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

December 9, 2019

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

Abstract

TBC

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 point forecasts. In doing so, we follow the geometric interpretation introduced by Panagiotelis et al. (2019), since this formulation naturally generalises 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 a collection 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,

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

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

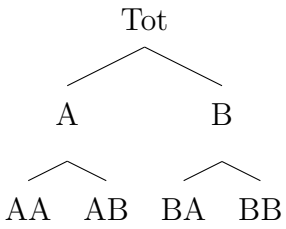

Figure 1: An example of a two level hierarchical structure.

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 $s : \mathbb{R}^m \to \mathbb{R}^n$ is used 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 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. The mapping ψ may be considered as the composition of two mappings $\psi = s \circ g$. Here, $g : \mathbb{R}^n \to \mathbb{R}^m$ combines incoherent base forecasts of all series to produce new bottom-level forecasts, which are then aggregated via s. When g is a linear mapping this corresponds to pre-multiplying 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. For instance, OLS reconciliation (Hyndman et al. 2011) projects along a direction perpendicular to $\boldsymbol{S}$, in which case $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$. Several other choices of $\boldsymbol{G}$ currently extant in the literature, including the bottom-up (Dunn et al. 1976) WLS (Hyndman et al. 2016, Athanasopoulos et al. 2017) and MinT (Wickramasuriya et al. 2019) methods, are also special cases where $\boldsymbol{SG}$ is a projection. The are presented in Table 2 and discussed in Section 6.

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 σ-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 $(\mathfrak{s}, \mathscr{F}_{\mathfrak{s}}, \breve{\nu})$, 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, Taylor & Hyndman (2017) who define them in terms of convolutions. While these definitions do not contradict one another our definition has two advantages. First it can more naturally be extended to problems with non-linear constraints with the coherent subspace $\mathfrak{s}$ replaced with a manifold. Second, it 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, a set of points is mapped to another set of points by a transformation, with the same probabilities assigned to each set under the base and reconciled measure respectively. Recall that the mapping 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.

Definition 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, $\mathcal{B}$ is mapped to $s(\mathcal{B})$ shown in red. Under our definition of 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.

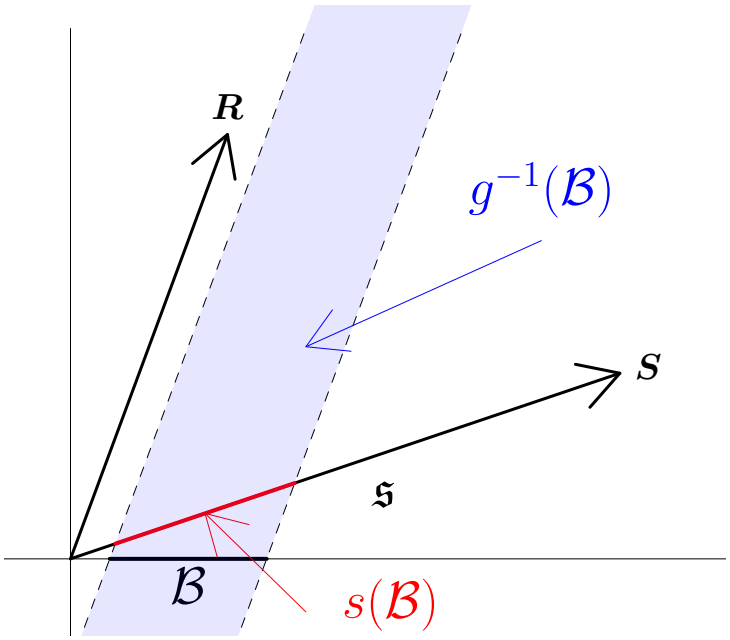

Figure 2: Summary of probabilistic forecast reconciliation. The probability that $\boldsymbol{y}_{t+h}$ lies in the red line segment under the reconciled probabilistic forecast is defined to be equal to the probability that $\boldsymbol{y}_{t+h}$ lies in the shaded blue area under the unreconciled probabilistic 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 a new density for the bottom-level. 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 $n \times m$ generalised 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 $\boldsymbol{G}\boldsymbol{G}_{\perp} = \boldsymbol{0}$, $\boldsymbol{G}^* = \begin{pmatrix} \boldsymbol{G}^{-} \vdots \boldsymbol{G}_{\perp} \end{pmatrix}$, 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 A. □

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 $\mathbb{1}\{.\}$ equals 1 when the statement in braces is true and 0 otherwise and,

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

and $\boldsymbol{S}^{-}$ is an $m \times n$ generalised 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 $\boldsymbol{S}'_{\perp}\boldsymbol{S} = \boldsymbol{0}$.

Proof. See Appendix A. □

Example: Gaussian Distribution

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

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

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

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

where

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

Noting that

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

where $\boldsymbol{G}_{\perp}^{-}$ is an $(n-m)\times n$ matrix such that $\boldsymbol{G}_{\perp}^{-}\boldsymbol{G}_{\perp} = \boldsymbol{I}$, q can be rearranged as

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

After the change of variables, the density can be recognised as a multivariate Gaussian in $\boldsymbol{b}$ and $\boldsymbol{a}$. The mean and covariance matrix for the margins of 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 describe how the true predictive distribution can be recovered for elliptical distributions via linear reconciliation and a translation. Let the base probabilistic forecast be from the elliptical class with location parameter $\hat{\boldsymbol{\mu}}$ and scale matrix $\hat{\boldsymbol{\Sigma}}$. The objective is to

obtain a mean vector and covariance matrix for the reconciled distribution that are equal to those from the true DGP which we denote as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively.

