Probabilistic Forecasts for Hierarchical Time Series

Puwasala Gamakumara
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: puwasala.gamakumara@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

November 7, 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 are generally referred to as hierarchical time series. To ensure aligned decision making, it is important that forecasts across all levels of aggregation add 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 then aggregating these up or disaggregate these 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 associated 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 has been 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 show that reconciliation improves forecast accuracy compared to incoherent forecasts in both parametric and non-parametric approaches and furthermore, MinT reconciliation performs best.

The remainder of the paper is structured as follows. In Section 2.1 notation and some preliminary work on point forecast reconciliation is discussed. Section 2 contains the definitions and interpretation of coherent probabilistic forecasts and reconciliation. In Section 5 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 3 while the non-parametric approach is introduced in Section 4. An empirical application on tourism forecasting is contained in Section 7. Finally Section 8 concludes with some discussion and thoughts on future research.

2 Hierarchical probabilistic forecasts

Before introducing coherence and reconciliation to the probabilistic setting we first briefly refresh these concepts in the case of the point forecasts. In do so, we follow the more geometric intepretation introduced by (Panagiotelis et al. 2019), since this formulation provides a natural framework for generalisation to probabilistic forecasting.

2.1 Point Forecasting

A *hierarchical time series* is a collection of time series adhering to some known linear constraints. Stacking the value of each series at time t into an n-vector $\boldsymbol{y}_t$, the constraints imply that $\boldsymbol{y}_t$ lies in an m-dimensional linear subspace of $\mathbb{R}^n$ for all t. This subspace is referred to as the *coherent subspace* and is denoted as $\mathfrak{s}$. A typical (and the original) motivating example is collections of time series some of which are aggregates of other series. In this case $\boldsymbol{b}_t \in \mathbb{R}^m$ can be defined as the values of the most disaggregated or *bottom-level series* at time t and the aggregation constraints can be formulated as,

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

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

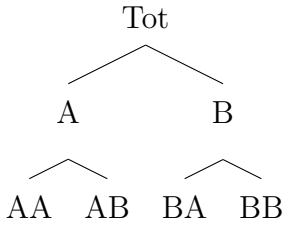

Figure 1: An example of a two level hierarchical structure.

An example of a hierarchy is shown in Figure 1. There are $n = 7$ series of which $m = 4$ are bottom-level series. Also, $\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $\boldsymbol{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, \boldsymbol{b}_t']'$, 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 connection between this characterisation and the coherent subspace is that the columns of $\boldsymbol{S}$ span $\mathfrak{s}$. Below, the notation $s : \mathbb{R}^m \to \mathbb{R}^n$ when premultiplication by $\boldsymbol{S}$ is though of as a mapping. Finally, while $\boldsymbol{S}$ is defined in terms of m bottom level series here, in general any m series can be chosen with the $\boldsymbol{S}$ matrix redefined accordingly. The columns of all appropriately defined $\boldsymbol{S}$ matrices span the same coherent subspace $\mathfrak{s}$.

When forecasts of all n series are produced, they may not adhere to constraints. In this case forecasts are called *incoherent* or *base* forecasts and are denoted $\hat{\boldsymbol{y}}_{t+h}$, with the subscript $t + h$ implying a h-step ahead forecast at time t. To exploit the fact that the target of the forecast adheres to known linear constraints, these forecasts can be adjusted in a process known as *forecast reconciliation*. At its most general, this involves selecting a mapping $\psi : \mathbb{R}^n \to \mathfrak{s}$ and then setting $\tilde{\boldsymbol{y}}_{t+h} = \psi(\hat{\boldsymbol{y}}_{t+h})$, where $\tilde{\boldsymbol{y}}_{t+h} \in \mathfrak{s}$ is called the *reconciled* forecast. This mapping itself may be considered as the composition of two mappings $\psi = s \circ g$. Here, $g : \mathbb{R}^n \to \mathbb{R}^m$ combines forecast of all series to produce new bottom level forecasts, which are then aggregated via s. When both mappings are linear this corresponds to premultiplying base forecasts by a matrix $\boldsymbol{SG}$.

Panagiotelis et al. (2019) discuss how a number of important properties arise when $\boldsymbol{SG}$ is a projection matrix. In this case, $\boldsymbol{G}$ can be written as $(\boldsymbol{R}_\perp'\boldsymbol{S})^{-1}\boldsymbol{R}_\perp'$, where $\boldsymbol{R}_\perp$ is an $n \times m$ orthogonal complement to the $n \times (n - m)$ matrix $\boldsymbol{R}$, with $\boldsymbol{R}$ defining a direction

of projection. For instance, OLS reconciliation (Hyndman et al. 2011) projects along a direction perpendicular to $\boldsymbol{S}$, in which case $\boldsymbol{R} = \boldsymbol{S}_{\perp}$, $\boldsymbol{R}_{\perp} = \boldsymbol{S}$ and $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$. Several other choices of $\boldsymbol{G}$ currently extant in the literature, including the bottom-up (Dunn et al. 1976) WLS and MinT (Wickramasuriya et al. 2019) methods, are also special cases where $\boldsymbol{SG}$ is a projection. These methods are summarised in Table 5 in Appendix.

2.2 Coherent probabilistic forecasts

We now turn our attention towards a novel definition of coherence in a probabilistic setting. First 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 $\mathbb{R}^m$. This triple can be thought of as a probabilistic forecast for the bottom level series. A sigma-algebra $\mathscr{F}_{\mathfrak{s}}$ can then be constructed as the collection of sets $s(\mathcal{B})$ for all $\mathcal{B} \in \mathscr{F}_{\mathbb{R}^m}$, where $s(\mathcal{B})$ denotes the image of $\mathcal{B}$ under the mapping $s(.)$.

Definition 2.1 (Coherent Probabilistic Forecasts)**.** Given the triple, $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m}, \nu)$, a coherent probability triple is given by $\mathfrak{s}$, the sigma algebra $\mathscr{F}_{\mathfrak{s}}$ and a measure $\breve{\nu}$ such that

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

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. While these definitions do not contradict one another we believe our definition has two advantages. First it can more naturally be easily be extended to problems with non-linear constraints with the coherent subspace $\mathfrak{s}$ replaced with a manifold. Second, the geometric understanding of coherence facilitates a definition of probabilistic forecast reconciliation to which we now turn our attention.

2.3 Probabilistic forecast reconciliation

Let $(\mathbb{R}^n, \mathscr{F}_{\mathbb{R}^n}, \hat{\nu})$ be a probability triple characterising a probabilistic forecast for all n series. The hat is used for $\hat{\nu}$ analogously with $\hat{\boldsymbol{y}}$ in the point forecasting case. The objective is to derive a reconciled measure $\tilde{\nu}$, assigning probability to each element of the σ-algebra $\mathscr{F}_{\mathfrak{s}}$.

