Reconciliation of probabilistic forecasts with an application to wind power

Jooyoung Jeon^{a,b,*}, Anastasios Panagiotelis^c, Fotios Petropoulos^a

^aSchool of Management, University of Bath ^bGraduate School of Engineering Practice, Seoul National University ^cDepartment of Econometrics & Business Statistics, Monash University

Abstract

New methods are proposed for adjusting probabilistic forecasts to ensure coherence with the aggregation constraints inherent in temporal hierarchies. The different approaches nested within this framework include methods that exploit information at all levels of the hierarchy as well as a novel method based on cross-validation. The methods are evaluated using real data from two wind farms in Crete, an application where it is imperative for optimal decisions related to grid operations and bidding strategies to be based on coherent probabilistic forecasts of wind power. Empirical evidence is also presented showing that probabilistic forecast reconciliation improves the accuracy of both point forecasts and probabilistic forecasts.

Keywords: Forecasting, Temporal hierarchies, Cross-validation, Aggregation, Renewable energy generation

1. Introduction

Data are often arranged in cross-sectional or temporal hierarchies characterised by an aggregation structure that holds for all realised values; for example, the annual sum of monthly data series will be equivalent to annual data series. When forecasts are independently produced for different series or levels within a hierarchy these aggregation constraints will not hold, a property known as *incoherence*. To ensure that operational decisions are aligned, a rich literature has emerged on forecast reconciliation (Athanasopoulos, Ahmed, and Hyndman, 2009; Hyndman et al., 2011;

^{*}Correspondence: Jooyoung Jeon

Email addresses: j.jeon@bath.ac.uk (Jooyoung Jeon), Anastasios.Panagiotelis@monash.edu (Anastasios Panagiotelis), f.petropoulos@bath.ac.uk (Fotios Petropoulos)

Athanasopoulos et al., 2017; Wickramasuriya, Athanasopoulos, and Hyndman, 2017). These methods not only ensure that forecasts are coherent but also lead to improvements in forecast accuracy. However, a shortcoming of these methods is their focus on point forecasting despite the increasing importance of probabilistic forecasts on decision-making (Gneiting and Katzfuss, 2014). This paper proposes a new methodology for the reconciliation of probabilistic forecasts.

Our proposed methodology can be described according to its three novel features. First, this study is the first to combine information about the full probabilistic forecast of each series in the reconciliation process. Second, this study is the first to focus on producing coherent probabilistic forecasts in the temporal rather than in the cross-sectional hierarchical setting, although we note that our methodology is general enough to handle both of these settings. Third, this study is the first to consider training reconciliation weights via a cross-validation procedure in either the point or probabilistic forecasting setting. Indeed to the best of our knowledge, the only other paper to tackle the issue of coherent probabilistic forecasts is that of Ben Taieb, Taylor, and Hyndman (2017) and our approach can be distinguished from theirs by each of the above-mentioned features. Crucially, with the exception of the mean and variance, the construction of a coherent probabilistic forecast by Ben Taieb, Taylor, and Hyndman (2017) relies on a bottom up approach. In contrast our entire reconciled probabilistic forecast is based on probabilistic forecasts of series from all hierarchical levels.

The methods we propose are evaluated using wind power data measured at various frequencies ranging from hourly to daily. This application is chosen for two main reasons. First, due to the highly volatile nature of wind power generation, informed decision-making depends not only on point forecasts but on probabilistic considerations. For instance, dispatch and risk management decisions may be based on the probability that a wind farm supplies at least 300kWh between midnight and 6am the following day. Second, wind farm operators, grid system operators and electricity traders are each required to make decisions based on different forecast horizons and sampling intervals. As such coherent probabilistic forecasts are crucial to ensure aligned decision-making. Our empirical results demonstrate that the proposed reconciliation methods improve the accuracy of point and probabilistic forecasts, with more substantial improvements at higher aggregation levels.

In the next section, we review the literature on hierarchical forecast reconciliation. Section 3 presents the methods to produce coherent and reconciled density forecasts. Section 4 introduces our

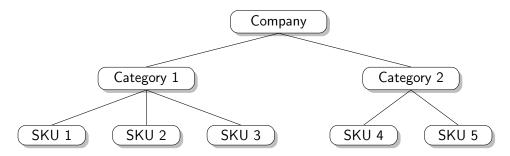


Figure 1: A cross-sectional hierarchy.

wind power data, and describes the density forecasting models for wind power generation. Section 5 describes the empirical results of the various reconciliation methods considered in Section 3. The final section provides a summary and conclusion.

2. Background

2.1. Cross-sectional Hierarchical Reconciliation of Point Forecasts

Data within companies are organised in hierarchical structures. For example, a company may organise its five stock keeping units (SKUs) into two categories, as depicted in Figure 1. If the historical data at the bottom level (SKU) are available, then data at every other level can be calculated using appropriate aggregations. Forecasts may be produced at any of the three levels of the hierarchy. However, if forecasts are independently produced at all levels they will not be coherent. For example, the sum of the forecasts of SKUs 1, 2 and 3 in Figure 1 is not guaranteed to be the same as the forecast of Category 1.

One way to tackle this issue is to simply produce forecasts on a single hierarchical level. For example, forecasts can be produced only on the very bottom level, and then aggregated to the higher levels in the hierarchical structure, an approach known as the bottom-up approach (see for example Dangerfield and Morris, 1992; Zellner and Tobias, 2000; Athanasopoulos, Ahmed, and Hyndman, 2009). In some cases, the bottom-level data may be too granular or noisy, rendering the forecasting task difficult. Alternatively, forecasts may be produced at the very top-level and then appropriately disaggregated to lower level forecasts, an approach known as top-down (Lütkepohl, 1984; Fliedner, 1999; Gross and Sohl, 1990). Disaggregation of the forecasts to lower levels may be based on historical or predicted proportions of the lower level data (Athanasopoulos, Ahmed, and Hyndman, 2009). The top-down approach has the disadvantage of information loss, as aggregated

series may not reflect the individual characteristics of their descendants. Finally, forecasts can also be produced at a middle level; forecasts for higher/lower levels nodes can be calculated by appropriate aggregation/disaggregation of the middle-level forecasts. This approach is known as *middle-out*, a conceptual combination of the bottom-up and top-down approaches.

A shortcoming of the methods above is that forecasts are only based on information at a single level of the hierarchy. The optimal combination method introduced by Athanasopoulos, Ahmed, and Hyndman (2009) and Hyndman et al. (2011) overcome this problem by tackling hierarchical forecasting in two stages. In the first stage, forecasts are produced for all series at all levels independently. In the second stage, these forecasts are adjusted in a reconciliation step to ensure coherence with aggregation constraints. More specifically the reconciled forecast for each node is formed as a weighted combination of the original - or so-called 'base' - forecasts of all nodes, in a way that ensures coherence for the hierarchy overall. The key advantage of reconciliation is that information is used at all levels of the hierarchy in contrast to the approaches described in the previous paragraph that focus on a single level. More recently, Hyndman, Lee, and Wang (2016) propose algorithms for fast computation of coherent hierarchical forecasts, and Wickramasuriya, Athanasopoulos, and Hyndman (2017) suggest calculating coherent forecasts through trace minimisation.

2.2. Temporal Hierarchical Reconciliation of Point Forecasts

A time series can be aggregated or disaggregated to create alternative frequency (or resolution) as needed. Time series at different frequencies will exhibit different characteristics. Seasonality and noise will be amplified in lower aggregation levels (higher frequencies), while the long-term trend can be more easily estimated using higher aggregation levels (lower frequencies) (Kourentzes, Petropoulos, and Trapero, 2014; Spithourakis et al., 2012). Similar to the case of cross-sectional aggregation, forecasts produced using data at different frequencies will not generally agree. For example, the sum of the forecasts for the next three months produced using data measured at the monthly frequency will not equal to the one-step-ahead quarterly forecast based on data measured at a quarterly frequency. This problem is particularly relevant for aligning decisions across the different departments within a company (operations, sales, finance, marketing, strategy), which usually operate at different data frequencies.

Similarly to cross-sectional aggregation, the issue of non-coherent forecasts at different temporal aggregation levels can be addressed either by combining (reconciling) the forecasts from multiple

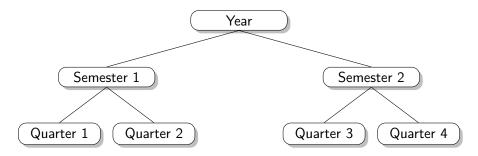


Figure 2: A temporal hierarchy.

aggregation levels or by producing forecasts for a single temporal aggregation level and then deriving the forecasts at the other levels as discussed previously.

