Probabilistic Forecasts for Hierarchical Time Series

Puwasala Gamakumara
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: puwasala.pamakumara@monash.edu
and
Anastasios Panagiotelis*
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: anastasios.panagiotelis@monash.edu
and
George Athanasopoulos
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: george.athanasopoulos@monash.edu
and
Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: rob.hyndman@monash.edu

October 9, 2019

Abstract

TBC

*The authors gratefully acknowledge the support of Australian Research Council Grant DP140103220. We also thank Professor Mervyn Silvapulle for valuable comments.

1 Introduction

Large collections of time series often follow some aggregation structure. For example, tourism flows of a country can be disaggregated along a geographic hierarchy of states, zones, and cities. Such collections of time series generally referred to as hierarchical time series. To ensure aligned decision making, it is important that forecasts across all levels of aggregation are adding up. This property is called "coherence". If the forecasts are not coherent, then these can be adjusted so that they become coherent. Earlier approaches for obtaining coherent forecasts involve generating first-stage forecasts for series in a single level of the hierarchy and aggregate them up or disaggregate them down to obtain forecasts for the remaining series. These are often call "bottom-up" and "top-down" forecasts respectively. For example see Dunn et al. (1976), Gross & Sohl (1990) and references therein.

An alternative approach to these single level forecasting methods is to do forecast "reconciliation". Reconciliation starts with a set of incoherent forecasts for the entire hierarchy and then revises these so that they are coherent with the aggregate constraints, see for example Athanasopoulos et al. (2009), Hyndman et al. (2011), Van Erven & Cugliari (2015), Shang & Hyndman (2017). From this literature we see that coherency and reconciliation has been extensively developed for the point forecasting case. Generalising both of these concepts, particularly the latter, to probabilistic forecasting is a gap that we seek to address in this chapter.

In contrast to the point forecasts, the entire probability distribution of future values provides a full description of the uncertainty associate with the predictions (Abramson & Clemen 1995, Gneiting & Katzfuss 2014). Therefore probabilistic forecasting has become of great interest in many disciplines such as, economics (Zarnowitz & Lambros 1987, Rossi

2014), meteorological studies (Pinson et al. 2009, McLean Sloughter et al. 2013), energy forecasting (Wytock & Kolter 2013, Ben Taieb, Huser, Hyndman & Genton 2017) and retail forecasting (Böse et al. 2017). However, the attention on probabilistic forecasts in the hierarchical literature is very limited. Indeed to the best of our knowledge, Ben Taieb, Taylor & Hyndman (2017) and Jeon et al. (2019) are the only papers to deal with probabilistic forecasts in the hierarchical time series. Although Ben Taieb, Taylor & Hyndman (2017) reconcile the means of predictive distributions, the overall distributions are constructed in a bottom-up fashion rather than using a reconciliation approach. Jeon et al. (2019) propose a novel method for probabilistic forecast reconciliation based on cross-validation which is particularly applied to temporal hierarchies. In contrast to these studies, the main objective of this chapter is to generalise both the concepts of coherence and reconciliation from point to probabilistic forecasting.

Extending the geometric interpretation related to point forecast reconciliation derived in (Panagiotelis et al. 2019) we provide new definitions of coherence and forecast reconciliation in the probabilistic setting. We also cover the topic of forecast evaluation of probabilistic forecasts via scoring rules. In particular, we prove that for a coherent data generating process, the log score is not proper with respect to incoherent forecasts. Therefore we recommend the use of the energy score or variogram score for comparing reconciled to unreconciled forecasts. Two or more reconciled forecasts can be compared using log score, energy score or variogram score, although we show that comparisons should be made on the full hierarchy for the latter two scores.

When parametric density assumptions are made we describe how the probabilistic forecast definitions lead to a reconciliation procedure that merely involves a change of basis and marginalisation. We show that probabilistic reconciliation via linear transformations can recover the true predictive distribution as long as the latter is in the elliptical class.

We provide conditions for which this linear transformation is a projection, and although this projection cannot be feasibly estimated in practice, we provide a heuristic argument in favour of MinT reconciliation.

Further we propose a new method to generate coherent forecasts when the parametric distributional assumptions are not applicable. This method uses a non-parametric bootstrap based approach to generate future paths for all series in the hierarchy and then reconcile each sample path using projections. This will provide a possible sample from the reconciled predictive density of the hierarchy. An extensive simulation study was carried out to find the optimal reconciliation of bootstrap future paths with respect to a proper scoring rule. This has shown that the MinT method is at least as good as the optimal method for reconciling future paths.

Finally we applied both parametric and non-parametric approaches to generate probabilistic forecasts for domestic tourism flow in Australia. The results shows that reconciliation improves forecast accuracy compared to incoherent forecasts in both parametric and non-parametric approaches and further, MinT reconciliation is performing best.

The remainder of the paper is structured as follows. In Section 2 notation and some preliminary work on point forecast reconciliation is discussed. Section 3 contains the definitions and interpretation of coherent probabilistic forecasts and reconciliation. In Section 4 we consider the evaluation of probabilistic hierarchical forecasts via scoring rules. Parametric forecast reconciliation and some theoretical results related to elliptical distributions are discussed in Section 5 while the non-parametric approach is introduced in Section 6. An empirical application on tourism forecasting is contained in Section 7. Finally Section 8 concludes with some discussion and thoughts on future research.

2 Notations and preliminaries

Following (Panagiotelis et al. 2019) we define notation and some preliminary work on hierarchical time series and forecasting.

2.1 Notations

Hierarchical time series are a collection of n time series some of which are aggregates of other series. In a conventional hierarchy, m number of *bottom-level series* are aggregated to form series in upper levels. These bottom-level series can be considered as linearly independent since they cannot be formed as linear combination of other series. Let us denote with $\boldsymbol{y}_t$, a n dimensional vector containing observations of all series in the hierarchy at time t. Further, let $\boldsymbol{b}_t \in \mathbb{R}^m$ contain the observations of only the *bottom-level series* of the hierarchy at time t. Due to the aggregation constraints of the hierarchy we can write,

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t, \tag{1}$$

where $\boldsymbol{S}$ is an $n \times m$ constant matrix for a given hierarchical structure.

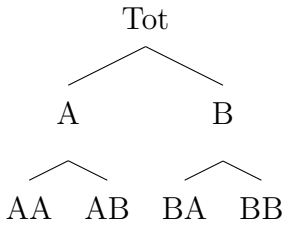

Figure 1: An example of a two level hierarchical structure.

To make explicit this notation, we refer to the hierarchy in Figure 1. This hierarchy consists of $m = 4$ bottom-level series with $n = 7$ total number of series. Further, $\boldsymbol{y}_t =$

$[y_{Tot,t}, y_{A,t}, y_{B,t}, y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$ and

$$\boldsymbol{S} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \boldsymbol{I}_4 & & \end{pmatrix},$$

where $\boldsymbol{I}_4$ is the 4×4 identity matrix.

The columns of $\boldsymbol{S}$ define the aggregation constraints of the hierarchy and they span a subspace $\mathfrak{s} \subset \mathbb{R}^n$ which all $\boldsymbol{y}_t$ lie on. This subspace is referred to as the *coherent subspace*. Any point that lies in this subspace is referred to as a *coherent point*. In particular, all observations are coherent by construction. Further, any point forecast $\breve{\boldsymbol{y}}_{t+h}$ that is conditional on past information upto and including time t, is said to be coherent if it lies in the coherent subspace.

It will sometimes be useful to think of pre-multiplication by $\boldsymbol{S}$ in equation (1) as a mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$, in which case we use the notation $s(.)$. Although the codomain of $s(.)$ is $\mathbb{R}^n$, its image is the coherent space $\mathfrak{s}$ as depicted in Figure 2.

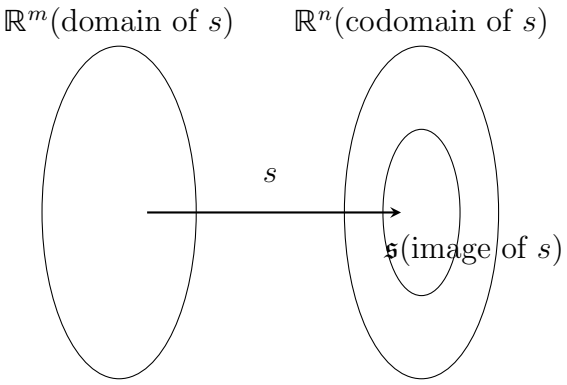

Figure 2: The domain, codomain and image of the mapping s.

2.2 Hierarchical point forecasts

(Panagiotelis et al. 2019) provides insights to coherent point forecasts and forecast reconciliation in terms of geometric concepts. We briefly discuss these to facilitate extensions into the probabilistic framework presented in Section 3.

Suppose we have a vector of point forecasts at time $t+h$ derived by using information up to and including time t. Let these forecasts be stacked in a vector with the same order as $\boldsymbol{y}_t$ and denoted by $\hat{\boldsymbol{y}}_{t+h} \in \mathbb{R}^n$. These are referred to as *incoherent point forecasts* as they do not satisfy the aggregation constraints. Assume a linear function that maps these incoherent forecasts into new bottom-level forecasts. Let $\boldsymbol{G}$ and $\boldsymbol{d}$ be an $m \times n$ matrix and $m \times 1$ vector respectively, and let $g : \mathbb{R}^n \to \mathbb{R}^m$ be the mapping $g(\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{y} + \boldsymbol{d}$. A composition of $g(.)$ and $s(.)$ gives the reconciled point forecasts as,

$$\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}(\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h} + \boldsymbol{d})\,. \tag{2}$$

Several choices of $g(.)$ are currently extant in the literature, including the bottom-up (Dunn et al. 1976), OLS, WLS and MinT (Hyndman et al. 2011, Wickramasuriya et al. 2019) methods. These are special cases where $s \circ g$ is a projection. These can be defined so that $\boldsymbol{G} = (\boldsymbol{R}'_\perp \boldsymbol{S})^{-1}\boldsymbol{R}'_\perp$ and $\boldsymbol{d} = \boldsymbol{0}$, where, $\boldsymbol{R}_\perp$ is a $n \times m$ orthogonal complement to an $n \times (n-m)$ matrix $\boldsymbol{R}$, where the columns of the latter span the null space of $\mathfrak{s}$. For example, a straightforward choice of $\boldsymbol{R}$ for the most simple three variable hierarchy where $y_{1,t} = y_{2,t} + y_{3,t}$, is the vector $(1, -1, -1)$ which is orthogonal (in the Euclidean sense) to the columns of $\boldsymbol{S}$. In this case, the matrix $\boldsymbol{R}$ can be interpreted as a 'restrictions' matrix since it has the property that $\boldsymbol{R}'\boldsymbol{y} = \boldsymbol{0}$ for coherent $\boldsymbol{y}$. For this three variable hierarchy, $\boldsymbol{R}'_\perp = \boldsymbol{S}$ and reconciliation corresponds to the OLS method. For the case where $\boldsymbol{R}'_\perp \neq \boldsymbol{S}$, for example WLS and MinT, $\boldsymbol{R}'_\perp$ is usually written in the form $\boldsymbol{S}'\boldsymbol{W}^{-1}$. These projections

can be thought of as orthogonal projections after pre-multiplying by $\boldsymbol{W}^{-1/2}$. More detailed explanation on this can be found in Sections **??** and **??** of (Panagiotelis et al. 2019). Table 1 summarises existing reconciliation methods.

Table 1: Summary of reconciliation methods that are projections. Here, $\hat{\boldsymbol{W}}^{sam}$ is the variance covariance matrix of one-step ahead in-sample forecast errors, $\hat{\boldsymbol{W}}^{shr}$ is a shrinkage estimator more suited to large dimensions proposed by Schäfer & Strimmer (2005), $\hat{\boldsymbol{W}}^{wls}$ is the diagonal matrix with diagonal elements w_{ii}, and $\tau = \frac{\sum_{i\neq j}\hat{\text{Var}}(\hat{w}_{ij})}{\sum_{i\neq j}\hat{w}_{ij}^2}$, where w_{ij} denotes the (i,j)th element of $\hat{\boldsymbol{W}}^{sam}$.

Method	$\boldsymbol{W}$	$\boldsymbol{R}_{\perp}'$
OLS	$\boldsymbol{I}$	$\boldsymbol{S}'$
MinT(Sample)	$\hat{\boldsymbol{W}}^{sam}$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{sam})^{-1}$
MinT(Shrink)	$\tau\text{Diag}(\hat{\boldsymbol{W}}^{sam}) + (1-\tau)\hat{\boldsymbol{W}}^{sam}$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{shr})^{-1}$
WLS	$\text{Diag}(\hat{\boldsymbol{W}}^{sam})$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{wls})^{-1}$

The columns of $\boldsymbol{S}$ and $\boldsymbol{R}$ provide a basis for $\mathbb{R}^n$. Therefore any incoherent set of point forecasts $\hat{\boldsymbol{y}}_{t+h}$ can be expressed in terms of coordinates in the basis defined by $\boldsymbol{S}$ and $\boldsymbol{R}$. Let $\tilde{\boldsymbol{b}}_{t+h}$ and $\tilde{\boldsymbol{a}}_{t+h}$ be the coordinates corresponding to $\boldsymbol{S}$ and $\boldsymbol{R}$ respectively, after a change of basis. The process of reconciliation involves setting the values of the reconciled bottom-level forecasts to be $\tilde{\boldsymbol{b}}_{t+h}$, and ignoring $\tilde{\boldsymbol{a}}_{t+h}$ to ensure coherence. From properties of linear algebra it follows that

$$\hat{\boldsymbol{y}}_{t+h} = (\boldsymbol{S}\ \boldsymbol{R})\begin{pmatrix}\tilde{\boldsymbol{b}}_{t+h}\\ \tilde{\boldsymbol{a}}_{t+h}\end{pmatrix} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{a}}_{t+h},$$

while the reconciled point forecast is

$$\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h}.$$

In order to find $\tilde{\boldsymbol{b}}_{t+h}$ we require the inverse $(\boldsymbol{S}\ \boldsymbol{R})^{-1}$ which is given by

$$(\boldsymbol{S}\ \boldsymbol{R})^{-1} = \begin{pmatrix} (\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp \\ (\boldsymbol{S}'_\perp \boldsymbol{R})^{-1} \boldsymbol{S}'_\perp \end{pmatrix}, \tag{3}$$

where $\boldsymbol{S}_\perp$ is the orthogonal complements of $\boldsymbol{S}$. Thus it follows that $\tilde{\boldsymbol{b}}_{t+h} = (\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp \hat{\boldsymbol{y}}_{t+h}$ and $\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}(\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp \hat{\boldsymbol{y}}_{t+h}$. Here $(\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp$ corresponds to $\boldsymbol{G}$ as defined previously.