Definition 2.2. The reconciled probability measure of $\hat{\nu}$ with respect to the mapping $\psi(.)$ is a probability measure $\tilde{\nu}$ on $\mathfrak{s}$ with σ-algebra $\mathscr{F}_{\mathfrak{s}}$ such that

$$\tilde{\nu}(\mathcal{A}) = \hat{\nu}(\psi^{-1}(\mathcal{A})) \qquad \forall \mathcal{A} \in \mathscr{F}_{\mathfrak{s}}\,,$$

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

This definition naturally extends forecast reconciliation to the probabilistic setting. In the point forecasting case, the reconciled forecast is obtained by passing an incoherent forecast through a transformation. Similarly, for probabilistic forecasts, to determine the probability assigned to a region of points by the reconciled forecast, we consider all probability assigned by the base forecasts to all points mapped to that region by a transformation. The transformation ψ can also be expressed as a composition of two transformations $s \circ g$. In this case, an m-dimensional reconciled probabilistic distribution ν can be obtained such that $\nu(\mathcal{B}) = \hat{\nu}(g^{-1}(\mathcal{B}))$ for all $\mathcal{B} \in \mathscr{F}_{\mathbb{R}^m}$ and a probabilistic forecast for the full hierarchy can then be obtained via Definition 2.1. This construction will be used in Section 3.

Defintion 2.2 can use any continuous mapping ψ, where continuity is required to ensure that open sets in $\mathbb{R}^n$ used to construct $\mathscr{F}_{\mathbb{R}^n}$ are mapped to open sets in $\mathfrak{s}$. However, hereafter, we restrict our attention to ψ as a linear mapping. This is depicted in Figure 2 when ψ is a projection. This figure is only a schematic, since even the most trivial hierarchy is 3-dimensional. The arrow labelled $\boldsymbol{S}$ spans an m-dimensional coherent subspace $\mathfrak{s}$, while

the arrow labelled $\boldsymbol{R}$ spans an $n-m$-dimensional direction of projection. The mapping g collapses all points in the blue shaded region $g^{-1}(\mathcal{B})$ to the black interval $\mathcal{B}$. Under s, $\mathcal{B}$ as an $s(\mathcal{B})$ shown in red. Under our definitions of coherence and reconciliation, the same probability is assigned to the red region under the reconciled measure as is assigned to the blue region under the incoherent measure.

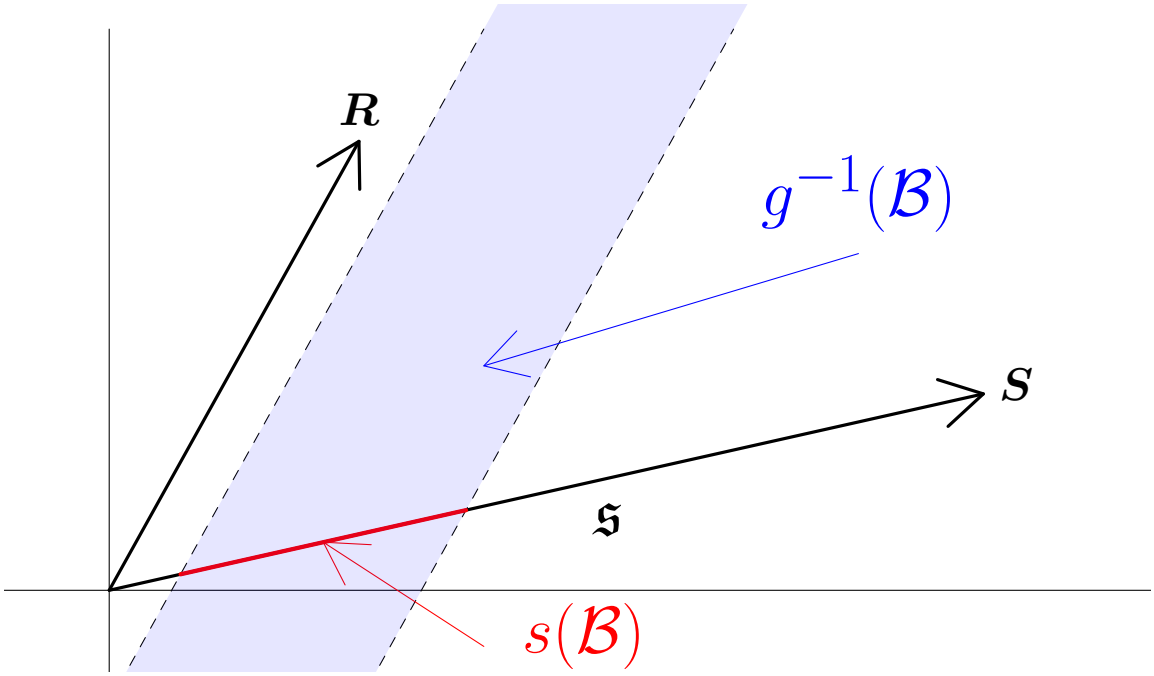

Figure 2: Summary of probabilistic forecast reconciliation. The probability that $\boldsymbol{y}_{t+h}$ lies in the red line segment under the reconciled probabilistic forecast is defined to be equal to the probability that $\boldsymbol{y}_{t+h}$ lies in the shaded blue area under the unreconciled probabilistic forecast. Note that since the smallest possible hierarchy involves three dimensions, this figure is only a schematic.

3 Analytical solution

3.1 Density of a reconciled distribution

In this section we describe how a reconciled distribution can be derived analytically, from an incoherent (or base) probabilistic forecast. We restrict our attention to linear s and g, and show that reconciliation involves changes of coordinates and marginalisation.

Theorem 3.1 (Reconciled density of bottom level)**.** *Consider the case where reconciliation is carried out using a composition of linear mappings $s \circ g$ where g combines information from all levels of the base forecast into the bottom levels. The density of the bottom level series under the reconciled distribution is*

$$\tilde{f}_{\boldsymbol{b}}(\boldsymbol{b}) = |\boldsymbol{G}^*| \int \hat{f}(\boldsymbol{G}^{-}\boldsymbol{b} + \boldsymbol{G}_{\perp}\boldsymbol{a}) d\boldsymbol{a}\,,$$

where $\hat{f}$ is the density of the incoherent base probabilistic forecast, $\boldsymbol{G}^{-}$ is an $n \times m$ pseudo inverse of $\boldsymbol{G}$ such that $\boldsymbol{G}\boldsymbol{G}^{-} = \boldsymbol{I}$, $\boldsymbol{G}_{\perp}$ is an $n \times (n-m)$ orthogonal complement to $\boldsymbol{G}$ such that $\boldsymbol{G}\boldsymbol{G}_{\perp} = \boldsymbol{0}$ and $\boldsymbol{G}_{\perp}'\boldsymbol{G}_{\perp} = \boldsymbol{I}$, $\boldsymbol{G}^* = \begin{pmatrix} \boldsymbol{G}^{-} \vdots \boldsymbol{G}_{\perp} \end{pmatrix}$, and $\boldsymbol{b}$ and $\boldsymbol{a}$ are obtained via the change of variables

$$\boldsymbol{y} = \boldsymbol{G}^* \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{a} \end{pmatrix}\,.$$

Proof. See appendix. □

Theorem 3.2 (Reconciled density of full hierarchy)**.** *Consider the case where a reconciled density for the bottom level series has been obtained using Theorem 3.1. The density of the full hierarchy under the reconciled distribution is*

$$\tilde{f}_{\boldsymbol{y}}(\boldsymbol{y}) = |\boldsymbol{S}^*| \tilde{f}_{\boldsymbol{b}}(\boldsymbol{S}^{-}\boldsymbol{y}) \mathbb{1}\left\{\boldsymbol{y} \in \mathfrak{s}\right\}\,,$$

where

$$\boldsymbol{S}^* = \begin{pmatrix} \boldsymbol{S}^- \\ \boldsymbol{S}'_\perp \end{pmatrix},$$

and $\boldsymbol{S}^-$ *is an* $m \times n$ *pseudo inverse of* $\boldsymbol{S}$ *such that* $\boldsymbol{S}^-\boldsymbol{S} = \boldsymbol{I}$, $\boldsymbol{S}_\perp$ *is an* $n \times (n-m)$ *orthogonal complement to* $\boldsymbol{S}$ *such that* $\boldsymbol{S}'_\perp \boldsymbol{S} = \boldsymbol{0}$ *and* $\boldsymbol{S}'_\perp \boldsymbol{S}_\perp = \boldsymbol{I}$.