Nikolopoulos et al. (2011) show empirically that in the context of intermittent demand there exists an optimal aggregation level, unique to each series, and proposed the Aggregate-Disaggregate Intermittent Demand Approach (ADIDA), where forecasts are produced at a (single) higher aggregation level and the lower level forecast is subsequently produced by disaggregation. This approach is particularly relevant for slow moving data, as temporal aggregation will result in series with a lower degree of intermittence (Petropoulos, Kourentzes, and Nikolopoulos, 2016). Rostami-Tabar et al. (2013) derive analytical results that improvement in forecasting performance is a function of the aggregation level, under specific data generation processes.

The idea of using aggregation/disaggregation for forecasting was further extended to derive the combined forecasts from forecasts simultaneously produced at multiple temporal aggregation (MTA) levels by Kourentzes, Petropoulos, and Trapero (2014) and Petropoulos and Kourentzes (2014). MTA was also applied to the context of intermittent demand (Petropoulos and Kourentzes, 2015), and Kourentzes and Petropoulos (2016) propose an extension of incorporating the effects of external variables. More recently, Athanasopoulos et al. (2017) express the MTA approach as a hierarchical concept using a temporal hierarchy for forecasting. A simple temporal hierarchy is depicted in Figure 2, where the bottom-level data are at a quarterly frequency (1 quarter per node), middle-level data are at a semesterly frequency (2 quarters per node), and the top-level represents the yearly frequency (4 quarters for the top-level node).

The representation of multiple temporal aggregation as temporal hierarchies allows for the application of the approaches designed for cross-sectional hierarchies, such as bottom-up, top-down, middle-out and optimal combination. Moreover, Athanasopoulos et al. (2017) provide three

approximations of the sample covariance estimator of the covariance matrix of the base forecast errors. These approximations are based on hierarchy variance scaling, series variance scaling and structural scaling, an order that reflects on their increasing simplicity in terms of implementation. Athanasopoulos et al. (2017) show empirically that simpler scaling approximations provide better results, especially as the frequency of the bottom level increases. Note that in contrast to cross-sectional hierarchies where forecasts are produced separately for each node, forecasts within a temporal hierarchy are typically produced by fitting one model per aggregation level to model dependencies over time and the unique behaviour of each frequency. For example, a single model fitted to the quarterly time series produces multi-step ahead forecasts for quarters 1 to 4 at the lowest level in Figure 2.

2.3. Hierarchical Reconciliation of Probabilistic Forecasts

Decision-making based on probabilistic forecasts has received increasing attention recently (Gneiting and Katzfuss, 2014, and references therein). In a similar way to point forecasts, probabilistic forecasts could be produced independently for each level in the hierarchy, but independent series cannot be said to be coherent since the aggregation constraint induces dependence between the variables. The first approach to tackling hierarchical forecasting in the probabilistic setting is the paper of Ben Taieb, Taylor, and Hyndman (2017). After carrying out reconciliation on the mean, they construct a coherent probabilistic forecast in a bottom up fashion where the dependency between nodes at each level is modelled by reordering quantile forecasts as suggested by Arbenz, Hummel, and Mainik (2012). The method we propose is distinct from Ben Taieb, Taylor, and Hyndman (2017) in two ways. First, our proposed method is a true reconciliation method, where each probabilistic forecast is based on information from all nodes in the hierarchy. Second, our problem focuses on temporal aggregation of density forecasts which provides a distinct case since dependence within each level can be obtained directly rather than through copula modelling. Recently, Athanasopoulos et al. (2017) propose methods to reconcile temporal point forecasts in the hierarchy, but none has yet focused on temporal hierarchical reconciliation of density forecasts. To the best of our knowledge, this study is the first to consider reconciliation of probabilistic forecasts for temporal hierarchies.

3. Methodology

Let us introduce the following notation. We let x_{j,f_l}^t be the realisation of a variable recorded on cycle t during the j^{th} period of the cycle, where f_l is the sampling interval for level l of the hierarchy. Cycle may refer, for instance, to a full year. For example, in the case of the hierarchy in Figure 3, we let f = [4, 2, 1]. Subsequently, $x_{1,f_1}^1 = x_{1,4}^1$ is the demand for the first year (first four quarters), $x_{2,2}^3$ is the demand for the second semester of the third year and $x_{3,1}^5$ is the demand for the third quarter of the fifth year. The same notation can be used for any other temporal hierarchy. Assuming, for example, a daily cycle and hourly data granularity, f = [24, 12, 8, 6, 4, 3, 2, 1] with $x_{5,4}^{10}$ referring to the 4 hourly demand of the fifth observation (16:00-20:00) of the tenth day. In the rest of Section 3, we will illustrate the methods of our interest using the temporal hierarchy depicted in Figure 3.

Let the scaled vectors $\mathbf{z}_l^t := (\mathbf{f}_L/\mathbf{f}_l)(x_{1,\mathbf{f}_l}^t,\dots,x_{(\mathbf{f}_1/\mathbf{f}_l),\mathbf{f}_l}^t)'$ for all l, where L is the number of levels of the hierarchy (L=3 for the example hierarchy in Figure 3). Then, \mathbf{z}_l^t is the vector of the realisations of all the nodes at the level l, scaled to be in the same units as the lowest level L, i.e. the highest resolution. This allows us to avoid the complex scale conversion in the density reconciliation between any levels and to interpret reconciliation as forecast combination between levels. Afterwards, the probabilistic forecasts can be rescaled back to the original units for each level. Finally, let $\mathbf{y}^t := (\mathbf{z}_1^{\prime t},\dots,\mathbf{z}_L^{\prime t})'$. The notation y_i^t will be used to denote the i^{th} scalar element of \mathbf{y}^t for $i=1,\dots,M$, where M is the number of nodes in the hierarchy (e.g. M=7 in Figure 3).

3.1. Coherent and Reconciled Probabilistic Forecasts

Some care must be taken in extending concepts such as coherent forecasts and reconciled forecasts to the probabilistic setting. Formal definitions of coherent probabilistic forecasts are provided in Ben Taieb, Taylor, and Hyndman (2017). In brief, a coherent probabilistic forecast is an Mdimensional multivariate distribution which, due to the degeneracy induced by the aggregation constraints, is only supported on an m-dimensional linear subspace of \mathbb{R}^M where $m = f_1/f_L < M$ (e.g. m = 4 in Figure 3, simply referring to the number of nodes on the bottom level).

As discussed in Section 2, reconciliation in the point forecasting context refers to a process by which a vector of incoherent forecasts is made coherent. We now provide some detail. Letting \hat{y} be a vector of unreconciled or 'base' forecasts, then a reconciled point forecast is given by $\tilde{y} = SP\hat{y}$. The matrix P is a $m \times M$ matrix that forms point forecasts for the bottom level of the hierarchy as

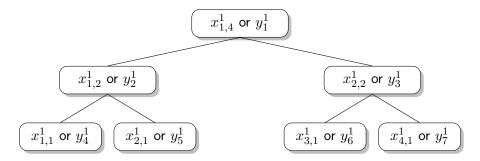


Figure 3: An illustration of notation for a temporal hierarchy.

linear combinations of the base point forecasts of all nodes. The matrix S is a $M \times m$ matrix that encodes the aggregation constraints and recovers a full set of coherent forecasts from the bottom level forecasts. For the simple hierarchy in Figure 3, S is given by

$$S = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$
 (1)

Taken together, the matrix SP is a projection matrix which takes any vector in \mathbb{R}^M and projects it to an m-dimensional subspace spanned by the vectors of S, a linear subspace where all aggregation constraints hold.

A common way to build probabilistic forecasts - that we follow here - is to generate a sample of size N from the distribution $f\left(\boldsymbol{y}^{t+h}|\mathcal{F}_{l}^{t};\hat{\boldsymbol{\theta}}\right)$, where \mathcal{F}_{l}^{t} represents all the information up to time t in the level l and $\hat{\boldsymbol{\theta}}$ indicates that the probabilistic forecast is based on parameter estimates. Denoting the i^{th} vector from this sample as $\hat{\boldsymbol{y}}_{i}^{t+h|t}$, we can store these in a matrix as $\hat{\boldsymbol{Y}} = \left(\hat{\boldsymbol{y}}_{1}^{t+h|t}, \ldots, \hat{\boldsymbol{y}}_{N}^{t+h|t}\right)$. Typically there is no guarantee that the aggregation constraints will hold for each (or in fact any) of the columns of $\hat{\boldsymbol{Y}}$. However, if $\hat{\boldsymbol{Y}}$ is pre-multiplied by a projection matrix to give $\hat{\boldsymbol{Y}} = \boldsymbol{SP}\hat{\boldsymbol{Y}}$, the columns of the resulting matrix will respect the aggregation constraints and can therefore be thought of as observations sampled from the reconciled probabilistic forecast. In this way existing

reconciliation methods for the mean can be extended to a probabilistic setting. To summarise, the process for forming probabilistic forecasts consists of two stages, in the first a sample is obtained from an estimate of the joint density $f\left(\boldsymbol{y}^{t+h}|\mathcal{F}_{l}^{t};\hat{\boldsymbol{\theta}}\right)$, and in the second each sampled vector is premultiplied by a projection matrix. At the first stage there are alternative approaches to constructing a joint sample, while at the second stage there are alternative projection matrices that can be used. We now discuss each of these stages in detail.