3 Hierarchical probabilistic forecasts

The geometric intuition in hierarchical point forecasts provides a solid basis for extending the idea into the probabilistic framework. We start with providing definitions for coherent probabilistic forecasts and probabilistic forecast reconciliation.

3.1 Coherent probabilistic forecasts

Let $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m}, \nu)$ be a probability triple, where $\mathscr{F}_{\mathbb{R}^m}$ is the usual Borel σ-algebra on $\mathbb{R}^m$. Let $\breve{\nu}$ be a probability measure on $\mathfrak{s}$ with σ-algebra $\mathscr{F}_{\mathfrak{s}}$. Here $\mathscr{F}_{\mathfrak{s}}$ is a collection of sets $s(\mathcal{B})$, where $s(\mathcal{B})$ denotes the image of the set $\mathcal{B} \in \mathscr{F}_{\mathbb{R}^m}$ under the mapping $s(.)$.

Definition 3.1 (Coherent Probabilistic Forecasts)**.** The measure $\breve{\nu}$ is coherent if it has the property

$$\breve{\nu}(s(\mathcal{B})) = \nu(\mathcal{B}) \quad \forall \mathcal{B} \in \mathscr{F}_{\mathbb{R}^m},$$

A probabilistic forecast for time $t+h$ is coherent if uncertainty in $\boldsymbol{y}_{t+h}$ conditional on all information up to time t is characterised by the probability triple $(\mathfrak{s}, \mathscr{F}_{\mathfrak{s}}, \breve{\nu})$.

To the best of our knowledge, the only other definition of coherent probabilistic forecasts is given by Ben Taieb, Huser, Hyndman & Genton (2017) who define coherent probabilistic

forecasts in terms of convolutions. According to their definition, probabilistic forecasts are coherent when a convolution of forecast distributions of disaggregate series is identical to the forecast distribution of the corresponding aggregate series. Their definition is consistent with our definition; our reason for providing a different definition is that the geometric understanding of coherence will facilitate our definition of probabilistic forecast reconciliation to which we now turn our attention.

3.2 Probabilistic forecast reconciliation

We now extend the methodology of point forecast reconciliation to probabilistic forecasts. Let $(\mathbb{R}^n, \mathscr{F}_{\mathbb{R}^n}, \hat{\nu})$ be a probability triple that is not coherent and which characterises forecast uncertainty for all variables in the hierarchy at time $t+h$ conditional on all information up to time t. This is obtained from the first stage of the forecasting process; by modelling and forecasting each series in the hierarchy. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a linear function, and let $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m}, \nu)$ be a probability triple defined on $\mathbb{R}^m$.

Definition 3.2. The reconciled probability measure of $\hat{\nu}$ with respect to the mapping $g(.)$ is a probability measure $\tilde{\nu}$ on $\mathfrak{s}$ with σ-algebra $\mathscr{F}_{\mathfrak{s}}$ such that

$$\tilde{\nu}(s(\mathcal{B})) = \nu(\mathcal{B}) = \hat{\nu}(g^{-1}(\mathcal{B})) \qquad \forall \mathcal{B} \in \mathscr{F}_{\mathbb{R}^m}\,, \tag{4}$$

where $g^{-1}(\mathcal{B}) := \{\breve{\boldsymbol{y}} \in \mathbb{R}^n : g(\breve{\boldsymbol{y}}) \in \mathcal{B}\}$ is the pre-image of $\mathcal{B}$, that is the set of all points in $\mathbb{R}^n$ that $g(.)$ maps to a point in $\mathcal{B}$.

This definition extends the notion of forecast reconciliation to the probabilistic setting. Under point reconciliation methods, the reconciled point forecast is equal to the unreconciled point forecast after the latter is passed through two linear functions. Similarly, probabilistic forecast reconciliation assigns the same probability to two sets where

the points in one set are obtained by passing all points in the other set through two linear functions. This is depicted in Figure 3 schematically when $s \circ g$ is a projection.

Following these definitions we can derive probabilistic forecasts for parametric distributions which we will elaborate in Section 5. Now we turn our attention to the evaluation of hierarchical probabilistic forecasts.

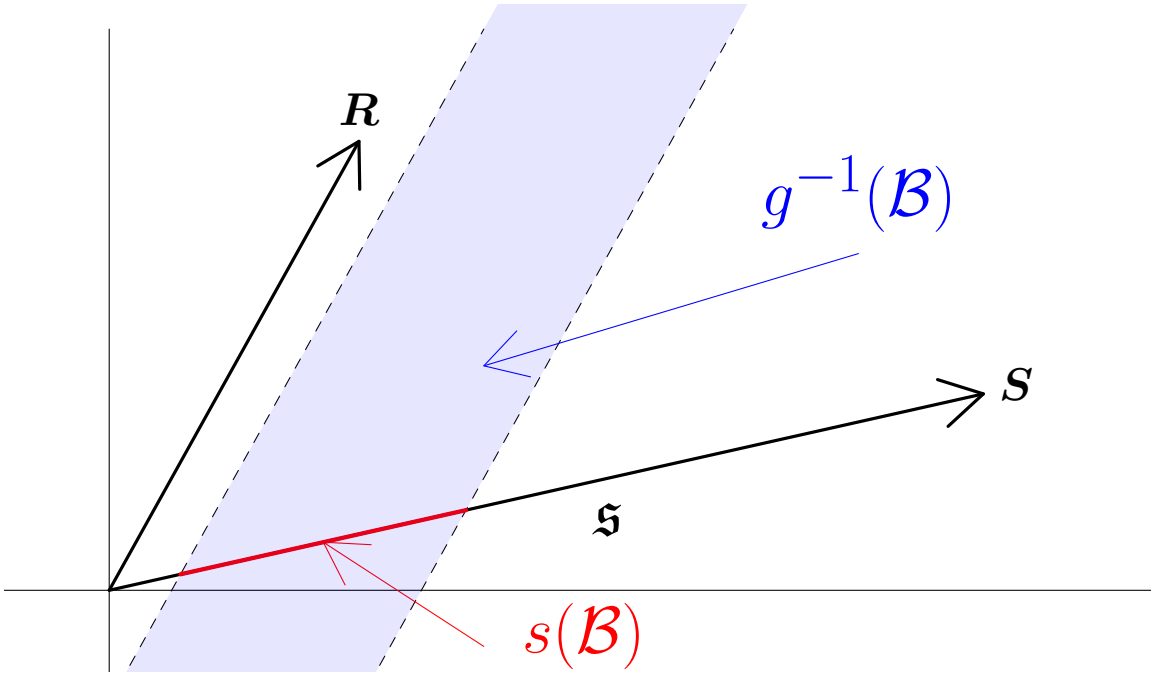

Figure 3: Summary of probabilistic forecast reconciliation. The probability that $\boldsymbol{y}_{t+h}$ lies in the red line segment under the reconciled probabilistic forecast is defined to be equal to the probability that $\boldsymbol{y}_{t+h}$ lies in the shaded blue area under the unreconciled probabilistic forecast. Note that since the smallest possible hierarchy involves three dimensions, this figure is only a schematic.

4 Evaluation of hierarchical probabilistic forecasts

The necessary final step in hierarchical forecasting is to make sure that our forecast distributions are accurate. In general, forecasters prefer to maximize the sharpness of the forecast distribution subject to calibration (Gneiting & Katzfuss 2014). Therefore the probabilistic forecasts should be evaluated with respect to these two properties.

Calibration refers to the statistical compatibility between probabilistic forecasts and realizations. In other words, random draws from a perfectly calibrated forecast distribution should be equivalent in distribution to the realizations. On the other hand, sharpness refers to the spread or the concentration of the predictive distributions and it is a property of the forecasts only. The more concentrated the forecast distributions, the sharper the forecasts (Gneiting et al. 2008). However, independently assessing the calibration and sharpness will not help to properly evaluate the probabilistic forecasts. Therefore we need to assess these properties simultaneously using scoring rules.

Scoring rules are summary measures obtained based on the relationship between the forecast distributions and the realizations. In some studies, researchers take the scoring rules to be positively oriented, in which case the scores should be maximized (Gneiting & Raftery 2007). However, scoring rules have also been defined to be negatively oriented, and then the scores should be minimized (Gneiting & Katzfuss 2014). We follow the latter convention here.

Let P be a forecast distribution and let Q be the true data generating process respectively. Furthermore let ω be a realization from Q. Then a scoring rule is a function $S(P, \omega)$ that maps P, ω to $\mathbb{R}$. It is a "proper" scoring rule if

$$\mathrm{E}_{\boldsymbol{Q}}[S(Q, \omega)] \leq \mathrm{E}_{\boldsymbol{Q}}[S(P, \omega)], \tag{5}$$

where $\mathrm{E}_Q[S(P, \omega)]$ is the expected score under the true distribution Q (Gneiting et al. 2008,

Gneiting & Katzfuss 2014). When this inequality is strict, the scoring rule is said to be strictly proper.

In the context of probabilistic forecast reconciliation there could be two motivations for using scoring rules. The first is to compare unreconciled densities to reconciled densities. Reconciliation itself is a valuable goal since it can be important in aligning decision making across, for example, different units of an enterprise. In the point forecasting literature, forecast reconciliation has also been shown to improve forecast performance (Athanasopoulos et al. 2017, Wickramasuriya et al. 2019). It will be worthwhile to see whether the same holds in the probabilistic forecasting case. The second motivation for using scoring rules is to compare two or more sets of reconciled probabilistic forecasts to one another. The objective here is to evaluate which reconciliation mapping $g(.)$ works best in practice.

4.1 Univariate scoring rules

One way to evaluate hierarchical probabilistic forecasts is via the application of univariate scoring rules to each time series in the hierarchy. A summary can be taken of the expected scores across each margin, for example a mean or median. We consider two such scoring rules. The log score is given by the log density, in this case for each margin of the probabilistic forecast. The continuous rank probability score generalises mean square error and is given by

$$\text{CRPS}(\breve{F}_i, y_i) = \int \left(\breve{F}_i(\breve{Y}_i) - \mathbb{1}(\breve{Y}_i < y_i)\right) d\breve{Y}_i \tag{6}$$

$$= \text{E}_{\breve{Y}_i}|\breve{Y}_i - y_i| - \frac{1}{2}\text{E}_{\breve{Y}_i}|\breve{Y}_i - \breve{Y}_i^*|\,, \tag{7}$$

where $\breve{F}_i$ is the cumulative distribution function of the i^{th} margin of the probabilistic forecast, $\breve{Y}_i$ and $\breve{Y}_i^*$ are independent copies of a random variable with distribution $\breve{F}_i$, and

y_i is the outcome of the i^{th} margin. The expectations in the second line can be approximated by Monte Carlo when a sample from the predictive distribution is available.

An advantage of this approach is that it allows the forecaster to separately evaluate different levels and individual series of the hierarchy to determine where the gains from reconciliation are greatest. For this reason this approach has been used in the limited literature on probabilistic forecasting for hierarchies (Ben Taieb, Huser, Hyndman & Genton 2017, Jeon et al. 2019) to date. A major shortcoming of this approach however, is that evaluating univariate scores on the margins does not account for the dependence in the hierarchy.

4.2 Multivariate scoring rules

While a number of alternative proper scoring rules are available for univariate forecasts, the multivariate case is somewhat more limited. Here we focus on three scoring rules: the log score (LS), the energy score (ES) and the variogram score (VS).

The log score can be approximated using a sample of values from the probabilistic forecast density (Jordan et al. 2017); however it is more commonly used when a parametric form for the density is available for the probabilistic forecast.

The energy score on the other hand can be defined in terms of the characteristic function of the probabilistic forecast, but the following representation in terms of expectations

$$\mathrm{ES}(\breve{\boldsymbol{Y}}_{T+h}, \boldsymbol{y}_{T+h}) = \mathrm{E}_{\breve{\boldsymbol{Y}}}||\breve{\boldsymbol{Y}}_{T+h} - \boldsymbol{y}_{T+h}||^{\alpha} - \frac{1}{2}\mathrm{E}_{\breve{\boldsymbol{Y}}}||\breve{\boldsymbol{Y}}_{T+h} - \breve{\boldsymbol{Y}}^{*}_{T+h}||^{\alpha},\ \alpha \in (0, 2]\,, \tag{8}$$

lends itself to easy computation when samples from the probabilistic forecast are available and given as,

$$\mathrm{ES}(\breve{\boldsymbol{Y}}_{T+h}, \boldsymbol{y}_{T+h}) \approx \frac{1}{M}\sum_{i=1}^{M}||\boldsymbol{SG}\breve{\boldsymbol{y}}_{T+h,i} - \boldsymbol{y}_{T+h}|| - \frac{1}{2(M-1)}\sum_{i=1}^{M-1}||\boldsymbol{SG}(\breve{\boldsymbol{y}}_{T+h,i} - \breve{\boldsymbol{y}}_{T+h,i+1})||, \tag{9}$$

where, $\breve{\boldsymbol{y}}_{T+h,i}$ is the i^{th} Monte-Carlo sample from the forecast distribution. An interesting limiting case is where $\alpha = 2$, where it can be easily shown that energy score simplifies to mean squared error around the mean of the predictive distribution. In this limiting case, the energy score is proper but not strictly proper. Pinson & Tastu (2013) also argue that the energy score has low discriminative ability for incorrectly specified covariances, even though it discriminates the misspecified means well.

In contrast, Scheuerer & Hamill (2015) have shown that the variogram score has a higher discrimination ability for misspecified means, variances and correlation structures than the energy score. When $\breve{\boldsymbol{y}}$ is a random variable from probabilistic forecast $\breve{F}$, the empirical variogram score is defined as

$$\text{VS}(\breve{F}, \boldsymbol{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(|y_i - y_j|^p - E_{\breve{Y}_i,\breve{Y}_j}|\breve{Y}_i - \breve{Y}_j|^p\right)^2. \tag{10}$$

Scheuerer & Hamill (2015) recommend using $p = 0.5$.

