.0.CO;2

Möller, A., Thorarinsdottir, T. L., Lenkoski, A., & Gneiting, T. (2015). Spatially adaptive, bayesian estimation for probabilistic temperature forecasts. http://arxiv.org/pdf/1507.05066v3

Nielsen, H. A., Madsen, H., Nielsen, T. S., Badger, J., Giebel, G., Landberg, L., Sattler, K., & Feddersen, H. (2004). Wind power ensemble forecasting, In *Proceedings of the 2004 global windpower conference and exhibition.*

Nowotarski, J., & Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, *81*, 1548–1568. https://doi.org/10.1016/j.rser.2017.05.234

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Schefzik, R. (2017). Ensemble calibration with preserved correlations: Unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 999–1008. https://doi.org/10.1002/qj.2984

Schefzik, R., Thorarinsdottir, T. L., & Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, *28*(4), 616–640. https://doi.org/10.1214/13-STS443

Scheuerer, M., & Buermann, L. (2014). Spatially adaptive post-processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(3), 405–422. https://doi.org/10.1111/rssc.12040

Taieb, S. B., Taylor, J. W., & Hyndman, R. J. (2020). Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*, 1–17. https://doi.org/10.1080/01621459.2020.1736081

Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, *54*(8), 799–805. https://doi.org/10.1057/palgrave.jors.2601589

Taylor, J. W., & Buizza, R. (2002). Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems*, *17*(3), 626–632. https://doi.org/10.1109/TPWRS.2002.800906

Taylor, J. W., & Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, *19*(1), 57–70. https://doi.org/10.1016/S0169-2070(01)00123-6

Thorey, J., Chaussin, C., & Mallet, V. (2018). Ensemble forecast of photovoltaic power with online crps learning. *International Journal of Forecasting*, *34*(4), 762–773. https://doi.org/10.1016/j.ijforecast.2018.05.007

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

van der Meer, D. W., Shepero, M., Svensson, A., Widén, J., & Munkhammar, J. (2018). Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using gaussian processes. *Applied Energy*, *213*, 195–207. https://doi.org/10.1016/j.apenergy.2017.12.104

Wang, Y., Xia, Q., & Kang, C. (2011). Unit commitment with volatile node injections by using interval optimization. *IEEE Transactions on Power Systems*, *26*(3), 1705–1713. https://doi.org/10.1109/TPWRS.2010.2100050

Wilks, D. S., & Hamill, T. M. (2007). Comparison of ensemble-mos methods using gfs reforecasts. *Monthly Weather Review*, *135*(6), 2379–2390. https://doi.org/10.1175/MWR3402.1

Zamo, M., Mestre, O., Arbogast, P., & Pannekoucke, O. (2014). A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. part ii: Probabilistic forecast of daily production. *Solar Energy*, *105*, 804–816.

Appendix A. Supplementary Material

Table A1. For the CRPS, pairwise Diebold-Mariano test statistic and corresponding p-values for all models for 1 day-ahead prediction. We use a one-sided test with the null hypothesis that the performance of the model named in the column heading is at most as accurate as the model named in the row. Hypothesis rejection ($p < 0.05$) is indicated in bold.

Models	Actuals	ECC Ensembles	EMOS Ensembles	Raw Ensembles	No Weather
Actuals		-4.32 (1.00)	-5.85 (1.00)	-6.45 (1.00)	-6.83 (1.00)
ECC Ensembles	**4.32 (0.00)**		-3.89 (1.00)	-4.16 (1.00)	-5.49 (1.00)
EMOS Ensembles	**5.85 (0.00)**	**3.89 (0.00)**		-1.94 (0.97)	-3.57 (1.00)
Raw Ensembles	**6.45 (0.00)**	**4.16 (0.00)**	**1.94 (0.03)**		-3.2 (1.00)
No Weather	**6.83 (0.00)**	**5.49 (0.00)**	**3.57 (0.00)**	**3.2 (0.00)**	

Table A2. List of all variables with a non-zero coefficient in the LASSO regression with $\lambda = 9.46$, where $\times$ denotes interaction terms.

1	April	16	Lag Year ($t - 365$)	31	Spring Bank Holiday
2	Christmas	17	March	32	Summer Bank Holiday
3	day count $\times$ days	18	May	33	Sunday
4	(day count $\times$ days)^3	19	May Day	34	Temperature at 2m
5	Cooling power of wind	20	New Years Day	35	Temperature $\times$ Weekend
6	day count	21	November	36	Total Cloud Cover
7	(day count)^2	22	October	37	Whit Monday
8	days per year (days)[a]	23	Proximity Days (PD)	38	Winter
9	Easter	24	PD $\times$ Friday	39	Wind Speed
10	February	25	PD $\times$ Saturday		
11	Friday	26	PD $\times$ Sunday		
12	Heating[b]	27	PD $\times$ Winter		
13	July	28	Special Days (SD)		
14	June	29	SD $\times$ Winter		
15	Lag Week ($t - 7$)	30	Smooth[c]		

[a] Days per year is an indicator for leap years and either 365 or 366.

[b] Heating is a dummy variable indicating the heating season from October to March.

[c] Smooth is a moving average over the past seven days of load.

Table A3. List of all variables with a coefficient of zero in the LASSO regression with $\lambda = 9.46$, where $\times$ denotes interaction terms.

1	(days per year)^2	6	Saturday	11	New Years Eve
2	(days per year)^3	7	September	12	SD $\times$ Summer
3	(day count)^3	8	December	13	Proximity Day $\times$ Summer
4	(count $\times$ days)^2	9	Summer	14	Effective Temperature
5	Eta	10	Ascension		

Table A4. Oktas and the corresponding intervals used for quantization of total cloud cover.

Okta	Interval
0	$[0, 0.01[$
1	$[0.01, 0.1875[$
2	$[0.1875, 0.3125[$
3	$[0.3125, 0.4375[$
4	$[0.4375, 0.5625[$
5	$[0.5625, 0.6875[$
6	$[0.6875, 0.8125[$
7	$[0.8125, 0.99[$
8	$[0.99, 1]$