MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Coherent Probabilistic Forecasts for Hierarchical Time Series

Souhaib Ben Taieb, James W. Taylor, Rob J. Hyndman

April 2017

Working Paper 03/17

Coherent Probabilistic Forecasts for Hierarchical Time Series

Souhaib Ben Taieb
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Souhaib.BenTaieb@monash.edu

James W Taylor
Saïd Business School
University of Oxford
Oxford, OX1 1HP, UK.
Email: James.Taylor@sbs.ox.ac.uk

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Rob.Hyndman@monash.edu

14 April 2017

JEL classification: C53, Q47, C32

Coherent Probabilistic Forecasts for Hierarchical Time Series

Abstract

Many applications require forecasts for a hierarchy comprising a set of time series along with aggregates of subsets of these series. Although forecasts can be produced independently for each series in the hierarchy, typically this does not lead to coherent forecasts — the property that forecasts add up appropriately across the hierarchy. State-of-the-art hierarchical forecasting methods usually reconcile the independently generated forecasts to satisfy the aggregation constraints. A fundamental limitation of prior research is that it has considered only the problem of forecasting the mean of each time series. We consider the situation where probabilistic forecasts are needed for each series in the hierarchy. We define forecast coherency in this setting, and propose an algorithm to compute predictive distributions for each series in the hierarchy. Our algorithm has the advantage of synthesizing information from different levels in the hierarchy through a sparse forecast combination and a probabilistic hierarchical aggregation. We evaluate the accuracy of our forecasting algorithm on both simulated data and large-scale electricity smart meter data. The results show consistent performance gains compared to state-of-the art methods.

Keywords: forecast combination, probabilistic forecast, copula, machine learning

1 Introduction

Producing forecasts that support decision-making in a hierarchical structure is a central problem for many organizations. For example, retail sales forecasts typically form a hierarchy, with the inventory control system of a retail outlet relying on forecasts for store-level demand, while forecasts of regionally aggregated demand are needed for managing inventory at a distribution centre (Kremer, Siemsen, and Thomas, 2016). Another context where a hierarchy naturally arises is electricity demand, where the bottom level might consist of time series of the electricity consumption of individual customers, while the top level could be the total load on the grid. Forecasts of electricity consumption are needed at various levels of aggregation in order to operate the power grid efficiently and securely (Ben Taieb et al., 2017).

Producing accurate forecasts for these hierarchical structures is particularly challenging. First, the many time series involved can interact in varying and complex ways. In particular, time series at different levels of the hierarchy can contain very different patterns (see, for example, Figure 3); time series at the bottom level are typically very noisy sometimes exhibiting intermittency, while aggregated series at higher levels are much smoother. As a result, a naive bottom-up approach whereby forecasts of aggregates are generated by summing the forecasts of the corresponding series in the lower levels is unlikely to deliver accurate results when the aggregation involves a large number of series (Hyndman, Ahmed, et al., 2011).

Second, in order to ensure coherent decision-making at the different levels of a hierarchy, it is essential that the forecast of each aggregated series should equal the sum of the forecasts of the corresponding disaggregated series. Unfortunately, independently forecasting each time series within each level is very unlikely to deliver coherent forecasts. Finally, the bottom level can consist of several thousand or even millions of time series, which can induce a massive computational load.

Recent work in this area (Erven and Cugliari, 2015; Wickramasuriya, Athanasopoulos, and Hyndman, 2015) has focused on a two-stage approach in which base forecasts are first produced independently for each series in the hierarchy; these are then combined to generate coherent revised forecasts (see Section 2). The rationale behind this approach is to both improve forecast accuracy due to the synthesis of information from different forecasts, as well as produce coherent forecasts. A fundamental limitation of actual research is that it has looked only at the problem of forecasting the mean of each time series. This contrasts with the shift in the forecasting literature over the past two decades towards probabilistic forecasting (Gneiting and Katzfuss, 2014). This form of prediction quantifies the uncertainty, which enables improved decision making and risk management (see, for example, Berrocal et al. (2010)).

We address the key problem of generating probabilistic forecasts for large-scale hierarchical time series. This problem is particularly challenging since it is required to forecast the entire distribution of future observations, not only the mean (Hothorn, Kneib, and Bühlmann, 2014; Kneib, 2013). Furthermore, because of the hierarchical structure, this problem also involves computing the distribution of hierarchical sums of random variables in high dimensions. Finally, another challenge is the possible variety of distributions in the hierarchy. In fact, although the distributions become more normally distributed with the aggregation level as a consequence of the central limit theorem, the series at lower levels often exhibit non-normality including multi-modality and high levels of skewness.

We propose an algorithm that computes predictive distributions under the form of random samples for each series in the hierarchy. First, probabilistic forecasts are independently computed for all series in the hierarchy, and samples are computed from the associated predictive distributions. Then, a sequence of permutations extracted from estimated copulas are applied to the multivariate samples in a hierarchical manner to restore the dependencies between the variables before computing the sums (see Section 3). Finally, the algorithm computes sparse forecast combinations for all series in the hierarchy, where the combination weights are estimated by solving a possibly high-dimensional LASSO problem (see Section 3.2). The result is a set of coherent probabilistic forecasts for each series in the hierarchy.

Our algorithm has multiple advantages compared to the state-of-the art hierarchical forecasting methods: (1) it quantifies the uncertainty in the predictions for the entire hierarchy while satisfying the aggregation constraints; (2) it is scalable to high-dimensional hierarchies since the problem is decomposed into multiple lower-dimensional sub-problems; and (3) it synthesizes information from different levels in the hierarchy to estimate the marginal forecasts and the dependence structures through the mean forecast combination and the hierarchical aggregation, respectively.

We evaluate our algorithm using both simulated data sets (see Section 4.2) and a large scale electricity smart meter data set (see Section 4.3).

2 Mean Hierarchical Forecasting

An hierarchical time series is a multivariate time series with an hierarchical structure. Figure 1 gives an example with five bottom series and three aggregate series. The different observations in the hierarchy satisfy the following aggregation constraints:

$$y_t = y_{A,t} + y_{B,t}, y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t} \text{ and } y_{B,t} = y_{BA,t} + y_{BB,t}$$

for all time periods $t = 1, \ldots, T$.

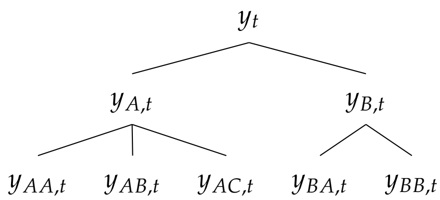

Figure 1: *Example of a hierarchical time series .*

Let $\boldsymbol{a}_t$ be an r-vector containing the observations at the different levels of aggregation at time t, $\boldsymbol{b}_t$ be an m-vector with the observations at the bottom level only, and $\boldsymbol{y}_t = (\boldsymbol{a}_t \; \boldsymbol{b}_t)'$ be an n-vector that contains the observations of all series in the hierarchy with $n = r + m$. We can then write

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t,$$

where $\boldsymbol{S} = \left[\boldsymbol{S}_a' \; \boldsymbol{I}_m\right]' \in \{0,1\}^{n \times m}$ is the summing matrix.