3.2. Construction of Unreconciled Forecasts

The first stage of our procedure, namely to obtain a matrix \hat{Y} is itself broken down into two steps. In the first step, each level will be modelled independently with details of these models provided in section 4.2. Let \hat{Z}_l be a $(f_1/f_l) \times N$ matrix defined similarly to \hat{Y} . Then, its columns are observations sampled from the joint predictive distribution but only using nodes in the level l, i.e. $f\left(z_l^{t+h}|\mathcal{F}_l^t;\hat{\theta}\right)$. A sample from this joint density can be produced by forming multi-step ahead forecasts in the usual recursive fashion and, as a consequence, the dependence within level is preserved. In the second step, we consider three alternatives for forming a sample \hat{Y} using all \hat{Z}_l . Each of these alternatives can be thought of as capturing the dependence between the elements of \hat{Y} in a different way - the appeal of these methods is that they avoid the challenge of modelling for the dependence explicitly.

3.2.1. Stacked Sample

The most straightforward way to form \hat{Y} is to simply concatenate the matrices $\hat{Z}_l^{t+h|t}$ which we refer to as the 'stacked' sample.

$$\hat{\mathbf{Y}}^{S} = \begin{bmatrix} \hat{\mathbf{Z}}_{1} \\ \hat{\mathbf{Z}}_{2} \\ \vdots \\ \hat{\mathbf{Z}}_{L} \end{bmatrix}$$

$$(2)$$

Using this approach leads to a joint distribution that preserves the dependence within each level but effectively assumes independence between levels.

3.2.2. Ranked Sample

An alternative to the stacked sample involves ordering the elements in each row of $\hat{\mathbf{Y}}^S$ in ascending (or descending) order after concatenation. We refer to this as the 'ranked sample' denoted

 $\hat{\mathbf{Y}}^R$. The rows of $\hat{\mathbf{Y}}^R$ will have a comonotonic dependence structure with respect to one another, and this approach can therefore be expected to work well in applications where dependence is high. Furthermore, the i^{th} column of $\hat{\mathbf{Y}}^R$ can be thought of as a vector of the $(i/N)^{th}$ quantiles, each element corresponding to a different node. As such, this approach also has an interpretation as a method that reconciles quantiles. This approach also has similarities to the combination of probabilistic forecasts by Lichtendahl, Grushka-Cockayne, and Winkler (2013). Whereas they focus on combining probabilistic forecasts that come from different models, the same idea can easily be applied to appropriately rescaled temporal hierarchies since the probabilistic forecast at each node can be understood as coming from a different model for modelling the wind power over a one hour period. Lichtendahl, Grushka-Cockayne, and Winkler (2013) also propose an approach that averages cumulative probabilities, but find this approach to be inferior to a quantile averaging approach. Our own application of probability averaging to the reconciliation of temporal hierarchies leads to the same conclusion and these results are omitted.

3.2.3. Permuted Sample

A final alternative would be to randomly shuffle the elements within each row of $\hat{\mathbf{Y}}^S$. We refer to this as the 'permuted sample' $\hat{\mathbf{Y}}^P$. The shuffling has the effect of decoupling the dependence within each level, making the rows of $\hat{\mathbf{Y}}^P$ independent with respect to one another. Although this may seem to be an unreasonable approach, it provides an interesting contrast with the ranked sample and may be a useful method that guards against over-fitting when dependence is low.

3.3. Reconciliation Methods

Once the matrix \hat{Y} has been formed either as the stacked, ranked or permuted sample, it is premultiplied by a projection matrix SP to yield a reconciled sample. We consider several alternatives for P in this section.

3.3.1. Bottom Up (BU)

A simple choice for P is to simply ignore information above the bottom level of the hierarchy and simply aggregate the unreconciled bottom level forecasts. For the example, in Figure 3 this

implies:

$$\mathbf{P}_{BU} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

or more generally $\mathbf{P}_{BU} = \begin{bmatrix} \mathbf{0}_{m \times (M-m)} & \mathbf{I}_m \end{bmatrix}$, where $\mathbf{0}_{a \times b}$ denotes a $a \times b$ matrix of zeroes and \mathbf{I}_a is an identity matrix of order a,

3.3.2. Bottom Average (BA)

Another straightforward method that only uses bottom level information is to average the bottom level. In this case, the P matrix is given by

$$\mathbf{P}_{BA} = \begin{bmatrix} 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}, \tag{4}$$

in Figure 3 and by $P_{BA} = \begin{bmatrix} \mathbf{0}_{m \times (M-m)} & (1/m)\mathbf{1}_{m \times m} \end{bmatrix}$ in general, where $\mathbf{1}_{a \times b}$ denotes a $a \times b$ matrix of ones.

3.3.3. Global Average (GA)

Another method is to use information at all nodes of the hierarchy via a simple average, or

$$\tilde{\mathbf{Y}}_{i,.} = \frac{1}{M} \sum_{j=1}^{M} \hat{\mathbf{Y}}_{j,.} \quad \forall i.$$

This is equivalent to assuming that the matrix P is a matrix of ones scaled by (1/M), that is $P_{GA} = (1/M)\mathbf{1}_{m\times M}$. We note that each of the bottom average and global average lead to probabilistic forecasts that are the same for every node, before being transformed back to the original scale.

3.3.4. Lineal Average (LA)

An alternative to reconciliation based on an average of all nodes is to build an average based on a set of nodes constructed in the following way. Supposing we are interested in Node i at the bottom level, take the parent nodes of Node i recursively as well as Node i, and calculate the average over

the nodes, referring to this as the 'lineal average'. The P matrix for the lineal average, where Row i allows to take the average of Node i and its ancestor nodes, is defined for the hierarchy in Figure 3 as

$$\mathbf{P}_{LA} = \begin{bmatrix} 1/L & 1/L & 0 & 1/L & 0 & 0 & 0 \\ 1/L & 1/L & 0 & 0 & 1/L & 0 & 0 \\ 1/L & 0 & 1/L & 0 & 0 & 1/L & 0 \\ 1/L & 0 & 1/L & 0 & 0 & 0 & 1/L \end{bmatrix}.$$
 (5)

This method does not use information from forecasts of sibling nodes to reconcile probabilistic forecasts. A motivation for this is that in the temporal forecasting context, the dependence within each level can be easily preserved.

3.3.5. Weighted Least Squares (WLS)

In the context of point forecasts, Athanasopoulos et al. (2017) derive unbiased optimal point forecasts as $\tilde{Y} = S(S'\Sigma^{-1}S)^{-1}S'\Sigma^{-1}\hat{Y}$, where Σ is the variance covariance matrix of the so-called reconciliation errors. Since Σ is unidentified (Wickramasuriya, Athanasopoulos, and Hyndman, 2017) it is replaced with a one of three diagonal matrices W. Our choice of W is similar to the structural scaling approach discussed in Athanasopoulos et al. (2017). The only difference between the structural scaling approach and our own is that for the former, the element on the diagonal of W corresponding to a node in level l is set to f_l while we prefer f_l^2 reflecting the fact that W is a proxy for a variance covariance matrix and that standard deviations rather than variances scale proportionally when the underlying random variable is rescaled. Furthermore, our choice of W leads to results that are equivalent to OLS on the rescaled data while the structural scaling of Athanasopoulos et al. (2017) does not have this property.

3.3.6. Cross-Validated (CV)

A shortcoming of all the approaches above is that the weights are fixed. In this section we propose a class of data-driven weights that are determined via cross-validation to maximise the sharpness of the reconciled predictive distributions, subject to calibration. The notions of sharpness and calibration are discussed by Gneiting and Katzfuss (2014). To the best of our knowledge, such a use of cross-validation weights has not been considered in hierarchical reconciliation, either in point forecasting, nor probabilistic forecasting.

