

# Hierarchical Probabilistic Forecasting of Electricity Demand with Smart Meter Data

Souhaib Ben Taieb

Department of Econometrics and Business Statistics, Monash University

James W. Taylor

Saïd Business School, University of Oxford

and

Rob J. Hyndman

Department of Econometrics and Business Statistics, Monash University

June 12, 2017

## Abstract

Electricity smart meters record consumption, on a near real-time basis, at the level of individual commercial and residential properties. From this, a hierarchy can be constructed consisting of time series of demand at the smart meter level, and at various levels of aggregation, such as substations, cities and regions. Forecasts are needed at each level to support the efficient and reliable management of consumption. A limitation of previous research in this area is that it considered only deterministic prediction. To enable improved decision-making, we introduce an algorithm for producing a probability density forecast for each series within a large-scale hierarchy. The resulting forecasts are coherent in the sense that the forecast distribution of each aggregate series is equal to the convolution of the forecast distributions of the corresponding disaggregate series. Our algorithm has the advantage of synthesizing information from different levels in the hierarchy through forecast combination. Distributional assumptions are not required, and dependencies between forecast distributions are imposed through the use of empirical copulas. Scalability to large hierarchies is enabled by decomposing the problem into multiple lower-dimension sub-problems. Results for UK electricity smart meter data show performance gains for our method when compared to benchmarks.

*Keywords:* Aggregates of time series; Empirical copulas; Forecast combinations; Predictive distributions; Smart meters

# 1 Introduction

Forecasts of electricity demand are needed for the efficient and secure management of power grids. Although interest has traditionally been in forecasting the demand for the whole system, this is changing as the installation of smart meters is becoming widespread (Borenstein & Bushnell 2015), providing an accurate record of electricity consumption within intraday periods at the level of individual commercial and residential properties. Along with detailed data regarding power generation, smart meters provide the basis of smart grids, which enable electricity consumption to be managed in a dynamic, efficient, reliable and sustainable way (Ramchurn et al. 2012, Mirowski et al. 2014). Demand forecasts are also important for various levels of aggregation of the individual consumers. For example, grid management can benefit from the availability of predictions at the level of individuals, transformers, feeders, substations, balancing areas, cities, regions, and countries (Smith et al. 2010, Cho et al. 2013, Haben et al. 2014, Sun et al. 2016, López Cabrera & Schulz 2017). In this context, the forecasting problem involves a hierarchy of time series, with levels consisting of differing degrees of aggregation.

Forecasting a hierarchy of time series presents a number of challenges. First, the series at different levels of the hierarchy, corresponding to different degrees of aggregation, can exhibit notably different features. While smart meter data is characterized by intermittency, relatively high volatility, and skewness, aggregated series will naturally be more smoothly evolving, less skewed, and show clearer signs of seasonality and weather-dependency. These differences have implications for the choice of forecasting method to use at different levels of the hierarchy. Forecasts for the aggregates could be produced by simply summing forecasts of the corresponding series at the lower levels. However, this bottom-up approach generally delivers poor results (Hyndman et al. 2011). This links to a second significant challenge, which is the requirement that the forecast of each aggregated series should equal the sum of the forecasts of the corresponding disaggregated series, in order to support coherent decision-making at different levels of the hierarchy. This aggregation constraint is unlikely to be satisfied if the forecasts for each series in the hierarchy are produced independently. Lastly, in many applications, such as the smart meter context, the hierarchy can consist of thousands, or even millions, of time series at the bottom level, which has implications

for the choice of forecasting method in terms of computational load.

Generating forecasts to support decision-making in a hierarchical structure is important in many other applications, such as retail (Kremer et al. 2016) and tourism (Wickramasuriya et al. 2015). Recently proposed methods involve two stages, with forecasts first produced independently for each series in the hierarchy, and then combined to enable a synthesis of available information, and to ensure compliance with the aggregation constraint (see, for example, van Erven & Cugliari 2015). A notable feature of the hierarchical forecasting literature is the lack of probabilistic prediction. This is a significant limitation, because optimal decision making, in a variety of applications, requires an assessment of forecast uncertainty (see, for example, Berrocal et al. 2010, Jeon & Taylor 2012). In the electricity demand context, probabilistic predictions of the total system load are used for stochastic unit commitment models, power supply planning, setting operating reserve, price forecasting, and electricity market trading (Hong & Fan 2016). With regard to disaggregated load, probabilistic forecasts are used to ensure a balance between generation and consumption at regional (López Cabrera & Schulz 2017) and distributional levels (Sun et al. 2016). At the smart meter level, suppliers can use probabilistic forecasts to update pricing and incentive schemes for individual customers (Arora & Taylor 2016, Ben Taieb et al. 2016). An advantage of generating probabilistic forecasts for the entire hierarchy is that it allows the possibility of predictions for each series in the hierarchy to benefit from the information elsewhere in the hierarchy.

We introduce an approach for producing a probability density forecast for each series within a large-scale hierarchy. Our aim is “coherent” forecasts; i.e., where the forecast distribution of each aggregate series is equal to the convolution of the forecast distributions of the corresponding disaggregate series. Such forecasts naturally satisfy the aggregation constraints of the hierarchy. Our method allows for different types of distributions, and accounts for dependencies to enable the computation of the predictive distribution of the aggregates. It proceeds by independently generating a density forecast for each series in the hierarchy. Then, a state-of-the art hierarchical forecast combination method is applied to produce coherent revised mean forecasts. Samples are then taken from the densities. Drawing on the work of Arbenz et al. (2012) and Schefzik et al. (2013), a sequence of permutations, derived from empirical copulas, are applied to the multivariate

samples to restore dependencies before computing the sums corresponding to the aggregates within the hierarchy. The result is coherent density forecasts for the entire hierarchy.

Our approach has four distinct advantages. First, for every series in the hierarchy, a density forecast is produced, and these together satisfy the coherency constraints. Second, each estimated density and dependency structure is the result of a synthesis of information from different levels of the hierarchy. Third, distributional assumptions are not required. Fourth, our procedure decomposes the problem into lower-dimension sub-problems, enabling the approach to be scalable to large hierarchies, such as the electricity demand case.

Section 2 introduces the data that we analyze in this paper. Section 3 discusses hierarchical forecasting for the mean, while Section 4 presents our new approach to probabilistic hierarchical forecasting. Section 5 describes the forecasting methods that we implement for individual series in our hierarchy, and Section 6 presents forecasting results. Section 7 concludes the paper.

## **2 Smart Electricity Meter Data**

Our empirical analysis involves time series of electricity consumption, recorded by smart meters at residential properties in Great Britain. The data was collected, along with geographic and demographic information, by four energy supply companies as part of a research project aimed at understanding domestic energy usage (AECOM Building Engineering 2014, AECOM 2011). Although more than 14000 series were available, for simplicity, we chose to avoid relatively short series with a lot of missing observations, and this led us to select data recorded at 1578 meters for the period of approximately 15 months from 20 April 2009 to 31 July 2010, inclusive. With the data recorded half-hourly, each series consisted of 22464 observations. For each series, we used the approximately 12 months up to 30 April 2010, for model estimation, and the final three calendar months for post-sample evaluation. Our interest is in predictions made each day from a forecast origin of 23:30 for each half-hour of the next day, which implies 48 different lead times.

Using the geographical information, we constructed a hierarchy consisting of five levels with the following numbers of series: 1578 (Level 5, the bottom level), 40 (Level 4), 11 (Level 3), 3 (Level 2), and 1 (Level 1, the top level). In total, therefore, the hierarchy consists of 1633 series,

with 1578 in the bottom level and 55 aggregates in the other levels. The hierarchy’s architecture is summarized in Figure 1, where each node corresponds to one of the aggregates in Levels 1 to 4, and its size is in proportion to the number of smart meters aggregated from the bottom level. To reflect reality, we have allowed quite large variation in the number of smart meters making up the different aggregates. In the following, we use the terms ‘node’ and ‘series’ interchangeably.

In Figure 2, we present a one-week period for a series taken from each level of the hierarchy. The values on the left hand side of the figure indicate the number of smart meter series that have been summed to give each aggregated series. The series from the higher levels show daily and weekly patterns, which become increasingly evident with greater aggregation. For example, the intraday patterns for Saturday and Sunday can be seen to be similar to each other, but different to the weekdays, for the top level series and the series formed from the aggregation of the 150 series. However, this is much less clear in the three other series in Figure 2. Indeed, daily and weekly patterns are not so apparent in the series from the bottom level, which exhibits relatively volatile behavior, with occasional spikes. For the same five series, Figure 3 shows the intraday pattern for each day of the week, averaged over the full series. It is interesting to see from Figure 3 the existence of cyclical patterns in the daily profiles for the individual smart meter series. For this particular household, the average daily pattern varies little across the seven days of the week.

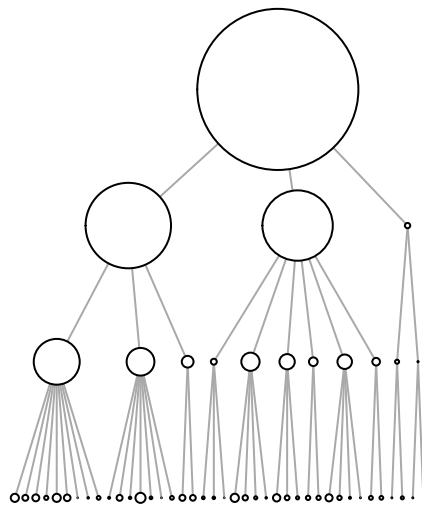


Figure 1: The hierarchical structure of the electricity network, where the size of each node is proportional to the number of aggregated meters.