Suppose we have access to T historical observations, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, of a hierarchical time series. Under mean squared error (MSE) loss, the optimal h-period-ahead forecasts are given by the conditional mean (Gneiting, 2011), i.e.

$$\mathbb{E}[\boldsymbol{y}_{T+h} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T] = \boldsymbol{S}\, \mathbb{E}[\boldsymbol{b}_{T+h} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T], \tag{1}$$

where $h = 1, 2, \ldots, H$.

It is possible to compute forecasts for all series at all levels independently, which we call *base* forecasts. For example, we can estimate $\mathbb{E}[y_{i,T+h} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$ for $i = 1, \ldots, n$, i.e. for all nodes in the hierarchy. This approach is very flexible since we can use different forecasting methods for each series and aggregation level. However, the aggregation constraints will not necessarily be satisfied.

Definition 1 *Let $\hat{\boldsymbol{r}}_{T+h} = \hat{\boldsymbol{a}}_{T+h} - \boldsymbol{S}_a\hat{\boldsymbol{b}}_{T+h}$ denote the coherency errors of the h-period-ahead base forecasts $\hat{\boldsymbol{y}}_{T+h} = (\hat{\boldsymbol{a}}_{T+h}\ \hat{\boldsymbol{b}}_{T+h})'$. In other words, $\hat{\boldsymbol{r}}_{T+h}$ is a vector containing the magnitude of constraint violations for each aggregate series. Then, the forecasts $\hat{\boldsymbol{y}}_{T+h}$ are coherent if $\hat{\boldsymbol{r}}_{T+h} = \boldsymbol{0}$, i.e. if there are no coherency errors.*

Since the optimal mean forecasts in (1) are coherent by definition, it is necessary to impose the aggregation constraints when generating hierarchical mean forecasts. Also, from a decision-making perspective, coherent forecasts will guarantee coherent decisions over the entire hierarchy.

2.1 Best Linear Unbiased Mean Revised Forecasts

Hyndman, Ahmed, et al. (2011) proposed to compute coherent hierarchical mean forecasts of the following form:

$$\tilde{\boldsymbol{y}}_{T+h} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h}, \tag{2}$$

for some appropriately chosen matrix $\boldsymbol{P} \in \mathbb{R}^{m\times n}$, and where $\hat{\boldsymbol{y}}_{T+h}$ are some base forecasts.

This approach has multiple advantages: (1) the forecasts are coherent by construction; (2) the forecasts are generated by combining forecasts from all levels; and (3) multiple hierarchical forecasting methods can be represented as particular cases, including bottom-up forecasts with $\boldsymbol{P} = \left[\boldsymbol{0}_{m\times r}|\boldsymbol{1}_{m\times m}\right]$, and top-down forecasts with $\boldsymbol{P} = \left[\boldsymbol{p}_{m\times 1}|\boldsymbol{0}_{m\times(n-1)}\right]$ where $\boldsymbol{p}$ is a vector of proportions that sum to one.

Theorem 2 *(Adapted from Wickramasuriya, Athanasopoulos, and Hyndman, 2015) Let $\boldsymbol{W}_h$ be the positive definite covariance matrix of the h-period-ahead base forecast errors, $\hat{\boldsymbol{e}}_{T+h} = \boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h}$, i.e. $\boldsymbol{W}_h = \mathbb{E}[\hat{\boldsymbol{e}}_{T+h}\hat{\boldsymbol{e}}'_{T+h}]$.*

Then, assuming unbiased base forecasts, the best (i.e. having minimum sum of variances) linear unbiased revised forecasts are given by (2) with $\boldsymbol{P} = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}$. We will denote this method MinT.

In practice, the error covariance matrix $\boldsymbol{W}_h$ needs to be estimated using historical observations of the base forecast errors. Wickramasuriya, Athanasopoulos, and Hyndman (2015) estimated $\boldsymbol{W}_1$, and assumed that $\boldsymbol{W}_h \propto \boldsymbol{W}_1$, since the estimation of $\boldsymbol{W}_h$ is challenging for $h > 1$. To trade off bias and estimation variance, structural assumptions on the entries of the sample covariance matrix have also been considered in Hyndman, Lee, and Wang (2016).

2.2 Optimal Mean Combination and Reconciliation

The approach presented in the previous section applies both combination and reconciliation of the forecasts at the same time. Erven and Cugliari (2015) proposed to split the problem into two independent steps: "first one comes up with the best possible forecasts for the time series without worrying about ... coherency; and then a reconciliation procedure is used to make the forecasts ... coherent".

Given some possibly incoherent base forecasts $\hat{\boldsymbol{y}}_{T+h}$, and a weight matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, they proposed a method called GTOP which solves the following quadratic optimization problem:

$$\underset{\boldsymbol{x}_a \in \mathbb{R}^r, \boldsymbol{x}_b \in \mathbb{R}^m}{\text{minimize}} \left\| \boldsymbol{A}\hat{\boldsymbol{y}}_{T+h} - \boldsymbol{A} \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \right\|^2 \quad (3)$$

$$\text{subject to } (\boldsymbol{x}_a \ \boldsymbol{x}_b)' \in \mathcal{A} \cap \mathcal{B},$$

where $\mathcal{A} = \{(\boldsymbol{x}_a \ \boldsymbol{x}_b)' : \boldsymbol{x}_a = \boldsymbol{S}_a \boldsymbol{x}_b\}$ is the set of coherent vectors, and $\mathcal{B}$ is an additional set that allows the specification of additional constraints.

The solution of the previous problem is also equivalent to an optimal strategy in a minimax problem where the goal is to minimize the minimax error between the loss of the reconciled and the base forecasts. When $\boldsymbol{A} = \boldsymbol{I}$ and $\mathcal{B} = \emptyset$, the problem reduces to finding the closest reconciled forecasts to the base forecasts in terms of sum of squared errors (SSE).

A distinctive advantage of the GTOP approach compared to MinT is the guarantee to produce revised forecasts $\tilde{\boldsymbol{y}}_{T+h} = (\boldsymbol{x}_a^* \ \boldsymbol{x}_b^*)'$ with the same or smaller SSE than the base forecasts $\hat{\boldsymbol{y}}_{T+h}$. Furthermore, compared to MinT, the base forecasts are not required to be unbiased. Also, by separating forecast combination and reconciliation, the GTOP approach allows the inclusion of regularization in the forecast combination step. One comparative weakness of GTOP is that it does not have a closed-form solution in the general case.

3 Probabilistic Hierarchical Forecasting

Given some possibly incoherent h-period-ahead base forecasts, GTOP allows the computation of coherent *mean* forecasts, but do not provide any quantification of the uncertainty in the predictions. MinT allows for both coherent mean forecasts, and the calculation of the associated forecast variances, although Wickramasuriya, Athanasopoulos, and Hyndman (2015) do not discuss the variances in any detail.