The cross-validation procedure involves splitting the sample into three non-overlapping samples, the training sample \mathcal{T}_{train} , the validation sample \mathcal{T}_{val} and the test sample \mathcal{T}_{test} . Before cross-validation, model parameters are estimated using only training data. We denote these estimates as $\hat{\boldsymbol{\theta}}_{train}$. Then for all t+h in the validation sample, a sample is produced from $\hat{F}(\boldsymbol{y}^{t+h}|\mathcal{F}_l^t; \hat{\boldsymbol{\theta}}_{train})$, where \hat{F} is used to denote the unreconciled predictive cumulative distribution function (CDF). After pre-multiplication by some matrix \boldsymbol{SP} , a sample from the reconciled CDF $\tilde{F}(\boldsymbol{y}^{t+h}|\mathcal{F}_l^t; \hat{\boldsymbol{\theta}}_{train})$ is obtained. Let \tilde{F}_{j,f_l}^{t+h} be the CDF of the margin corresponding to the j^{th} node in the level l of the hierarchy. Finally let R(F,z) be a strictly proper scoring rule where F is a predictive CDF, and z is a scaled realisation.

The objective function for our cross validation is given by

$$CV(\mathbf{P}) = L^{-1} \sum_{l=1}^{L} CV_l(\mathbf{P}), \qquad (6)$$

where

$$CV_l(\mathbf{P}) = (\mathbf{f}_1/\mathbf{f}_l)^{-1} \sum_{j=1}^{(\mathbf{f}_1/\mathbf{f}_l)} \sum_{t+h \in \mathcal{T}_{val}} R(\tilde{F}_{j,\mathbf{f}_l}^{t+h}, z_{j,\mathbf{f}_l}^{t+h}).$$
(7)

In this paper, the scoring function used is the continuous ranked probability score (CRPS) given in general by

$$R(F,z) = \int_{\mathcal{U}} (F(u) - \mathbb{1} \{z \le u\})^2 du, \qquad (8)$$

where $\mathbbm{1}\{.\}$ is an indicator function equal to 1 if the statement in braces is true and 0 otherwise. We note that the same scoring rule is used in our empirical evaluation with the notable difference that after determination of cross validation weights, a new \tilde{F} based on both training and validation samples can be obtained.

The quantity $CV(\mathbf{P})$ is optimised with respect to \mathbf{P} . Since the \mathbf{P} matrix can be quite large we propose the following sparse structure

$$\mathbf{P}_{CV} = \begin{bmatrix} v_{1,1} & v_{2,1} & 0 & v_{3,1} & 0 & 0 & 0 \\ v_{1,1} & v_{2,1} & 0 & 0 & v_{3,2} & 0 & 0 \\ v_{1,1} & 0 & v_{2,2} & 0 & 0 & v_{3,3} & 0 \\ v_{1,1} & 0 & v_{2,2} & 0 & 0 & 0 & v_{3,4} \end{bmatrix},$$
(9)

where $v_{l,r}$ corresponds to the weight on the r^{th} node in the level l. The bottom-up method in Section 3.3.1 is a special case of this method, where only $v_{L,.}$ are 1, and the other weights are zero. The lineal average method in Section 3.3.4 is also a special case of this method, where all v are 1/L.

If the temporal hierarchy of interest is not too large and the study involves a sufficient cross-validation period, all weights of P_{CV} could be determined with cross-validation. Where cross-validation is not feasible, further constraints can be placed on the CV weights. One such restriction is to force the same value of the weight within each level, which gives the following P matrix for the hierarchy in Figure 3:

$$\mathbf{P}_{CVR} = \begin{bmatrix} v_1 & v_2 & 0 & v_3 & 0 & 0 & 0 \\ v_1 & v_2 & 0 & 0 & v_3 & 0 & 0 \\ v_1 & 0 & v_2 & 0 & 0 & v_3 & 0 \\ v_1 & 0 & v_2 & 0 & 0 & 0 & v_3 \end{bmatrix}, \tag{10}$$

where v_l corresponds to the weight on all the nodes in the level l. This sparse form reduces the number of weights to optimize over to L, with an additional constraint that each row sum is equal to the sum of all the weights. Thus, we use this simpler and practical form of matrix for the case study in Sections 4 and 5. To allow for the possibility of poorly calibrated basic forecasts, we tried different restrictions on the weights in cross-validation in Expression 10. In particular we consider the following cases: (1) all weights in a row sum to one and are positive; (2) all weights in a row sum to one; and (3) all weights are unconstrained.

4. Empirical design

4.1. Temporal Probabilistic Hierarchy of Wind Power

As a case study of the methods we propose in Section 3, we use hourly time series of wind power from the Rokas and Aeolos wind farms in Crete, the largest island in the Aegean Sea. The island has an autonomous electricity grid and high wind energy potential. The generation capacities of the Rokas and Aeolos wind farms were 16.3MW and 11.6MW, respectively, in 2006. The wind speed and direction observations were recorded at the turbine hub height of the two wind farms and plotted with the corresponding wind power observations in Figures 4 and 5, respectively, for

each hour in 2006, which amounted to 8,760 observations in each series. Each time series was split to \mathcal{T}_{train} , the training period of the first 6 months, 1 January 2006 to 30 June 2006, used for training our wind speed density forecasting models; \mathcal{T}_{val} , the validation period of the next 3 months, 1 July 2006 to 30 September 2006, used for choosing the most accurate wind speed density forecast model for the time series in each temporal hierarchical level and for selecting the cross-validation weights in Section 3.3.6; and \mathcal{T}_{test} , the test period of the last 3 months, 1 October 2006 to 31 December 2006, reserved for evaluation of the models we proposed. As in Figures 4 and 5, wind power is more volatile than wind speed, and the volatilities tend to be clustered.

It is a major challenge for grid operators to maximise the utilisation of wind power due to the intermittency nature of the supply. Due to the inherent uncertainty in the wind power forecasting, probabilistic approaches have received increasing attention recently (Taylor, 2017; Roulston and Smith, 2003; Gneiting et al., 2006; Jeon and Taylor, 2012; Hering and Genton, 2010; Taylor and Jeon, 2015; Dowell and Pinson, 2015), as these enable more informed decision-making by allowing for the optimal design of bidding strategies and power balance by wind farm operators, grid system operators and electricity traders (Pinson, 2013). One of the most extensive approaches to probabilistic forecasting is to estimate density forecasts, and we estimate these multi-step ahead. Spot power exchange markets are typically a day-ahead auction, and the market price is calculated for each hour of the following day. Pinson (2013) also explains that although forecasts up to 2 hours ahead are crucial for dispatch and control problems, much longer lead times are also relevant to decision-making for transmission operations, load-balancing and scheduling for spinning reserve and planning for optimal trading strategies. Therefore, in this paper we focus on enhancing temporal hierarchical probabilistic forecasts up to 24 hours ahead. The overlapping hierarchy consists of 1 \times 24 hour forecast, 2 \times 12 hourly forecasts, 3 \times 8 hourly forecasts, 4 \times 6 hourly forecasts, 6 \times 4 hourly forecasts, 8×3 hourly forecasts, 12×2 hourly forecasts and 24×1 hourly forecasts. This amounts to L=8 levels, M=60 nodes and m=24 bottom-level nodes in the hierarchy. Thus, S

has the following structure:

$$S = \begin{bmatrix} 24^{-1} \iota'_{24} \\ 12^{-1} I_2 \otimes \iota'_{12} \\ 8^{-1} I_3 \otimes \iota'_8 \\ 6^{-1} I_4 \otimes \iota'_6 \\ 4^{-1} I_6 \otimes \iota'_4 \\ 3^{-1} I_8 \otimes \iota'_3 \\ 2^{-1} I_{12} \otimes \iota'_2 \\ I_{24} \end{bmatrix},$$
(11)

where ι_a is an column of a ones and \otimes denotes the Kronecker product. Figure 6 illustrates the 24 hourly, 8 hourly and 2 hourly time series of Rokas, aggregated from the 1 hourly time series. A lower frequency time series exhibits more smoothed movements.

4.2. Probabilistic Forecasting Models for Wind Power

To construct a temporal hierarchy of wind power forecasts, we produced density forecasts for each level of the hierarchy. Although a separate density forecast model could be considered for each of the 60 different nodes in the hierarchy, we feel it is unlikely to do so in practice as it demands high computation cost, and significant autocorrelations between nodes in the same level attract modelling them together. In this paper, for each level, a separate multi-step ahead density forecast model is selected after comparing various wind power density forecast models. This procedure is consistent with the study by Athanasopoulos et al. (2017) that focuses on the point forecasts of temporal hierarchies.