4.2.1 Comparing unreconciled forecasts to reconciled forecasts

For both reconciled and unreconciled densities it is possible to obtain a density from the probability measures defined in Section 3. Therefore it may seem sensible to compare unreconciled densities to reconciled densities on the basis of log score. However, the following theorem shows that using the log score may fail in the case of multivariate distributions with a degeneracy.

Theorem 4.1 (Impropriety of log score)**.** *When the true data generating process is coherent, then the log score is improper with respect to the class of incoherent measures.*

Proof. Consider a rotated version of hierarchical time series, $\boldsymbol{z}_t = \boldsymbol{U}\boldsymbol{y}_t$, so that the first m elements of $\boldsymbol{z}_t$ denoted $\boldsymbol{z}_t^{(1)}$ are unconstrained, while the remaining $n-m$ elements denoted

$\boldsymbol{z}_t^{(2)}$ equal 0 when the aggregation constraints hold. An example of the $n \times n$ $\boldsymbol{U}$ is the matrix of left singular vectors of $\boldsymbol{S}$.

Consider the case where the true predictive density is $f_1(\boldsymbol{z}_t^{(1)})\mathbb{1}\left(\boldsymbol{z}_t^{(2)} = \mathbf{0}\right)$, and we evaluate an incoherent density given by $f_1(\boldsymbol{z}_t^{(1)})f_2(\boldsymbol{z}_t^{(2)})$, where f_2 is highly concentrated around 0 but still non-degenerate. For example, f_2 may be Gaussian with variance $\sigma^2\boldsymbol{I}$ with $\sigma^2 < (2\pi)^{-1}$. The log score under the true data generating process is

$$LS\left(f, \boldsymbol{z}_t^{(1)}\right) = -\log f_1\left(\boldsymbol{z}_t^{(1)}\right),$$

while that of the unreconciled density is

$$\begin{aligned} LS\left(\hat{f}, \boldsymbol{z}_t^{(1)}\right) &= -\log f_1(\boldsymbol{z}_t^{(1)}) - \log f_2(\boldsymbol{z}_t^{(1)}) && (11)\\ &= -\log f_1(\boldsymbol{z}_t^{(1)}) + \frac{n-m}{2}\log(2\pi\sigma^2) && (12)\\ &< -\log f_1(\boldsymbol{z}_t^{(1)}) = LS\left(f, \boldsymbol{z}_t^{(1)}\right). && (13) \end{aligned}$$

After taking expectations $E\left[LS(f, f)\right] > E\left[LS(\hat{f}, f)\right]$, violating the condition in Equation (5) for a proper scoring rule. □

A similar issue also arises when discrete random variables are modelled as if they were continuous, an issue discussed in Section 4.1 of Gneiting & Raftery (2007). This implies that the log score should be avoided when comparing reconciled and unreconciled probabilistic forecasts.

4.2.2 Comparing reconciled forecasts to one another

Coherent probabilistic forecasts can be completely characterised in terms of basis series; if a probabilistic forecast is available for the basis series, then a probabilistic forecast can be recovered for the entire hierarchy via Definition 3.1. This may suggest that it is adequate

to merely compare two coherent forecasts to one another using the basis series only. We now show how this depends on the specific scoring rule used.

For the log score, suppose the coherent probabilistic forecast has density $f(\boldsymbol{b})$. The density for the full hierarchy is given by $f(\boldsymbol{y}) = f(\boldsymbol{S}\boldsymbol{b}) = f(\boldsymbol{b})J^{-1}$, where $J = \prod_{j=1}^{m} \lambda_j$ is a pseudo-determinant of the non-square matrix $\boldsymbol{S}$ and λ_j are the non-zero singular values of $\boldsymbol{S}$. Therefore for any coherent density, the log score of the full hierarchy differs from the log score for the bottom-level series by the term $log(J)$. This term depends only on the structure of the hierarchy and is fixed across different reconciliation methods. Therefore if one method achieves a lower expected log score compared to an alternative method using the bottom-level series only, the same ordering is preserved when an assessment is made on the basis of the full hierarchy.

The same property does not hold for all scores in general. For example, the energy score can be expressed in terms of expectations of norms. In general, since norms are invariant under orthogonal rotations, the energy score is also invariant under orthogonal transformations (Székely & Rizzo 2013, Gneiting & Raftery 2007). In the context of two coherent forecasts, the same is true of a semi-orthogonal transformation from a lower dimensional basis series to the full hierarchy. However, when $\boldsymbol{S}$ is the usual summing matrix, it is not semi-orthogonal. Therefore the energy score computed on the bottom-level series will differ from the energy score computed using the full hierarchy and the ordering of different reconciliation methods may change depending on the basis series used. In this case we recommend computing the energy score using the full hierarchy. Although the discussion here is related to energy score, the same logic holds for other multivariate scores, for example the variogram score.

The properties of multivariate scoring rules in the context of evaluating reconciled probabilistic forecasts are summarised in Table 2.

Table 2: Summary of properties of scoring rules in the context of reconciled probabilistic forecasts.

	Coherent v Incoherent	Coherent v Coherent
Log Score	Not proper	Ordering preserved if compared using bottom-level only
Energy/	Proper	Full hierarchy should be used
Variogram Score	Proper	Full hierarchy should be used

5 Parametric reconciliation

Recall that when $s \circ g$ is a projection, the case of point forecast reconciliation can be broken down into three steps. In what follows we drop the subscript $t+h$ from conditional forecasts for brevity.

1. $\hat{\boldsymbol{y}}$ is transformed into coordinates $\tilde{\boldsymbol{b}}$ and $\tilde{\boldsymbol{a}}$ via a change of basis.
2. $\tilde{\boldsymbol{a}}$ is discarded and $\tilde{\boldsymbol{b}}$ are kept as the bottom-level reconciled forecasts.
3. Reconciled forecasts for the entire hierarchy are recovered via $\tilde{\boldsymbol{y}} = \boldsymbol{S}\tilde{\boldsymbol{b}}$.

We now outline the analogous steps for probabilistic forecasts when predictive densities are available.

While $\hat{\nu}$ is a probability measure for an n-vector $\hat{\boldsymbol{y}}$, probability statements in terms of a different coordinate system can be made via an appropriate change of basis. Letting $f(.)$ be generic notation for a probability density function, and following the notation from point forecast reconciliation where $\hat{\boldsymbol{y}} = \boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}}$, we obtain

$$f(\hat{\boldsymbol{y}}) = f(\boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}})|(\boldsymbol{S}\ \boldsymbol{R})| \tag{14}$$

The expression $\hat{\nu}(g^{-1}(\mathcal{B}))$ in Definition 3.2 is equivalent to the probability statement $\Pr(\hat{\boldsymbol{y}} \in$

$g^{-1}(\mathcal{B})$). After the change of basis, this is equivalent to $\Pr(\tilde{\boldsymbol{b}} \in \mathcal{B})$, which implies

$$\Pr(\hat{\boldsymbol{y}} \in g^{-1}(\mathcal{B})) = \int_{g^{-1}(\mathcal{B})} f(\hat{\boldsymbol{y}}) d\hat{\boldsymbol{y}} \tag{15}$$

$$= \int_{\mathcal{B}} \int f(\boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}}) |(\boldsymbol{S}\ \boldsymbol{R})| d\tilde{\boldsymbol{a}} d\tilde{\boldsymbol{b}}. \tag{16}$$

After integrating out over $\tilde{\boldsymbol{a}}$, a step analogous to setting $\tilde{\boldsymbol{a}} = 0$ for point forecasting, we obtain an expression that gives the probability that the reconciled bottom-level series lies in the region $\mathcal{B}$. This corresponds to $\nu(\mathcal{B})$ in Definition 3.2. To make a valid probability statement about the entire hierarchy we simply use the bottom-level probabilistic forecasts together with Definition 3.1.

Example: Gaussian Distributions

Suppose an unreconciled probabilistic forecast is Gaussian with mean $\hat{\boldsymbol{\mu}}$ and variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$. Let the unreconciled density be given by

$$f(\hat{\boldsymbol{y}}) = (2\pi)^{-n/2} |\hat{\boldsymbol{\Sigma}}|^{-1/2} \exp\left\{ -\frac{1}{2} \left[(\hat{\boldsymbol{y}} - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{y}} - \hat{\boldsymbol{\mu}}) \right] \right\}. \tag{17}$$

In an alternative basis,

$$f(\tilde{\boldsymbol{b}}, \tilde{\boldsymbol{a}}) = (2\pi)^{-\frac{n}{2}} \left| \hat{\boldsymbol{\Sigma}} \right|^{-\frac{1}{2}} \left| (\boldsymbol{S}\ \boldsymbol{R}) \right| \exp\{-\frac{1}{2} q\}, \tag{18}$$

where

$$q = (\boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}} - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}} - \hat{\boldsymbol{\mu}}). \tag{19}$$

The quadratic form q can be rearranged as

$$\begin{aligned} q &= \left((\boldsymbol{S}\ \boldsymbol{R}) \begin{pmatrix} \tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{a}} \end{pmatrix} - \hat{\boldsymbol{\mu}} \right)' \hat{\boldsymbol{\Sigma}}^{-1} \left((\boldsymbol{S}\ \boldsymbol{R}) \begin{pmatrix} \tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{a}} \end{pmatrix} - \hat{\boldsymbol{\mu}} \right), \\ &= \left(\begin{pmatrix} \tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{a}} \end{pmatrix} - (\boldsymbol{S}\ \boldsymbol{R})^{-1} \hat{\boldsymbol{\mu}} \right)' \left[(\boldsymbol{SR})^{-1} \hat{\boldsymbol{\Sigma}} \left((\boldsymbol{S}\ \boldsymbol{R})^{-1} \right)' \right]^{-1} \left(\begin{pmatrix} \tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{a}} \end{pmatrix} - (\boldsymbol{S}\ \boldsymbol{R})^{-1} \hat{\boldsymbol{\mu}} \right). \end{aligned}$$

Recall that

$$(\boldsymbol{S}\ \boldsymbol{R})^{-1} = \begin{pmatrix} (\boldsymbol{R}_\perp'\boldsymbol{S})^{-1}\boldsymbol{R}_\perp' \\ (\boldsymbol{S}_\perp'\boldsymbol{R})^{-1}\boldsymbol{S}_\perp' \end{pmatrix} := \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}.$$

Then q can be rearranged further as

$$\begin{aligned} q &= \left[\begin{pmatrix} \tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{a}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}\hat{\boldsymbol{\mu}}\right]' \left[\begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}\hat{\boldsymbol{\Sigma}}\begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}'\right]^{-1} \left[\begin{pmatrix} \tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{a}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}\hat{\boldsymbol{\mu}}\right] \\ &= \begin{pmatrix} \tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{a}} - \boldsymbol{H}\hat{\boldsymbol{\mu}} \end{pmatrix}' \left[\begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}\hat{\boldsymbol{\Sigma}}\begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}'\right]^{-1} \begin{pmatrix} \tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{a}} - \boldsymbol{H}\hat{\boldsymbol{\mu}} \end{pmatrix}. \end{aligned}$$

Similar manipulations on the determinant of the covariance matrix lead to the following expression for the density:

$$f(\tilde{\boldsymbol{b}}, \tilde{\boldsymbol{a}}) = (2\pi)^{-\frac{n}{2}} \left|\begin{pmatrix} \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' & \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}' \\ \boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' & \boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}' \end{pmatrix}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\begin{pmatrix} \tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{a}} - \boldsymbol{H}\hat{\boldsymbol{\mu}} \end{pmatrix}' \begin{pmatrix} \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' & \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}' \\ \boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' & \boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}' \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{a}} - \boldsymbol{H}\hat{\boldsymbol{\mu}} \end{pmatrix}\right\}.$$

Marginalising out $\tilde{\boldsymbol{a}}$ leads to the following bottom-level reconciled forecasts:

$$f(\tilde{\boldsymbol{b}}) = (2\pi)^{-\frac{m}{2}} \left|\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\right|^{\frac{1}{2}} \exp\Big\{-\frac{1}{2}(\tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}})'(\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}')^{-1}(\tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}})\Big\}. \tag{20}$$

This implies that the reconciled probabilistic forecast for the bottom-level series is $\tilde{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{G}\hat{\boldsymbol{\mu}}, \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}')$. The reconciled probabilistic forecasts for the whole hierarchy follow a degenerate Gaussian distribution with mean $\boldsymbol{SG}\hat{\boldsymbol{\mu}}$ and rank deficient covariance matrix $\boldsymbol{SG}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\boldsymbol{S}'$.

5.1 Elliptical distributions

We now show that the true predictive distribution can be recovered for elliptical distributions by linear reconciliation via pre-multiplication and translation respectively by a matrix we denote $\boldsymbol{G}_{opt}$ and vector we denote $\boldsymbol{d}_{opt}$. Here, for any square matrix $\boldsymbol{C}$, $\boldsymbol{C}^{1/2}$ and $\boldsymbol{C}^{-1/2}$ are defined to satisfy $\boldsymbol{C}^{1/2}(\boldsymbol{C}^{1/2})' = \boldsymbol{C}$ and $\boldsymbol{C}^{-1/2}(\boldsymbol{C}^{-1/2})' = \boldsymbol{C}^{-1}$, for example $\boldsymbol{C}^{1/2}$ may be obtained via the Cholesky or eigenvalue decompositions.