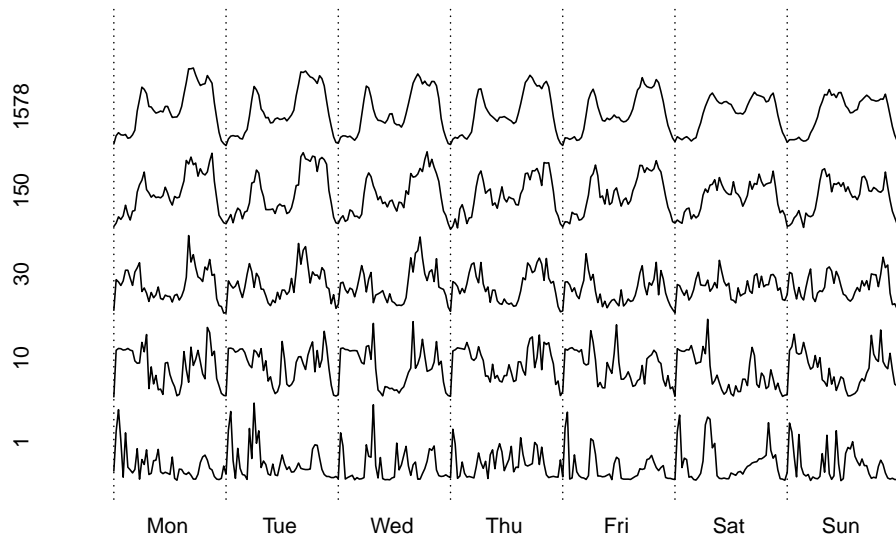


Figure 2: A one-week period of demand corresponding to different numbers of aggregated meters.

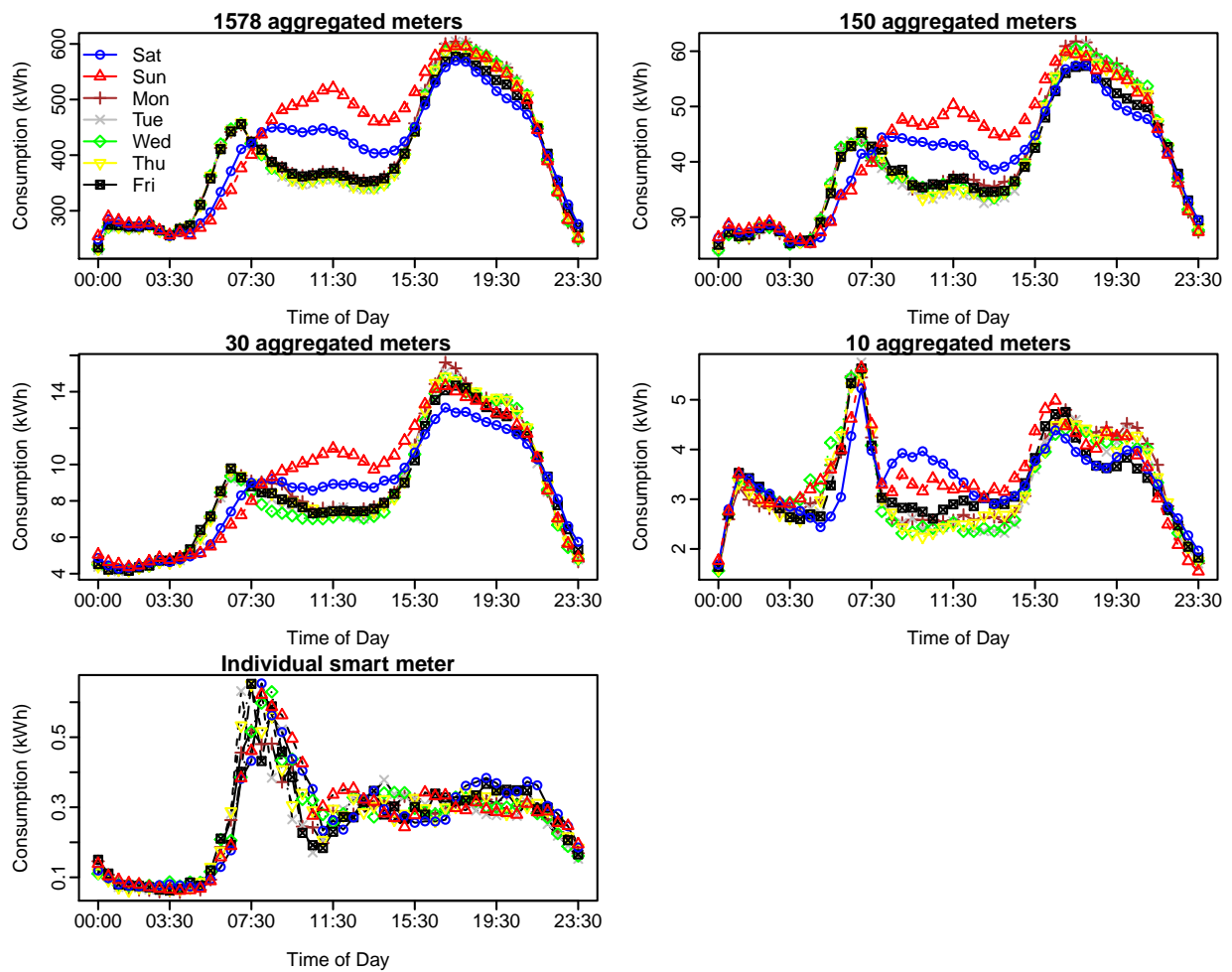


Figure 3: Intraday pattern for each day of the week and different numbers of aggregated meters.

### 3 Hierarchical Mean Electricity Demand Forecasting

In this section, we first introduce terminology and our notation relating to hierarchical forecasting. We then describe the method of Wickramasuriya et al. (2015), which has been introduced for mean hierarchical forecasting, but has the potential to be used for probabilistic hierarchical forecasting.

#### 3.1 Mean coherent hierarchical forecasts

Electricity consumption collected by smart meters, and the associated aggregated consumption form a multivariate time series with a hierarchical structure, also called a hierarchical time series. Building on the notation in Hyndman et al. (2016), we let  $\mathbf{b}_t$  be an  $m$ -vector with the observations at time  $t$  in the bottom level. Then, the  $r$ -vector with the observations at the different levels of aggregation is given by  $\mathbf{a}_t = \mathbf{A}\mathbf{b}_t$  where the matrix  $\mathbf{A} \in \{0, 1\}^{r \times m}$ , and  $t = 1, \dots, T$ . Each entry of  $\mathbf{A}$  is equal to 1 if the associated bottom-level observation is included in the aggregation. Finally, we let  $\mathbf{y}_t = (\mathbf{a}_t \ \mathbf{b}_t)'$  be an  $n$ -vector containing observations both at the bottom and aggregate levels, given by  $\mathbf{y}_t = \mathbf{S}\mathbf{b}_t$ , where  $\mathbf{S}' = [\mathbf{A}' \ \mathbf{I}_m] \in \{0, 1\}^{n \times m}$  and  $\mathbf{I}_m$  is an identity matrix of order  $m$ . To avoid pathological hierarchies, we will assume that  $m > 0$ ,  $r > 0$  and  $\sum_{j=1}^r s_{i,j} > 1$  where  $s_{ij}$  is the entry in the  $i$ th row and  $j$ th column of matrix  $\mathbf{S}$ . Figure 4 gives a small example with  $m = 5$  bottom level series with  $\mathbf{b}_t = (y_{aa,t}, y_{ab,t}, y_{ba,t}, y_{bb,t}, y_{bc,t})'$ , and  $r = 3$  aggregate series with  $\mathbf{a}_t = (y_t, y_{a,t}, y_{b,t})'$ .

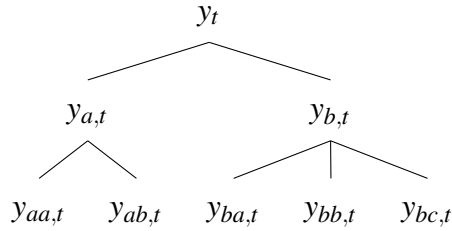


Figure 4: Example of a hierarchical time series.

Given historical observations  $\mathbf{y}_1, \dots, \mathbf{y}_T$ , forecasting the mean electricity consumption for each half-hour  $h$  of the next day reduces to estimating the  $h$  period-ahead conditional expectation:

$$\mathbb{E}[\mathbf{y}_{T+h} \mid \mathbf{y}_1, \dots, \mathbf{y}_T] = \mathbf{S} \mathbb{E}[\mathbf{b}_{T+h} \mid \mathbf{b}_1, \dots, \mathbf{b}_T], \quad (1)$$

where  $h = 1, 2, \dots, 48$ . Recall that our interest is in predictions made each day from a forecast origin of 23:30 for each half-hour of the next day, implying 48 lead times.

A natural plug-in estimator for expression (1), known as *bottom-up*, is given by

$$\hat{\mathbf{y}}_{T+h} = \mathbf{S}\hat{\mathbf{b}}_{T+h}, \quad (2)$$

where  $\hat{\mathbf{b}}_{T+h} = (\hat{b}_{1,T+h}, \dots, \hat{b}_{m,T+h})'$ ,  $\hat{b}_{j,T+h}$  is an estimate of  $\mathbb{E}[b_{j,T+h} \mid b_{j,1}, \dots, b_{j,T}]$ , and  $b_{j,t}$  denotes the  $j$ th variable of the vector  $\mathbf{b}_t$  with  $j = 1, \dots, m$ . In other words, the mean forecasts for each aggregate node are computed by summing the mean forecasts of the associated bottom level series. This approach is computationally efficient since it only requires summing bottom level forecasts. However, the series at the most disaggregated level often have a low signal-to-noise ratio, which makes them hard to predict. Indeed, this is the case for our electricity consumption hierarchy. Therefore, summing bottom level forecasts is unlikely to provide good forecast accuracy at upper levels in the hierarchy (Hyndman et al. 2011).

Another approach, known as *base*, forecasts all nodes at all levels independently. More precisely, we compute

$$\hat{\mathbf{y}}_{T+h} = (\hat{y}_{1,T+h}, \dots, \hat{y}_{n,T+h})' = (\hat{\mathbf{a}}'_{T+h} \hat{\mathbf{b}}'_{T+h})' \quad (3)$$

where  $\hat{y}_{i,T+h}$  is an estimate of  $\mathbb{E}[y_{i,T+h} \mid y_{i,1}, \dots, y_{i,T}]$  and  $y_{i,t}$  denotes the  $i$ th variable of the vector  $\mathbf{y}_t$  with  $i = 1, \dots, n$ . This approach provides flexibility since we can use different forecasting algorithms for each series and aggregation level. However, the base forecasts  $\hat{\mathbf{y}}_{T+h}$  will generally not satisfy the aggregation constraints, i.e.  $\hat{\mathbf{a}}_{T+h} = \mathbf{A}\hat{\mathbf{b}}_{T+h}$ , with the constraint violations for each aggregate series at horizon  $h$  given by

$$\tilde{\mathbf{e}}_{T+h} = \hat{\mathbf{a}}_{T+h} - \mathbf{A}\hat{\mathbf{b}}_{T+h}. \quad (4)$$

The forecasts  $\hat{\mathbf{y}}_{T+h}$  in expression (3) are *mean coherent* if  $\tilde{\mathbf{e}}_{T+h} = \mathbf{0}$ , and incoherent otherwise.