Statistical models are considerably cheaper than a numerical weather prediction (NWP) system (see, for example, Sloughter, Gneiting, and Raftery, 2010), and are considered to be very competitive for short lead times (Pinson, 2013). The statistical models we can consider include direct modelling of wind power using historical simulations, but, as discussed by Jeon and Taylor (2012), it is somewhat challenging due to the non-linear evolution of the time series, and did not perform better than indirect modelling of wind power, which models wind speed first and then converts to wind power. The indirect models we used included univariate autoregressive moving average – generalized autoregressive conditional heteroskedasticity (ARMA-GARCH) models for wind speed density forecasting, and VEC-type bivariate vector autoregressive moving average – generalised

autoregressive conditional heteroskedasticity (VARMA-GARCH) models proposed by Bollerslev, Engle, and Wooldridge (1988) for wind speed and wind direction density forecasting (see Jeon and Taylor (2012) for further details). We further considered long memory dependence in the mean of the wind speed time series captured by the autoregressive fractionally integrated moving average (ARFIMA) model of Granger and Joyeux (1980) and Hosking (1981) and in the volatility modelled by the fractionally integrated generalized autoregressive conditionally heteroskedastic (FIGARCH) model introduced by Baillie, Bollerslev, and Mikkelsen (1996). These ARMA-GARCH type models were fitted to \mathcal{T}_{train} with Gaussian, Student t and skew t distribution assumptions for the noise term.

For each hierarchical level of each data set, the wind speed (and wind direction) density forecast model producing the smallest average of the CRPS values evaluated for 1 to 24 hourly ahead wind speed forecasts in \mathcal{T}_{val} was chosen and presented in Table 1. As evidenced in the paper of Taylor, McSharry, and Buizza (2009), the fractional integration in level and volatility was found to be useful for daily wind speed forecasts. For the time series of higher frequency, the bivariate VARMA-GARCH with Student t was chosen the most frequently for both wind farms. Overall, the frequent selection of Student t or skew t rather than Gaussian distribution indicates the conditional distribution of wind speed follows non-Gaussianity.

The wind speed and direction density forecasts are then converted to wind power density forecasts using the conditional kernel density estimation, as described by Jeon and Taylor (2012) to model the conversion uncertainty in the power curve, which relates wind speed and wind direction to wind power. The noise in the power curve could be brought about by changes in air pressure, temperature, precipitation and wind direction, the complexity of the terrain, different behaviour between speed up and down, turbulence in the turbines, the maintenance of them, and errors in measurement amongst other things.

Based on the models that are individually chosen the best for each level of the hierarchical density forecasting for wind power, 1,000 Monte-Carlo simulated sample paths were generated to construct 1 to 24 hour ahead density forecasts for each node of the hierarchy from each forecast origin in \mathcal{T}_{test} . We did not re-estimate density forecasting method parameters for wind power as we rolled the forecast origin forward, because it would be unlikely to be done in practice due to high computational cost, and because the focus of the paper is more about hierarchical reconciliation.

The density forecasts used for the cross-validated method described in Section 3.3.6 are produced using a similar approach but in \mathcal{T}_{val} . The weights for the cross-validated method with the four different constraints using the P_{CVR} matrix discussed in Section 3.3.6 are presented in Tables 2 and 3. It is interesting to see that the highest weight for each method is mostly on the 1 or 2 hourly hierarchical level for Rokas and 1 hourly hierarchical level for Aeolos. This is sensible as wind power data is fast moving with a high degree of intermittence and the lower level (higher frequency) forecasts contain more useful information.

Table 1: Models chosen for each wind farm and each hierarchical level. Univariate models produce wind speed density forecasts only. Bivariate models produce density forecasts of wind speed and wind direction.

Method	Rokas	Aeolos
24 hourly	Univariate ARFIMA-FIGARCH with Gaussian	Univariate ARMA-FIGARCH with Student t
12 hourly	Bivariate VARMA-GARCH with Student t	Bivariate VARMA-GARCH with Student t
8 hourly	Bivariate VARMA-GARCH with Student t	Univariate ARMA-FIGARCH with Student t
6 hourly	Bivariate VARMA-GARCH with Student t	Bivariate VARMA-GARCH with Student t
4 hourly	Bivariate VARMA-GARCH with Student t	Bivariate VARMA-GARCH with Student t
3 hourly	Bivariate VARMA-GARCH with Student t	Bivariate VARMA-GARCH with Student t
2 hourly	Univariate ARMA-GARCH-skew t	Bivariate VARMA-GARCH with Student t
1 hourly	Univariate ARMA-FIGARCH with skew t	Univariate ARMA-GARCH with skew t

Table 2: Weights(v) of the CV method in Section 3.3.6 derived for Rokas, determined by minimising the average of the level-wise average CRPS values in the hierarchy. The sum of v is the row sum.

Method	Hierarchical level (in mean)									
	24h	12h	8h	6h	4h	3h	2h	1h	Sum	
Permuted Sample										
$\sum v_i = 1 \& \forall v_i \ge 0$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	
$\sum v_i = 1$	-0.37	0.05	0.38	-0.15	0.19	0.10	0.87	-0.07	1.00	
Unconstrained	-0.28	-0.03	0.44	-0.19	0.20	0.09	0.87	-0.05	1.05	
Stacked Sample										
$\sum v_i = 1 \& \forall v_i \ge 0$	0.00	0.00	0.01	0.00	0.02	0.00	0.98	0.00	1.00	
$\sum v_i = 1$	-0.34	0.29	0.23	-0.07	0.49	-0.38	0.64	0.14	1.00	
Unconstrained	-0.25	-0.08	0.62	-0.11	0.53	-0.61	0.94	-0.04	1.00	
Stacked Sample										
$\sum v_i = 1 \& \forall v_i \ge 0$	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.78	1.00	
$\sum v_i = 1$	-0.01	-0.04	0.04	-0.02	0.03	0.07	0.38	0.56	1.00	
Unconstrained	-0.01	0.00	0.05	-0.03	0.08	0.24	0.01	0.35	0.69	

Table 3: Weights(v) of the CV method in Section 3.3.6 derived for Aeolos, determined by minimising the average of the level-wise average CRPS values in the hierarchy. The sum of v is the row sum.

Method	Hierarchical level (in mean)									
	24h	12h	8h	6h	4h	3h	2h	1h	Sum	
Permuted Sample										
$\sum v_i = 1 \& \forall v_i \ge 0$	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.25	1.00	
$\sum v_i = 1$	-0.01	-0.00	0.53	-0.34	-0.00	0.34	-0.19	0.68	1.00	
Unconstrained	0.06	-0.23	0.46	-0.43	-0.09	0.18	0.00	0.86	0.82	
Stacked Sample										
$\sum v_i = 1 \& \forall v_i \ge 0$	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.84	1.00	
$\sum v_i = 1$	0.17	0.14	0.16	-0.34	-0.40	0.48	-0.03	0.82	1.00	
Unconstrained	0.18	-0.13	0.24	-0.34	-0.59	0.72	-0.00	0.78	0.86	
Stacked Sample										
$\sum v_i = 1 \& \forall v_i \ge 0$	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.99	1.00	
$\sum v_i = 1$	0.01	-0.01	0.03	-0.05	-0.01	0.02	0.02	1.00	1.00	
Unconstrained	-0.01	-0.02	0.03	-0.07	0.13	0.04	0.32	0.34	0.77	

5. Empirical results

In this section, the methods suggested in Section 3 were compared in terms of the accuracy in the density forecast and the point forecast, for the hierarchy in the case study defined in Section 4.

5.1. Density Forecast Evaluation

For the stacked, ranked and permuted samples and for each reconciliation method, the CRPS value of the wind power density forecast from each of 60 nodes are evaluated for each forecast origin in \mathcal{T}_{test} . These values are averaged across \mathcal{T}_{test} , and then averaged again across all the nodes in each hierarchical level to be presented as each column of Table 4. The final column of the table is the average of all the previous columns in the same row, equivalent to the average of the level-wise average of the CRPS values in the hierarchy. The unit of the CRPS values is Mega Watt (MW), and lower values of this measure are preferred.

If we look at Table 4, the CRPS values are presented first by the sampling scheme defined in Section 3.2, namely the permuted, stacked or ranked sample, and then by the reconciliation methods, defined in Section 3.3, as the results are more influenced by the choice of the sampling scheme than the choice of the reconciliation method. For example, if we look at the final column, the permuted sample, which would make sense for independent data, and the stacked sampling scheme, which uses the level-wise dependency given by Monte-Carlo simulations of underlying

density forecast generation process, mostly did not improve the CRPS of the benchmark, noreconciliation even though we applied various reconciliation methods. The only exception was
the cross-validated reconciliation methods of the permuted sample, which were outperforming noreconciliation in the levels from 24 hourly to 4 hourly, but not for 2 hourly and 1 hourly. It
is particularly disappointing to see the stacked sample demonstrated little improvement over the
permuted sample in bottom-up, bottom average, global average, lineal average and WLS, while it
was much worse in cross-validated. On the other hand, if we look at the CRPS values in the ranked
sample, all the reconciliation methods clearly improved the results of no-reconciliation. The strong
performance of the ranked sample may be explained by its interpretation as a forecast combination
method.