This contrasts with the shift, in the forecasting literature, over the past two decades, towards probabilistic forecasting (Gneiting and Katzfuss, 2014). This form of prediction quantifies the uncertainty, which enables improved decision making and risk management. Probabilistic forecasts require the estimation of the conditional predictive cumulative distribution function for all series in the hierarchy:

$$F_{i,T+h}(y|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T) = \mathbb{P}(y_{i,T+h} \leq y|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T),$$

and not only the conditional mean $\mathbb{E}[y_{i,T+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$ or conditional variance $\mathbb{V}[y_{i,T+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$, with $i = 1, \ldots, n$.

As with mean forecasts, it is possible to compute probabilistic forecasts for each series in the hierarchy, but, again, these forecasts will not necessarily be coherent as defined below.

Definition 3 *Let $X_i \sim \hat{F}_i$ for $i = 1, \ldots, n$, and let $i_1, \ldots, i_{n_k}$ denote the n_k children of series i. The forecasts $\hat{F}_i$ are probabilistically coherent if $X_i \stackrel{d}{=} X_{i_1} + \cdots + X_{i_{n_k}}$ for $i = 1, \ldots, r$, where $\stackrel{d}{=}$ denotes equality in distribution.*

In other words, the predictive distribution of each aggregate series must be equal to the distribution of the sum of the children series.

3.1 Bottom-Up Probabilistic Forecasting

With mean forecasts, it was possible to compute coherent bottom-up forecasts for the ith aggregated series by simply summing the associated lowest level mean forecasts, i.e. $\tilde{y}_{it} = \boldsymbol{s}_i' \hat{\boldsymbol{y}}_t$ where $\boldsymbol{s}_i$ is the ith row of the $\boldsymbol{S}$ matrix, and $i = 1, \ldots, r$. Now, given some base probabilistic forecasts for all the bottom series, how do we compute the bottom-up coherent probabilistic forecasts for all aggregated series? Since each aggregate series is the sum of a subset of bottom series, bottom-up probabilistic forecasting are harder to compute than mean forecasts because we need to compute the joint distribution of the component random variables. The marginal predictive distributions are not enough.

Definition 4 *Let $X_1, \ldots, X_d$ be a set of continuous random variables with joint distribution function $\boldsymbol{F}$. Then, the distribution of $Z = \sum_{i=1}^d X_i$ is given by*

$$F_{X_1+\cdots+X_d}(z) = \int_{\mathbb{R}^d} \mathbf{1}\{x_1 + \cdots + x_d \leq z\}\, d\boldsymbol{F}(x_1, \ldots, x_d). \tag{4}$$

To model the joint distribution, we can resort to the copula framework (Nelsen, 2007). Copulas originate from Sklar's theorem (Sklar, 1959), which states that for any continuous distribution function $\boldsymbol{F}$ with marginals $F_1, \ldots, F_d$, there exists a unique function $\boldsymbol{C} : [0,1]^d \to [0,1]$ such that $\boldsymbol{F}$ can be written as $\boldsymbol{F}(x_1, \ldots, x_n) = \boldsymbol{C}(F_1(x_1), \ldots, F_d(x_d))$. In other words, starting from marginal predictive distributions for each series, and using a copula for the dependence structure, we can first compute the joint distribution, and then compute the distribution of the sum using (4).

Although it is convenient to decompose the estimation of the joint distribution into the estimation of multiple marginal predictive distributions and one copula, the number of bottom series can be large in practice, which implies a high-dimensional copula. Furthermore, in highly disaggregated time series data, the bottom series are often very noisy, and as a result, the estimation of the dependence structure between all bottom series will be hard to estimate.

Since we are only interested in specific aggregations, we can avoid explicitly modelling the (often) high-dimensional copula that describes the dependence between all bottom series. Building on the approach proposed by Arbenz, Hummel, and Mainik (2012), we propose to decompose the possibly high-dimensional copula into multiple lower-dimensional copulas for all child series of each aggregate series.

Example 1 *Let us consider the hierarchy given in Figure 1. A classical bottom-up approach would require modelling the joint distribution of* $(y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t})$. *Then, the distribution of all aggregate series* $y_{A,t}$, $y_{B,t}$ *and* y_t *can be computed using (4).*

However, since the marginals and the copula completely specify the joint distribution, the following procedure allows us to compute the marginal predictive distributions of all aggregates using three lower-dimensional copulas in an hierarchal manner:

1. *Compute* $F_{AA,t}$, $F_{AB,t}$, $F_{AC,t}$, $F_{BA,t}$, *and* $F_{BB,t}$.
2. *Compute* $F_{A,t}$ *using* $\boldsymbol{C}_1(F_{AA,t}, F_{AB,t}, F_{AC,t})$.
3. *Compute* $F_{B,t}$ *using* $\boldsymbol{C}_2(F_{BA,t}, F_{BB,t})$.
4. *Compute* F_t *using* $\boldsymbol{C}_3(F_{A,t}, F_{B,t})$.

Except in some special cases where the distribution of the sum can be computed analytically, we would typically resort to Monte Carlo simulations. Let us assume that $\boldsymbol{F}(x_1, \dots, x_d) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d) = \boldsymbol{C}(F_1(x_1), \dots, F_d(x_d))$. Suppose we have samples $x_k^i \sim F_i$, and $\boldsymbol{u}_k = (u_k^1, \dots, u_k^d) \sim \boldsymbol{C}$, $k = 1, \dots, K$, then we can compute

$$\hat{\boldsymbol{F}}(x_1, \dots, x_d) = \hat{\boldsymbol{C}}(\hat{F}_1(x_1), \dots, \hat{F}_d(x_d)),$$

where $\hat{F}_i$ are the empirical margins and $\hat{\boldsymbol{C}}$ is the empirical copula (see Rüschendorf, 2009, and the references therein), given respectively by

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

and

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

for $\boldsymbol{u} = (u_1, \dots, u_d) \in [0, 1]^d$, where $\mathrm{rk}(u_k^i)$ is the rank of u_k^i within the set $\{u_1^i, \dots, u_K^i\}$.

The procedure of applying empirical copulas to empirical margins can be efficiently represented in terms of sample reordering. In fact, 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 $\{1, \dots, K\}$, defined by $p_i(k) = \mathrm{rk}(u_k^i)$ for $k = 1, \dots, K$. If we then 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 $\hat{\boldsymbol{C}}$. We can then compute the samples for the sum $\{x_1, \dots, x_K\}$ where $x_k = \sum_{i=1}^d x_k^i$.

Introducing a dependence structure into originally independent marginal samples goes back to Iman and Conover (1982) who considered the special case of normal copulas. A similar idea has been considered more recently in Schefzik, Thorarinsdottir, and Gneiting (2013) to specify multivariate dependence structure with applications to weather forecasting.