Proof. See appendix. □

Example: Gaussian Distribution

Let the base forecast be Gaussian with mean $\hat{\boldsymbol{\mu}}$, covariance matrix $\hat{\boldsymbol{\Sigma}}$ and density,

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

Using Theorem 3.1, the reconciled density for the bottom level series is given by,

$$\tilde{f}_{\boldsymbol{b}}(\boldsymbol{b}) = \int (2\pi)^{-\frac{n}{2}} \left|\hat{\boldsymbol{\Sigma}}\right|^{-\frac{1}{2}} \left|\boldsymbol{G}^*\right| \exp\{-\frac{1}{2}q\} d\boldsymbol{a}\,,$$

where $q = (\boldsymbol{G}^-\boldsymbol{b} + \boldsymbol{G}_\perp\boldsymbol{a} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{G}^-\tilde{\boldsymbol{b}} + \boldsymbol{G}_\perp\tilde{\boldsymbol{a}} - \hat{\boldsymbol{\mu}})$, can be rearranged as

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

Noting that

$$\boldsymbol{G}^{*-1} = \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{G}_\perp \end{pmatrix}.$$

Then q can be rearranged further as

$$q = \left[\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{a} \end{pmatrix} - \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{G}'_\perp \end{pmatrix} \hat{\boldsymbol{\mu}}\right]' \left[\begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{G}'_\perp \end{pmatrix} \hat{\boldsymbol{\Sigma}} \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{G}'_\perp \end{pmatrix}'\right]^{-1} \left[\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{a} \end{pmatrix} - \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{G}'_\perp \end{pmatrix} \hat{\boldsymbol{\mu}}\right]$$

This can be recognised as a multivariate density in $\boldsymbol{b}$ and $\boldsymbol{a}$. The mean and covariance matrix for the margin corresponding to the first m elements are $\boldsymbol{G}\hat{\boldsymbol{\mu}}$ and $\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'$ respectively. Marginalising out $\boldsymbol{a}$, the reconciled forecast for the bottom-level is $\tilde{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{G}\hat{\boldsymbol{\mu}}, \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}')$.

3.2 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 3.3 (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 the bottom level series $\boldsymbol{y}$ also belong to the elliptical class with location parameter $\boldsymbol{\beta}$ and scale matrix $\boldsymbol{\Omega}$. Also, let $\boldsymbol{A}$ be any $m \times n$ matrix such that $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{\Omega}$. Then the linear reconciliation mapping $g(\boldsymbol{y}) = \boldsymbol{G}_{opt}\breve{\boldsymbol{y}} + \boldsymbol{d}_{opt}$ with $\boldsymbol{G}_{opt} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}^{-1/2}$ and $\boldsymbol{d}_{opt} = \boldsymbol{\beta} - \boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}}$ recovers the true predictive density.*

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). By similar working to the Gaussian example shown above, the scale matrix of the reconciled forecast for the bottom level series will be given by $\boldsymbol{G}_{opt}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'_{opt}$, while the location matrix is given by $\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} + \boldsymbol{d}_{opt}$. Substituting the values of $\boldsymbol{G}_{opt}$ and $\boldsymbol{d}_{opt}$, given in the theorem, the reconciled scale matrix is

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

and the reconciled location vector is

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

□

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 3.4 (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:*

$$sp(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \subset sp(\boldsymbol{R}) \tag{2}$$

$$sp(\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}^{1/2}) \subset sp(\boldsymbol{R}) \tag{3}$$

$$\tag{4}$$

Proof. The reconciled location vector will be given by

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

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.

4 Sample based solution: 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.

4.1 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 $i = 1, \ldots, n$ and $t = 1, \ldots, T$ where $\hat{y}_{i,t} = E(y_{i,t}|y_{i,1:t-1})$. The training errors are stored in a matrix $\boldsymbol{\Gamma}_{(T\times n)} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_T)'$ where $\boldsymbol{e}_t = (e_{1,t}, \ldots, e_{n,t})$ is stored in the same order as $\boldsymbol{y}_t$ for $t = 1, \ldots, T$. 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 $\boldsymbol{\Gamma}^b_{(H\times n)} = (\boldsymbol{e}^b_1, \ldots, \boldsymbol{e}^b_H)'$ and repeat this for $b = 1, \ldots, B$.

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.,

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

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

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

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

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

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

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

4.2 Reconciliation of incoherent future paths

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

$$\tilde{\boldsymbol{y}}^b_{T+h} = \boldsymbol{SG}\hat{\boldsymbol{y}}^b_{T+h}, \tag{5}$$

consequently we have,

$$\tilde{\boldsymbol{\Upsilon}}'_{T+h} = \boldsymbol{SG}\hat{\boldsymbol{\Upsilon}}'_{T+h}, \tag{6}$$

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 5) 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.

4.3 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,

$$\underset{\boldsymbol{G}_h}{\text{argmin}} \quad \mathrm{E}_{\boldsymbol{Q}}[S(\boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h}, \boldsymbol{y}_{T+h})], \tag{7}$$

where S is a proper scoring rule. The proper scoring rules and their properties will be discussed in Section 5. The subscript h on $\boldsymbol{G}$ is used to emphasise distinct $\boldsymbol{G}$ matrices for different forecast horizons. The expectation $E_{\boldsymbol{Q}}$ over the true forecast distribution $\boldsymbol{Q}$ can be approximated through sample mean and we can rewrite the objective function as follows.

$$\underset{\boldsymbol{G}_h}{\text{argmin}} \quad \frac{1}{N}\sum_{j=1}^{N}[S(\boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h,j}, \boldsymbol{y}_{T+h,j})], \tag{8}$$

We now discuss our algorithm for generating optimally reconciled bootstrapped future paths.

1. Consider a rolling window of L observations. We call this the "outer" rolling window. Inside this outer rolling window consider an "inner" rolling window of T observations.

2. For this inner rolling window, fit univariate models to each series in the hierarchy.

3. Based on these fitted models, generate B number of $h = 1$ to H steps-ahead incoherent future paths incorporating bootstrap errors as described in Subsection 4.1. Thus we get $\{\hat{\boldsymbol{\Upsilon}}_{T+1,j=1}, \hat{\boldsymbol{\Upsilon}}_{T+2,j=1}, \hat{\boldsymbol{\Upsilon}}_{T+H,j=1}\}$.