Theorem 5.1 (Reconciliation for Elliptical Distributions)**.** *Let an unreconciled probabilistic forecast come from the elliptical class with location parameter $\hat{\boldsymbol{\mu}}$ and scale matrix $\hat{\boldsymbol{\Sigma}}$. Let the true predictive distribution of $\boldsymbol{y}$ also belong to the elliptical class with location parameter $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$. Then the linear reconciliation mapping $g(\breve{\boldsymbol{y}}) = \boldsymbol{G}_{opt}\breve{\boldsymbol{y}} + \boldsymbol{d}_{opt}$ with $\boldsymbol{G}_{opt} = \boldsymbol{a}\hat{\boldsymbol{\Sigma}}^{-1/2}$ and $\boldsymbol{d}_{opt} = \boldsymbol{\mu} - \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}}$ recovers the true predictive density where $\boldsymbol{a}$ is any $m \times n$ matrix such that $\boldsymbol{a}\boldsymbol{a}' = \boldsymbol{\Omega}$ and $\boldsymbol{\Omega}$ is a sub-matrix of $\boldsymbol{\Sigma}$ corresponding to the bottom-level series.*

Proof. Since elliptical distributions are closed under affine transformations, and are closed under marginalisation, reconciliation of an elliptical distribution yields an elliptical distribution (although the unreconciled and reconciled distributions may be different members of the class of elliptical distributions). The scale matrix of the reconciled forecast is given by $\boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'_{opt}\boldsymbol{S}'$, while the location matrix is given by $\boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} + \boldsymbol{d}_{opt}$. The reconciled scale matrix is

$$\tilde{\boldsymbol{\Sigma}}_{opt} = \boldsymbol{S}\boldsymbol{a}\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}^{-1/2}\right)'\boldsymbol{a}'\boldsymbol{S}' = \boldsymbol{S}\boldsymbol{\Omega}\boldsymbol{S}' = \boldsymbol{\Sigma}.$$

For the choices of $\boldsymbol{G}_{opt}$ and $\boldsymbol{d}_{opt}$ given above, the reconciled location vector is

$$\tilde{\boldsymbol{\mu}}_{opt} = \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} + \boldsymbol{\mu} - \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}.$$

□

A number of insights can be drawn from this theorem. First, although a linear function $g(.)$ can be used to recover the true predictive in the elliptical case, the same does not hold in general. Second, $g(.)$ is not, in general, a projection matrix. The conditions for which a the true predictive density can be recovered by a projection are given below.

Theorem 5.2 (True predictive via projection)**.** *Assume that the true predictive distribution is elliptical with location $\boldsymbol{\mu}$ and scale $\boldsymbol{\Sigma}$. Consider reconciliation via a projection $g(\boldsymbol{y}) = (\boldsymbol{R}_\perp' \boldsymbol{S})^{-1} \boldsymbol{R}_\perp' \boldsymbol{y}$. The true predictive distribution can be recovered via reconciliation of an elliptical distribution with location $\hat{\boldsymbol{\mu}}$ and scale $\hat{\boldsymbol{\Sigma}}$ when the following conditions hold:*

$$sp(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \subset sp(\boldsymbol{R}) \tag{21}$$

$$sp(\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}^{1/2}) \subset sp(\boldsymbol{R}) \tag{22}$$

$$\tag{23}$$

Proof. The reconciled location vector will be given by

$$\begin{aligned}
\tilde{\boldsymbol{\mu}} &= \boldsymbol{S}(\boldsymbol{R}_\perp' \boldsymbol{S})^{-1} \boldsymbol{R}_\perp' \hat{\boldsymbol{\mu}} \\
&= \boldsymbol{S}(\boldsymbol{R}_\perp' \boldsymbol{S})^{-1} \boldsymbol{R}_\perp' \left(\hat{\boldsymbol{\mu}} + \boldsymbol{\mu} - \boldsymbol{\mu}\right) \\
&= \boldsymbol{S}(\boldsymbol{R}_\perp' \boldsymbol{S})^{-1} \boldsymbol{R}_\perp' \boldsymbol{\mu} + \boldsymbol{S}(\boldsymbol{R}_\perp' \boldsymbol{S})^{-1} \boldsymbol{R}_\perp' \left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right).
\end{aligned}$$

Since $\boldsymbol{S}(\boldsymbol{R}_\perp' \boldsymbol{S})^{-1} \boldsymbol{R}_\perp'$ is a projection onto $\mathfrak{s}$ and $\boldsymbol{\mu} \in \mathfrak{s}$, the first term simplifies to $\boldsymbol{\mu}$. If $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}$ lies in the span of $\boldsymbol{R}$, then multiplication by $\boldsymbol{R}_\perp'$ reduces the second term to $\mathbf{0}$. By a similar argument it can be shown that $\tilde{\boldsymbol{\Sigma}}^{1/2} = \boldsymbol{\Sigma}^{1/2}$. The closure property of elliptical distributions under affine transformations ensures that the full true predictive distribution can be recovered. □

Although these conditions will rarely hold in practice and only apply to a limited class of distributions, they do provide some insight into selecting a projection for reconciliation. If

the value of $\hat{\boldsymbol{\mu}}$ were equi-probable in all directions, then a projection orthogonal to $\mathfrak{s}$ would be a sensible choice for $\boldsymbol{R}$ since it would in some sense represent a 'median' direction for $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}$. However, the one-step-ahead in-sample errors are usually correlated suggesting that $\hat{\boldsymbol{\mu}}$ is more likely to fall in some directions than others. Therefore an orthogonal projection after transformation by the inverse of the one-step-ahead in-sample error covariance matrix may be more intuitively appealing. This is exactly what the MinT projection provides, and we demonstrate this in a simulation setting in the following subsection.

5.2 Simulations

To compare different reconciliation methods in parametric densities we assume a Gaussian predictive distribution for the hierarchy. We choose the Gaussian case due to its analytical tractability which allows for evaluation using all scoring rules (including the log score).

5.2.1 Data generating process (DGP)

The data generating process, we consider is the hierarchy given in Figure 1, comprising two aggregation levels with four bottom-level series. Each bottom-level series will be generated first, and then summed to obtain the data for the upper-level series.

First $\{w_{AA,t}, w_{AB,t}, w_{BA,t}, w_{BB,t}\}$ are generated from ARIMA(p, d, q) processes, where (p, q) and d take integers from $\{1, 2\}$ and $\{0, 1\}$ respectively with equal probability. The parameters for the AR and MA components are randomly and uniformly generated from $[0.3, 0.5]$ and $[0.3, 0.7]$ respectively. The errors driving the ARIMA processes were generated from Gaussian and non-Gaussian distributions separately. This will allow us to demonstrate impact of true DGP for the parametric reconciliation approach.

Gaussian errors:

Errors were jointly generated from a normal distribution, and denoted by $\{\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}\} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})\ \forall t$, where,

$$\mathbf{\Sigma} = \begin{pmatrix} 5.0 & 3.1 & 0.6 & 0.4 \\ 3.1 & 4.0 & 0.9 & 1.4 \\ 0.6 & 0.9 & 2.0 & 1.8 \\ 0.4 & 1.4 & 1.8 & 3.0 \end{pmatrix}. \tag{24}$$

Non-Gaussian errors:

Non-Gaussian errors were generated from a Gumbel copula model with beta margins. Using a copula model helps to impose a non-linear dependence structure among the series. A two dimensional Gumbel copula is given by,

$$C_\theta(u_1, u_2) = exp\{-[(-ln(u_1))^\theta + (-ln(u_2))^\theta]^{1/\theta}\}.$$

We generate random variates $\{u_{AA}, u_{AB}\}$ from $C_{\theta=10}(.)$ and $\{u_{BA}, u_{BB}\}$ from $C_{\theta=8}(.)$ for series $\{AA, AB\}$ and $\{BA, BB\}$ respectively. Next we generate the errors, $\{\varepsilon_{AA}, \varepsilon_{AB}, \varepsilon_{BA}, \varepsilon_{BB}\}$ as the quantiles from beta distributions with shape parameters $\alpha = 1$ and $\beta = 3$ correspond to $\{u_{AA}, u_{AB}, u_{BA}, u_{BB}\}$.

Although copulas go beyond the concepts of linear dependence, we will apply a Gaussian assumption to the data to investigate model misspecification. To give some idea of the covariance matrix of $\{\varepsilon_{AA}, \varepsilon_{AB}, \varepsilon_{BA}, \varepsilon_{BB}\}$, the sample estimate is,

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.0388 & 0.0385 & 0.0010 & 0.0010 \\ 0.0385 & 0.0390 & 0.0008 & 0.0008 \\ 0.0010 & 0.0008 & 0.0387 & 0.0377 \\ 0.0010 & 0.0008 & 0.0377 & 0.0381 \end{pmatrix}. \tag{25}$$

Signal-to-noise ratio:

In practice, hierarchical time series are likely to have relatively noisier series at lower levels of aggregation. Following the method proposed by Wickramasuriya et al. (2019), we replicate this feature in our simulations by generating the bottom-level series $\{y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}\}$ as follows:

$$\begin{aligned} y_{AA,t} &= w_{AA,t} + u_t - 0.5v_t, \\ y_{AB,t} &= w_{AB,t} - u_t - 0.5v_t, \\ y_{BA,t} &= w_{BA,t} + u_t + 0.5v_t, \\ y_{BB,t} &= w_{BB,t} - u_t + 0.5v_t, \end{aligned}$$

where $u_t \sim \mathcal{N}(0, \sigma_u^2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$. The aggregate series in the middle-level are given by:

$$\begin{aligned} y_{A,t} &= w_{AA,t} + w_{AB,t} - v_t, \\ y_{B,t} &= w_{BA,t} + w_{BB,t} + v_t, \end{aligned}$$

and the total series is given by

$$y_{Tot,t} = w_{AA,t} + w_{AB,t} + w_{BA,t} + w_{BB,t}.$$

To ensure the disaggregate series are noisier than the aggregate series, we choose σ_u^2 and σ_v^2 such that

$$\text{Var}(\varepsilon_{AA,t} + \varepsilon_{AB,t} + \varepsilon_{BA,t} + \varepsilon_{BB,t}) \leq \text{Var}(\varepsilon_{AA,t} + \varepsilon_{AB,t} - v_t) \leq \text{Var}(\varepsilon_{AA,t} + u_t - 0.5v_t).$$

Similar inequalities hold when $\varepsilon_{AA,t}$ is replaced by $\varepsilon_{AB,t}$, $\varepsilon_{BA,t}$ and $\varepsilon_{BB,t}$ in the third term.

Thus for the Gaussian DGP we choose $\sigma_u^2 = 24$ and $\sigma_u^2 = 18$ whereas for non-Gaussian DGP we choose $\sigma_u^2 = 10$ and $\sigma_v^2 = 7$.

We generate 2000 observations for each series from this Gaussian and non-Gaussian DGP. We ignore the first 500 observations from each series to avoid the impact from initial values. Using a rolling window of $T = 500$ observations, we fit univariate ARIMA models for each series using the `auto.arima()` function in the `forecast` package (Hyndman 2019) in R (R Core Team 2018). Using the fitted models we generate 1 to 3 steps ahead base (incoherent) Gaussian probabilistic forecasts. We estimate the mean and the variance of this incoherent Gaussian density as the h-steps ahead point forecasts $\hat{\boldsymbol{y}}_{t+h}$ and shrinkage estimator for variance covariance matrix of one-step ahead forecast errors $\hat{\boldsymbol{W}}^{\text{shr}}$ respectively. These were then reconciled using different projections summarised in Table 1. This process was replicated for 1000 times by rolling the window one step at a time.

To assess the predictive performance of different forecasting methods, we use scoring rules as discussed in Section 4. To facilitate comparisons, we report skill scores (Gneiting & Raftery 2007). For a given forecasting method, evaluated by a particular scoring rule, the skill score gives the percentage improvement of the preferred forecasting method relative to a reference method. A negative valued skill score indicates that a method is worse than the reference method, whereas any positive value indicates that the method is superior to the reference method.

Table 3 summarises the forecasting performance of incoherent, bottom-up, OLS, WLS and two MinT reconciliation methods using log score, energy score and variogram score.

The top panel refers to the Gaussian DGP whereas the bottom panel refers to the non-Gaussian DGP. Recall that the log score is improper with respect to incoherent forecasts. Therefore we calculate the skill scores with reference to the bottom-up forecasts instead of incoherent forecasts in all cases and leave blank the cell for log score of the incoherent forecasts. Further, all log scores are evaluated on the basis of bottom-level series only, however these only differ from the log scores for the full hierarchy by a fixed constant. Overall, the MinT methods provide the best performance irrespective of the scoring rule, and all methods that reconcile using information at all levels of the forecast improve upon incoherent forecasts. Bottom-up forecasts perform even worse than incoherent forecasts in some cases. These results hold for both the Gaussian as well as the non-Gaussian DGP.

Tables 4 and 5 break down the forecasting performance of the different methods by considering univariate scores on each individual margin. Table 4 summarises the results for the top and middle levels, Table 4 does the same for bottom-level. Univariate log score and CRPS are considered, while skill scores are computed with the incoherent forecasts as a reference. When broken down in this fashion, irrespective of DGP, the methods based on MinT perform best for most series and outperform bottom-up forecasts in almost all cases.