Since we are interested in multivariate forecasting, we will need another version of Sklar's theorem for conditional joint distributions proposed by Patton (2006):

$$
\begin{aligned}
&\text{If } \boldsymbol{y}_t|\mathcal{F}_{t-1} \sim \boldsymbol{F}(\cdot|\mathcal{F}_{t-1}),\\
&\text{with } y_{it}|\mathcal{F}_{t-1} \sim F_i(\cdot|\mathcal{F}_{t-1}), \quad i = 1, \dots, n,\\
&\text{then}\\
&\boldsymbol{F}(\boldsymbol{y}|\mathcal{F}_{t-1}) = \boldsymbol{C}(F_1(y_1|\mathcal{F}_{t-1}), \dots, F_n(y_n|\mathcal{F}_{t-1})|\mathcal{F}_{t-1}).
\end{aligned}
$$

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

$$
y_{it} = \mu_i(\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \dots) + \sigma_i(\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \dots)\varepsilon_{it}, \tag{5}
$$

where $\varepsilon_{it}|\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \cdots \sim F_i(0, 1)$. In other words, each series can have a potentially time-varying conditional mean and variance, but the standardized residual, ε_{it}, has a constant conditional distribution for simplicity. See Fan and Patton (2014) for a review on copulas in econometrics.

The following algorithm describes how to compute the bottom-up samples using the reordering procedure for a complete hierarchy:

Algorithm 5 *(Bottom-up Probabilistic Forecasting)*

1. *For all series in the hierarchy, as defined in* (5)*, model the conditional marginal distributions; i.e. compute* $\hat{\mu}_i$ *and* $\hat{\sigma}_i$ *for* $i = 1, \dots, n$*.*
2. *Then, compute the standardized residuals* $\hat{\varepsilon}_{it} = (y_{i,t} - \hat{\mu}_{i,t})/\hat{\sigma}_{i,t}$*, and define the permutations* $p_i(t) = rk(\hat{\varepsilon}_{it})$*, where* $i = 1, \dots, n$ *and* $t = 1, \dots, T$*.*
3. *For all bottom series* $i = r + 1, \dots, n$*:*
   (a) *Compute h-period ahead conditional marginal predictive distributions* $\hat{F}_{i,T+h}$*.*
   (b) *Extract a discrete sample of size* $K = T$*, say* $x^i_1, \dots, x^i_K$*, where* $x^i_k = \hat{F}^{-1}_{i,T+h}(k/K+1)$*.*
4. *For all aggregate series* $i = 1, \dots, r$*:*
   (a) *Let* $i_1, \dots, i_{n_k}$ *be the* n_k *children series of the aggregate series* i*.*
   (b) *Recursively compute*
   $$x^i_k = x^{i_1}_{(p_{i_1}(k))} + \cdots + x^{i_{n_k}}_{(p_{i_{n_k}}(k))},$$
   where $x^i_{(k)}$ *denotes the kth order statistics of* $\{x^i_1, \dots, x^i_K\}$*, i.e.* $x^i_{(1)} \leq x^i_{(2)} \leq \cdots \leq x^i_{(K)}$*.*

Similarly to the classical bottom-up algorithm, Algorithm 5 produce coherent samples by construction. Furthermore, the samples of each aggregate are computed using only the predictive distributions of the bottom series. However, Algorithm 5 has two main advantages compared to a classical bottom-up algorithm: (1) instead of estimating a high-dimensional copula for the dependence between all the bottom series, we only need to specify the joint dependence between the child series of each aggregate series, and (2) since each copula is estimated at different aggregate levels, we can benefit from better estimation since the series are smoother, and easier to model and forecast.

3.2 Mean Forecast Combination and Reconciliation

Algorithm 5 allows the computation of coherent samples for all series in the hierarchy. Although the algorithm learns the permutations by estimating the copula dependence functions using data from different levels, the mean forecasts are computed using a classical bottom-up approach. In order to exploit possibly better forecasts from higher levels, we add a mean forecast combination step in our algorithm. Forecast combination is known to improve forecasts in many cases (Genre et al., 2013; Timmermann, 2006). We could adjust the means of our predictive distributions using the MinT revised forecasts. However, as Erven and Cugliari (2015), we propose to first combine the mean forecasts, and then apply a reconciliation step.

Let $\hat{\boldsymbol{y}}_{T+h}$ be the means of our predictive distributions. We compute the following forecast combination:

$$\check{\boldsymbol{y}}_t = \boldsymbol{Q}\hat{\boldsymbol{y}}_t, \tag{6}$$

where $\boldsymbol{Q} = \left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\right]' \in \mathbb{R}^{n \times n}$ is a weight matrix.

Since the combined mean forecasts $\check{\boldsymbol{y}}_t$ are not necessarily coherent, we also apply a reconciliation step using the GTOP approach described in Section 2.2. More precisely, we solve the quadratic optimization problem in (3), and obtain reconciled forecasts $\tilde{\boldsymbol{y}}_t$.

Since the total number of series in the hierarchy, n, can be very large compared to the number of observations T, it is necessary to use some regularization for the weights. Therefore, we will estimate the weights by solving the following L_1 optimization problem:

$$\underset{\boldsymbol{Q}}{\text{minimize}} \; \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{y}_t - \boldsymbol{Q}\hat{\boldsymbol{y}}_t\|^2 + \sum_{i=1}^{n} \lambda_i \|\boldsymbol{q}_i\|_1 ,$$

where $\lambda_i \geq 0$ is a regularization parameter for the ith weight vector $\boldsymbol{q}_i$. The previous problem can be rewritten as

$$\underset{\boldsymbol{q}_1,\ldots,\boldsymbol{q}_n}{\text{minimize}} \sum_{i=1}^{n} \frac{1}{T}\sum_{t=1}^{T} (y_{it} - \hat{\boldsymbol{y}}_t'\boldsymbol{q}_i)^2 + \sum_{i=1}^{n} \lambda_i \|\boldsymbol{q}_i\|_1 ,$$

which is decomposable in the vectors $\boldsymbol{q}_i$. As a result, we can solve the n problems independently. Our implementation of the LASSO is based on a cyclical coordinate descent algorithm (Friedman et al., 2007), and the regularization parameters are selected by minimizing time series cross-validated errors (Hyndman and Athanasopoulos, 2014, Section 2.5).

The forecast combination we are considering in (6) has multiple advantages compared to the MinT forecast combination in (2). First, since $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$, all series in the hierarchy can benefit directly from the forecast combination, not only the bottom series as in MinT with $\boldsymbol{P} \in \mathbb{R}^{m \times n}$. Second, we do not assume the base forecast are unbiased, and we do not seek to compute unbiased revised forecasts as in MinT. We rather seek to learn the weights to compute combined forecasts with low forecast errors; i.e. with the right trade-off between bias and estimation variance. Finally, even if we start with coherent base forecasts, we can still apply a forecast combination, and eventually reconcile them later. In contrast with MinT, no forecast combination will be applied in that case. Of course, MinT has the advantage of having a closed-form solution which does not require the solution of n possibly high-dimensional regression problems. Finally, our reconciled forecasts are guaranteed to have smaller or equal SSE than the combined forecasts which is guaranteed by the GTOP method as discussed in Section 2.2. Our final algorithm can be summarized as follows:

Algorithm 6 *(Mean Combined and Reconciled Probabilistic Forecasting)*

1. *Run Algorithm 5 to obtain bottom-up samples for all series in the hierarchy, say* $x_1^i, \dots, x_K^i$ *with* $i = 1, \dots, n$.
2. *Extract mean forecasts* $\hat{\boldsymbol{y}}_{T+h}$ *from all base predictive distributions* $\hat{F}_{i,T+h}$, *and compute combined forecasts* $\check{\boldsymbol{y}}_{T+h}$ *by applying the mean forecast combination described above.*
3. *Given a weight matrix* $\boldsymbol{A}$, *and using the combined forecasts* $\check{\boldsymbol{y}}_{T+h}$ *as base forecasts, solve the optimization problem in (3) to obtain reconciled forecasts* $\tilde{\boldsymbol{y}}_{T+h}$.
4. *Compute revised samples* $\tilde{x}_1^i, \dots, \tilde{x}_K^i$ *where* $\tilde{x}_k^i = x_k^i + \boldsymbol{\theta}_i$ *and* $\boldsymbol{\theta}_i = (\check{y}_{i,t} - \hat{y}_{i,t}) + (\tilde{y}_{i,t} - \check{y}_{i,t}) = \check{y}_{i,t} - \hat{y}_{i,t}$ *is an adjustment term, with* $i = 1, \dots, n$.