4. Repeat step (2) and (3) for $j = 1, \ldots, N$ by rolling the inner window one step ahead at a time.

5. Collect $\{\hat{\boldsymbol{\Upsilon}}_{T+h,j=1}, \ldots, \hat{\boldsymbol{\Upsilon}}_{T+h,j=N}\}$ for $h = 1, \ldots, H$ into separate arrays of matrices.

6. For each forecast horizon h, estimate the optimal $\boldsymbol{G}_h$ that will reconcile $\{\hat{\boldsymbol{\Upsilon}}_{T+h,j=1}, \ldots, \hat{\boldsymbol{\Upsilon}}_{T+h,j=N}\}$, by minimising the objective function in equation (8). Denote this optimum as $\boldsymbol{G}_h^{Opt}$.

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

8. Compute $\tilde{\boldsymbol{\Upsilon}}'_{T+h} = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h}$ for $h = 1, \ldots, H$ using $\boldsymbol{G}_h^{Opt}$.

We implement this algorithm in Section 6.3 to reconcile probabilistic forecasts in a simulation setting. Now we turn to the discussion of evaluation criteria for hierarchical probabilistic forecasts.

5 Evaluation of hierarchical probabilistic forecasts

An important issue in all forecasting problems is evaluating forecast accuracy. In the probabilistic setting, it is common to evaluate forecasts using proper scoring rules (see Gneiting & Raftery 2007, Gneiting & Katzfuss 2014, and references therein). Throughout we follow the convention of negatively-oriented scoring rules such that smaller values of the score indicate more accurate forecasts. A scoring rule $S(P, \boldsymbol{\omega})$ is a function that takes

two inputs, a probabilistic forecast P and a realisation $\boldsymbol{\omega}$ from the true data generating Q. A scoring rule is *proper* if $\mathrm{E}_{\boldsymbol{Q}}[S(Q,\omega)] \leq \mathrm{E}_{\boldsymbol{Q}}[S(P,\boldsymbol{\omega})]$ where $\mathrm{E}_Q[S(P,\boldsymbol{\omega})]$ is the expected score under the true distribution Q. When this inequality is strict, the scoring rule is said to be *strictly proper*.

Here we focus on three scoring rules: the log score (LS), the energy score (ES) and the variogram score (VS). Since hierarchical forecasting is by its very nature a multivariate problem (the linear constraints affect all variables), our focus is on multivariate scoring rules. However, in our simulations we will at times restrict our attention to evaluating individual margins of interest using the univariate counterparts of the log score and energy score (the latter is called the CRPS).

The log score simply involves evaluating the negative log density function at the value of the realisation, $LS(P,\boldsymbol{\omega}) = -\log p(\boldsymbol{\omega})$, where p is the density associated with a distribution P. The log score is more commonly used when a parametric form for the density is available. However, even in cases only a sample of values from the probabilistic forecast are available, this density can be approximated (see Jordan et al. 2017).

The energy score is defined in terms of the characteristic function of the probabilistic forecast. It is, however more commonly calculated using the following representation

$$\mathrm{ES}(P,\omega) = \mathrm{E}_P||\breve{\boldsymbol{y}} - \boldsymbol{\omega}||^{\alpha} - \frac{1}{2}\mathrm{E}_P||\boldsymbol{y} - \boldsymbol{y}^*||^{\alpha},\ \alpha \in (0,2]\,, \tag{9}$$

where $\boldsymbol{y}$ and $\boldsymbol{y}^*$ are independent copies drawn from the distribution P. These expectations can be easily approximated via Monte Carlo when samples from the probabilistic forecast are available. An interesting limiting case is $\alpha = 2$, where energy score simplifies to mean squared error around the mean of the predictive distribution. In this case, the energy score is proper but not strictly proper.

The energy score has been criticised by Pinson & Tastu (2013) for its low discriminative ability for incorrectly specified covariances. The variogram score (Scheuerer & Hamill 2015),

overcomes this issue and is defined as

$$\mathrm{VS}(\breve{P}, \boldsymbol{\omega}) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(|\omega_i - \omega_j|^p - E_P |y_i - y_j|^p \right)^2 , \tag{10}$$

where y_i and y_j are the i^{th} and j^{th} elements of $\boldsymbol{y} \sim P$ and p is usually set to 0.5.

In the context of probabilistic forecast reconciliation there could be two motivations for using scoring rules. The first is to compare incoherent base probabilistic forecasts to their reconciled counterparts, to see whether reconciliation improves forecast accuracy. The second motivation is to compare two reconciled probabilistic forecasts to one another, to see which choice of $\boldsymbol{G}$ performs best in practice.

5.1 Comparing base forecasts to reconciled forecasts

Section 2 describes how the density of a reconciled forecast can be derived from the density of the incoherent base forecast. Therefore it may seem sensible to compare base forecasts to reconciled forecasts using of log score. However, the following theorem shows that using the log score in improper in this setting.

Theorem 5.1 (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. See Appendix. □

As a result of Theorem 5.1 we recommend avoiding the log score when comparing reconciled and unreconciled probabilistic forecasts.

5.1.1 Comparing reconciled forecasts to one another

If probabilistic forecasts are available for any m series, then a probabilistic forecast for the full hierarchy can be derived. Definition 2.1 provides an example using the bottom level

series. This suggests that it may be adequate to merely compare two coherent forecasts to one another using the bottom level series only. We now show that this is not true in general and depends on the specific scoring rule used.

For the log score, suppose a coherent probabilistic forecast P has density $f_{\boldsymbol{y}}$ for the full hierarchy and a density $f_{\boldsymbol{b}}$ for the bottom level series. By Theorem 3.2, $f_{\boldsymbol{y}}(\boldsymbol{y}) = |\boldsymbol{S}^*| f_{\boldsymbol{b}}(S^- \boldsymbol{y}) \mathbb{1} y \in \mathfrak{s}$. Any realisation $\boldsymbol{y}^*$ will lie on the coherent subspace and can be written as $\boldsymbol{S}\boldsymbol{b}^*$. The expression for the log score is therefore

$$\begin{aligned} LS(f, \boldsymbol{y}^*) &= -\log\left(|\boldsymbol{S}^*| f_{\boldsymbol{b}}(\boldsymbol{S}^- \boldsymbol{S}\boldsymbol{b}^*)\right) \\ &= -log|\boldsymbol{S}^*| - \log f_{\boldsymbol{b}}(\boldsymbol{b}^*)\,. \end{aligned}$$

For coherent densities, the log score for the full hierarchy differs from the log score for the bottom-level series only by the term $-log(|\boldsymbol{S}^*|)$. This term is fixed for different choices of $\boldsymbol{G}$. As such, the density of only m series in the hierarchy are needed to compute the ranking of forecast accuracy for different reconciliation methods.

The same property does not hold for all scores in general. For example, the energy score which can be expressed in terms of norms is invariant under orthogonal transformations (Székely & Rizzo 2013, Gneiting & Raftery 2007). The same is not true of linear transformations in general. The ordering of energy scores for different reconciliation methods computed using bottom level series may change if the energy scores are recomputed after premultiplying bottom level series by $\boldsymbol{S}$. We recommend computing the energy score using the full hierarchy as well as for the variogram score, for which the same reasoning holds.

The properties of multivariate scoring rules in the context of evaluating reconciled probabilistic forecasts are summarised in Table 1.

Table 1: 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

6 Simulations

We now present the simulation study carried-out to evaluate the performance of probabilistic forecasts in both parametric and non-parametric setting. Let us first discuss the data generating process.