Since the optimal mean forecasts in expression (1) are coherent by definition, it seems sensible to impose the aggregation constraints when generating hierarchical mean forecasts. Furthermore, coherent forecasts will allow coherent decisions over the entire hierarchy.



Starting from a set of (probably incoherent) base forecasts  $\hat{\mathbf{y}}_{T+h}$ , Hyndman et al. (2011) proposes the computation of revised coherent hierarchical forecasts of the following form:

$$\tilde{\mathbf{y}}_{T+h} = \mathbf{S}\tilde{\mathbf{b}}_{T+h} \text{ and } \tilde{\mathbf{b}}_{T+h} = \mathbf{P}_h\hat{\mathbf{y}}_{T+h}, \quad (5)$$

for some appropriately chosen matrix  $\mathbf{P}_h \in \mathbb{R}^{m \times n}$ . In other words, forecasts are obtained by summing bottom level revised forecasts  $\tilde{\mathbf{b}}_{T+h}$ , which have been computed by combining forecasts from all levels  $\hat{\mathbf{y}}_{T+h}$ . An advantage of the forecast in expression (5) is that it involves the combination of the forecasts from all levels of the hierarchy, applied through the weight matrix  $\mathbf{P}_h$ . Furthermore, many hierarchical forecasting methods are represented as particular cases, including the bottom-up forecasts, for which  $\mathbf{P}_h = \left[ \mathbf{0}_{m \times r} \mid \mathbf{1}_{m \times m} \right]$ . Finally, the forecasts are coherent by construction.

### 3.2 The MinT approach to hierarchical forecasting

Instead of using an arbitrary weight matrix  $\mathbf{P}_h$ , Wickramasuriya et al. (2015) propose the minimization of the sum of the error variances, and derive a closed-form expression, as in the following:

**Theorem 1.** (Adapted from Wickramasuriya et al. 2015) Let  $\hat{\mathbf{e}}_{t+h} = \mathbf{y}_{t+h} - \hat{\mathbf{y}}_{t+h}$  be the  $h$  period-ahead base forecast errors for  $t = 1, \dots, T$ , and  $\mathbf{W}_h = \mathbb{E}[\hat{\mathbf{e}}_{t+h}\hat{\mathbf{e}}'_{t+h}]$ , the corresponding positive definite covariance matrix. Then, assuming unbiased base forecasts,

1. the covariance matrix of the  $h$  period-ahead revised forecast errors is given by

$$\tilde{\mathbf{V}}_h = \mathbb{V}[\mathbf{y}_{t+h} - \tilde{\mathbf{y}}_{t+h}] = \mathbf{S}\mathbf{P}_h\mathbf{W}_h\mathbf{P}'_h\mathbf{S}', \quad (6)$$

2. the weight matrix  $\mathbf{P}_h$  of the unbiased revised forecasts, which minimizes the trace of  $\tilde{\mathbf{V}}_h$ , is

$$\mathbf{P}_h^* = (\mathbf{S}'\mathbf{W}_h^{-1}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}_h^{-1}, \quad (7)$$

3. the revised forecasts of the mean are given by

$$\tilde{\mathbf{y}}_{T+h} = \mathbf{S}\mathbf{P}_h^*\hat{\mathbf{y}}_{T+h}. \quad (8)$$

We refer to this method as *MinT*, as it involves the minimization of the trace. Computing the matrix  $\mathbf{P}_h^*$  in expression (7) requires the estimation of the possibly high-dimensional covariance matrix  $\mathbf{W}_h$  or its inverse. However, since  $\mathbf{W}_h$  is hard to estimate for  $h > 1$ , we will assume  $\mathbf{W}_h \propto \mathbf{W}_1$  (as in Wickramasuriya et al. 2015). For  $h = 1$ , the sample covariance matrix is given by

$$\hat{\mathbf{W}}_1 = \frac{1}{T} \sum_{t=1}^T \hat{\mathbf{e}}_t \hat{\mathbf{e}}_t',$$

where  $\hat{\mathbf{e}}_t$  are the residuals (the in-sample one-step forecast errors) from the base models applied to each node in the hierarchy. However, since electricity demand volatility varies with the time of day, we will allow  $\mathbf{W}_1$  to change for each half-hour  $s$ ,  $s = 1, \dots, 48$ ; i.e. we compute

$$\hat{\mathbf{W}}_{1,s} = \frac{\sum_t \mathbb{1}\{\lfloor t/48 \rfloor = s\} \hat{\mathbf{e}}_t \hat{\mathbf{e}}_t'}{\mathbb{1}\{\lfloor t/48 \rfloor = s\}}$$

where  $\mathbb{1}\{\cdot\}$  is the indicator function, and  $\lfloor \cdot \rfloor$  denotes the floor operator.

When  $T = \mathcal{O}(n)$ , the sample covariance matrix is singular, and hence not invertible. Consequently, Wickramasuriya et al. (2015) proposed using either a diagonal estimate (ignoring all covariances), or a regularized estimate (shrinking the off-diagonal terms of the sample covariance estimate towards zero). We call these *MinT*Diag and *MinT*Shrink, respectively.

## 4 Hierarchical Probabilistic Electricity Demand Forecasting

In this section, we first introduce our definition for probabilistically coherent hierarchical forecasts. We then discuss how copulas provide a convenient framework for summing random variables within a hierarchical probabilistic forecasting approach. Following that, we introduce our proposed hierarchical probabilistic forecasting algorithm.

### 4.1 Probabilistically coherent hierarchical forecasts

Let  $y_{i,t}$  denote the  $i$ th variable of the vector  $\mathbf{y}_t$  with  $i = 1, \dots, n$ . Recall that  $n$  is the number of nodes in the hierarchy, and is, therefore, the number of time series that need predicting. Instead of

estimating the conditional mean given in expression (1), we want to estimate, for each node in the hierarchy, the  $h$  period-ahead conditional cumulative distribution function:  $F_{i,T+h}(y | \mathbf{y}_1, \dots, \mathbf{y}_T) = \mathbb{P}(y_{i,T+h} \leq y | \mathbf{b}_1, \dots, \mathbf{b}_T)$  for each half-hour  $h = 1, \dots, 48$ .

Let  $\hat{\mathbf{F}}_{\mathbf{b},T+h}$  be the joint predictive distribution for  $\mathbf{b}_{T+h}$ ;  $\hat{F}_{a_j,T+h}$  the predictive distribution for  $a_{j,T+h}$ ;  $\mathbf{s}_j$  the  $j$ th row vector of matrix  $\mathbf{S}$ , for  $j = 1, \dots, r$ ; and let  $\stackrel{d}{=}$  denote equality in distribution. When  $a_{j,T+h} \stackrel{d}{=} \mathbf{s}_j \mathbf{b}_{T+h}$ , we describe the forecasts as *probabilistically coherent*. Note that if the forecasts are probabilistically coherent, they are also mean coherent.

It is possible to compute independently probabilistic base forecasts for each series:  $F_{i,T+h}(y | y_{i,1}, \dots, y_{i,T}) = \mathbb{P}(y_{i,T+h} \leq y | y_{i,1}, \dots, y_{i,T})$ . However, producing a forecast for each node, using only historical observations for that node, does not exploit the possible dependencies between the time series for different nodes. Furthermore, the probabilistic forecasts will not necessarily be probabilistically coherent due to estimation errors and possible model differences.

The MinT method of Theorem 1 allows for both coherent mean forecasts  $\tilde{\mathbf{y}}_{T+h}$  given in expression (8), and the calculation of the associated forecast variances  $\tilde{\mathbf{V}}_h$  given in expression (6). Therefore, for the situation where the distribution is completely specified by its first two moments, it is possible to produce coherent probabilistic forecasts. Forecast distributions for electricity demand tend to be highly skewed, and so one option is to assume a log-normal distribution, as this is completely specified by its first two moments. In this paper, we take the view that it is preferable to avoid distributional assumptions, as we do not want to restrict the choice of forecasting method used for any individual series. Our proposal uses a bottom-up approach based on Monte Carlo sampling of the bottom level predictive distributions. We incorporate the MinT approach within the estimation of the bottom level distributions, and account for dependencies using empirical copulas (Rüschendorf 2009).

## 4.2 Copula-based distribution of sums of random variables

Since each aggregate node is the sum of a subset of bottom level nodes, computing bottom-up predictive distributions reduces to computing the distribution of sums of random variables. Let us first focus on a simple hierarchy with  $d$  bottom level nodes and one aggregate node, and only

consider unconditional distributions. Let  $\mathbf{X} = (X_1, \dots, X_d)$  be a random vector with continuous marginals  $F_i$ , and joint distribution  $\mathbf{F}$ . The distribution of  $Z = \sum_{i=1}^d X_i$  at value  $z$  is given by

$$F_Z(z) = \int_{\mathbb{R}^d} \mathbb{1}\{x_1 + \dots + x_d \leq z\} d\mathbf{F}(x_1, \dots, x_d), \quad (9)$$

where  $(x_1, \dots, x_d) \in \mathbb{R}^d$ . Clearly, computing the distribution of the sum requires the joint distribution. In our application, the bottom level marginal predictive distributions  $\hat{F}_i$  are available. To take advantage of this, instead of directly estimating the joint distribution  $\mathbf{F}$ , it seems appealing to employ a forecasting algorithm that uses the predictive marginals  $\hat{F}_i$  to estimate the joint distribution  $\mathbf{F}$ . A natural and convenient framework to enable this is provided by copulas (Nelsen 2007).