Algorithm 6 computes coherent forecasts since both the bottom-up samples (computed using Algorithm 5) and the reconciled means are coherent.

4 Experiments

We compare the following forecasting methods: (1) `BASE`: the base predictive distributions; (2) `NAIVEBU`: the naive bottom-up forecasts computed by summing *independent* samples from the bottom predictive distributions (without forecast combination); (3) `PERMBU`: the bottom-up forecasts computed using Algorithm 5 (without forecast combination); (4) `PERMBU-MINT`: similar to `PERMBU` with mean forecasts computed using MinT; (5) `PERMBU-GTOP1`: the forecasts are computed using Algorithm 6 with $\boldsymbol{A} = \boldsymbol{I}$; and (6) `PERMBU-GTOP2`: similar to `PERMBU-GTOP1` but with $\boldsymbol{A} = \text{diag}(\underbrace{0, \dots, 0}_{r}, \underbrace{1, \dots, 1}_{m})$; i.e. bottom-up instead of reconciled combined mean forecasts.

4.1 Probabilistic Forecast Evaluation

We evaluate our predictive distributions using the continuous ranked probability score (CRPS), which is a proper scoring rule, i.e. the score is maximized when the true distribution is reported (Gneiting and Raftery, 2007). Given an h-period-ahead cumulative predictive distribution function $\hat{F}_{t+h}$ and an observation y_{t+h}, the CRPS is defined equivalently as follows (Gneiting, Balabdaoui, and Raftery, 2007;

Gneiting and Ranjan, 2011):

$$\begin{aligned}\mathrm{CRPS}(\hat{F}_{t+h}, y_{t+h}) &= \int_{-\infty}^{\infty} \left(\hat{F}_{t+h}(z) - \mathbb{1}\{y_{t+h} \leq z\}\right)^2 dz \\ &= \int_0^1 \mathrm{QS}_\tau \left(\hat{F}^{-1}_{t+h}(\tau), y_{t+h}\right) d\tau,\end{aligned}$$

where QS_τ is the quantile score, defined as

$$\begin{aligned}&\mathrm{QS}_\tau \left(\hat{F}^{-1}_{t+h}(\tau), y_{t+h}\right) \\ &= 2\left(\mathbb{1}\{y_{t+h} \leq \hat{F}^{-1}_{t+h}(\tau)\} - \tau\right)\left(\hat{F}^{-1}_{t+h}(\tau) - y_{t+h}\right),\end{aligned}$$

which is also known as the pinball or check loss (Koenker and Bassett, 1978).

In order to quantify the gain/loss of the different methods with respect to the base forecasts, we compute the *Skill Score* defined as $(\mathrm{SCORE}_{\mathrm{BASE}} - \mathrm{SCORE})/\mathrm{SCORE}_{\mathrm{BASE}}$ where SCORE is the considered evaluation score. In the following experiments, SCORE will be computed by averaging the the CRPS or QS over all observations in the test set. Finally, as proposed by Laio and Tamea (2007), we will plot the (skill) QS_τ versus τ as a diagnostic tool in the comparison of the different methods.

4.2 Simulated Data

We begin with simulated time series, implemented using the same processes as Wickramasuriya, Athanasopoulos, and Hyndman (2015) to evaluate different hierarchical forecasting methods. However, we focus on distributional forecasts rather than mean forecasts. We used a hierarchy with four bottom series, where the two pairs of bottom series are aggregated in two aggregate series, which are then aggregated in a top series. Hence, the hierarchy is composed of $n = 7$ series, $m = 4$ bottom series and $r = 3$ aggregate series.

Each series in the bottom level is generated from an $\mathtt{ARIMA}(p, d, q)$ process, with p and q taking values of 0, 1 and 2 with equal probability and d taking values of 0 and 1 with equal probability. The parameters are chosen randomly from a uniform distribution from a specific parameter space for each each component of the ARIMA process (see Table 3.2 in Wickramasuriya, Athanasopoulos, and Hyndman (2015)). The error terms of the bottom-level ARIMA processes have a multivariate Gaussian distribution with a covariance structure that allows a strongly positive correlation among series with the same parents, but a moderately positive correlation among series with different parents.

For each series, we generate $T = 100, 300$ or 500 observations, with an additional $H = 10$ observations as a test set. We fit an ARIMA model by minimizing the AIC, and compute 10-period ahead Gaussian predictive distributions as base forecasts. The whole process is repeated $2,000$ times.

Figure 2 shows the results for $T = 100$. The first panel gives the skill CRPS for each horizon; the second and third panels show the skill QS averaged over horizons $h = 1$–6 and $h = 7$–10, respectively; the last panel gives the skill CRPS for the bottom level.

In the first panel, we can see that `PERM-BU` has a better skill than `NAIVE-BU` until horizon 6, and vice versa for the subsequent horizons. The second panel shows that `PERM-BU` outperforms `NAIVE-BU` especially in the lower and upper tails. In other word, the independence assumption of `NAIVE-BU` is not valid, and modelling the dependence structure between the children series of each aggregated series provides better tail forecasts for the aggregate series. The third panel shows that `NAIVE-BU` has consistently better skill QS compared to `PERM-BU` for horizons 7–10. This suggests that using one-period ahead dependence structure for 7 to 10-period ahead forecasts (i.e. using a misspecified dependence structure) is worse than assuming independence.

The first panel also shows that the methods using forecast combinations have significantly increased the skill CRPS compared to `PERM-BU`. This suggests that the mean forecast combination step is particularly useful in further improving the distributional forecasts. Furthermore, we can see that `PERM-GTOP2` has better skill than `PERM-MINT` until horizon 6. This shows the benefit of our forecast combination which learns the best combination weights, without making an unbiasedness assumption. The better skill of `PERM-GTOP2` compared to `PERM-GTOP1` suggests an advantage in splitting the forecast combination and reconciliation steps. The same observations can be made in the last panel for the bottom level.

Finally, with a larger training set size ($T = 300$ and $T = 500$), the forecast combination methods have similar skills, as can be seen in Figures A1 and A2 (see appendix). With more observations, the fitted ARIMA model becomes more accurate, and therefore, forecast combination is less likely to improve the base forecasts. However, even with a large training set, modeling the dependence structure is still important as shown by the better skill of `PERM-BU` compared to `NAIVE-BU`.