6.1 Data generating process (DGP)

The data generating process we consider corresponds to 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 the distinct impact of true DGP for parametric and non-parametric reconciliation approaches.

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,

$$\mathbf{\Sigma} = \begin{pmatrix} 5.0 & 3.1 & 0.6 & 0.4 \\ 3.1 & 4.0 & 0.9 & 1.4 \\ 0.6 & 0.9 & 2.0 & 1.8 \\ 0.4 & 1.4 & 1.8 & 3.0 \end{pmatrix}. \tag{11}$$

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 $\{u_{AA}, u_{AB}, u_{BA}, u_{BB}\}$.

Signal-to-noise ratio:

In practice, hierarchical time series are likely to have relatively noisier series at lower levels of aggregation. Following the simulation setup in Wickramasuriya et al. (2019) we replicate this feature in our simulated data by adding some noise to $\{w_{AA,t}, w_{AB,t}, w_{BA,t}, w_{BB,t}\}$ (See

Section C.1 in Appendix for more detailed explanation). We denote these nosier bottom-level series by $\{y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}\}$ which will be summed to obtain the data for upper levels.

6.2 Simulation set up for analytical solution

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).

We generate 2000 observations for each series from the 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 et al. 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 5 in appendix. 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 5. Results are reported in terms of skill scores which are the percentage improvement of a probabilistic forecast P over a reference method P_{ref} such that a positive value indicates that a method is more accurate than the reference method.

6.2.1 Results and discussion

Table 2 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 3 break down the forecasting performance of the different methods by considering univariate scores on each individual margin. We have only presented the results for forecast horizon $h = 1$ and the results for rest of the forecasts horizons are presented in table C.3 in Appendix. 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 2: 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 3: Comparison of incoherent vs coherent forecasts based on the univariate forecast distribution of each 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 $h = 1$ step-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

	Total		A		B		AA		AB		BA		BB	
R.method	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	3.91	1.30	**12.04**	**3.82**	**10.07**	**3.12**	1.47	0.47
MinT(Sample)	-0.08	-0.04	**9.37**	**3.13**	5.24	1.67	**4.12**	**1.38**	11.99	**3.82**	9.90	3.10	**1.57**	**0.51**
WLS	-2.91	-1.24	8.78	2.86	**5.49**	**1.73**	1.10	0.42	10.37	3.21	9.12	2.78	-1.14	-0.26
OLS	-19.22	-6.86	6.28	2.06	4.86	1.58	0.80	0.25	8.47	2.59	7.91	2.39	-1.52	-0.49
Bottom up	-140.27	-33.67	-13.75	-3.89	-11.10	-3.17	0.01	0.00	0.04	0.00	0.15	0.00	-0.09	0.00
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
Non-Gaussian DGP														
MinT(Shrink)	-1.16	-0.26	**0.92**	0.27	**11.90**	4.91	**3.40**	**1.31**	-0.11	-0.11	**13.22**	**5.00**	**2.37**	0.87
MinT(Sample)	-1.16	-0.72	**0.92**	**0.28**	**11.90**	**4.94**	**3.40**	1.29	-0.11	-0.13	**13.22**	4.99	**2.37**	**0.91**
WLS	**0.01**	**0.35**	-1.02	-0.52	9.95	4.07	2.92	1.20	-1.90	-0.74	8.50	3.13	-0.96	-0.26
OLS	-96.77	-84.90	0.55	0.08	6.48	2.57	2.70	1.07	-1.46	-0.55	6.19	2.25	-0.81	-0.21
Bottom up	-541.40	-246.37	-4.60	-1.87	-8.99	-3.11	-0.10	0.00	-0.12	0.00	-0.02	0.00	-0.08	0.00
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

6.3 Simulation setup for non-parametric solution

We now implement the non-parametric approach introduced in Section 4. Our main focus in this study is to compare different reconciliation methods to optimal reconciliation. We narrow down the study to find optimality with respect to energy score. Following equation (9) for samples from the probability distribution and for $\alpha = 1$, we can rewrite the objective

function in (8) as,

$$\underset{\boldsymbol{G}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N}\left\{\frac{1}{B}\sum_{b=1}^{B}||\boldsymbol{S}\boldsymbol{G}_h\boldsymbol{y}^{b}_{T+h,j}-\boldsymbol{y}_{T+h,j}||-\frac{1}{2(B-1)}\sum_{b=1}^{B-1}||\boldsymbol{S}\boldsymbol{G}_h(\boldsymbol{y}^{b}_{T+h,j}-\boldsymbol{y}^{b+1}_{T+h,j})||\right\} \tag{12}$$

First we generate 2500 data points for each series in the hierarchy using the same DGP discussed in Section 6.1. Following the algorithm explained in Section 4.3, we choose $L = 600$ observations for outer rolling window and $T = 500$ observations for inner rolling window. We fit univariate ARIMA models using `auto.arima()` function and generate $B = 1000$ of $h = 1$ to 3 steps-ahead incoherent bootstrapped future paths following steps 2 and 3. We repeat these steps for $N = 100$ times by moving the inner window one step at a time. Following step 6, we then estimate $\boldsymbol{G}_h^{Opt}$. We use numerical optimisation methods to attain the optimum. Next, we generate $\tilde{\boldsymbol{\Upsilon}}_{T+h}$ following step 7 and compute $\tilde{\boldsymbol{\Upsilon}}'_{T+h} = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h}$ for $h = 1, 2, 3$ as described in step 8. We use $\boldsymbol{G}_h^{Opt}$ as well as other $\boldsymbol{G}$ matrices given in Table 5 (in appendix) for reconciliation in step 8. Finally we repeat this whole process for 1000 times by moving 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.

We also note that different parameterisation methods were used for estimation $\boldsymbol{G}_h^{Opt}$. This was discussed in detail in Section C.4 in Appendix.

6.3.1 Results and discussion

We use energy score and variogram score to assess the predictive performance from different reconciliation methods. Results following from Non-Gaussian and Gaussian DGP are presented in left and right panels of Table 4 respectively.

Mann-Whitney test was used to compare the difference of scores between reconciliation methods. The results 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 method have similar prediction accuracy as there is no significant difference between the scores from these reconciliation methods. Although the scores are relatively larger for Gaussian than non-Gaussian data, the overall conclusions are consistent.

The simulation results from reparameterisation methods are presented in Table 8 in Appendix. From these we note that the different parameterisation of $\boldsymbol{G}$ for optimal reconciliation give equivalent results irrespective to the forecast horizon or the DGP.

However we also 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. First, it is computationally efficient relative to the optimal method, and second, it produces accurate probabilistic forecasts that are at least as good as the Optimal method with respect to the energy score.

Table 4: 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*	5.36	1.21	5.51	1.27	5.83	1.38	9.59	4.86	11.50	5.38	13.80	6.13
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 geographic 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 et al. 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 $h = 1, \ldots, 12$. 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 3. For the non-parametric forecasts, we generate the bootstrapped future paths and then reconcile each sample path as described in Section 4. 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 3 shows the skill scores with respect to the multivariate predictive distributions across the entire hierarchy from the different methods. Figure 4 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 5 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.