Table 3: Comparison of coherent forecasts in forecast for $h = 1$ to 3 steps-ahead. All entries shows the percentage skill score with reference to the bottom-up method. The top panel shows results from the Gaussian DGP and bottom panel shows the results from the non-Gaussian DGP. "ES" and "VS" columns give scores based on the joint forecast distribution across the entire hierarchy. The "LS" column gives the log scores of the joint forecast distribution of the bottom-level.

	h=1			h=2			h=3		
Method	ES(%)	VS(%)	LS(%)	ES(%)	VS(%)	LS(%)	ES(%)	VS(%)	LS(%)
Gaussian DGP									
MinT(Shrink)	**19.48**	**9.78**	**3.16**	**19.57**	**14.16**	**6.53**	**16.47**	**16.56**	**8.34**
MinT(Sample)	**19.48**	9.74	3.09	19.50	14.16	6.51	16.28	16.42	8.09
WLS	18.08	7.21	0.64	17.68	10.97	2.31	14.99	13.17	3.76
OLS	16.01	5.80	-0.79	15.38	8.43	0.05	13.03	10.26	0.82
Bottom up	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Incoherent	11.65	-0.12		10.58	1.71		8.75	3.64	
Non-Gaussian DGP									
MinT(Shrink)	**15.04**	**0.69**	**4.52**	**16.98**	**1.34**	**4.55**	**18.00**	**0.66**	**4.01**
MinT(Sample)	15.02	0.59	4.40	16.94	1.02	4.30	17.88	0.64	3.42
WLS	12.72	0.00	0.93	14.22	0.41	1.34	15.20	-0.42	0.89
OLS	11.26	0.17	0.65	12.27	0.48	0.47	13.12	-0.24	0.10
Bottom up	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Incoherent	8.47	-2.79		8.94	-2.09		9.20	-3.62	

Table 4: Comparison of incoherent vs coherent forecasts based on the univariate forecast distribution of the aggregate series. Each entry represents the percentage skill score with reference to the incoherent forecasts based on "CRPS" and "LS". These entries show the percentage increase in score for different forecasting methods relative to the incoherent forecasts for $h = 1$ to 3 steps-ahead forecast. Results from the Gaussian DGP are presented in the top panel whereas the results from the non-Gaussian DGP are presented in the bottom panel

	h=1						h=2						h=3					
	Total		A		B		Total		A		B		Total		A		B	
R.method	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS
Gaussian DGP																		
MinT(Shrink)	-0.13	-0.01	**9.37**	3.12	5.42	1.67	**0.34**	-0.08	10.67	3.32	5.79	1.59	0.08	-0.17	8.13	1.58	4.17	1.04
MinT(Sample)	-0.08	-0.04	**9.37**	**3.13**	5.24	1.67	0.27	-0.10	**10.76**	**3.39**	5.67	1.62	0.04	-0.23	**8.14**	**1.69**	4.19	1.10
WLS	-2.91	-1.24	8.78	2.86	**5.49**	**1.73**	-0.41	1.02	9.99	2.97	**6.01**	**1.73**	**0.10**	2.97	7.72	1.57	**4.46**	1.26
OLS	-19.22	-6.86	6.28	2.06	4.86	1.58	-5.99	**4.05**	7.17	2.03	5.83	1.70	-2.13	13.11	5.30	0.97	4.41	**1.30**
Bottom up	-140.27	-33.67	-13.75	-3.89	-11.10	-3.17	-60.07	2.72	-13.82	-3.86	-9.04	-2.37	-30.95	**30.34**	-13.57	-3.37	-8.47	-1.87
Incoherent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Non-Gaussian DGP																		
MinT(Shrink)	-1.16	-0.26	**0.92**	0.27	**11.90**	4.91	-0.70	-1.16	**0.71**	0.29	**16.53**	6.83	-0.30	-1.35	**0.60**	0.25	**19.53**	8.28
MinT(Sample)	-1.16	-0.72	**0.92**	**0.28**	**11.90**	**4.94**	-0.70	-1.53	**0.71**	**0.31**	**16.53**	**6.87**	-0.30	-1.61	**0.60**	**0.31**	**19.53**	**8.35**
WLS	**0.01**	**0.35**	-1.02	-0.52	9.95	4.07	-0.11	-0.15	-2.50	-1.09	13.87	5.55	-0.02	-0.40	-3.96	-1.60	16.14	6.78
OLS	-96.77	-84.90	0.55	0.08	6.48	2.57	-44.06	-14.27	-0.18	-0.12	8.72	3.44	-22.75	**19.51**	-0.71	-0.32	10.48	4.27
Bottom up	-541.40	-246.37	-4.60	-1.87	-8.99	-3.11	-273.80	-68.47	-4.20	-1.67	-8.78	-2.84	-159.47	10.59	-4.71	-1.77	-8.06	-2.27
Incoherent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5: Comparison of incoherent vs coherent forecasts based univariate forecast distribution of bottom-level series. Each entry represents the skill score with reference to Incoherent forecasts based on "CRPS" and "LS"

	h=1								h=2								h=3							
	AA		AB		BA		BB		AA		AB		BA		BB		AA		AB		BA		BB	
R.method	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS
Gaussian DGP																								
MinT(Shrink)	3.91	1.30	**12.04**	**3.82**	**10.07**	**3.12**	1.47	0.47	3.80	1.28	**19.37**	**6.55**	**13.61**	**4.44**	-0.08	-0.14	**2.54**	**0.83**	**19.58**	**6.92**	**13.94**	**4.83**	-2.64	-0.96
MinT(Sample)	**4.12**	**1.38**	11.99	**3.82**	9.90	3.10	**1.57**	**0.51**	**4.06**	**1.33**	19.23	6.53	13.21	4.34	-0.24	-0.19	2.24	0.68	19.33	6.73	13.44	4.62	-3.44	-1.21
WLS	1.10	0.42	10.37	3.21	9.12	2.78	-1.14	-0.26	-0.55	-0.09	15.52	5.05	12.17	3.91	-3.48	-1.25	-1.52	-0.46	15.75	5.32	12.63	4.34	-5.74	-2.04
OLS	0.80	0.25	8.47	2.59	7.91	2.39	-1.52	-0.49	-1.23	-0.30	12.09	3.83	10.19	3.25	-4.42	-1.57	-2.04	-0.60	11.96	3.93	10.44	3.45	-6.72	-2.36
Bottom up	0.01	0.00	0.04	0.00	0.15	0.00	-0.09	0.00	-0.01	0.00	0.17	-0.01	0.12	0.00	**0.17**	0.00	-0.23	0.00	-0.01	-0.02	0.01	-0.01	-0.13	0.00
Base	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Non-Gaussian DGP																								
MinT(Shrink)	**3.40**	**1.31**	-0.11	-0.11	**13.22**	**5.00**	**2.37**	0.87	**3.67**	**1.30**	-0.11	-0.18	**16.19**	**6.12**	**2.29**	**0.80**	3.38	1.22	-1.21	-0.47	**17.90**	**6.92**	**2.09**	**0.76**
MinT(Sample)	**3.40**	1.29	-0.11	-0.13	**13.22**	4.99	**2.37**	**0.91**	**3.67**	1.25	-0.11	-0.24	**16.19**	6.01	**2.29**	0.76	3.38	1.16	-1.21	-0.58	**17.90**	6.71	**2.09**	0.64
WLS	2.92	1.20	-1.90	-0.74	8.50	3.13	-0.96	-0.26	3.34	1.24	-2.51	-1.00	10.79	3.96	-1.27	-0.38	**3.47**	**1.23**	-3.11	-1.12	12.60	4.78	-1.21	-0.41
OLS	2.70	1.07	-1.46	-0.55	6.19	2.25	-0.81	-0.21	2.95	1.15	-1.85	-0.73	7.61	2.74	-1.19	-0.32	3.22	1.18	-2.13	-0.76	8.85	3.25	-1.13	-0.35
Bottom up	-0.10	0.00	-0.12	0.00	-0.02	0.00	-0.08	0.00	-0.17	0.00	0.09	0.00	0.03	-0.01	-0.10	0.00	-0.20	0.00	-0.03	0.00	-0.05	-0.01	-0.22	0.00
Base	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6 A novel non-parametric bootstrap approach

Often in practice we come across hierarchical time series that have high level of disaggregation and/or contain even discrete data. For these time series, parametric distributional assumptions are misleading. An alternative for such cases is to apply non-parametric approaches. Hence we propose a novel non-parametric bootstrap based approach for obtaining coherent probabilistic forecasts.

Our proposed method initially involves obtaining probabilistic forecasts without considering the aggregation constraints. These incoherent probabilistic forecasts are then reconciled to make them coherent. We first focus on the methodology for obtaining base forecasts.

6.1 Incoherent probabilistic forecasts

First we fit appropriate univariate models for each series in the hierarchy based on the training data $\boldsymbol{y}_{1:T}$. We then compute 1-step-ahead training errors as $e_{i,t} = y_{i,t} - \hat{y}_{i,t}$ for $i = 1, ..., n$ and $t = 1, ..., T$ where $\hat{y}_{i,t} = E(y_{i,t}|y_{i,1:t-1})$. The training errors are stored in a matrix $\boldsymbol{\Gamma}_{(T\times n)} = (\boldsymbol{e}_1, ..., \boldsymbol{e}_T)'$ where $\boldsymbol{e}_t = (e_{1,t}, ..., e_{n,t})$ is stored in the same order as $\boldsymbol{y}_t$ for $t = 1, ..., T$. Next we block bootstrap a sample of size H from $\boldsymbol{\Gamma_{(T\times n)}}$. That is, we randomly select H consecutive rows from $\boldsymbol{\Gamma}$ and store in a matrix $\boldsymbol{\Gamma}^b_{(H\times n)} = (\boldsymbol{e}^b_1, ..., \boldsymbol{e}^b_H)'$ and repeat this for $b = 1, ..., B$.

Finally we generate the h-step-ahead future paths using the fitted univariate models conditioning on the past observations. We also incorporate the bootstrapped training errors as the error series for generating these future paths. By doing so we implicitly model the contemporaneous correlation structure of the hierarchy. Further the use of consecutive (block) training errors will ensure that the serial correlation of the series is accounted for. To explain this process more explicitly consider the following example.

Example: Suppose we fit an $ARMA(p, q)$ model for the i^{th} series of the hierarchy. i.e.,

$$y_{i,t} = \alpha_1 y_{i,t-1} + \alpha_2 y_{i,t-2} + ... + \alpha_p y_{i,t-p} + \beta_1 \epsilon_{i,t-1} + \beta_1 \epsilon_{i,t-2} + ... + \beta_q \epsilon_{i,t-q} + \epsilon_{i,t},$$
$$y_{i,t} = (\alpha_1 + \alpha_2 L + ... + \alpha_p L^{p-1}) y_{i,t-1} + (\beta_1 + \beta_1 L + ... + \beta_q L^{q-1}) \epsilon_{i,t-1} + \epsilon_{i,t}$$

where L is the usual lag operator. Then the h-step-ahead b^{th} future path conditional on past information upto and including time t, for the i^{th} series is produced as,

$$\hat{y}^b_{i,t+h} = (\hat{\alpha}_1 + \hat{\alpha}_2 L + ... + \hat{\alpha}_p L^{p-1}) y_{i,t+h-1} + (\hat{\beta}_1 + \hat{\beta}_1 L + ... + \hat{\beta}_q L^{q-1}) \epsilon_{i,t+h-1} + e^b_{i,h}$$

where, $e^b_{i,h}$ is the $(h \times i)^{\text{th}}$ element from $\mathbf{\Gamma}^b$,

$$y_{i,t+h-1} = \begin{cases} y_{i,1} : y_{i,T} & \text{for } t+h-1 \leq T \\ \hat{y}^b_{i,T+1} : \hat{y}^b_{i,T+h-1} & \text{for } t+h-1 > T \end{cases}$$

and

$$\epsilon_{i,t+h-1} = \begin{cases} \epsilon_{i,1} : \epsilon_{i,T} & \text{for } t+h-1 \leq T \\ e^b_{i,1} : e^b_{i,h-1} & \text{for } t+h-1 > T \end{cases}.$$

Once we obtain the h-step-ahead sample path for all n series in the hierarchy, we stack them in the same order as $\hat{\boldsymbol{y}}_{t+h}$. Repeating the same process for $b = 1, ..., B$ we obtain a set of h-step-ahead bootstrapped future paths of size B. We denote this as $\hat{\mathbf{\Upsilon}}_{T+h} = (\hat{\boldsymbol{y}}^1_{T+h}, ..., \hat{\boldsymbol{y}}^B_{T+h})'$ where the b^{th} row of $\hat{\mathbf{\Upsilon}}_{T+h}$ represents the h-step-ahead b^{th} sample path for all series in the hierarchy.

We note that $\hat{\mathbf{\Upsilon}}_{T+h}$ is an empirical sample from the incoherent probability distribution of the hierarchy. Since the aggregation constraints are not imposed while generating $\hat{\mathbf{\Upsilon}}_{T+h}$, it is very unlikely that they lie on the coherent subspace. Thus it requires reconciliation to which we now turn our attention.

6.2 Reconciliation of incoherent future paths

To reconcile the incoherent sample paths, we follow the definition of reconciliation. We project each sample path in $\hat{\mathbf{\Upsilon}}_{T+h}$ to the coherent subspace via the projection $\boldsymbol{SG}$. i.e. for any $\boldsymbol{G}$ we can write,

$$\tilde{\boldsymbol{y}}^b_{T+h} = \boldsymbol{SG}\hat{\boldsymbol{y}}^b_{T+h}, \tag{26}$$

consequently we have,

$$\tilde{\boldsymbol{\Upsilon}}'_{T+h} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\Upsilon}}'_{T+h}, \tag{27}$$

where, each row in $\tilde{\boldsymbol{\Upsilon}}_{T+h}$ represent a single reconciled sample path. Further $\tilde{\boldsymbol{\Upsilon}}_{T+h}$ form an empirical sample from the reconciled forecast distribution of the hierarchy. Any $\boldsymbol{G}$ matrix introduced in point forecast reconciliation (also given in Table 1) can be used for this sample path reconciliation. However, in the following subsection we discuss a method to find $\boldsymbol{G}$ that is optimal for probabilistic forecasts with respect to a proper scoring rule.

6.3 Optimal reconciliation of incoherent future paths

Let us now propose to find an optimal $\boldsymbol{G}$ for reconciling future paths by minimising a proper multivariate scoring rule. The respective objective function can be written as,

$$\underset{\boldsymbol{G}_h}{\operatorname{argmin}} \quad \mathrm{E}_{\boldsymbol{Q}}[S(\boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h}, \boldsymbol{y}_{T+h})], \tag{28}$$

where S is a proper scoring rule that follows equation (5). We use the subscript h on $\boldsymbol{G}$ to emphasis distinct $\boldsymbol{G}$ matrices for different forecast horizons. Recall that the energy score given in equation (8) is a proper scoring rule. Let $\alpha = 1$ and following equation (9) we can write,

$$\mathrm{ES}(\boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h}, \boldsymbol{y}_{T+h}) \approx \frac{1}{B}\sum_{b=1}^{B}||\boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{y}}^{b}_{T+h,j} - \boldsymbol{y}_{T+h}|| - \frac{1}{2(B-1)}\sum_{b=1}^{B-1}||\boldsymbol{S}\boldsymbol{G}_h(\hat{\boldsymbol{y}}^{b}_{T+h,j} - \hat{\boldsymbol{y}}^{b+1}_{T+h,j})||. \tag{29}$$

where B is the empirical sample size from the coherent forecast distribution. Now we can rewrite the objective function in (28) as,

$$\underset{\boldsymbol{G}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N}\left\{\frac{1}{B}\sum_{b=1}^{B}||\boldsymbol{S}\boldsymbol{G}_h\boldsymbol{y}^{b}_{T+h,j}-\boldsymbol{y}_{T+h,j}||-\frac{1}{2(B-1)}\sum_{b=1}^{B-1}||\boldsymbol{S}\boldsymbol{G}_h(\boldsymbol{y}^{b}_{T+h,j}-\boldsymbol{y}^{b+1}_{T+h,j})||\right\} \tag{30}$$

where, the expectation $E_{\boldsymbol{Q}}$ over true forecast distribution $\boldsymbol{Q}$ is approximated through the sample mean over $\{\text{ES}(\boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h,1},\boldsymbol{y}_{T+h,1}),...,\text{ES}(\boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h,N},\boldsymbol{y}_{T+h,N})\}$. We can use numerical optimization methods to estimate the matrix $\boldsymbol{G}_h$ that minimises the above objective function and thus obtain the optimally reconciled future paths.