Copulas originate from Sklar's theorem (Sklar 1959), which states that for any continuous distribution function  $\mathbf{F}$  with marginals  $F_1, \dots, F_d$ , there exists a unique "copula" function  $\mathbf{C} : [0, 1]^d \rightarrow [0, 1]$  such that  $\mathbf{F}$  can be written as  $\mathbf{F}(x_1, \dots, x_d) = \mathbf{C}(F_1(x_1), \dots, F_d(x_d))$ . This allows us to decouple the computation of the marginal predictive distributions from the computation of the dependence structure. With the exception of some special cases where the distribution of the sum in expression (9) can be computed analytically, we would typically resort to Monte Carlo simulation (Gijbels & Herrmann 2014). Suppose we have samples  $x_k^i \sim F_i$ , and  $\mathbf{u}_k = (u_k^1, \dots, u_k^d) \sim \mathbf{C}$  with  $i = 1, \dots, d$  and  $k = 1, \dots, K$ , then we can compute the empirical margins

$$\hat{F}_i(x) = \frac{1}{K} \sum_{k=1}^K \mathbb{1}\{x_k^i \leq x\},$$

and the empirical copula

$$\hat{\mathbf{C}}(\mathbf{u}) = \frac{1}{K} \sum_{k=1}^K \mathbb{1}\left\{ \frac{\text{rk}(u_k^1)}{K} \leq u_1, \dots, \frac{\text{rk}(u_k^d)}{K} \leq u_d \right\}, \quad (10)$$

for  $\mathbf{u} = (u_1, \dots, u_d) \in [0, 1]^d$ , where  $\text{rk}(u_k^i)$  is the rank of  $u_k^i$  within the set  $\{u_1^i, \dots, u_K^i\}$ . Then, by Sklar's theorem, the joint distribution can be computed as

$$\hat{\mathbf{F}}(x_1, \dots, x_d) = \hat{\mathbf{C}}(\hat{F}_1(x_1), \dots, \hat{F}_d(x_d)). \quad (11)$$

The procedure of applying empirical copulas to empirical margins is convenient since it can be efficiently represented in terms of sample reordering. Indeed, this can be seen by rewriting expression (11) using expression (10), and by exploiting the fact that  $\hat{F}_i^{-1}(k/K) = x_{(k)}^i$  where  $x_{(k)}^i$  denotes the order statistics of  $\{x_1^i, \dots, x_K^i\}$ . In other words, the order statistics  $u_{(1)}^i, \dots, u_{(K)}^i$  of the samples  $u_1^i, \dots, u_K^i$  induce a permutation  $p_i$  of the integers  $\{1, 2, \dots, K\}$ , defined by  $p_i(k) = \text{rk}(u_k^i)$  for  $k = 1, \dots, K$ , where a permutation denotes a bijective mapping from  $\{1, 2, \dots, K\}$  to  $\{1, 2, \dots, K\}$ . If we apply the permutations to each independent marginal sample  $\{x_1^i, \dots, x_K^i\}$ , the reordered samples inherit the multivariate rank dependence structure from the copula  $\hat{\mathbf{C}}$ .

Introducing a dependence structure into originally independent marginal samples goes back to Iman & Conover (1982) who considered the special case of normal copulas. Under appropriate regularity conditions, Mainik (2015) showed that the estimate of the distribution of the sum, computed using the reordering procedure, is strongly uniformly consistent with convergence rate  $\mathcal{O}_p(T^{-1/2})$  where  $T$  is the sample size. The use of a sample reordering procedure to specify multivariate dependence structure has been considered in Arbenz et al. (2012) for high-dimensional risk aggregation, and by Schefzik et al. (2013) with applications to weather forecasting.

### 4.3 A new algorithm for hierarchical probabilistic forecasting

Consider the hierarchy of Figure 4. If we compute the marginals for all bottom level nodes, i.e.  $\hat{F}_{aa,t}, \hat{F}_{ab,t}, \hat{F}_{ba,t}, \hat{F}_{bb,t}, \hat{F}_{bc,t}$  and the associated five-dimensional copula  $\hat{\mathbf{C}}$ , then the joint distribution  $\hat{\mathbf{F}}$  can be computed using expression (11). Given independent samples from the margins, the reordering procedure allows the restoration of the dependencies from the copula  $\hat{\mathbf{C}}$ . We can compute, for each aggregate node, the sum of the reordered samples for the associated bottom level nodes.

Although it is convenient to decouple the computation of the marginals from the computation of the dependence function, it is still challenging to specify a  $d$ -dimensional copula when  $d$  is large, which is indeed the case in our application where  $d$  is the number of households. Furthermore, with highly disaggregated time series data, the bottom level series often have a small signal-to-noise ratio making it difficult to estimate the dependence structure between the series.

However, since we are only interested in specific aggregations, we can avoid explicitly mod-

eling the (often) high-dimensional copula that describes the dependence between all bottom level nodes. Similar to Arbenz et al. (2012), we decompose the problem into the estimation of multiple lower-dimensional copulas for all child nodes of each aggregate node (i.e. the nodes directly connected to the aggregate node), and compute the sum distribution in a hierarchical manner.

For the example in Figure 4, the following procedure allows us to compute the distributions of all aggregates in a hierarchical manner:

1. Compute  $\hat{F}_{a,t}$  using  $\hat{\mathbf{C}}_1(\hat{F}_{aa,t}(\cdot), \hat{F}_{ab,t}(\cdot))$ .
2. Compute  $\hat{F}_{b,t}$  using  $\hat{\mathbf{C}}_2(\hat{F}_{ba,t}(\cdot), \hat{F}_{bb,t}(\cdot), \hat{F}_{bc,t}(\cdot))$ .
3. Compute  $\hat{F}_t$  using  $\hat{\mathbf{C}}_3(\hat{F}_{a,t}(\cdot), \hat{F}_{b,t}(\cdot))$ .

Instead of computing a five-dimensional copula  $\hat{\mathbf{C}}$  for all bottom level nodes in Figure 4, the hierarchical procedure uses three lower-dimensional copulas  $\hat{\mathbf{C}}_1$ ,  $\hat{\mathbf{C}}_2$  and  $\hat{\mathbf{C}}_3$  for  $(y_{aa,t}, y_{ab,t})$ ,  $(y_{ba,t}, y_{bb,t}, y_{bc,t})$  and  $(y_{a,t}, y_{b,t})$ , respectively. Furthermore, by contrast with a traditional bottom-up procedure, which uses information from only the bottom level, our hierarchical procedure conditions on information at multiple levels, which can be beneficial, as the series at the higher levels are smoother, and easier to model.

So far we have considered unconditional distributions, but as we are interested in multivariate forecasting, we will need a version of Sklar's theorem for conditional distributions proposed by Patton (2006). Given some information set  $\mathcal{F}_{t-1}$ , if  $\mathbf{y}_t | \mathcal{F}_{t-1} \sim \mathbf{F}(\cdot | \mathcal{F}_{t-1})$  with  $y_{it} | \mathcal{F}_{t-1} \sim F_i(\cdot | \mathcal{F}_{t-1})$ ,  $i = 1, \dots, n$ , then

$$\mathbf{F}(\mathbf{y} | \mathcal{F}_{t-1}) = \mathbf{C}(F_1(y_1 | \mathcal{F}_{t-1}), \dots, F_n(y_n | \mathcal{F}_{t-1}) | \mathcal{F}_{t-1}).$$

As in Patton (2012), we will assume the following structure for our series:

$$y_{i,t} = \mu_i(\mathbf{x}_{t-1}) + \sigma_i(\mathbf{x}_{t-1})\varepsilon_{i,t}, \quad (12)$$

where  $\mu_i$  and  $\sigma_i$  are conditional mean and standard deviation functions,  $\mathbf{x}_{t-1} \in \mathcal{F}_{t-1}$ ,  $\varepsilon_{it} | \mathcal{F}_{t-1} \sim F_i(0, 1)$ , and  $i = 1, \dots, n$ . In other words, for each series, we allow time-varying conditional mean and variance, but we assume a constant conditional distribution for the standardized residuals.

In the following, we present the different steps of our algorithm, which produces probabilistic forecasts for a hierarchy with  $n$  half-hourly series for lead times from one to  $H$  steps ahead.

1. For each series  $i = 1, \dots, n$ , obtain base forecasts and permutations, as follows:
  - (a) Fit the base model in expression (12), and compute the in-sample conditional means  $\hat{y}_{i,t} = \hat{\mu}_i(\mathbf{x}_{t-1})$  and standard deviations  $\hat{\sigma}_{i,t} = \hat{\sigma}_i(\mathbf{x}_{t-1})$ .
  - (b) Compute the base standardized residuals  $\hat{e}_{i,t} = (y_{i,t} - \hat{y}_{i,t})/\hat{\sigma}_{i,t}$ .
  - (c) Define the permutations  $p_i(t) = \text{rk}(\hat{e}_{i,t})$  as explained in Section 4.2, where  $\text{rk}(\hat{e}_{i,t})$  is the rank of  $\hat{e}_{i,t}$  within the set  $\{\hat{e}_{i,1}, \dots, \hat{e}_{i,T}\}$ . Note that it is possible to allow the copula (and the associated permutations) to vary for each half-hour. However, we did not find this beneficial, and this can be explained by the fact that we are already conditioning by half-hour for the margins.
2. Apply the MinT approach to the means of the base predictive distributions:
  - (a) Compute the covariance matrix  $\hat{\mathbf{W}}_1$  of the base forecast errors  $\hat{\mathbf{e}}_t = \mathbf{y}_t - \hat{\mathbf{y}}_t$  as explained in Section 3, and compute the combination weight matrix  $\hat{\mathbf{P}}_h$  given by expression (7).
  - (b) Using the base models from step 1a, compute for each node  $i = 1, \dots, n$  and each half-hour  $h = 1, \dots, H$ , an  $h$  period-ahead base predictive distribution  $\hat{F}_{i,T+h}$ .
  - (c) For each half-hour  $h = 1, \dots, H$ , let  $\hat{\mathbf{y}}_{T+h} = (\hat{y}_{1,T+h}, \dots, \hat{y}_{n,T+h})'$ , where  $\hat{y}_{i,T+h}$  is the base predictive mean of  $y_{i,T+h}$ , and compute the revised mean forecasts  $\tilde{\mathbf{y}}_{T+h} = \mathbf{S}\hat{\mathbf{P}}_h\hat{\mathbf{y}}_{T+h}$ , where  $\hat{\mathbf{P}}_h$  has been previously computed in step 2a.
3. For each bottom level node  $i = r + 1, \dots, n$ , produce a sample of size  $K$ ,  $x_1^i, \dots, x_K^i$ , with

$$x_k^i = \tilde{F}_{i,T+h}^{-1}(k/K + 1), \quad (13)$$

where

$$\tilde{F}_{i,T+h}^{-1}(\tau) = \hat{F}_{i,T+h}^{-1}(\tau) - \hat{y}_{i,T+h} + \tilde{y}_{i,T+h}, \quad (14)$$

and  $k = 1, \dots, K$ . In other words,  $x_1^i, \dots, x_K^i$  is a sample from the base predictive distribution  $\hat{F}_{i,T+h}$  with revised predictive mean  $\tilde{y}_{i,T+h}$ .