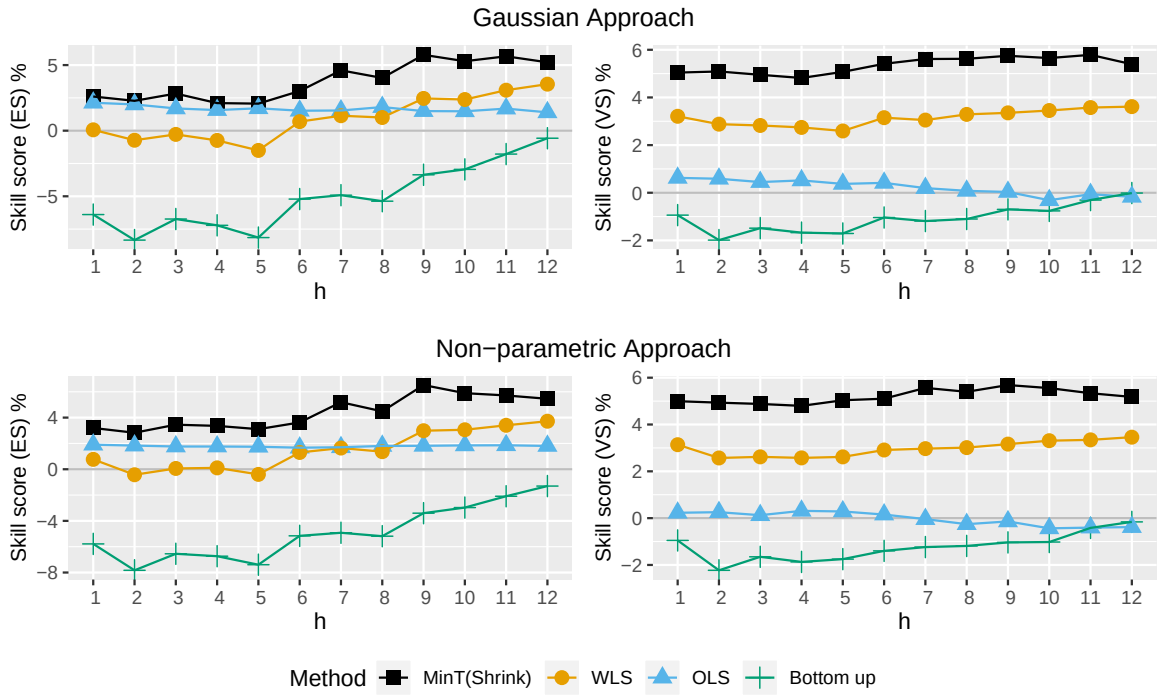

Figure 3: Skill scores with reference to incoherent forecasts for multivariate predictive distribution across the entire hierarchy from different methods. A positive (negative) skill score indicates a gain (loss) in forecast accuracy over the incoherent forecast distribution. The top panel shows the 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.

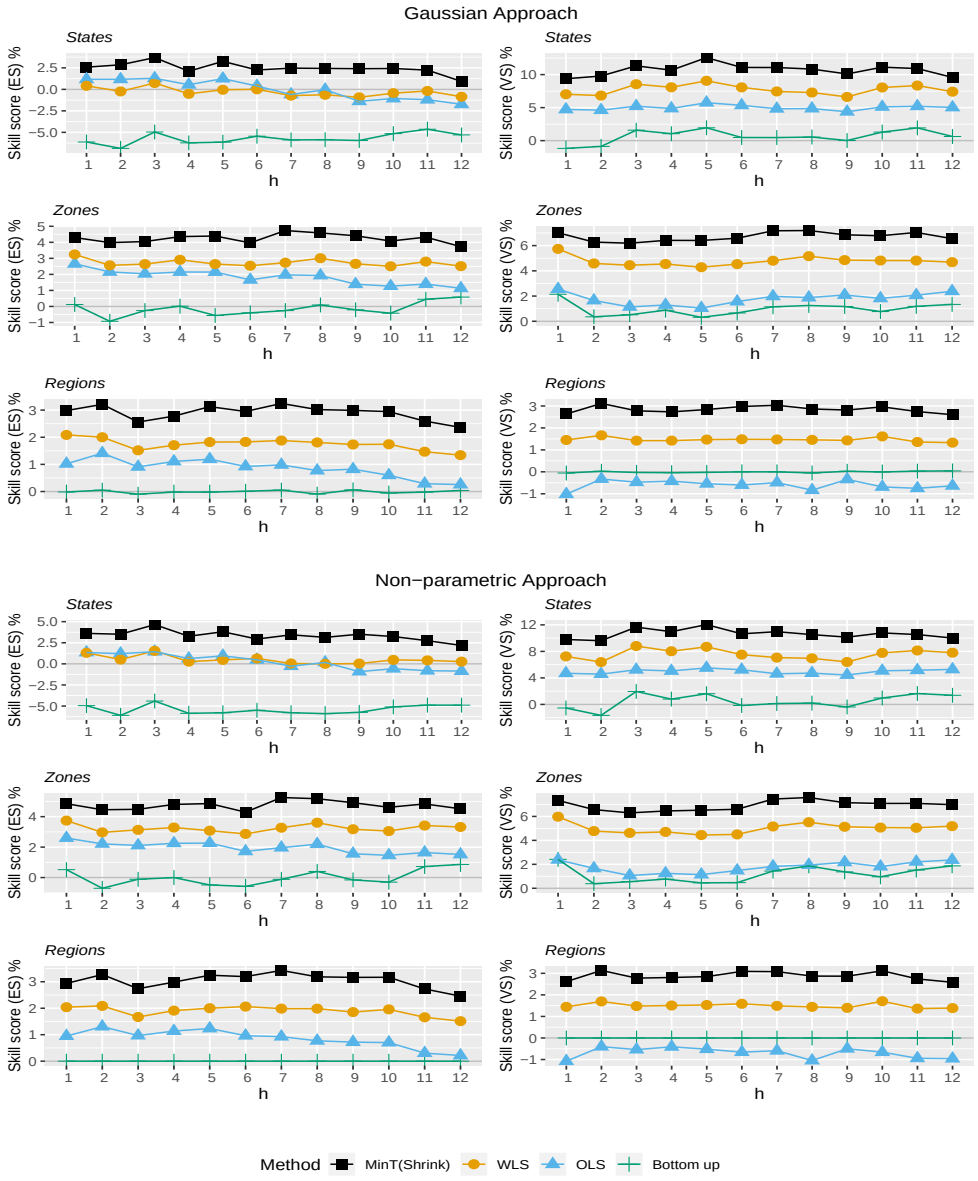

Figure 4: Skill scores for multivariate probabilistic forecasts across different levels of the hierarchy. A positive (negative) skill score indicates a gain (loss) in forecast accuracy over the incoherent forecast distribution. Results from the Gaussian approach are presented in the top three panels while results from the non-parametric approach are presented in the bottom three panels.