6.3.1 Reparameterisation of $\boldsymbol{G}$

We consider different reparameterisations when estimating the optimal $\boldsymbol{G}_h$ via the proposed optimisation process. Let,

$$\boldsymbol{G}_h=(\boldsymbol{S}'\boldsymbol{W}_h\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h. \tag{31}$$

This structure for $\boldsymbol{G}_h$ will ensure $\boldsymbol{S}\boldsymbol{G}_h$ is a projection matrix and it projects each sample path onto $\mathfrak{s}$.

Method 1 Minimising the objective function in (30) over symmetric $\boldsymbol{W}_h$. This solves an unconstrained optimisation problem

Method 2 Consider the Cholesky decomposition of $\boldsymbol{W}_h$. i.e. let $\boldsymbol{W}_h=\boldsymbol{U}'_h\boldsymbol{U}_h$ where $\boldsymbol{U}_h$ is an upper triangular matrix. Thus minimising (30) over $\boldsymbol{U}_h$

Method 3 Similar to method 2, minimising (30) over the Cholesky decomposition of $\boldsymbol{W}_h$, but imposing restrictions for scaling. i.e., $\boldsymbol{W}_h=\boldsymbol{U}'_h\boldsymbol{U}_h$ s.t $\boldsymbol{i}'\boldsymbol{W}_h\boldsymbol{i}=1$ where $\boldsymbol{i}=(1,0,..,0)'$

Method 4 Minimising (30) over $\boldsymbol{G}_h$ such that $\boldsymbol{G}_h\boldsymbol{S}=\boldsymbol{I}$. This constraint is an alternative way to ensure that $\boldsymbol{S}\boldsymbol{G}_h$ is a projection onto $\mathfrak{s}$

6.4 Simulation study

We now compare different reconciliation methods with optimal reconciliation in a simulation setting. We use the same Gaussian and non-Gaussian DGP explained in Subsection 5.2.1 corresponding to the hierarchy in Figure 1.

6.4.1 Simulation set up for optimal reconciliation

1. Generate time series with 2500 data points for each series in the hierarchy.
2. Consider a rolling window of 600 observations. We call this the "outer" rolling window.
    i. Inside this outer rolling window consider an inner rolling window of $T = 500$ observations.
    ii. For this inner rolling window, fit univariate ARIMA models to each series in the hierarchy.
    iii. Based on these fitted models, generate $B = 1000$ of $h = 1$ to 3 steps-ahead incoherent future paths incorporating bootstrap errors as described in Subsection 6.1. Thus we get $\{\hat{\boldsymbol{\Upsilon}}_{T+1,j=1}, \hat{\boldsymbol{\Upsilon}}_{T+2,j=1}, \hat{\boldsymbol{\Upsilon}}_{T+3,j=1}\}$.
    iv. Repeat step (iii) for $j = 1, ..., N$ where $N = 100$ by rolling the inner window one step ahead at a time.
    v. Collect $\{\hat{\boldsymbol{\Upsilon}}_{T+h,j=1}, ..., \hat{\boldsymbol{\Upsilon}}_{T+h,j=100}\}$ for $h = 1, ..., 3$ into separate arrays of matrices.
    vi. For each forecast horizon h, estimate the optimal $\boldsymbol{G}_h$ that will reconcile $\{\hat{\boldsymbol{\Upsilon}}_{T+h,j=1}, ..., \hat{\boldsymbol{\Upsilon}}_{T+h,j=100}\}$ by minimising the average energy score as explained in Subsection 6.3. Also use the different reparameterisations of $\boldsymbol{G}_h$ as explained in Subsection 6.3.1. Denote this as $\boldsymbol{G}_h^{Opt}$.

vii. Roll the inner rolling window another one step ahead and repeat steps (ii) and (iii). Denote these future paths by $\hat{\boldsymbol{\Upsilon}}_{T+h}$ for $h = 1, 2, 3$.

viii. Compute $\tilde{\boldsymbol{\Upsilon}}'_{T+h} = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h}$ for $h = 1, 2, 3$ using $\boldsymbol{G}_h^{Opt}$ as well as using other $\boldsymbol{G}$ matrices given in Table 1.

3. Repeat Step 2 1000 times by rolling the outer rolling window one step- ahead at a time. Collect 1000 reconciled future paths, $\tilde{\boldsymbol{\Upsilon}}_{T+h}$, from different reconciliation methods for $h = 1, 2, 3$ and evaluate the forecasting performances.

6.4.2 Evaluation

Following the simulation process, we generate reconciled non-parametric probabilistic forecasts separately for Gaussian data and non-Gaussian data. To assess their predictive performance we use energy and variogram scores as discussed in Section 4. Results are presented in Table 6.

We see that the scores for reconciled forecasts for different optimal methods are equivalent irrespective to the forecast horizon or the DGP. This implies that there is no difference in results due to different reparameterisations of $\boldsymbol{G}$. Further, Mann-Whitney tests for location comparison support that the ES and VS for all reconciled forecasts are significantly lower than those of incoherent forecasts. This implies that all reconciliation methods produce coherent probabilistic forecasts with improved predictive ability compared to the incoherent forecasts. In addition to that, the MinT(Shrink) and Optimal methods have similar prediction accuracy as there is no significant difference between the scores from these reconciliation methods. These results are consistent for both Gaussian and non-Gaussian data.

However we note that optimal reconciliation required a high computational cost for

larger hierarchies. Further, it requires sufficient data points to learn the $\boldsymbol{G}$ matrix. Thus we suggest using the MinT $\boldsymbol{G}$ for reconciling bootstrapped future paths for two reasons. Firstly it is computationally efficient relative to the optimal method and secondly, it produces accurate probabilistic forecasts that are at least as good as the Optimal method with respect to the energy score.

Table 6: Energy scores (ES) and variogram scores (VS) for probabilistic forecasts from different reconciliation methods are presented. Bottom row represent the scores for base forecasts which are not coherent. The smaller the scores, the better the forecasts are.

	Non-Gaussian DGP						Gaussian DGP					
Reconciliation	h=1		h=2		h=3		h=1		h=2		h=3	
method	ES	VS	ES	VS	ES	VS	ES	VS	ES	VS	ES	VS
Optimal(Method-1)*	5.36	1.21	5.51	1.27	5.83	1.38	9.59	4.86	11.50	5.38	13.80	6.13
Optimal(Method-2)*	5.37	1.21	5.53	1.27	5.83	1.37	9.58	4.85	11.50	5.37	13.80	6.14
Optimal(Method-3)*	5.37	1.21	5.53	1.27	5.83	1.37	9.58	4.85	11.50	5.37	13.80	6.14
Optimal(Method-4)*	5.38	1.21	5.54	1.27	5.83	1.38	9.58	4.85	11.50	5.37	13.80	6.14
MinT(Shrink)*	5.33	1.19	5.50	1.26	5.77	1.34	9.43	4.78	11.40	5.33	13.70	6.09
WLS	5.43	1.23	5.60	1.30	5.89	1.40	9.64	4.93	11.70	5.60	14.10	6.39
OLS	5.51	1.23	5.70	1.30	5.98	1.40	9.91	4.93	12.10	5.60	14.50	6.39
Incoherent	*5.71*	*1.28*	*5.94*	*1.37*	*6.27*	*1.49*	*10.40*	*5.31*	*12.70*	*6.22*	*15.20*	*7.14*

The differences in scores between methods noted by "" are statistically insignificant. The differences between these and the incoherent forecasts are statistically significant.*

7 Application: Forecasting Australian domestic tourism flow

In this section we illustrate how the probabilistic forecast reconciliation methods can be used in practice, by forecasting domestic tourism flows in Australia. Previous studies have shown that reconciliation for this data generate more accurate point forecasts compared to the bottom-up or incoherent forecasts. For example see Athanasopoulos et al. (2009), Hyndman et al. (2011) and Wickramasuriya et al. (2019). This study is the first to apply reconciliation methods for forecasting tourism in a probabilistic framework.

7.1 Data

As a measure of domestic tourism flows, we consider the "overnight trips" to different destinations across the country. Data are collected through the National Visitor Survey (NVS) managed by Tourism Research Australia based on an annual sample of $120,000$ Australian residents aged 15 years or more, through telephone interviews (Tourism Research Australia 2019).

The total number of overnight trips in Australia can be naturally disaggregated through a geographical hierarchy. This hierarchy consists of 7 states [1] in the 1st level of disaggregation, 27 zones in the 2nd level of disaggregation and 76 regions in the bottom-level and thus comprises 110 series in total. More details about the individual series are provided in Table **??**. We consider monthly overnight trips for all series spanning the period January 1998 to December 2018. This gives 152 observations per series.

[1] We have considered ACT as a part of New South Wales and Northern Territory as a state.

7.2 Forecasting methodology

We apply both the parametric and non-parametric reconciliation approaches as discussed in previous sections. We use a rolling window of 100 observations as the training sample where the first training sample will span the period Jan-1998 to Apr-2006. Based on this training set we fit univariate ARIMA and ETS models for each series in the hierarchy using automated functions `auto.arima()` and `ets()` from the `forecast` package (Hyndman 2019) in R software (R Core Team 2018). From the estimated models we generate parametric and non-parametric probabilistic forecasts for one year ahead, i.e for $h = 1, ..., 12$. For the parametric forecasts, we assume Gaussian densities and obtain the incoherent mean and variance forecasts. These are then reconciled using the methods described in Section 5. For the non-parametric forecasts, we generate the bootstrapped future paths and then reconcile each sample path as described in Section 6. We note that we do not implement the MinT(Sample) approach as the sample size of training data set is less than the dimension of the hierarchy. Using a rolling window, one month at a time, we replicate the process until the end of the sample. This yields, 152 1-step ahead, 151 2-steps ahead through to 141 12-step ahead probabilistic forecasts available for evaluation. We note that we only present the results for ARIMA models in the following section. The results for ETS models are similar and we present these in the Appendix.

7.3 Evaluation, results and discussion

We evaluate the predictive accuracy using scoring rules. More specifically we use energy and variogram scores to assess the predictive accuracy of multivariate forecast distributions across the entire hierarchy as well as for the different disaggregation levels. CRPS is used to assess the predictive accuracy of univariate forecast distributions for each series in the

hierarchy. We calculate average scores over the replications for each forecast horizon separately. In the results that follow we present skill scores for each of the coherent predictive distributions with reference to the incoherent distributions. A positive (negative) values in the skill score indicates a gain (loss) in forecast accuracy over the incoherent forecast distribution.

Figure 4 shows the skill scores with respect to the multivariate predictive distributions across the entire hierarchy from the different methods. Figure 5 shows the evaluation across each level. The top panels present the results from the Gaussian approach while the bottom panels present the results from the non-parametric approach. Both figures show that almost all reconciliation methods improve forecast accuracy irrespective of whether the parametric or non-parametric approaches are implemented. Furthermore, the bottom-up approach shows losses compared to the incoherent forecasts at all forecast horizons. This reflects the fact that bottom-level series are noisier and therefore more challenging to forecast. Finally and most importantly, MinT(Shrink) outperforms all probabilistic forecast reconciliation methods for both parametric and non-parametric approaches.

Figure 6 shows the predictive accuracy of the univariate forecast distributions for the Total overnight trips. OLS and MinT(Shrink) reconciliation methods show gains in accuracy for the top level of the hierarchy for both Gaussian and non-parametric approaches.

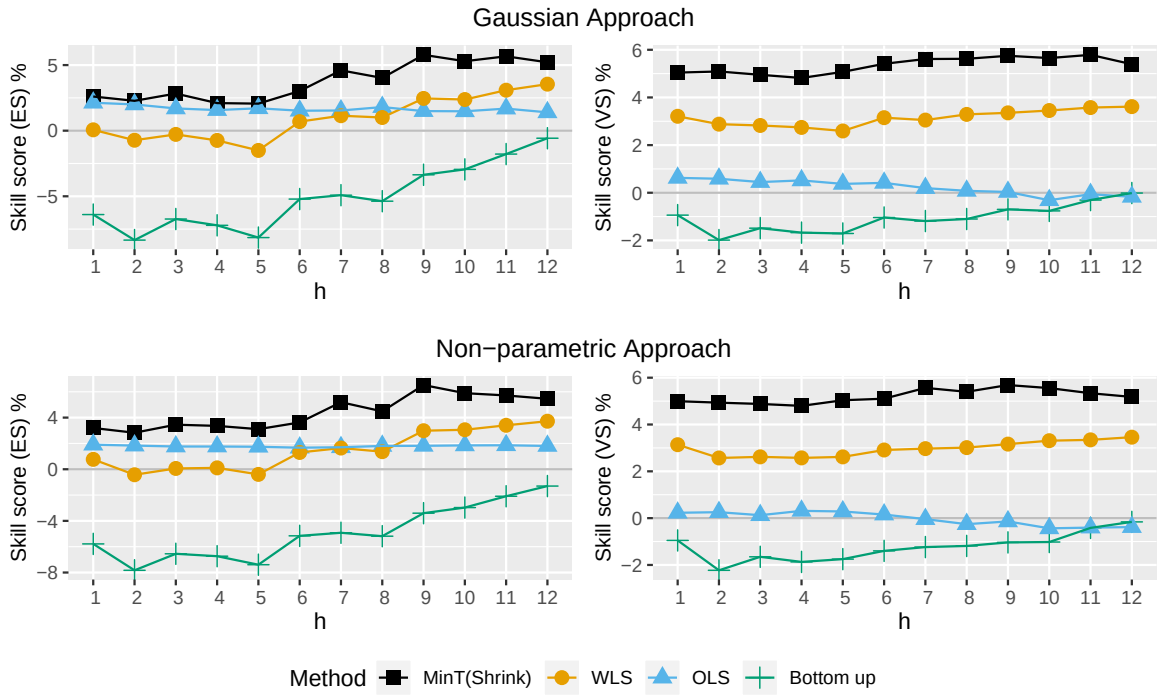

Figure 4: Skill scores with reference to incoherent forecasts for multivariate predictive distribution across the entire hierarchy from different methods. A positive (negative) skill score indicates a gain (loss) in forecast accuracy over the incoherent forecast distribution. The top panel shows the results from the Gaussian approach where the bottom panel shows the results from the non-parametric approach. Left and right panels show the skill scores based on energy and variogram scores respectively.