4. For each aggregate node  $i = 1, \dots, r$ , generate a predictive distribution:
  - (a) Let  $c(i, j)$  be the  $j$ th child node of the aggregate node  $i$ , where  $j = 1, \dots, n_i$ .
  - (b) For  $k = 1, \dots, K$ , recursively compute

$$x_k^i = x_{(p_{c(i,1)}(k))}^{c(i,1)} + \dots + x_{(p_{c(i,n_i)}(k))}^{c(i,n_i)}, \quad (15)$$

where  $x_{(k)}^i$  denotes the  $k$ th order statistic of  $\{x_1^i, \dots, x_K^i\}$ , i.e.  $x_{(1)}^i \leq \dots \leq x_{(K)}^i$ .

- (c)  $x_1^i, \dots, x_K^i$  is then a sample from  $y_{i,T+h} \sim \tilde{F}_{i,T+h}$ .

In other words, in expression (15), for each aggregate node, we first apply the permutations computed in step 1c to the independent samples of the child nodes generated using expression (13). A sample for the aggregate node is then obtained by summing the permuted samples. Note that we have used the  $T$  historical observations to compute the permutations in step 1c. Therefore, in order to apply the permutations to the independent sample of size  $K$ , we need  $K \leq T$ , i.e. the size of the independent sample must be less than or equal to the number of the historical observations. If a parametric copula is fitted, then this constraint can be relaxed since we can generate a sample of size  $K$  from the estimated copula.

Compared to a simple bottom-up approach, which uses information from only the bottom level, our algorithm synthesizes information from multiple levels. In particular, we apply the MinT approach to revise the mean forecasts, which allows us to take advantage of more accurate forecasts of the mean at higher levels. Furthermore, we estimate the permutations at different aggregate levels, which allows us to benefit from better estimation since the series are smoother, and hence easier to model. Finally, compared to MinT, our approach does not require distributional assumptions.

## 5 Forecasting Individual Electricity Demand Series

The first step of our proposed hierarchical probabilistic forecasting algorithm involves independently generating density forecasts for each series in the hierarchy. In this section, we present the density forecasting methods that we applied to each individual series. In the hierarchical forecast-



ing context, these are termed the *base* forecasting methods.

Hong & Fan (2016) describe how there are many different approaches available for probabilistic forecasting of the total electricity demand on a transmission system. For day-ahead prediction, a weather-based model is typically used (see, for example, Cottet & Smith 2003, Smith et al. 2010, Cho et al. 2013). For shorter lead times, univariate time series models have been proposed (see, for example, De Livera et al. 2011), and from these, density forecasting is straightforward. This contrasts with weather-based models, for which the generation of probabilistic demand forecasts requires density forecasts for the weather variables (Taylor & Buizza 2002). As our focus is on large-scale hierarchical forecasting, rather than the modeling of individual time series, for simplicity, we use univariate methods. Because the series in the bottom level of the hierarchy exhibit different features to the aggregate series, we use different forecasting methods for each.

## 5.1 Forecasting bottom level series

For our smart meter series, the non-Gaussian behavior and lack of clear autoregressive structure (see, for example, Figure 2) motivates a nonparametric approach to density forecasting. This is supported by the results of Arora & Taylor (2016), who found that, for residential smart meter series, several kernel density estimation (KDE) approaches outperformed exponential smoothing. We implemented the best performing of those KDE approaches (see Section 3.7 of Arora & Taylor 2016) for our smart meter series, which constitute the bottom level of our hierarchy.

This KDE approach has the appeal of simplicity, which is important because, in a smart meter application, it would typically need to be applied within an automated procedure. The approach treats the five weekdays as identical, and applies KDE separately to the 48 half-hours of the day. It treats Saturday and Sunday as different, and again performs KDE separately for each period of the day. The days of the week are, therefore, categorized into three different day types. Using historical observations  $y_1, \dots, y_T$ , the predictive density at value  $y$  for period  $T + h$  is expressed as:

$$\hat{f}_{T+h}(y) = \frac{\sum_{t \in S_{T+h}} \lambda^{\lfloor (T-t)/336 \rfloor} K_b(y_t - y)}{\sum_{t \in S_{T+h}} \lambda^{\lfloor (T-t)/336 \rfloor}},$$

where  $K_b(\cdot) = K(\cdot/b)/b$  denotes a kernel function with bandwidth  $b$ ;  $S_{T+h}$  is the set of past periods that fall on the same period of the day and the same day type as period  $T+h$ ; and  $\lambda$  ( $0 \leq \lambda \leq 1$ ) is a decay parameter. The decay is imposed in steps of size 336, which is the number of half-hours in a week. This ensures that observations belonging to the same week are given the same weighting. For each forecast origin, we implemented the KDE using the previous 3 months of data.

For each series, we selected  $b$  and  $\lambda$  by minimizing the continuous ranked probability score (CRPS) (see, for example, Gneiting & Raftery 2007) for a cross-validation period consisting of the one month immediately prior to our post-sample period. Since we used a Gaussian kernel function  $K$ , we computed the CRPS using the closed-form expression for a mixture of Gaussians (see Gritmit et al. 2006). We are not aware of the previous use of this expression for KDE. The selected values of  $b$  and  $\lambda$  tended to be quite small, indicating little kernel smoothing and fast decay.

## 5.2 Forecasting aggregated series

The autocorrelation in the time series of the aggregates (see, for example, Figure 2) motivates the use of a time series model, rather than KDE. For the modeling of the total demand on a transmission system, De Livera et al. (2011) generalize the HWT exponential smoothing model of Taylor (2010, Section 3.1), which models the intraday and intraweek cycles in intraday data, and incorporates an autoregressive model for the residuals. We chose to use this model for the aggregates, rather than the models of De Livera et al. (2011), as these require a selection procedure, which is challenging in our context, where the chosen method must be applied in an automated fashion to many series.

Let  $y_1, \dots, y_T$  be one of the aggregate series in the hierarchy. The HWT exponential smoothing model is given by the following expressions:

$$y_t = \ell_{t-1} + d_{t-48} + w_{t-336} + r_t, \quad (16)$$

$$\ell_t = \ell_{t-1} + \alpha r_t, \quad (17)$$

$$d_t = d_{t-1} + \delta r_t, \quad (18)$$

$$w_t = w_{t-1} + \omega r_t, \quad (19)$$

$$r_t = \phi r_{t-1} + \varepsilon_t, \quad (20)$$

where  $\ell_t$  is the smoothed level;  $d_t$  is the seasonal index for the intraday pattern;  $w_t$  is the seasonal index for the intraweek pattern that remains after the intraday pattern has been removed;  $\alpha$ ,  $\delta$  and  $\omega$  are smoothing parameters;  $r_t$  is a residual term that is modeled as a first order autoregressive process with parameter  $\phi$ ; and  $\varepsilon_t$  is a random noise term with mean zero and finite variance.

We estimated the initial smoothed values for the level and seasonal components of each series by averaging the observations from the first three weeks of data. For each series, we optimized the parameters by minimizing the sum of squared one step-ahead forecast errors for the first 12 months of data. Table 1 presents the optimized parameter values for the four series of aggregates considered in Figures 2 and 3. These values are similar to those found in previous applications of the model to electricity load (see, for example, Taylor & McSharry 2012), with relatively small values of  $\alpha$  and relatively high values of  $\phi$ .

An expression for the forecast of the mean is given by Taylor (2010). We constructed density forecasts using bootstrapping based on the in-sample one step-ahead forecast errors  $\{\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_T\}$ . Let  $\{\hat{\varepsilon}_{T+1}^{(b)}, \dots, \hat{\varepsilon}_{T+48}^{(b)}\}$  be the  $b$ th bootstrapped sample, with  $b = 1, \dots, B$ . Using this sample and the estimated parameters  $\hat{\alpha}$ ,  $\hat{\delta}$ ,  $\hat{\omega}$ ,  $\hat{\phi}$  in the model of expressions (16)-(20), we compute a sample path  $\{y_{T+1}^{(b)}, \dots, y_{T+48}^{(b)}\}$ . The empirical distribution of  $\{y_{T+h}^{(1)}, \dots, y_{T+h}^{(B)}\}$  is then used as the predictive distribution for  $y_{T+h}$  for each lead time  $h$ . We used  $B = 5000$  in our empirical work.

The first of the two plots in Figure 5 shows forecasts of the mean, and 50% and 95% prediction intervals produced using the exponential smoothing method for the aggregated series at the top of the hierarchy. The forecasts were produced from the end of the estimation sample of 12 months, for lead times from one half-hour up to 24 hours ahead. The second of the plots in Figure 5 gives the same information for the KDE approach for one individual smart meter series.

The simplest approach to hierarchical forecasting is to produce forecasts independently for each node of the hierarchy. In Section 3, in relation to expression (3), we referred to these as the base forecasts, and we discussed how they are unlikely to satisfy the aggregation constraint. This is illustrated by Figure 6, which shows boxplots of the resulting base forecast coherency errors, defined in (4), computed for each period of the day using the 92-day post-sample period. The two panels correspond to two of the aggregate series considered in Figures 2 and 3. Naturally, we can

Table 1: Parameters of the exponential smoothing method at different levels of aggregation.