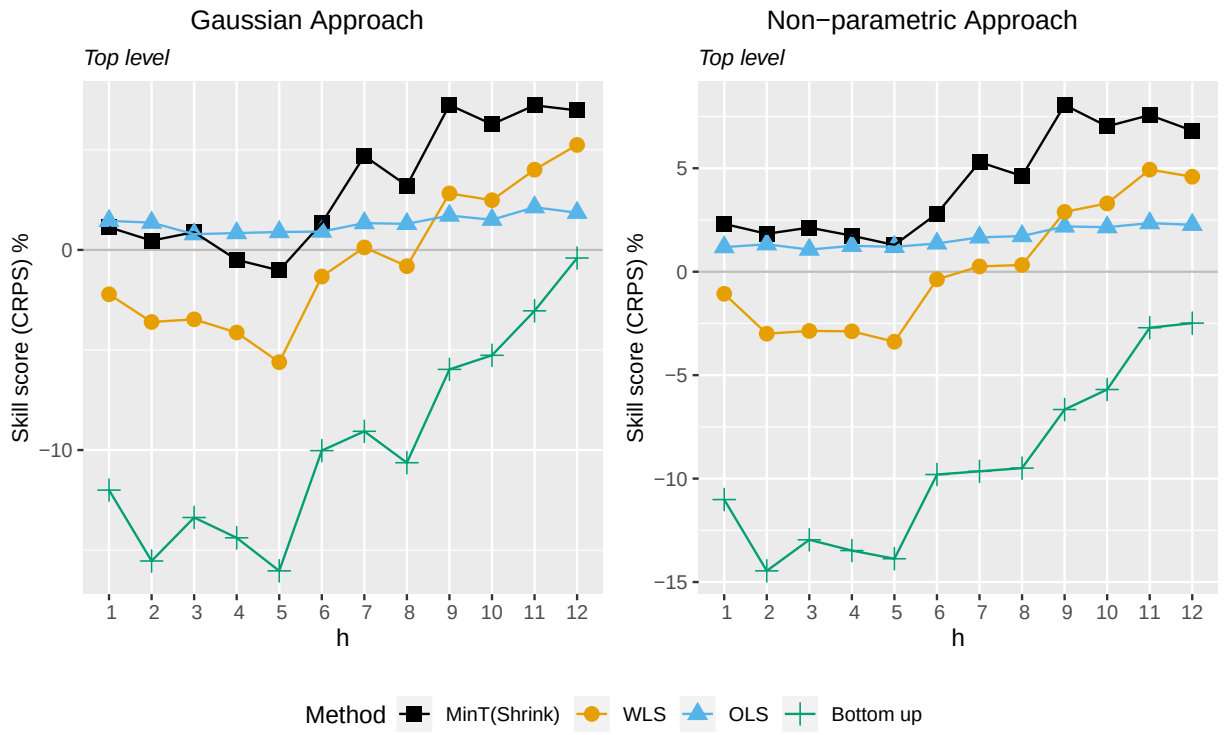

Figure 5: Skill score based on CRPS (with reference to the incoherent forecasts) for univariate probabilistic forecasts for the Total (top level) overnight trips. A positive (negative) skill score indicates a gain (loss) in forecast accuracy over the incoherent forecast distribution. Left panel shows 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 has been a relative lack of attention given to the probabilistic setting. We fill this gap in the literature by providing a mathematically rigorous formulation of coherence and reconciliation for probabilistic forecasts.

The geometric interpretation of point forecast reconciliation can be extended 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 forecast accuracy 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 $g(.)$.

A Proof of Theorem 3.1 and Theorem 3.2

Consider the interval $\mathcal{I}$ given by the Cartesian product of intervals $(l_1, u_1), (l_2, u_2), \ldots (l_m, u_m)$. We derive the probability that the bottom level series lie in the interval $\mathcal{I}$, which we denote as $\Pr(\boldsymbol{l} \succ \boldsymbol{b} \succ \boldsymbol{u})$ where $\boldsymbol{l} = (l_1, l_2, \ldots, l_m)$, $\boldsymbol{u} = (u_1, u_2, \ldots, u_m)$ and $\succ$ denotes element-wise inequality between vectors. The pre-image of $\mathcal{I}$ under g can similarly be denoted as all points $\boldsymbol{y}$ satisfying $\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{y} \succ \boldsymbol{u}$. By Definition 2.2

$$\Pr(\boldsymbol{l} \succ \boldsymbol{b} \succ \boldsymbol{u}) = \int\limits_{\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{y} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{y}) d\boldsymbol{y}\,,$$

where $\hat{f}$ is the density of the base probabilistic forecast. Now consider a change of variables to an n-dimensional vector $\boldsymbol{z}$ where $y = \boldsymbol{G}^* \boldsymbol{z}$. Recall, $\boldsymbol{G}^* = \left(\boldsymbol{G}^-, \vdots, \boldsymbol{G}_\perp\right)$ where $\boldsymbol{G}^-$ is an $n \times m$ pseudo inverse of $\boldsymbol{G}$ and $\boldsymbol{G}_\perp$ is an orthogonal complement of $\boldsymbol{G}$. By the change of variables

$$\begin{aligned}
\Pr(\boldsymbol{l} \succ \boldsymbol{b} \succ \boldsymbol{u}) &= \int\limits_{\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{y} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{y}) d\boldsymbol{y} \\
&= \int\limits_{\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{G}^*\boldsymbol{z} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{G}^*\boldsymbol{z}) |\boldsymbol{G}^*| d\boldsymbol{z} \\
&= \int\limits_{\boldsymbol{l} \succ \boldsymbol{z}_1 \succ \boldsymbol{u}} \hat{f}(\boldsymbol{G}^*\boldsymbol{z}) |\boldsymbol{G}^*| d\boldsymbol{z}\,,
\end{aligned}$$

where $\boldsymbol{z}_1$ denotes the first m elements of $\boldsymbol{z}$. Letting $\boldsymbol{a}$ denote the last $n - m$ elements of $\boldsymbol{z}$ the integral above can be written as

$$\Pr(\boldsymbol{b} \in \mathcal{I}) = \int\limits_{\boldsymbol{l} \succ \boldsymbol{z}_1 \succ \boldsymbol{u}} \int \hat{f}(\boldsymbol{G}^- \boldsymbol{z}_1 + \boldsymbol{G}_\perp \boldsymbol{a}) |\boldsymbol{\Gamma}| d\boldsymbol{a} d\boldsymbol{z}_1$$

Replacing $\boldsymbol{z}_1$ with $\boldsymbol{b}$ it can be seen that the term in outer integral is a density for the bottom level series. Therefore

$$\tilde{f}_{\boldsymbol{b}}(\boldsymbol{b}) = \int \hat{f}(\boldsymbol{G}^{-}\boldsymbol{b} + \boldsymbol{G}_{\perp}\boldsymbol{a})|\boldsymbol{\Gamma}|d\boldsymbol{a}\,, \tag{13}$$

is the density of $\boldsymbol{b}$. To obtain the density of the full hierarchy we first augment the density in Equation14 by $n - m$ variables denoted $\boldsymbol{u}$

$$f(\boldsymbol{b}, \boldsymbol{u}) = \tilde{f}(\boldsymbol{b})\mathbb{1}\left\{\boldsymbol{u} = 0\right\}\,, \tag{14}$$

such that the density $f(\boldsymbol{b}, \boldsymbol{u})$ is a density for n-dimensional vector that is degenerate across the dimensions corresponding to u. We then consider a change of variables,

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{S} \vdots \boldsymbol{S}_{\perp} \end{pmatrix} \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{u} \end{pmatrix}\,.$$