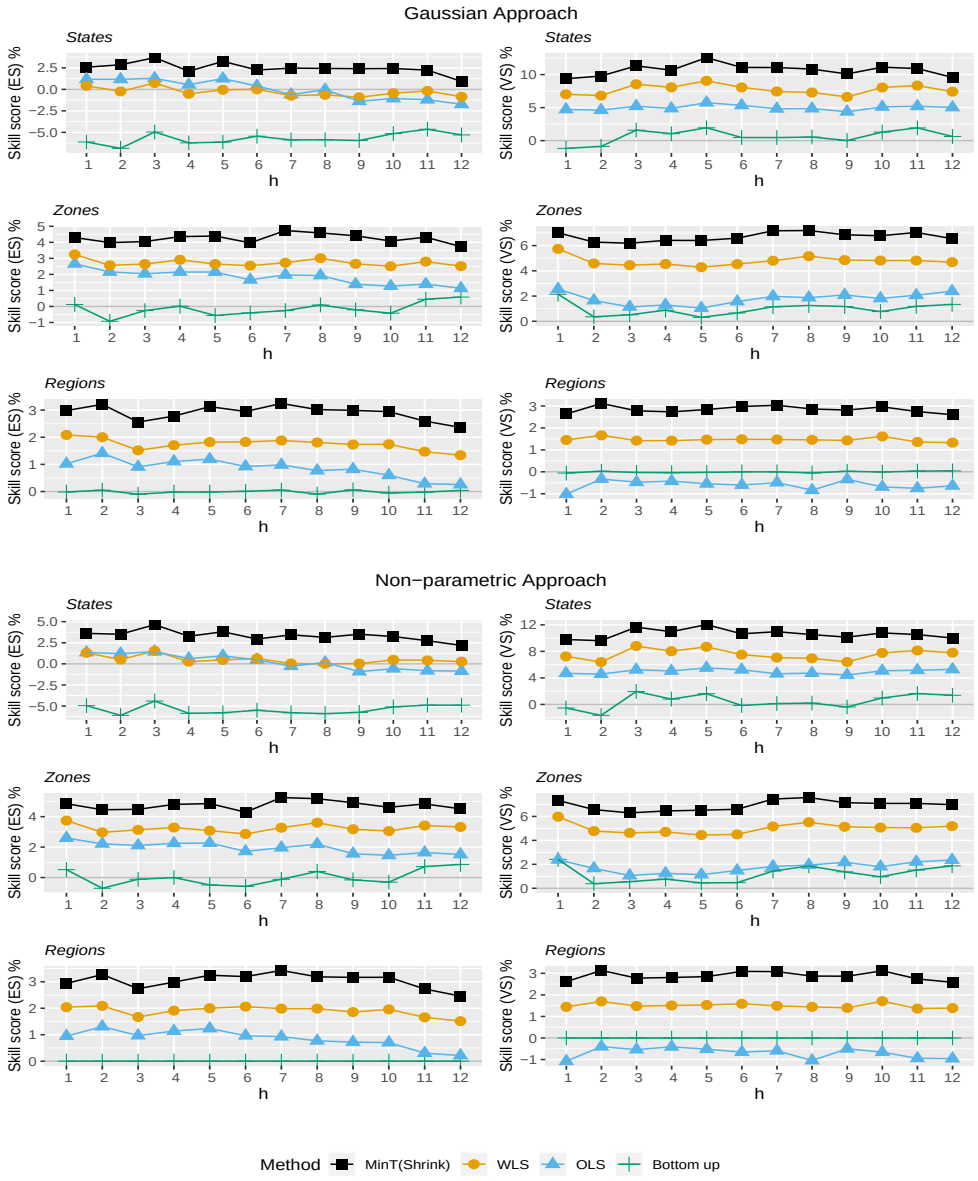

Figure 5: Skill scores for multivariate probabilistic forecasts across different levels of the hierarchy. A positive (negative) skill score indicates a gain (loss) in forecast accuracy over the incoherent forecast distribution. Results from the Gaussian approach are presented in the top three panels while results from the non-parametric approach are presented in the bottom three panels.

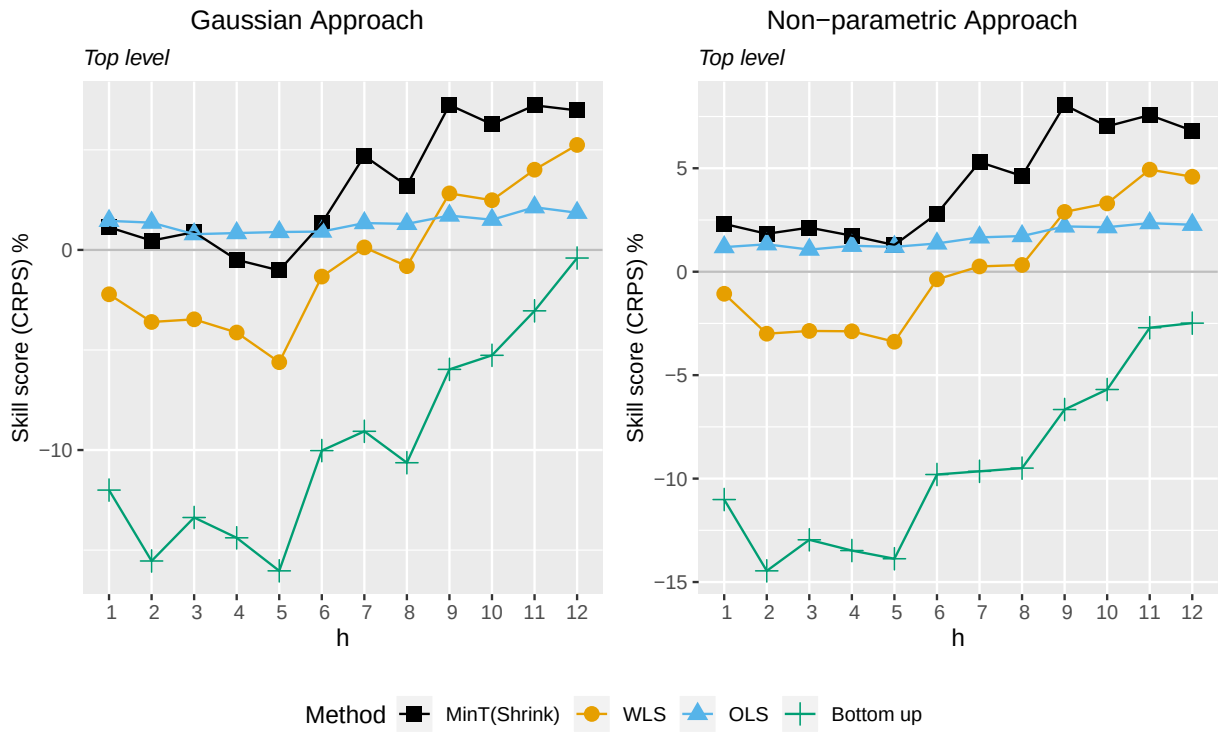

Figure 6: Skill score based on CRPS (with reference to the incoherent forecasts) for univariate probabilistic forecasts for the Total (top level) overnight trips. A positive (negative) skill score indicates a gain (loss) in forecast accuracy over the incoherent forecast distribution. Left panel shows the results from the Gaussian approach and right panel shows the results from the non-parametric approach.

8 Conclusions

Although hierarchical point forecasting is well studied in the literature, there is a lack of attention in the probabilistic setting. We fill this gap in the literature by providing substantial theoretical background to the problem.

The geometric interpretation of point forecast reconciliation allows us to extend these concepts to the probabilistic setting. We have also discussed strategies for evaluating probabilistic forecasts for hierarchical time series advocating the use of multivariate scoring rules on the full hierarchy, while establishing a key result that the log score is not proper with respect to incoherent forecasts.

We have shown that for elliptical distributions the true predictive density can be recovered by linear reconciliation and we have established conditions for when this is a projection. Although this projection cannot feasibly be obtained in practice, a projection similar to the MinT approach provides a good approximation in applications. This is supported by the results of a simulation study as well as the empirical application.

We have further proposed a novel non-parametric approach for obtaining coherent probabilistic forecasts for when the parametric densities are unavailable. Initially this method involves generating thousands of sample paths using bootstrapped forecast errors. Then each sample path is reconciled via projections. Using an extensive simulation setting we have shown that the MinT projection is at least as good as the optimal projection with respect to minimising Energy score. Further we have shown in an empirical application that reconciled probabilistic forecasts via MinT show gains in the forecasts over incoherent and bottom-up forecasts.

In many ways this chapter sets up a substantial future research agenda. For example, having defined what amounts to an entire class of reconciliation methods for probabilistic

forecasts it will be worthwhile investigating which specific projections are optimal. This is likely to depend on the specific scoring rule employed as well as the properties of the base forecasts. Another avenue worth investigating is to consider whether it is possible to recover the true predictive distribution for non-elliptical distributions via a non-linear function $g(.)$.

9 Appendix

9.1 Results from ETS base forecasts

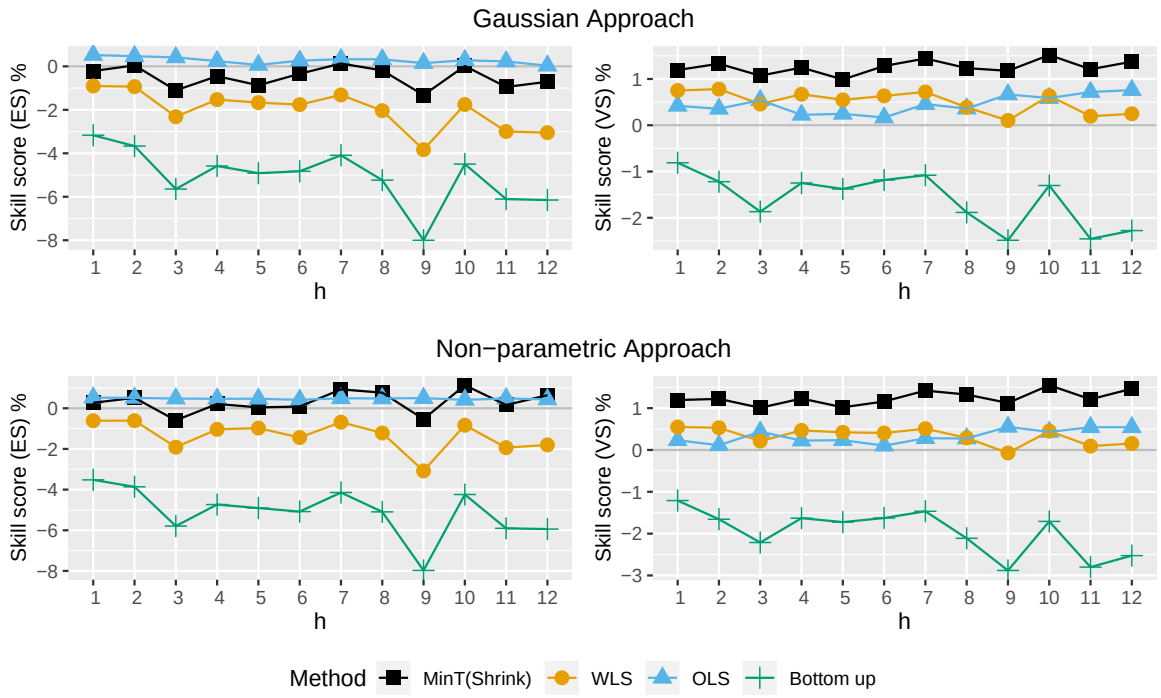

Figure 7: Skill scores with reference to ETS base forecasts for multivariate predictive distribution of the whole hierarchy from different reconciliation methods are presented. Top panel shows the results from Gaussian approach and the bottom panel shows the results from non-parametric approach. Left and right panels shows the skill scores based on energy score and variogram score respectively.