	Number of aggregated meters			
	1578	150	30	10
$\alpha$	0.007	0.019	0.012	0.017
$\delta$	0.209	0.113	0.083	0.060
$\omega$	0.187	0.090	0.081	0.081
$\phi$	0.863	0.597	0.554	0.398

see that the magnitude of the coherency errors increases with the amount of aggregation. We can also see that the sum of the bottom level forecasts tends to predict a lower electricity demand than the base aggregate forecasts, especially during the second peak of the day around 19:30.

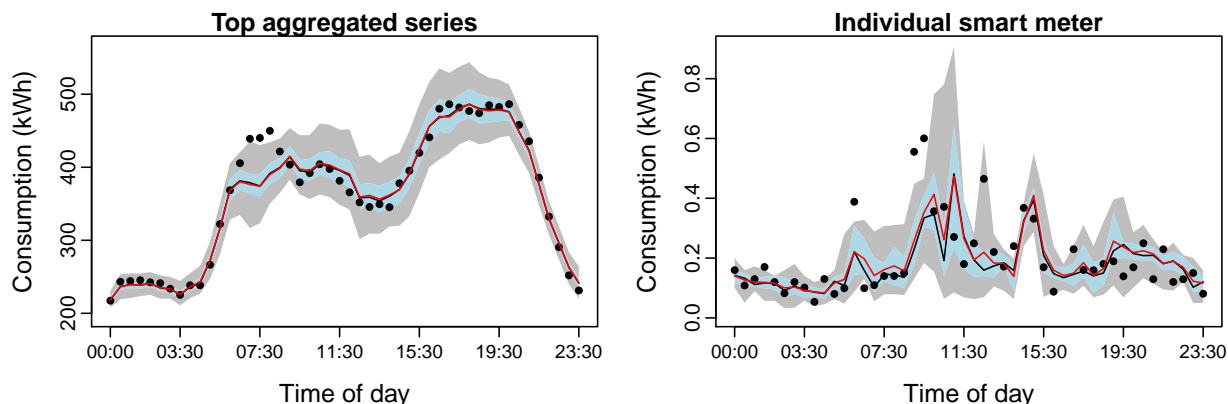


Figure 5: One day ahead probabilistic forecasts for the top aggregated series and one smart meter series, using the exponential smoothing and KDE method, respectively. The central line indicates the forecasts of the mean, and the shaded regions show 50% and 95% prediction intervals.

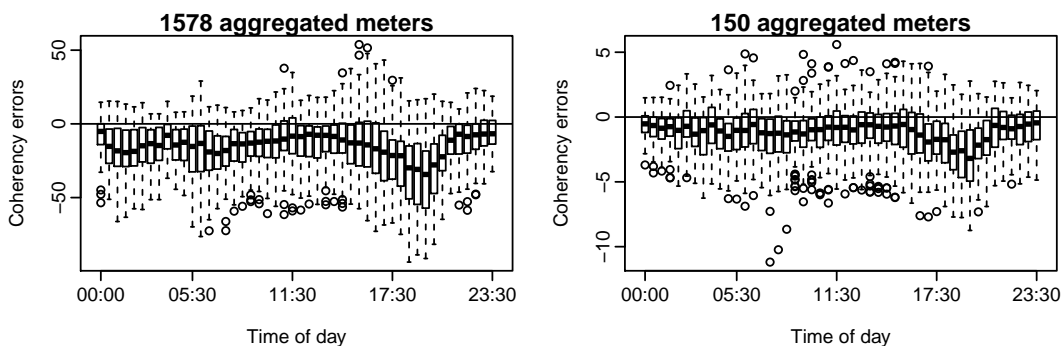


Figure 6: Box plots of coherency errors for the base forecasts of two aggregated series.

## 6 Empirical Evaluation

In this section, we empirically evaluate mean and probabilistic hierarchical forecasts. We first describe the evaluation measures that we use, and list the forecasting methods in our study.

### 6.1 Evaluation measures

To evaluate predictive distributions, we use the CRPS, which is a proper scoring rule, i.e. the score is maximized when the true distribution is reported (Gneiting & Raftery 2007). We also use the weighted form of the CRPS, which allows more emphasis to be put on probability levels of greater interest (Gneiting & Ranjan 2011). Given an  $h$  period-ahead predictive distribution  $\hat{F}_{t+h}$  and an observation  $y_{t+h}$ , the quantile-weighted version of the CRPS is

$$\text{CRPS}(\hat{F}_{t+h}, y_{t+h}) = \int_0^1 v(\tau) \text{QS}_\tau(\hat{F}_{t+h}^{-1}(\tau), y_{t+h}) d\tau, \quad (21)$$

where  $v(\tau)$  is a non-negative weight function on the unit interval, and  $\text{QS}_\tau$  is the quantile score at probability level  $\tau$ , defined as

$$\text{QS}_\tau(\hat{F}_{t+h}^{-1}(\tau), y_{t+h}) = 2(\mathbb{1}\{y_{t+h} \leq \hat{F}_{t+h}^{-1}(\tau)\} - \tau)(\hat{F}_{t+h}^{-1}(\tau) - y_{t+h}).$$

When closed-form expressions for the evaluation of expression (21) are not available, a discretized approximate version can be computed to any degree of accuracy. In our use of the weighted CRPS, we set  $v(\tau) = (2\tau - 1)^2$ , which gives more weight to the tails of the distribution.

We averaged the CRPS across the 92 days in our post-sample period. Although our main focus is probabilistic forecasting, we also evaluate forecasts of the mean. For this, we use the mean squared error, as this is a proper scoring rule for the mean (Gneiting 2011).

For conciseness, when presenting the results, we report the average of the scores for the periods of the day split into three intervals of eight successive half-hours. To compare more easily the scores, we present skill scores (see Section 2.3, Gneiting & Raftery 2007). These measure the performance relative to a reference method. In our application, a natural reference is the base

forecasts, which are produced independently for each series in the hierarchy. For each method, to calculate the skill score, we first compute the ratio of the score to that of the base forecasts, then subtract this ratio from one, and multiply the result by 100. The skill score, therefore, conveys the percentage by which a method’s score is superior to the score of the reference method.

## 6.2 Forecasting methods

Using a forecast origin of 23:30 for each of the 92 days in our post-sample period, we computed a predictive distribution for each half-hour of the next day, using the following seven methods:

1. BASE - We independently produce predictive distributions for each node in the hierarchy. For each bottom level series, we use the KDE method described in Section 5.1, and for each aggregate series, we use the exponential smoothing model of Section 5.2. This BASE method is used as the reference in the calculation of the skill scores.
2. MinTDiag - Forecasts of the means and variances for each node are computed using the MinT expressions (8) and (6), respectively. A diagonal covariance matrix  $\mathbf{W}_h$  is used within the method. Forecasts of the probability distributions are produced by assuming log-normality.
3. MinTShrink - This is identical to MinTDiag, except that a shrunken covariance matrix  $\mathbf{W}_h$  is used within the MinT method, as suggested by Wickramasuriya et al. (2015).
4. IndepBU - In this bottom-up approach, we compute the aggregate forecasts by Monte Carlo simulation. Predictive distributions for the aggregates are generated from the sum of independently sampled values of the predictive distributions, produced by the BASE method, for the bottom level nodes.
5. IndepBU-MinTShrink - This is identical to IndepBU, except that, prior to the Monte Carlo simulation, the MinTShrink method is used to revise the predictive distributions produced by the BASE method for the nodes in the bottom level.
6. DepBU - In this bottom-up approach, we impose dependency on the IndepBU method. Simulation is again used. We compute the aggregate forecasts by summing permuted indepen-

dently sampled values from the predictive distributions, produced by the BASE method, for the bottom level nodes. In other words, we use our algorithm with  $\tilde{F}_{i,T+h} = \hat{F}_{i,T+h}$  in expression (14). Note that the predictive means of IndepBU and DepBU are equal since they both reduce to simple bottom-up mean forecasts, given by expression (2).

7. DepBU-MinTShrink - This is identical to DepBU, except that, in an initial step, the MinTShrink method is used to revise the predictive distributions produced by the BASE method for the nodes in the bottom level.

Although we are not aware of previous literature that has considered any of these seven methods for probabilistic hierarchical forecasting, we should acknowledge that the BASE method is a simple benchmark, and that our MinTDiag and MinTShrink methods are founded on the work of Wickramasuriya et al. (2015), who consider hierarchical forecasting for the mean. The BASE method is typically neither mean or probabilistically coherent, and the MinT approaches require a distributional assumption. These issues motivate our use of a bottom-up approach based on Monte Carlo simulation, and we include four of these in our empirical study. While the IndepBU method can also be considered as a relatively simple benchmark, the IndepBU-MinTShrink method is more novel, with our aim being to incorporate, within the IndepBU method, the MinT approach to hierarchical forecasting of the mean. However, our main methodological contribution is the DepBU-MinTShrink method, which is the algorithm described in Section 4.3. A simplified version of this is the DepBU method, which is also new.

### 6.3 Hierarchical mean forecasting

For different times of the day, Figure 7 shows the MSE skill scores of different methods for aggregates constructed from different numbers of meters. For the x-axis, we use the log of the number of aggregated meters, as this enables us to show more easily how the extent of the aggregation affects the forecast accuracy. As we mentioned in the previous section, mean forecasts from the IndepBu and DepBU methods are the same. For these methods, Figure 7 shows negative skill scores at both intermediate and large aggregations, and this is consistent across different time periods. In other

words, aggregate forecasts obtained by simply summing the smart meter (bottom level) forecasts are poor compared to the base forecasts. This can be explained by the low signal-to-noise ratio at the smart meter level. By contrast, for the aggregate series, the BASE method has the advantage of using the observations for the aggregates, which are smoother. The lack of forecast accuracy for a bottom-up approach has also been observed in many other applications (Hyndman et al. 2011). Although BASE dominates IndepBU and DepBU, it has the disadvantage of not producing coherent forecasts, while the bottom-up forecasts are coherent by construction.

In Figure 7, both MinT methods have noticeably better MSE skill scores than IndepBU and DepBU. The positive skill scores for the MinT methods shows that they provide better forecasts than BASE at intermediate aggregations. Recall that all MinT methods are also bottom-up procedures, but with an additional forecast combination step. The improvement in forecast accuracy, in comparison with IndepBU and DepBU, shows the effectiveness of including forecast combination in a bottom-up procedure. The relative performance of MinTDiag and MinTShrink in Figure 7 depends on the time of day, and the level of aggregation. Recall that, in contrast to MinTShrink, MinTDiag ignores the correlation between the forecast errors at different nodes in the hierarchy. The relative performance of these two methods will depend on the correlation structure of the base forecast errors, and this will vary across the times of the day and levels of aggregation.