The inverse of $\begin{pmatrix} \boldsymbol{S} \vdots \boldsymbol{S}_{\perp} \end{pmatrix}$ is given by

$$\begin{pmatrix} \boldsymbol{S}^{-} \\ \boldsymbol{S}'_{\perp} \end{pmatrix}\,,$$

therefore $\boldsymbol{b} = \boldsymbol{S}^{-}\boldsymbol{y}$ and $\boldsymbol{u} = \boldsymbol{S}'_{\perp}\boldsymbol{y}$. Applying this change of variables we obtain the density

$$\tilde{f}_{\boldsymbol{y}}(\boldsymbol{y}) = \tilde{f}(\boldsymbol{S}^{-}\boldsymbol{y})\mathbb{1}\left\{\boldsymbol{S}'_{\perp}\boldsymbol{y} = 0\right\}\,.$$

Since $\boldsymbol{S}'_{\perp}$ is the orthogonal complement of $\boldsymbol{S}$ and since the columns of $\boldsymbol{S}$ span the coherent subspace, the statement $\boldsymbol{S}'_{\perp}\boldsymbol{y} = 0$ is equivalent to the statement $\boldsymbol{y} \in \mathfrak{s}$. As such, the reconciled density is given by

$$\tilde{f}_{\boldsymbol{y}}(\boldsymbol{y}) = \tilde{f}(\boldsymbol{S}^{-}\boldsymbol{y})\mathbb{1}\left\{\boldsymbol{y} \in \mathfrak{s}\right\}\,.$$

B Proof of Theorem 5.1

The proof below has some similarities to the case where discrete random variables are modelled as if they were continuous, an issue discussed in Section 4.1 of Gneiting & Raftery (2007). Consider a change of variables,

$$\boldsymbol{y} = \left(\boldsymbol{S} \vdots \boldsymbol{S}_{\perp}\right) \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{u} \end{pmatrix} .$$

The m-dimensional vector $\boldsymbol{b}$ corresponds to the bottom level series. For a coherent distribution the density is degenerate when the $n - m$ vector $\boldsymbol{u}$ equals 0 (i.e. the aggregation constraints hold). Also we define $\boldsymbol{S}^* := \left(\boldsymbol{S} \vdots \boldsymbol{S}_{\perp}\right)$

Consider the case where the true predictive density $f(\boldsymbol{y})$ is, after a change of variables written as $|\boldsymbol{S^*}| f_{\boldsymbol{b}}(\boldsymbol{b}) \mathbb{1}\{\boldsymbol{u} = \boldsymbol{0}\}$. Also consider an incoherent density, which after the same change of variables can be written as $|\boldsymbol{S^*}| \hat{f}_{\boldsymbol{b}}(\boldsymbol{b}) \hat{f}_{\boldsymbol{u}}(\boldsymbol{u})$. To prove the impropriety of the log score we will construct $f_{\boldsymbol{u}}$ to be highly concentrated around 0 but still non-degenerate. In particular we require $f_{\boldsymbol{u}} > 1$. As an example a Gaussian with mean zero and variance $\sigma^2 \boldsymbol{I}$ with $\sigma^2 < (2\pi)^{-1}$.

Consider any realisation from the true DGP $\boldsymbol{y}^*$. After a change of variables by premultiplying by $\boldsymbol{S}^*$, let the first m elements be given by $\boldsymbol{b^*}$. The remaining elements will all be equal to zero. The log score under the true data generating process is

$$\begin{aligned} LS\left(f, \boldsymbol{y}^*\right) &= -\log f(\boldsymbol{y}^*) \\ &= -\log |\boldsymbol{S^*}| - \log f_{\boldsymbol{b}}\left(\boldsymbol{b}^*\right) - \log\left(\mathbb{1}\left\{\boldsymbol{u}^* = \boldsymbol{0}\right\}\right) \\ &= -\log |\boldsymbol{S^*}| - \log f_{\boldsymbol{b}}\left(\boldsymbol{b}^*\right), \end{aligned} \tag{15}$$

where the third term in Equation 15 is equal to zero since $\boldsymbol{u}^* = \boldsymbol{0}$. The log score for the

incoherent density is

$$LS\left(\hat{f}, \boldsymbol{y}^*\right) = -\log|\boldsymbol{S}^*| - \log f_{\boldsymbol{b}}(\boldsymbol{b}^*) - \log f_{\boldsymbol{u}}(\boldsymbol{0})\,.$$

Since $f_{\boldsymbol{u}}(\boldsymbol{0}) > 1$ by construction, $-\log f_{\boldsymbol{u}}(\boldsymbol{0}) < 0$, therefore

$$LS\left(\hat{f}, \boldsymbol{y}^*\right) < -\log|\boldsymbol{S}^*| - \log f_{\boldsymbol{b}}(\boldsymbol{b}^*) = LS\left(f, \boldsymbol{y}^*\right)$$

Since this holds for any possible realisation of the true DGP, it will also hold after taking expectations. This violates the condition for a proper scoring rule.

C Simulations

C.1 Imposing higher signal-to-noise ratio in aggregate levels

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 $\sigma_v^2 = 7$.

C.2 Summary of hierarchical point forecasting methods

Table 5: 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	$\text{Diag}(\hat{\boldsymbol{W}}^{sam})$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{wls})^{-1}$

C.3 Simulation results from parametric solution for marginal forecast distributions

Table 6: 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 $h = 2$ and 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=2						h=3					
	Total		A		B		Total		A		B	
R.method	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS
Gaussian DGP												
MinT(Shrink)	**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)	0.27	-0.10	**10.76**	**3.39**	5.67	1.62	0.04	-0.23	**8.14**	**1.69**	4.19	1.10
WLS	-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	-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	-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
Non-Gaussian DGP												
MinT(Shrink)	-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)	-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	-0.11	-0.15	-2.50	-1.09	13.87	5.55	-0.02	-0.40	-3.96	-1.60	16.14	6.78
OLS	-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	-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

Table 7: 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" for forecast horizons $h = 2$ and $h = 3$.

	h=2								h=3							
	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
Gaussian DGP																
MinT(Shrink)	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)	**4.06**	**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	-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	-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	0.17	-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
Non-Gaussian DGP																
MinT(Shrink)	**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)	**3.67**	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	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	2.95	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	-0.17	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

C.4 Reparameterisation of G in optimal reconciliation of future paths and simulation results

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

$$\boldsymbol{G}_h = (\boldsymbol{S}'\boldsymbol{W}_h\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h. \tag{16}$$

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 (8) over symmetric $\boldsymbol{W}_h$. This solves an

Table 8: Energy scores (ES) and variogram scores (VS) for reconciled probabilistic forecasts from different parameterisation methods are presented.

	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

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 (8) over $\boldsymbol{U}_h$

Method 3 Similar to method 2, minimising (8) 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 (8) 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 $\mathfrak{s}$

D Application

D.1 Results from ETS base forecasts

Figure 6: 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 7: 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 8: 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.

D.2 Australian Tourism Data

Table 9: Geographic 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	Gippsland	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	Hobart 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.

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.

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

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.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., R Core Team, Ihaka, R., Reid, D., Shaub, D., Tang, Y. & Zhou, Z. (2019), *forecast: Forecasting Functions for Time Series and Linear Models*. Version 8.9.
URL: *https://CRAN.R-project.org/package=forecast*

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.