In the ranked sample, the greatest accuracy was obtained by cross-validated for every level. Cross-validated synthesises information from every level based on 'data-driven' cross-validation weights presented in Tables 2 and 3, which clearly improved the overall density forecasting performance over the other reconciliation methods. We could not find any consistent difference between various cross-validation conditions in the ranked sample. In terms of accuracy, cross-validated is followed by global average, WLS, bottom average, bottom-up and lineal average. Given that lineal average is a special case of cross-validated, where all v are 1/L, the poor performance of lineal average in comparison with cross-validated suggests that the optimal weights are far from such fixed weights. It is surprising to see that global average, bottom-up and bottom average performed well, given their simplicity.

To investigate more closely the performance for each of the forecast lead times in each level and for each wind farm, we plotted in Figure 7 the CRPS values of no-reconciliation, WLS using ranked sample and cross-validated using ranked sample with no-constraint. Although the three months in the evaluation period is not sufficient to obtain smooth lines of CRPS in the plots, there is a clear tendency for the CRPS values to increase with forecast lead times in each plot. The title of each plot in Figure 7 indicates the average improvement of cross-validated over no-reconciliation, in terms of CRPS, where lower values are preferred. For example, the 24 hourly density forecast of cross-validated produced the CRPS values that are 26.6% and 21.1% smaller than no-reconciliation in the Rokas and Aeolos wind farms, respectively. As we increase the forecast resolution by moving further down the hierarchical level in the following plots, this enhancement

Table 4: CRPS measured for each level of the hierarchy, averaged over the Rokas and Aeolos wind farms.

Sampling Scheme &	Forecast Resolution of Each Level									
Reconciliation Method	24h	12h	8h	6h	4h	3h	2h	1h	Mean	
No-reconciliation	1.69	1.75	1.72	1.78	1.76	1.76	1.70	1.74	1.74	
Permuted Sample										
Bottom-up	1.59	1.84	1.96	1.89	1.89	1.86	1.80	1.75	1.82	
Bottom Average	1.59	1.93	2.12	2.11	2.21	2.24	2.29	2.34	2.10	
Global Average	1.73	2.05	2.25	2.24	2.34	2.38	2.42	2.48	2.24	
Lineal Average	1.77	2.01	2.18	2.14	2.20	2.22	2.24	2.27	2.13	
WLS	1.73	1.99	2.13	2.07	2.07	2.02	1.95	1.79	1.97	
Cross-validated $\sum v_i = 1 \& \forall v_i \geq 0$	1.42	1.69	1.78	1.76	1.76	1.79	1.76	1.78	1.72	
$\sum v_i = 1$	1.29	1.59	1.70	1.68	1.72	1.73	1.73	1.75	1.65	
Unconstrained	1.29	1.57	1.69	1.67	1.71	1.72	1.72	1.73	1.64	
Ranked Sample										
Bottom-up	1.34	1.52	1.62	1.63	1.67	1.69	1.71	1.74	1.62	
Bottom Average	1.34	1.52	1.61	1.62	1.66	1.69	1.71	1.74	1.61	
Global Average	1.32	1.50	1.60	1.61	1.65	1.68	1.70	1.73	1.60	
Lineal Average	1.38	1.56	1.67	1.67	1.72	1.74	1.77	1.80	1.66	
WLS	1.32	1.50	1.61	1.61	1.65	1.68	1.70	1.73	1.60	
Cross-validated $\sum v_i = 1 \& \forall v_i \geq 0$	1.27	1.48	1.59	1.59	1.64	1.67	1.69	1.72	1.58	
$\sum v_i = 1$	1.28	1.49	1.59	1.59	1.64	1.67	1.69	1.72	1.58	
Unconstrained	1.28	1.49	1.60	1.59	1.65	1.67	1.70	1.73	1.59	
Stacked Sample										
Bottom-up	1.58	1.84	1.96	1.89	1.89	1.85	1.80	1.74	1.82	
Bottom Average	1.58	1.93	2.12	2.11	2.21	2.24	2.29	2.34	2.10	
Global Average	1.73	2.05	2.25	2.24	2.34	2.37	2.42	2.48	2.23	
Lineal Average	1.77	2.01	2.18	2.14	2.20	2.22	2.24	2.26	2.13	
WLS	1.73	1.99	2.13	2.07	2.07	2.02	1.95	1.79	1.97	
Cross-validated $\sum v_i = 1 \& \forall v_i \geq 0$	1.62	1.89	2.02	1.94	1.95	1.91	1.85	1.77	1.87	
$\sum v_i = 1$	1.57	1.86	2.00	1.93	1.94	1.91	1.85	1.78	1.86	
Unconstrained	1.98	2.13	2.22	2.19	2.19	2.18	2.15	2.11	2.14	

Note: Lower values are better. The best value in each column is in bold.

tended to be reduced. This indicates that wind power density forecasts of the higher resolution could be further enhanced by synthesizing forecasts of lower resolution. The reconciliation in some sense 'hedges' the misspecification errors by synthesizing information from all hierarchical nodes.

5.2. Point Forecast Evaluation

Although density forecast performance is our primary concern, a density forecast could produce a point forecast by calculating the expected value from the density, and this could be useful for evaluating the centre of the density forecast. Gneiting (2011a,b) explains the median of a density forecast being the optimal point forecast for symmetric piecewise linear loss functions such as MAE. Indeed, in terms of MAE, using the median showed slightly higher accuracy than the mean in our

empirical results. The MAE result using the median is presented in Table 5, produced in the same fashion as Table 4.

In Table 5, it is surprising to see that the various reconciliation methods we proposed provide more competitive results for MAE than for CRPS. For example, if we look at the overall mean of MAEs in the last column of Table 5, most of the combinations of sampling schemes and reconciliation methods produced smaller MAEs than no-reconciliation. This enhancement is clearer in Figure 8, which is plotted in similar fashion to Figure 7 but using MAE. In the plots for 24 hourly for Rokas and Aeolos, the enhancements of the best density forecast method against no-reconciliation, in terms of MAE, were 35.1% for Rokas and 27.2% for Aeolos, whereas the enhancements were 26.6% and 21.1% respectively in terms of CRPS. This supports the temporal hierarchical density reconciliation methods we propose produce further enhancement in the centre of the forecast distributions. If we go back to Table 5, we can observe that the sampling scheme produced the most accurate point forecasts overall was the ranked sample, which was consistent with the results of the density forecast evaluation. Among the ranked sample, the global average and unconstrained cross-validated reconciliation methods were the most accurate.

6. Concluding Comments

This paper focused on the reconciliation of probabilistic forecasts that are arranged in hierarchical structures, with a particular focus on temporal hierarchies. We propose three schemes for obtaining samples from the estimates of the joint densities, namely permuted, ranked and stacked sampling. These approaches correspond to the cases of no dependence between the hierarchical nodes, comonotonic dependence between nodes and temporal model driven dependencies within a level respectively. These sampling schemes are then applied to several reconciliation approaches, bottom-up, bottom/global/lineal average and WLS. Furthermore, we investigated for the first time the use of a cross-validation approach for obtaining the reconciliation weights. The performance of the various combinations of sampling schemes and reconciliation methods was subsequently measured by producing and evaluating probabilistic wind power forecasts reconciled from various frequencies.

The empirical results from two wind farms in Greece suggest that cross-validation reconciliation based on ranked samples offers the best performance compared to all other approaches. Performance

Table 5: MAE measured for each level of the hierarchy, averaged over the Rokas and Aeolos wind farms.