## 6.4 Hierarchical probabilistic forecasting

Figure 8 gives the CRPS skill scores for IndepBU and DepBU for different aggregations, and different times of the day. We observe similar results to those of Figure 7 for the mean forecasts. Indeed, both IndepBU and DepBU have negative skill scores at both intermediate and large aggregations, and this is consistent for different times of the day. As with the mean forecasts, it is hard to produce accurate aggregate predictive distributions from bottom level predictive distributions, as the latter are produced from individual smart meter series, which have a low signal-to-noise ratio.

Although IndepBU and DepBU give the same forecasts of the mean, they produce different aggregate predictive distributions. Figure 8 shows that, in this respect, DepBU is more accurate than IndepBU for large aggregations. This implies that modeling the dependences between nodes



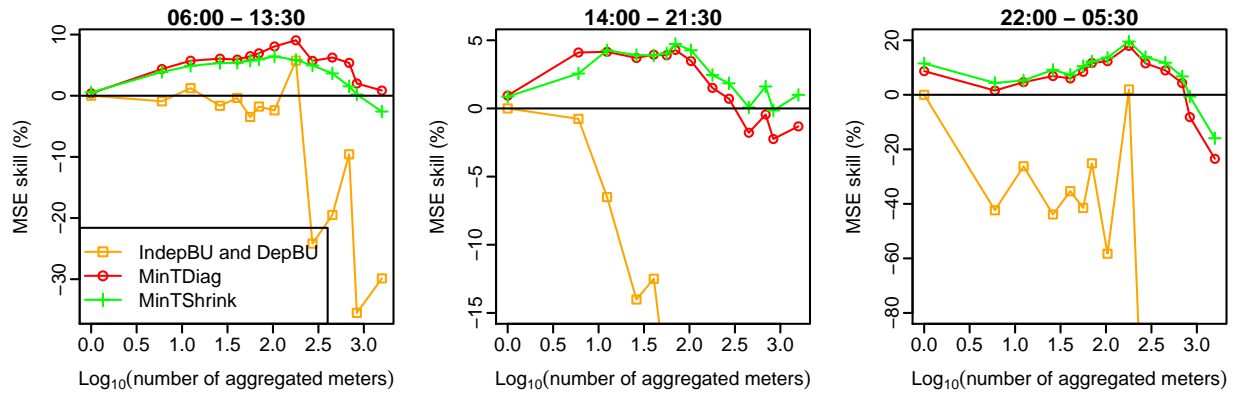


Figure 7: MSE skill score at different periods of the day for aggregates of different numbers of meters.

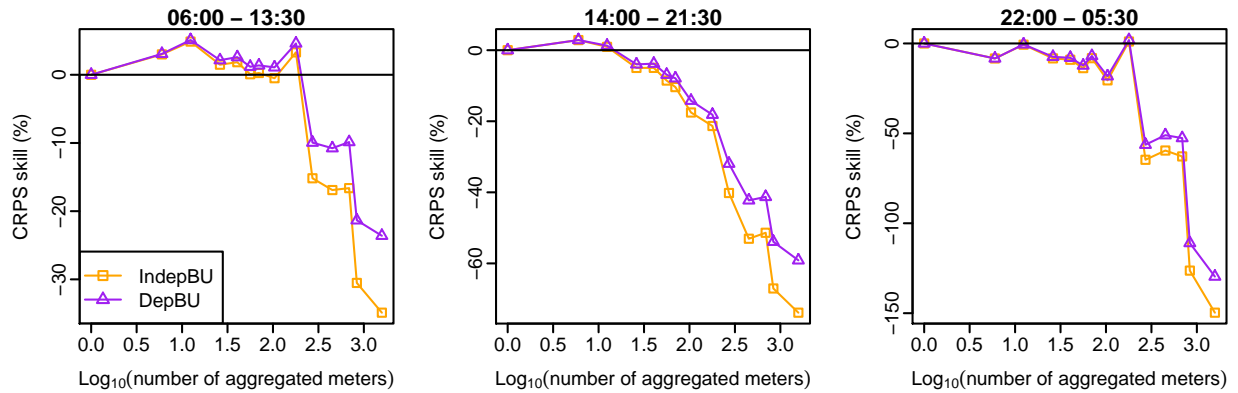


Figure 8: CRPS skill scores at different periods of the day for aggregates of different numbers of meters.

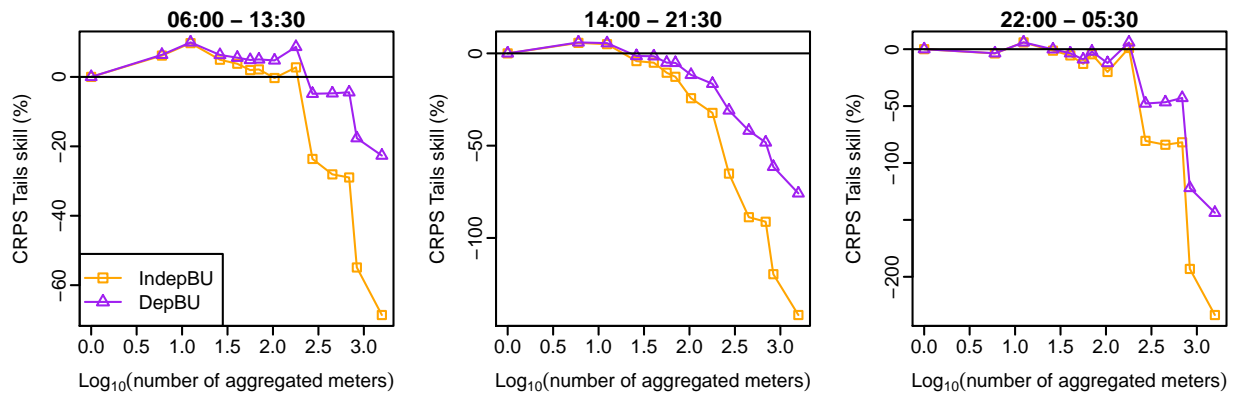


Figure 9: Weighted CRPS skill scores at different periods of the day for aggregates of different numbers of meters.

at the lower levels of the hierarchy improves accuracy for the higher levels, although it does not do so at the lower levels. The superiority of DepBU is more noticeable during day hours, which is when we would expect more dependence in the electricity consumption between different nodes.

Figure 9 reports the skill scores for the weighted CRPS. The difference between IndepBU and DepBU is larger in this figure than in Figure 8, showing that dependence modeling particularly improves the accuracy in the tails of the aggregate predictive distributions.

Figures 10 and 11 report the CRPS and weighted CRPS for MinTDiag, MinTShrink, IndepBU-MinTShrink and DepBU-MinTShrink. In Figure 10, MinTDiag is noticeably poorer than MinTShrink, especially for large aggregations. This contrasts with the results in Figure 7 for forecasts of the mean. This is not surprising, as the error correlations directly affect the variance, and hence the predictive distributions, but have only a second-order effect on the mean. The weighted CRPS results of Figure 11 show an even larger difference between MinTDiag and MinTShrink, indicating that capturing the correlation has a particularly strong impact on distributional tail accuracy.

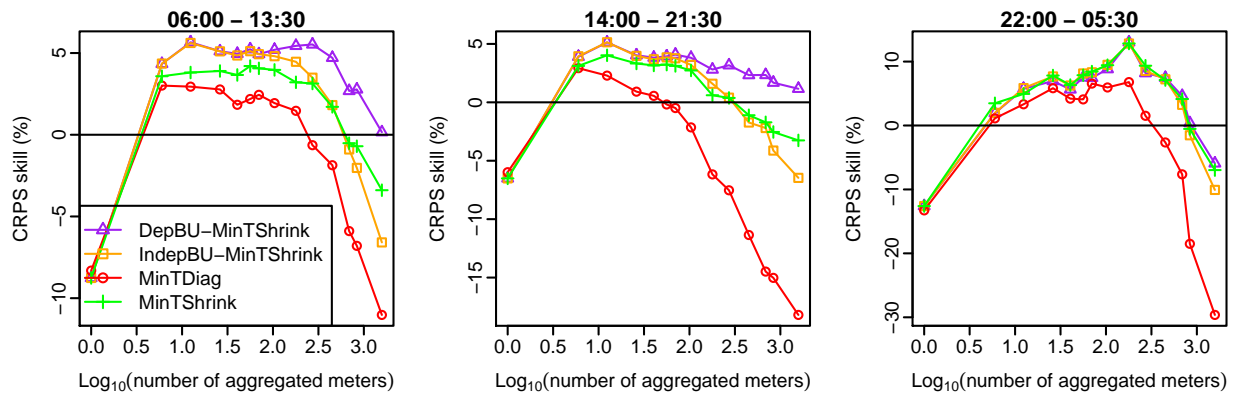


Figure 10: CRPS skill scores at different periods of the day for aggregates of different numbers of meters.

Comparing IndepBU-MinTShrink and DepBU-MinTShrink in Figure 10 with IndepBU and DepBU in Figure 11 shows, as with mean forecasting, the advantage of mean forecast combination on the accuracy of the aggregate predictive distributions. Indeed, the accuracy of any density forecast is typically heavily dependent on the accuracy of the forecast of the location of the density.