4.3 Electricity Smart Meter Data

We used smart meter electricity consumption data collected by four energy supply companies in Great Britain (AECOM, 2011). Consumption was recorded at half-hourly intervals for more than 14,000 households, along with geographic and demographic information. In our study, we were interested only in relatively long time series without missing values, and this led us to use data recorded at 1,578 meters for the period 20 April 2009 to 31 July 2010, inclusive. Each series, therefore, consisted of $T = 22,464$ half-hourly observations. We constructed a hierarchy based on geographical information comprising four levels of aggregation with $r = 55$ and $m = 1578$ series in the aggregate and bottom levels, respectively. Figure 3 presents observations for a one-week period for series taken from each of the four levels of the hierarchy.

We considered the problem of one-day-ahead (i.e. the next $H = 48$ half-hours) probabilistic demand forecasting, with a forecast origin at `23:30` for each day. We split each time series into training, validation

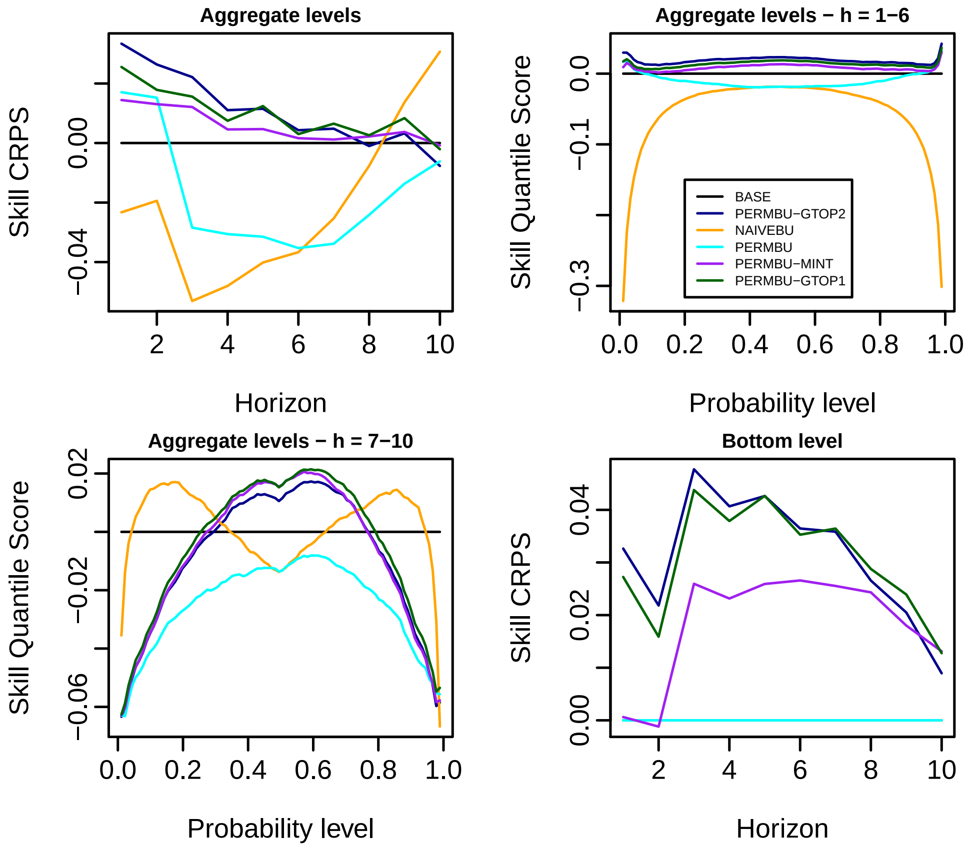

Figure 2: *Skill CRPS and skill QS for aggregate and bottom levels for $T = 100$.*

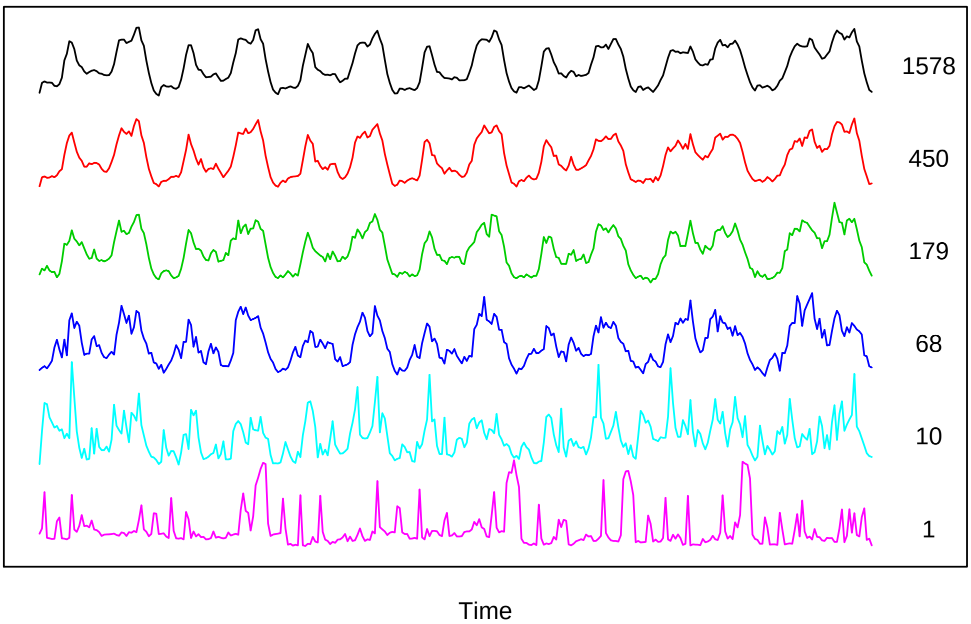

Figure 3: *One week of electricity demand with different number of aggregated series.*