Sampling Scheme &	Forecast Resolution of Each Level									
Reconciliation Method	24h	12h	8h	$6\mathrm{h}$	4h	3h	2h	1h	Mean	
No-reconciliation	2.55	2.58	2.50	2.59	2.58	2.59	2.49	2.59	2.56	
Permuted Sample										
Bottom-up	1.91	2.29	2.47	2.46	2.55	2.58	2.59	2.59	2.43	
Bottom Average	1.91	2.27	2.46	2.45	2.55	2.59	2.64	2.70	2.45	
Global Average	1.95	2.28	2.48	2.48	2.57	2.61	2.66	2.72	2.47	
Lineal Average	2.08	2.37	2.58	2.58	2.67	2.70	2.75	2.80	2.57	
WLS	1.95	2.30	2.51	2.50	2.58	2.61	2.64	2.63	2.46	
Cross-validated $\sum v_i = 1 \& \forall v_i \geq 0$	1.82	2.25	2.43	2.40	2.49	2.51	2.53	2.57	2.38	
$\sum v_i = 1$	1.74	2.21	2.39	2.37	2.47	2.49	2.52	2.55	2.34	
Unconstrained	1.74	2.20	2.37	2.36	2.45	2.47	2.50	2.54	2.33	
Ranked Sample		. – – –								
Bottom-up	1.84	2.22	2.40	2.39	2.47	2.51	2.55	2.59	2.37	
Bottom Average	1.84	2.15	2.31	2.31	2.40	2.43	2.48	2.52	2.31	
Global Average	1.78	2.11	2.29	2.28	2.37	2.42	2.46	2.51	2.28	
Lineal Average	1.96	2.26	2.45	2.43	2.52	2.56	2.60	2.65	2.43	
WLS	1.78	2.17	2.35	2.34	2.43	2.47	2.51	2.56	2.33	
Cross-validated $\sum v_i = 1 \& \forall v_i \geq 0$	1.71	2.19	2.35	2.33	2.42	2.45	2.50	2.54	2.31	
$\sum v_i = 1$	1.74	2.18	2.34	2.33	2.42	2.45	2.49	2.53	2.31	
Unconstrained	1.76	2.16	2.31	2.29	2.39	2.42	2.46	2.50	2.29	
Stacked Sample										
Bottom-up	1.91	2.29	2.47	2.46	2.55	2.57	2.58	2.59	2.43	
Bottom Average	1.91	2.27	2.46	2.45	2.55	2.59	2.64	2.70	2.45	
Global Average	1.95	2.28	2.48	2.48	2.57	2.61	2.67	2.72	2.47	
Lineal Average	2.08	2.37	2.58	2.57	2.67	2.70	2.74	2.80	2.56	
WLS	1.95	2.30	2.51	2.49	2.58	2.61	2.64	2.63	2.46	
Cross-validated $\sum v_i = 1 \& \forall v_i \geq 0$	1.92	2.31	2.50	2.48	2.57	2.60	2.61	2.61	2.45	
$\sum v_i = 1$	1.86	2.27	2.46	2.44	2.54	2.56	2.58	2.60	2.41	
Unconstrained	2.20	2.43	2.57	2.59	2.64	2.68	2.71	2.74	2.57	

Note: Point forecasts are medians of density forecasts. Lower values are better. The best value in each column is in bold.

enhancement is up to 25% and up to 35% relatively to no-reconciliation for density and point forecast evaluation respectively. Lower resolutions (higher levels of aggregation) enjoyed the most performance benefits, providing direct managerial benefits for transmission operations and planning for optimal trading strategies. The results also show that comonotonic aggregation of quantiles worked better than modelling level-wise dependencies.

While our study focused on the application of the various approaches in temporal hierarchies, these can be applied equally to the case of cross-sectional hierarchies, thus extending the work by Ben Taieb, Taylor, and Hyndman (2017) who investigate the construction of coherent probabilistic

forecasts in a bottom-up fashion. Furthermore, our study investigates for the very first time the performance of cross-validated derived weights for the construction of coherent forecasts. We suggest that cross-validation can also be applied to the construction of coherent point forecasts.

Looking forward, our research also poses new research questions that lie outside the scope of the current paper. For example, although an advantage of the stacked sample, ranked sample and permuted sample is their ease of construction, it may be worthwhile developing more complicated merging schemes based on the dependence structure of in sample forecast errors and investigating whether such schemes lead to better reconciled probabilistic forecasts. It may also be worthwhile investigating whether the sparse structure of the P matrix can be selected in a more data driven way, especially for cross sectional hierarchies where a different pattern of sparsity may be required to compensate for the base level forecasts are produced at each node rather than at each level. Finally, it would be interesting to see if methods based on ensemble or physics that can generate density forecasts also produce benefits using (temporal) hierarchical reconciliation.

Acknowledgements

Jooyoung Jeon was supported by the EPSRC grant (EP/N03466X/1). We are grateful to George Sideratos of the National Technical University of Athens and the EU SafeWind Project for providing the data. We are also grateful for the insight comments of participants at the International Symposium on Energy Analytics in Cairns, Australia, 2017.

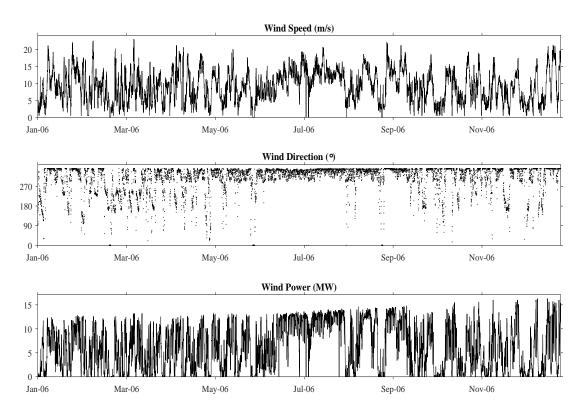


Figure 4: Hourly time series of wind speed, wind direction and wind power in the Rokas wind farm, Crete.

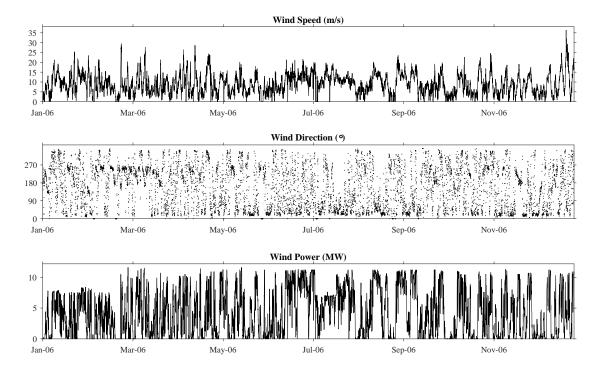


Figure 5: Hourly time series of wind speed, wind direction and wind power in the Aeolos wind farm, Crete.

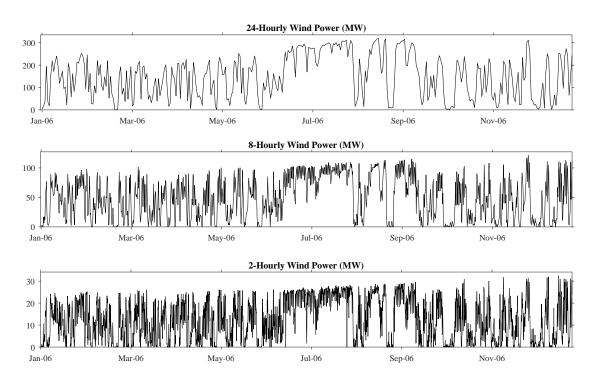


Figure 6: The 24 hourly, 8 hourly and 2 hourly time series of wind power in the Rokas wind farm, Crete.

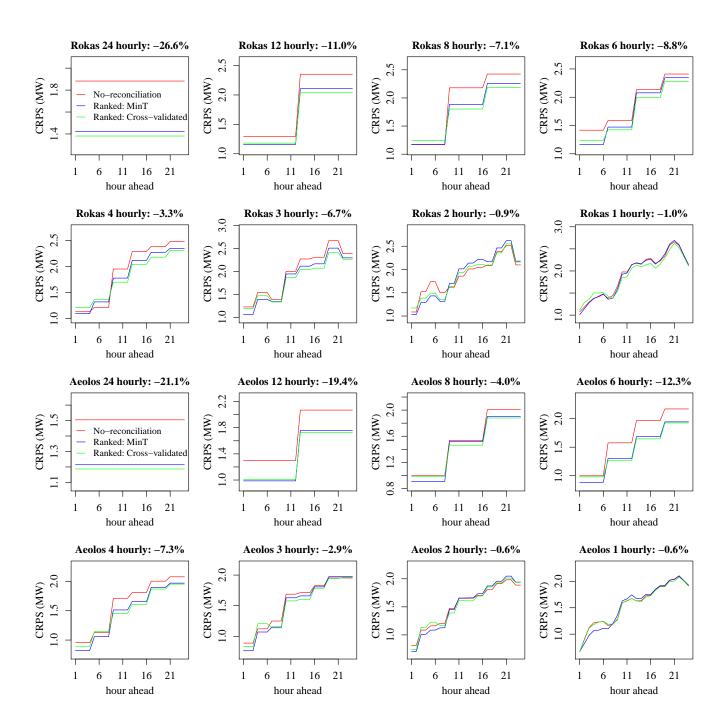


Figure 7: Probabilistic Evaluation of wind power forecasts in the evaluation period using CRPS for the Rokas and Aeolos wind farms, comparing (1) no-reconciliation and (2) bottom-up using the ranked sample and (3) Cross-validated using the ranked sample with constraint, $\sum v_i = 1 \& \forall v_i \geq 0$. The improvement of cross-validated over no-reconciliation is presented in average percentage for each level separately, on top of each plot. Lower values are better.