It is interesting to see that IndepBU-MinTShrink, which involves an independence assumption, sometimes has better skill scores than MinTShrink which models the full covariance matrix. While

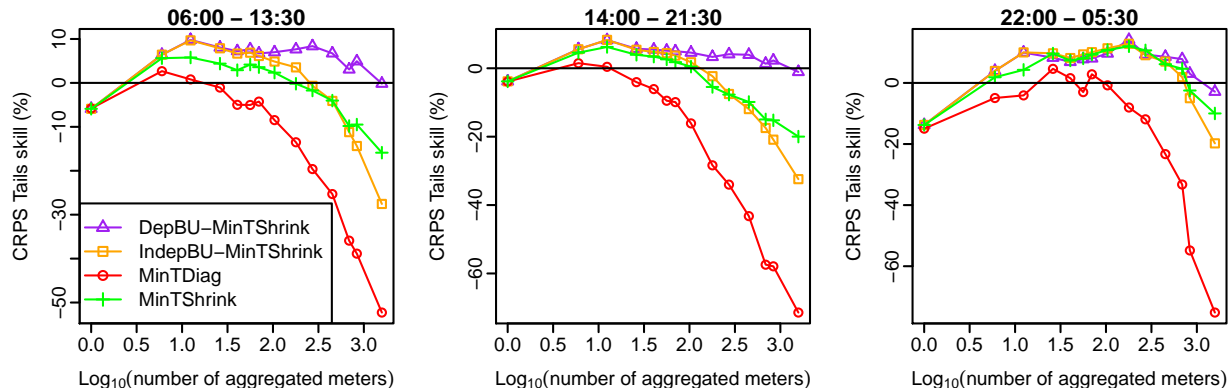


Figure 11: Weighted CRPS skill scores at different periods of the day for aggregates of different numbers of meters.

MinTShrink uses a log-normality assumption for the bottom level nodes, IndepBU-MinTShrink seemingly has the advantage of using a non-parametric approach for these nodes.

As with DepBU in Figure 8, DepBU-MinTShrink has better skill scores than IndepBU-MinTShrink during day hours, and similar or better skill scores during night hours. We observe the same behavior with MinTShrink. The negative skill scores for all methods at the bottom level suggests that the mean forecast combination does not improve the CRPS with respect to the base forecasts.

Finally, the fact that DepBU-MinTShrink provides better forecasts than BASE at almost all aggregations and times of the day, while being a bottom-up procedure, suggests that our algorithm is effective in providing both coherent and accurate probabilistic forecasts.

## 7 Conclusion

Modern electricity networks provide remarkably rich sources of data at a highly disaggregated level. As with many “big data” problems, the richness of the data set presents some new and interesting statistical challenges. We have focused on one such challenge, namely the problem of computing coherent probabilistic forecasts for a hierarchy of electricity consumption time series.

While previous considerations of hierarchical forecasting have focused on ensuring that the mean forecasts are coherent, we have tackled the much larger problem of obtaining probabilistically coherent forecast distributions. In fact, we have exploited recent results to ensure the means

of our probability distributions are not only coherent but efficiently use all available information. In principle, coherent forecast distributions require the estimation of the entire joint distribution of all the bottom-level time series. To circumvent this problem, we exploit the hierarchical structure of the data, allowing us to reduce the problem to smaller sub-problems, each of which is computationally tractable. Another key feature of our approach is that we make no assumptions about the underlying distributions, using conditional KDE for the smart meter time series, and using fast time series models that handle the multiple seasonal patterns observed in the aggregated data. In our empirical study, the best performing method, overall, combines mean forecast combination using a shrinkage estimator for the covariance matrix of the one-step forecast errors, with our copula-based approach for computing the predictive distributions of all nodes in the hierarchy.

Although we have focused on electricity demand, our work is equally applicable to other contexts, such as forecasting sales product hierarchies, or tourist numbers in geographical hierarchies.

## References

- AECOM (2011), Energy demand research project: Final analysis, Technical report, AECOM House, Hertfordshire, UK.
- AECOM Building Engineering (2014), ‘Energy demand research project: Early smart meter trials, 2007-2010’, <http://doi.org/10.5255/UKDA-SN-7591-1>.
- Arbenz, P., Hummel, C. & Mainik, G. (2012), ‘Copula based hierarchical risk aggregation through sample reordering’, *Insurance, Mathematics & Economics* **51**(1), 122–133.
- Arora, S. & Taylor, J. W. (2016), ‘Forecasting electricity smart meter data using conditional kernel density estimation’, *Omega* **59, Part A**, 47–59.
- Ben Taieb, S., Huser, R., Hyndman, R. J. & Genton, M. G. (2016), ‘Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression’, *IEEE Transactions on Smart Grid* **7**(5), 2448–2455.
- Berrocal, V. J., Raftery, A. E., Gneiting, T. & Steed, R. C. (2010), ‘Probabilistic weather forecasting for winter road maintenance’, *Journal of the American Statistical Association* **105**(490), 522–537.
- Borenstein, S. & Bushnell, J. (2015), ‘The US electricity industry after 20 years of restructuring’, *Annual Review of Economics* **7**(1), 437–463.
- 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.

- Cottet, R. & Smith, M. (2003), ‘Bayesian modeling and forecasting of intraday electricity load’, *Journal of the American Statistical Association* **98**(464), 839–849.
- De Livera, A. M., Hyndman, R. J. & Snyder, R. (2011), ‘Forecasting time series with complex seasonal patterns using exponential smoothing’, *Journal of the American Statistical Association* **106**(496), 1513–1527.
- Gijbels, I. & Herrmann, K. (2014), ‘On the distribution of sums of random variables with copula-induced dependence’, *Insurance, Mathematics & Economics* **59**, 27–44.
- Gneiting, T. (2011), ‘Making and evaluating point forecasts’, *Journal of the American Statistical Association* **106**(494), 746–762.
- Gneiting, T. & Raftery, A. E. (2007), ‘Strictly proper scoring rules, prediction, and estimation’, *Journal of the American Statistical Association* **102**(477), 359–378.
- Gneiting, T. & Ranjan, R. (2011), ‘Comparing density forecasts using threshold- and Quantile-Weighted scoring rules’, *Journal of Business & Economic Statistics* **29**(3), 411–422.
- Grimit, E. P., Gneiting, T., Berrocal, V. J. & Johnson, N. A. (2006), ‘The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification’, *Quarterly Journal of the Royal Meteorological Society* **132**(621C), 2925–2942.
- Haben, S., Ward, J., Vukadinovic Greetham, D., Singleton, C. & Grindrod, P. (2014), ‘A new error measure for forecasts of household-level, high resolution electrical energy consumption’, *International Journal of Forecasting* **30**(2), 246–256.
- Hong, T. & Fan, S. (2016), ‘Probabilistic electric load forecasting: A tutorial review’, *International Journal of Forecasting* **32**(3), 914–938.
- Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G. & Shang, H. L. (2011), ‘Optimal combination forecasts for hierarchical time series’, *Computational Statistics & Data Analysis* **55**(9), 2579–2589.
- Hyndman, R. J., Lee, A. J. & Wang, E. (2016), ‘Fast computation of reconciled forecasts for hierarchical and grouped time series’, *Computational Statistics & Data Analysis* **97**, 16–32.
- Iman, R. L. & Conover, W. J. (1982), ‘A distribution-free approach to inducing rank correlation among input variables’, *Communications in Statistics - Simulation and Computation* **11**(3), 311–334.
- Jeon, J. & Taylor, J. W. (2012), ‘Using conditional kernel density estimation for wind power density forecasting’, *Journal of the American Statistical Association* **107**(497), 66–79.
- Kremer, M., Siemsen, E. & Thomas, D. J. (2016), ‘The sum and its parts: Judgmental hierarchical forecasting’, *Management Science* **62**(9), 2745–2764.
- López Cabrera, B. & Schulz, F. (2017), ‘Forecasting generalized quantiles of electricity demand: A functional data approach’, *Journal of the American Statistical Association* **112**(517), 127–136.
- Mainik, G. (2015), ‘Risk aggregation with empirical margins: Latin hypercubes, empirical copulas, and convergence of sum distributions’, *Journal of Multivariate Analysis* **141**, 197–216.
- Mirowski, P., Chen, S., Ho, T. K. & Yu, C.-N. (2014), ‘Demand forecasting in smart grids’, *Bell Labs Technical Journal* **18**(4), 135–158.

- Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.
- Patton, A. J. (2006), ‘Modelling asymmetric exchange rate dependence’, *International Economic Review* **47**(2), 527–556.
- Patton, A. J. (2012), ‘Copula methods for forecasting multivariate time series’, *Handbook of Economic Forecasting* (April), 1–76.
- Ramchurn, S. D., Vytelingum, P., Rogers, A. & Jennings, N. R. (2012), ‘Putting the ‘smarts’ into the smart grid: a grand challenge for artificial intelligence’, *Communications of the ACM* **55**(4), 86–97.
- Rüschendorf, L. (2009), ‘On the distributional transform, sklar’s theorem, and the empirical copula process’, *Journal of Statistical Planning and Inference* **139**(11), 3921–3927.
- Schefzik, R., Thorarinsdottir, T. L. & Gneiting, T. (2013), ‘Uncertainty quantification in complex simulation models using ensemble copula coupling’, *Statistical Science: a review journal of the Institute of Mathematical Statistics* **28**(4), 616–640.
- Sklar, M. (1959), *Fonctions de répartition à n dimensions et leurs marges*, Université Paris 8.
- Smith, M., Min, A., Almeida, C. & Czado, C. (2010), ‘Modeling longitudinal data using a Pair-Copula decomposition of serial dependence’, *Journal of the American Statistical Association* **105**(492), 1467–1479.
- Sun, X., Luh, P. B., Cheung, K. W., Guan, W., Michel, L. D., Venkata, S. S. & Miller, M. T. (2016), ‘An efficient approach to Short-Term load forecasting at the distribution level’, *IEEE Transactions on Power Systems* **31**(4), 2526–2537.
- Taylor, J. & McSharry, P. (2012), Univariate methods for Short-Term load forecasting, in M. El-Hawary, ed., ‘Advances in Electric Power and Energy; Power Systems Engineering, Power Engineering Series of IEEE Press’, Wiley.
- Taylor, J. W. (2010), ‘Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles’, *International journal of forecasting* **26**(4), 627–646.
- Taylor, J. W. & Buizza, R. (2002), ‘Neural network load forecasting with weather ensemble predictions’, *IEEE Transactions on Power Systems* **17**(3), 626–632.
- van Erven, T. & Cugliari, J. (2015), Game-Theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts, in ‘Modeling and Stochastic Learning for Forecasting in High Dimensions’, Lecture Notes in Statistics, Springer International Publishing, pp. 297–317.
- Wickramasuriya, S. L., Athanasopoulos, G. & Hyndman, R. J. (2015), Forecasting hierarchical and grouped time series through trace minimization, Technical Report 15/15, Monash University.