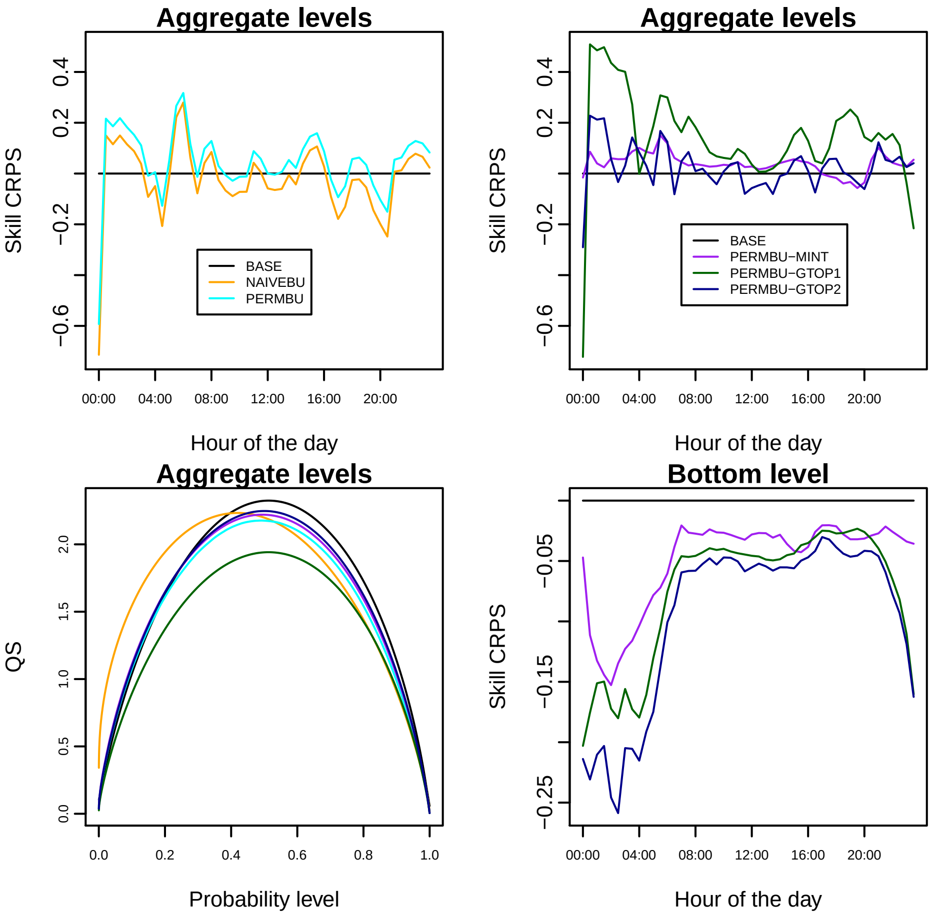

Figure 4: *Skill CRPS and QS for aggregate and bottom levels.*

and test sets; the first 12 months for training, the next month for validation and the remaining months for testing. Each model is re-estimated before forecasting each day in the test set using a rolling window of the historical observations.

We used different forecasting methods for the aggregate and bottom series. For the aggregate series, we capture the yearly cycle, the within-day and within-week seasonalities using seasonal Fourier terms with coefficients estimated by LASSO. After extracting the trend and seasonalities, we fitted an ARIMA model and computed Gaussian predictive distributions. This is justified by the fact that aggregate series are often smoother and easier to forecast, and by the central limit theorem. For the base forecasts, we implemented the same approach proposed by Arora and Taylor (2016), based on kernel density estimation.

In the first panel of Figure 4, we can see that `PERMBU` has better skill than `NAIVEBU` consistently over the horizon. The third panel shows that `PERMBU`, by modelling the dependence structure, has contributed to significantly increasing the QS skill in the lower tail. By analyzing the forecasts (not shown here), we noticed that `NAIVEBU` is penalized both for not being able to capture the trend at the top (i.e. a bad mean forecasts), and for having too sharp predictive distributions (i.e. bad dependence structure). The fact that

NAIVEBU seems competitive at moderately large quantiles can be explained by the unnecessarily wide prediction intervals which are penalized by the QS.

Overall, the second panel shows that the mean forecast combination methods have better skill than the base forecasts. We found that 75% of the series have less than 100 non-zero weights (see appendix); i.e. many forecast combinations were very sparse — an advantage of our approach compared to MinT which produces dense combination weights. Furthermore, we can see that PERMBU-GTOP1 is dominating the other methods consistently over the horizon. This suggests that computing bottom-up mean combined forecasts is better than reconciling the aggregate and bottom combined mean forecasts. This can be explained by the fact that PERMBU already produces competitive forecasts with the base forecasts, and so reconciling the bottom combined forecasts with the aggregate combined forecasts is unlikely to improve the final forecasts.

Finally, the last panel shows that all the mean forecast combination methods have lower skill than the base forecasts for the bottom series, especially in the first few horizons. One explanation could be that in order to reduce computational load, we used the same combination matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ for the entire test set, while the base forecasts use the most recent observations to generate the next-day-ahead forecasts. However, the forecast improvement at the aggregate levels are magnitudes larger than the decrease in accuracy at the bottom level.

5 Conclusion

We have proposed an algorithm to compute coherent probabilistic forecasts for hierarchical time series. The algorithm provides samples from coherent predictive distributions for each series in the hierarchy. To do so, we first generate independent samples from all series in the hierarchy. Then a sequence of permutations are applied to the samples in order to restore the dependencies between the children series of all aggregate series. Finally, a sparse forecast combination is applied using the base mean forecasts of all series in the hierarchy. Our algorithm has the advantage of synthesizing information from multiple levels in the hierarchy. Using simulated data, and a large scale electricity demand data set, we showed that restoring the dependencies of the children series consistently improves the forecast accuracy, especially in the tails, while the mean forecast combination provides an additional improvement by exploiting the more accurate base mean forecasts in the upper levels. Our algorithm can be used to produce coherent probabilistic forecasts for hierarchical time series in many applications.

References

AECOM (2011). *Energy Demand Research Project: Final Analysis*. Tech. rep. Hertfordshire, UK: AECOM House.

Arbenz, Philipp, Christoph Hummel, and Georg Mainik (2012). Copula based hierarchical risk aggregation through sample reordering. *Insurance, Mathematics & Economics* **51**(1), 122–133.

Arora, Siddharth and James W Taylor (2016). Forecasting electricity smart meter data using conditional kernel density estimation. *Omega* **59, Part A**, 47–59.

Ben Taieb, Souhaib, Jiafan Yu, Mateus Neves Barreto, and Ram Rajagopal (2017). Regularization in Hierarchical Time Series Forecasting With Application to Electricity Smart Meter Data. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press, 2017*.

Berrocal, Veronica J, Adrian E Raftery, Tilmann Gneiting, and Richard C Steed (2010). Probabilistic Weather Forecasting for Winter Road Maintenance. *Journal of the American Statistical Association* **105**(490), 522–537.

Erven, Tim van and Jairo Cugliari (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.

Fan, Yanqin and Andrew J Patton (2014). Copulas in Econometrics. *Annual Review of Economics* **6**(1), 179–200.

Friedman, Jerome, Trevor Hastie, Holger Höfling, and Robert Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**(2), 302–332.

Genre, Véronique, Geoff Kenny, Aidan Meyler, and Allan Timmermann (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting* **29**(1), 108–121.

Gneiting, Tilmann (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**(494), 746–762.

Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **69**(2), 243–268.

Gneiting, Tilmann and Matthias Katzfuss (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application* **1**(1), 125–151.

Gneiting, Tilmann and Adrian E Raftery (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**(477), 359–378.

Gneiting, Tilmann and Roopesh Ranjan (2011). Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics* **29**(3), 411–422.

Hothorn, Torsten, Thomas Kneib, and Peter Bühlmann (2014). Conditional transformation models. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **76**(1), 3–27.

Hyndman, Rob J, Roman A Ahmed, George Athanasopoulos, and Han Lin Shang (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* **55**(9), 2579–2589.

Hyndman, Rob J and George Athanasopoulos (2014). *Forecasting: principles and practice*. en. OTexts.

Hyndman, Rob J, Alan J Lee, and Earo Wang (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis* **97**, 16–32.

Iman, Ronald L and W J Conover (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation* **11**(3), 311–334.