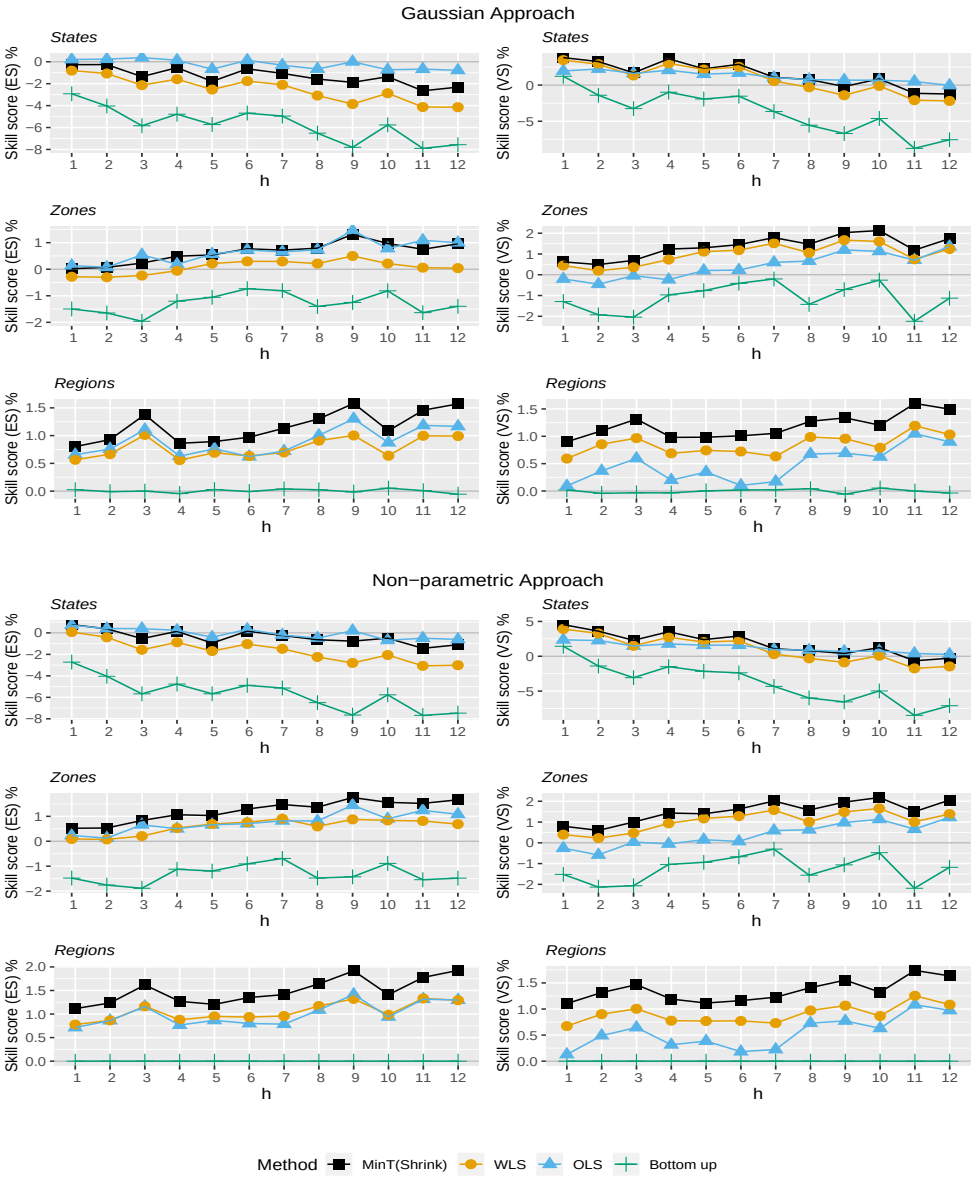

Figure 8: Skill score (with reference to ETS base forecasts) for multivariate probabilistic forecasts of different levels of the hierarchy are presented. Results from Gaussian approach are presented in the top three panels and results from the non-parametric approach are presented in the bottom three panels.

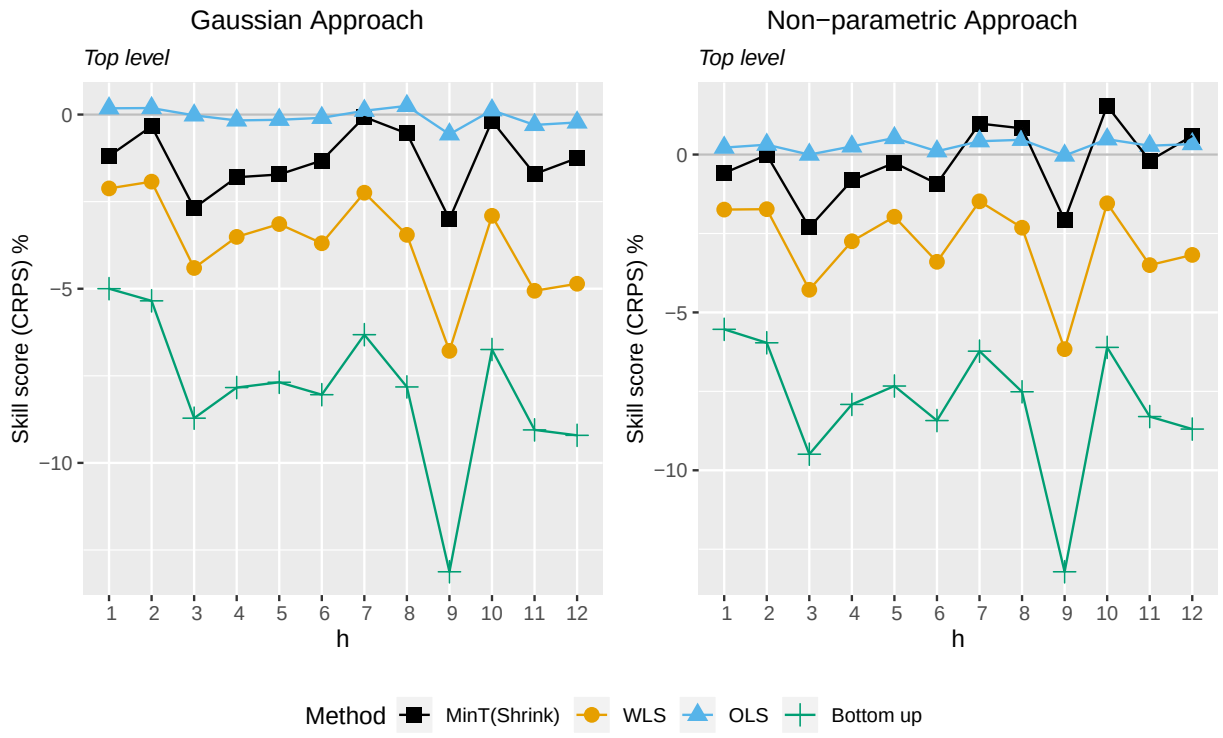

Figure 9: Skill score based on CRPS (with reference to the ETS base forecasts) for univariate probabilistic forecasts for the Total (top level) overnight trips are presented. Left panel shows the results from Gaussian approach and right panel shows the results from non-parametric approach.

9.2 Australian Tourism Data

Table 7: Geographical hierarchy of Australian tourism flow

Level 0 - Total			*Regions cont.*			*Regions cont.*		
1	Tot	Australia	37	AAB	Central Coast	75	CBD	Mackay
Level 1 - States			38	ABA	Hunter	76	CBE	Capricorn
2	A	NSW	39	ABB	North Coast NSW	77	CBF	Gladstone
3	B	Victoria	40	ACA	South Coast	78	CCA	Whitsundays
4	C	Queensland	41	ADA	Snowy Mountains	79	CCB	Townsville
5	D	South Australia	42	ADB	Capital Country	80	CCC	Tropical North Queensland
6	E	Western Australia	43	ADC	The Murray	81	CDA	Southern QLD country
7	F	Tasmania	44	ADD	Riverina	82	CDB	Outback QLD
8	G	Northern Territory	45	AEA	Central NSW	83	DAA	Adelaide
Level 2 - Zones			46	AEB	New England North West	84	DAB	Barossa
9	AA	Metro NSW	47	AEC	Outback NSW	85	DAC	Adelaide Hills
10	AB	North Coast NSW	48	AED	Blue Mountains	86	DBA	Limestone Coast
11	AC	South Coast NSW	49	AFA	Canberra	87	DBB	Fleurieu Peninsula
12	AD	South NSW	50	BAA	Melbourne	88	DBC	Kangaroo Island
13	AE	North NSW	51	BAB	Peninsula	89	DCA	Murraylands
14	AF	ACT	52	BAC	Geelong	90	DCB	Riverland
15	BA	Metro VIC	53	BBA	Western	91	DCC	Clare Valley
16	BB	West Coast VIC	54	BCA	Lakes	92	DCD	Flinders Range and Outback
17	BC	East Coast VIC	55	BCB	Grippsland	93	DDA	Eyre Peninsula
18	BD	North East VIC	56	BCC	Phillip Island	94	DDB	Yorke Peninsula
19	BE	North West VIC	57	BDA	Central Murray	95	EAA	Australia's Coral Coast
20	CA	Metro QLD	58	BDB	Goulburn	96	EAB	Experience Perth
21	CB	Central Coast QLD	59	BDC	High Country	97	EAC	Australia's South West
22	CC	North Coast QLD	60	BDD	Melbourne East	98	EBA	Australia's North West
23	CD	Inland QLD	61	BDE	Upper Yarra	99	ECA	Australia's Golden Outback
24	DA	Metro SA	62	BDF	Murray East	100	FAA	Hobert and South
25	DB	South Coast SA	63	BEA	Wimmera+Mallee	101	FBA	East Coast
26	DC	Inland SA	64	BEB	Western Grampians	102	FBB	Launceston, Tamar & North
27	DD	West Coast SA	65	BEC	Bendigo Loddon	103	FCA	North West
28	EA	West Coast WA	66	BED	Macedon	104	FCB	West coast
29	EB	North WA	67	BEE	Spa Country	105	GAA	Darwin
30	EC	South WA	68	BEF	Ballarat	106	GAB	Litchfield Kakadu Arnhem
31	FA	South TAS	69	BEG	Central Highlands	107	GAC	Katherine Daly
32	FB	North East TAS	70	CAA	Gold Coast	108	GBA	Barkly
33	FC	North West TAS	71	CAB	Brisbane	109	GBB	Lasseter
34	GA	North Coast NT	72	CAC	Sunshine Coast	110	GBC	Alice Springs
35	GB	Central NT	73	CBB	Bundaberg	111	GBD	MacDonnell
Level 2 - Regions			74	CBC	Fraser Coast			
36	AAA	Sydney						

References

Abramson, B. & Clemen, R. (1995), 'Probability forecasting', *International Journal of Forecasting* **11**(1), 1–4.

Athanasopoulos, G., Ahmed, R. A. & Hyndman, R. J. (2009), 'Hierarchical forecasts for Australian domestic tourism', *International Journal of Forecasting* **25**(1), 146 – 166.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N. & Petropoulos, F. (2017), 'Forecasting with temporal hierarchies', *European Journal of Operational Research* **262**(1), 60–74.

Ben Taieb, S., Huser, R., Hyndman, R. J. & Genton, M. G. (2017), 'Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression', *IEEE Transactions on Smart Grid* **7**(5), 2448–2455.

Ben Taieb, S., Taylor, J. W. & Hyndman, R. J. (2017), Coherent probabilistic forecasts for hierarchical time series, *in* 'Proceedings of the 34th International Conference on Machine Learning', Vol. 70, PMLR, pp. 3348–3357.

Böse, J.-H., Flunkert, V., Gasthaus, J., Januschowski, T., Lange, D., Salinas, D., Schelter, S., Seeger, M. & Wang, Y. (2017), 'Probabilistic demand forecasting at scale', *Proceedings of the VLDB Endowment* **10**(12), 1694–1705.

Dunn, D. M., Williams, W. H. & Dechaine, T. L. (1976), 'Aggregate Versus Subaggregate Models in Local Area Forecasting', *Journal of American Statistical Association* **71**(353), 68–71.

Gneiting, T. & Katzfuss, M. (2014), 'Probabilistic Forecasting', *Annual Review of Statistics and Its Application* **1**, 125–151.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly Proper Scoring Rules, Prediction, and Estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. & Johnson, N. A. (2008), Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds.

Gross, C. W. & Sohl, J. E. (1990), 'Disaggregation methods to expedite product line forecasting', *Journal of Forecasting* **9**(3), 233–254.

Hyndman, R. (2019), 'forecast: Forecasting Functions for Time Series and Linear Models, R package version 8.9', *URL: http://github.com/robjhyndman/forecast* .

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G. & Shang, H. L. (2011), 'Optimal combination forecasts for hierarchical time series', *Computational Statistics and Data Analysis* **55**(9), 2579–2589.

Jeon, J., Panagiotelis, A. & Petropoulos, F. (2019), 'Probabilistic forecast reconciliation with applications to wind power and electric load', *European Journal of Operational Research* **279**(2), 364–379.

Jordan, A., Krüger, F. & Lerch, S. (2017), 'Evaluating probabilistic forecasts with the R package scoringRules'.
URL: *http://arxiv.org/abs/1709.04743*

McLean Sloughter, J., Gneiting, T. & Raftery, A. E. (2013), 'Probabilistic wind vector forecasting using ensembles and bayesian model averaging', *Monthly Weather Review* **141**(6), 2107–2119.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G. & Hyndman, R. J. (2019), Forecast reconciliation: A geometric view with new insights on bias correction, Working paper 18/19, Monash University Econometrics & Business Statistics.

Pinson, P., Madsen, H., Papaefthymiou, G. & Klöckl, B. (2009), 'From Probabilistic Forecasts to Wind Power Production', *Wind Energy* **12**(1), 51–62.

Pinson, P. & Tastu, J. (2013), Discrimination ability of the Energy score, Technical report, Technical University of Denmark.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rossi, B. (2014), 'Density forecasts in economics, forecasting and policymaking'.

Schäfer, J. & Strimmer, K. (2005), 'A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics', *Statistical Applications in Genetics and Molecular Biology* **4**(1).

Scheuerer, M. & Hamill, T. M. (2015), 'Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities', *Monthly Weather Review* **143**(4), 1321–1334.

Shang, H. L. & Hyndman, R. J. (2017), 'Grouped functional time series forecasting: An application to age-specific mortality rates', *Journal of Computational and Graphical Statistics* **26**(2), 330–343.

Székely, G. J. & Rizzo, M. L. (2013), 'Energy statistics: A class of statistics based on distances', *Journal of Statistical Planning and Inference* **143**(8), 1249–1272.

Tourism Research Australia (2019), Tourism forecasts, Technical report, Tourism Research Australia, Canberra.

Van Erven, T. & Cugliari, J. (2015), Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts, *in* 'Modeling and Stochastic Learning for Forecasting in High Dimensions', Springer, pp. 297–317.

Wickramasuriya, S. L., Athanasopoulos, G. & Hyndman, R. J. (2019), 'Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization', *Journal of the American Statistical Association* **114**(526), 804–819.

Wytock, M. & Kolter, J. Z. (2013), Large-scale probabilistic forecasting in energy systems using sparse gaussian conditional random fields, *in* 'Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on', IEEE, pp. 1019–1024.

Zarnowitz, V. & Lambros, L. A. (1987), 'Consensus and uncertainty in economic prediction', *Journal of Political economy* **95**(3), 591–621.