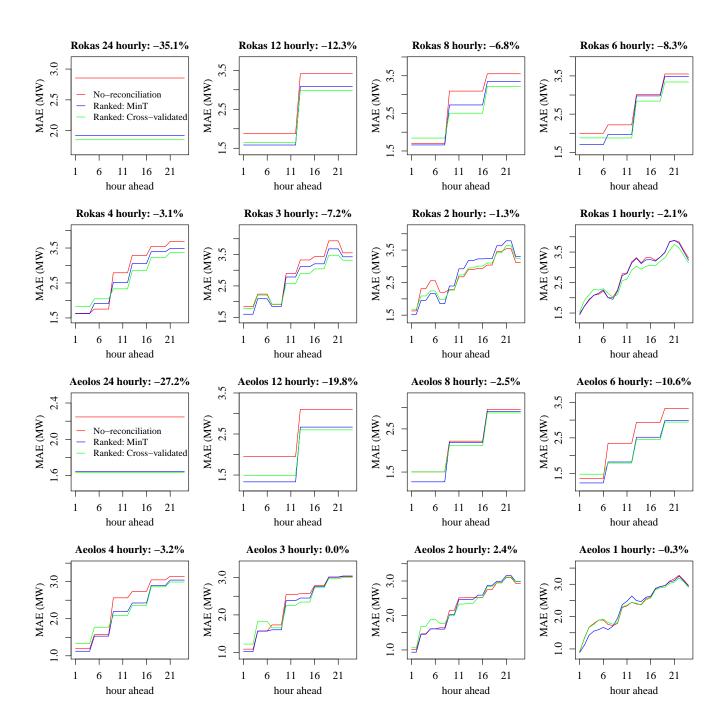


Figure 8: Point Evaluation of wind power density forecasts in the evaluation period using MAE for the Rokas and Aeolos wind farms, comparing (1) no-reconciliation and (2) bottom-up using the ranked sample and (3) Cross-validated using the ranked sample with constraint, $\sum v_i = 1 \& \forall v_i \geq 0$. The improvement of cross-validated over no-reconciliation is presented in average percentage for each level separately, on top of each plot. Lower values are better.

References

- Arbenz, Philipp, Christoph Hummel, and Georg Mainik. 2012. "Copula based hierarchical risk aggregation through sample reordering." *Insurance: Mathematics and Economics* 51 (1): 122–133.
- Athanasopoulos, George, Roman A Ahmed, and Rob J Hyndman. 2009. "Hierarchical forecasts for Australian domestic tourism." *International Journal of Forecasting* 25 (1): 146–166.
- Athanasopoulos, George, Rob J. Hyndman, Nikolaos Kourentzes, and Fotios Petropoulos. 2017. "Forecasting with temporal hierarchies." European Journal of Operational Research 262 (1): 60–74.
- Baillie, Richard T., Tim Bollerslev, and Hans Ole Mikkelsen. 1996. "Fractionally integrated generalized autoregressive conditional heteroskedasticity." *Journal of Econometrics* 74 (1): 3–30.
- Ben Taieb, Souhaib, James W. Taylor, and Rob J. Hyndman. 2017. "Coherent probabilistic forecasts for hierarchical time series." Proceedings of the 34th International Conference on Machine Learning 70: 3348–3357.
- Bollerslev, Tim, Robert F. Engle, and Jeffrey M. Wooldridge. 1988. "A capital asset pricing model with time-varying covariances." *Journal of Political Economy* 96 (1): 116–131.
- Dangerfield, Byron J, and John S Morris. 1992. "Top-down or bottom-up: Aggregate versus disaggregate extrapolations." *International Journal of Forecasting* 8 (2): 233–241.
- Dowell, Jethro, and Pierre Pinson. 2015. "Very-short-term probabilistic wind power forecasts by sparse vector autoregression." *IEEE Transactions on Smart Grid* 7 (2): 763–770.
- Fliedner, Gene. 1999. "An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation." Computers & Operations Research 26 (10): 1133–1149.
- Gneiting, Tilmann. 2011a. "Making and evaluating point forecasts." Journal of the American Statistical Association 106 (494): 746–762.
- Gneiting, Tilmann. 2011b. "Quantiles as optimal point forecasts." International Journal of Forecasting 27 (2): 197–207.
- Gneiting, Tilmann, and Matthias Katzfuss. 2014. "Probabilistic forecasting." Annual Review of Statistics and Its Application 1 (1): 125–151.
- Gneiting, Tilmann, Kristin Larson, Kenneth Westrick, Marc G Genton, and Eric Aldrich. 2006. "Calibrated probabilistic forecasting at the stateline wind energy center." *Journal of the American Statistical Association* 101 (475): 968–979
- Granger, C. W. J., and Roselyne Joyeux. 1980. "An introduction to long-memory time series models and fractional differencing." *Journal of Time Series Analysis* 1 (1): 15–29.
- Gross, Charles W, and Jeffrey E Sohl. 1990. "Disaggregation methods to expedite product line forecasting." *Journal of Forecasting* 9 (3): 233–254.
- Hering, Amanda S., and Marc G. Genton. 2010. "Powering up with space-time wind forecasting." *Journal of the American Statistical Association* 105 (489): 92–104.
- Hosking, J. R. M. 1981. "Fractional differencing." Biometrika 68 (1): 165.
- 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, 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.
- Jeon, Jooyoung, and James W. Taylor. 2012. "Using conditional kernel density estimation for wind power density forecasting." Journal of the American Statistical Association 107 (497): 66–79.
- Kourentzes, Nikolaos, and Fotios Petropoulos. 2016. "Forecasting with multivariate temporal aggregation: The case of promotional modelling." *International Journal of Production Economics* 181, Part A: 145–153.
- Kourentzes, Nikolaos, Fotios Petropoulos, and Juan Ramon Trapero. 2014. "Improving forecasting by estimating time series structural components across multiple frequencies." *International Journal of Forecasting* 30 (2): 291–302.
- Lichtendahl, Kenneth C., Yael Grushka-Cockayne, and Robert L. Winkler. 2013. "Is it better to average probabilities or quantiles?." *Management Science* 59 (7): 1594–1611.
- Lütkepohl, Helmut. 1984. "Forecasting contemporaneously aggregated vector ARMA processes." *Journal of Business & Economic Statistics* 2 (3): 201–214.
- Nikolopoulos, Konstantinos, Aris A Syntetos, John E Boylan, Fotios Petropoulos, and Vassilios Assimakopoulos. 2011. "An aggregate disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis." *Journal of the Operational Research Society* 62 (3): 544–554.
- Petropoulos, Fotios, and Nikolaos Kourentzes. 2014. "Improving forecasting via multiple temporal aggregation." Foresight: The International Journal of Applied Forecasting 34: 12–17.
- Petropoulos, Fotios, and Nikolaos Kourentzes. 2015. "Forecast combinations for intermittent demand." The Journal of the Operational Research Society 66 (6): 914–924.

- Petropoulos, Fotios, Nikolaos Kourentzes, and Konstantinos Nikolopoulos. 2016. "Another look at estimators for intermittent demand." *International Journal of Production Economics* 181, Part A: 154–161.
- Pinson, Pierre. 2013. "Wind energy: forecasting challenges for its operational management." Statistical Science 28 (4): 564–585.
- Rostami-Tabar, B., M.Z. Babai, A. Syntetos, and Y. Ducq. 2013. "Demand forecasting by temporal aggregation." Naval Research Logistics 60 (6).
- Roulston, M.S., and L.A. Smith. 2003. "Combining dynamical and statistical ensembles." *Tellus A: Dynamic Meteorology and Oceanography* 55 (1): 16–30.
- Sloughter, J. McLean, Tilmann Gneiting, and Adrian E. Raftery. 2010. "Probabilistic wind speed forecasting using ensembles and Bayesian model averaging." *Journal of the American Statistical Association* 105 (489): 25–35.
- Spithourakis, Georgios, Fotios Petropoulos, Konstantinos Nikolopoulos, and Vassilios Assimakopoulos. 2012. "A systemic view of ADIDA framework." *IMA Management Mathematics* (forthcoming).
- Taylor, J.W., P.E. McSharry, and R. Buizza. 2009. "Wind power density forecasting using ensemble predictions and time series models." *IEEE Transactions on Energy Conversion* 24: 775–782.
- Taylor, James W. 2017. "Probabilistic forecasting of wind power ramp events using autoregressive logit models." European Journal of Operational Research 259 (2): 703–712.
- Taylor, James W., and Jooyoung Jeon. 2015. "Forecasting wind power quantiles using conditional kernel estimation." Renewable Energy 80: 370–379.
- Wickramasuriya, Shanika L, George Athanasopoulos, and Rob J Hyndman. 2017. "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization." Working Paper .
- Zellner, Arnold, and Justin Tobias. 2000. "A note on aggregation, disaggregation and forecasting performance." Journal of Forecasting 19 (5): 457–465.