Kneib, Thomas (2013). Beyond mean regression. *Statistical Modelling* **13**(4), 275–303.

Koenker, Roger and Gilbert Bassett (1978). Regression Quantiles. *Econometrica: journal of the Econometric Society* **46**(1), 33–50.

Kremer, Mirko, Enno Siemsen, and Douglas J Thomas (2016). The Sum and Its Parts: Judgmental Hierarchical Forecasting. *Management Science* **62**(9), 2745–2764.

Laio, F and S Tamea (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences* **11**(4), 1267–1277.

Nelsen, Roger B (2007). *An introduction to copulas*. Springer Science & Business Media.

Patton, A J (2012). Copula methods for forecasting multivariate time series. *Handbook of economic forecasting* (April), 1–76.

Patton, Andrew J (2006). Modelling asymmetric exchange rate dependence. *International Economic Review* **47**(2), 527–556.

Rüschendorf, Ludger (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference* **139**(11), 3921–3927.

Schefzik, Roman, Thordis L Thorarinsdottir, and Tilmann Gneiting (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.

Timmermann, A (2006). "Forecast combinations". In: *Handbook of Economic Forecasting*. Vol. 1. Elsevier, pp.135–196.

Wickramasuriya, Shanika L, George Athanasopoulos, and Rob J Hyndman (2015). *Forecasting hierarchical and grouped time series through trace minimization*. Tech. rep. 15/15. Monash University.

Coherent probabilistic forecasts for hierarchical time series

Appendix

24 February 2017

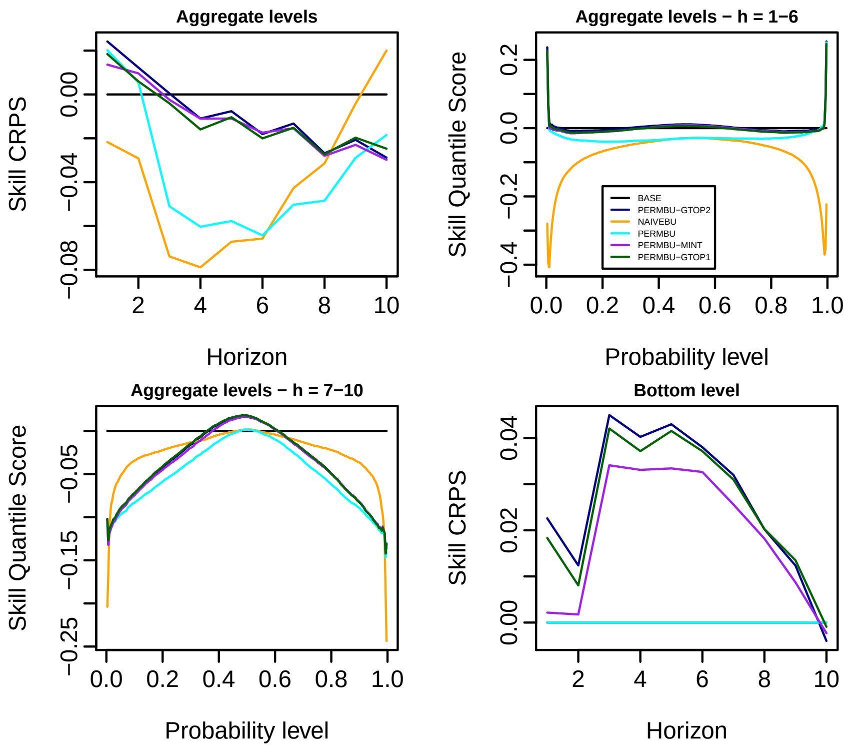

Figure A1: Skill CRPS and skill QS for aggregate and bottom levels for $T = 300$.

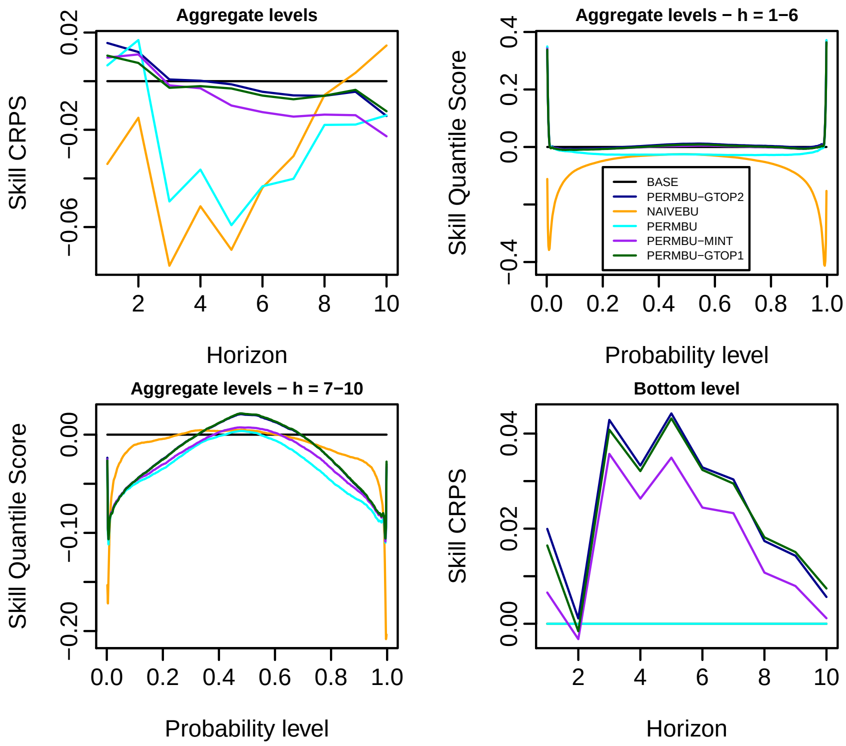

Figure A2: Skill CRPS and skill QS for aggregate and bottom levels for $T = 500$.

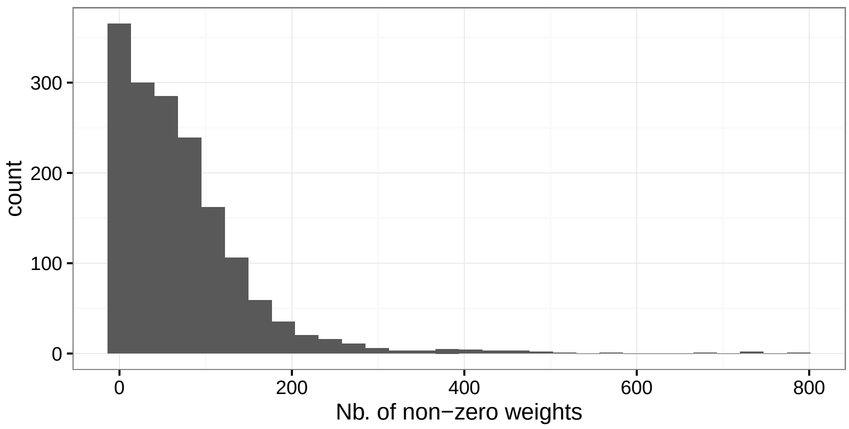

Figure A3: Histogram of the number of non-zero combination weights for 1633 series.