If reconciliation is carried out by pre-multiplying by an arbitrary matrix $\boldsymbol{SG}$ then the location and scale parameters of the reconciled distribution are $\boldsymbol{SG}\hat{\boldsymbol{\mu}}$ and $\boldsymbol{SG}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\boldsymbol{S}'$ respectively. If $\boldsymbol{\Omega}$ is the true scale matrix for the bottom-level series then the correct scale matrix can be recovered by setting $\boldsymbol{G}_0 = \boldsymbol{\Omega}^{1/2}\hat{\boldsymbol{\Sigma}}^{-1/2}$ where $\hat{\boldsymbol{\Sigma}}^{1/2}$ is any matrix such that $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}^{1/2}(\hat{\boldsymbol{\Sigma}}^{1/2})'$, for example a Cholesky factor. To ensure conformability of matrix multiplication, $\boldsymbol{\Omega}^{1/2}$ must be a $m \times n$ matrix so can be set to the Cholesky factor of $\boldsymbol{\Omega}$ augmented with an additional $n - m$ columns of zeros.

We note that in general $\boldsymbol{SG}_0$ is not a projection matrix. As a consequence, even if the base forecasts are unbiased (i.e. $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$) the reconciled forecasts will not be since $\boldsymbol{SG}_0\hat{\boldsymbol{\mu}} \neq \boldsymbol{\mu}$. To recover the correct mean, reconciliation should also include translation by $\boldsymbol{d}_0 = \boldsymbol{\mu} - \boldsymbol{SG}_0\hat{\boldsymbol{\mu}}$.

Although $\boldsymbol{SG}_0$ is not a projection matrix in general, there are some conditions under which it will be. These are described by the following theorem.

Theorem 3.3 (Optimal Projection for Reconciliation)**.** *Let $\hat{\boldsymbol{\Sigma}}$ be the scale matrix from an elliptical but incoherent base forecast and assume base forecasts are also unbiased. When the true predictive is also elliptical, then this can be recovered via reconciliation using a projection if rank$(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \leq n - m$.*

Proof. See Appendix B. □

4 Sample based solution

In practice it is often the case that samples are drawn from a probabilistic forecast since an analytical expression is either unavailable, or relies on unrealistic parametric assumptions. A useful result is the following

Theorem 4.1 (Reconciled samples)**.** *Suppose that $\left(\hat{\boldsymbol{y}}^{[1]}, \ldots, \hat{\boldsymbol{y}}^{[L]}\right)$ is a sample drawn from an incoherent probability measure $\hat{\nu}$. Then $\left(\tilde{\boldsymbol{y}}^{[1]}, \ldots, \tilde{\boldsymbol{y}}^{[L]}\right)$ where $\tilde{\boldsymbol{y}}^{[l]} := \psi(\hat{\boldsymbol{y}}^{[l]})$ for all*

$l = 1, \ldots, L$ is a sample drawn from the reconciled probability measure $\tilde{\nu}$ as defined in Definition 2.2

Proof. For any $\mathcal{A} \in \mathscr{F}_{\mathfrak{s}}$

$$\begin{aligned}
\Pr(\hat{\boldsymbol{y}} \in \psi^{-1}(\mathcal{A})) &= \lim_{L\to\infty} \sum_{l=1}^{L} \mathbb{1}\left\{\hat{\boldsymbol{y}}^{[l]} \in \psi^{-1}(\mathcal{A})\right\} \\
&= \lim_{L\to\infty} \sum_{l=1}^{L} \mathbb{1}\left\{\psi(\hat{\boldsymbol{y}}^{[l]}) \in (\mathcal{A})\right\} \\
&= \Pr(\tilde{\boldsymbol{y}} \in (\mathcal{A}))
\end{aligned}$$

□

This result implies that reconciling each member of a sample drawn from an incoherent distribution provides a sample from the reconciled distribution. Such a strategy has already been used by Jeon et al. (2019), without formal justification.

In the point forecasting setting it is common to generate forecasts from independent statistical models, allowing forecast reconciliation to be scaled up to large hierarchies. Since a single point forecast is generated for each series there is no ambiguity as to how these should be stacked into a vector of base forecasts $\hat{\boldsymbol{y}}$. This is no longer the case in the probabilistic setting where multiple draws are made for each series. Jeon et al. (2019) recommend some crude heuristics that can be used to form joint samples. In Section 4.1 we provide an alternative based on resampling forecast errors. This provides a sample from a base probabilistic forecast. A sample from the reconciled distribution can be obtained by premultiplying these draws by $\boldsymbol{SG}$. Strategies for choosing $\boldsymbol{G}$ are discussed in Section 6.3.1.

4.1 Drawing base probabilistic forecasts

The following method only requires that one-step ahead in-sample 'forecasts' $\hat{y}_{i,t}$ can be computed for each series i. These are not true forecasts since they are computed over the training data $t = 1, \ldots, T$; for instance $\hat{y}_{i,t}$ may be the fitted values of $E(y_{i,t}|y_{i,t-1})$ implied by some statistical model. Let the errors $e_{i,t} = y_{i,t} - \hat{y}_{i,t}$ be stacked in a $(n \times T)$ matrix $\boldsymbol{\mathcal{E}} := \{e_{i,t}\}_{i=1,\ldots,n,t=1,\ldots,T}$. When up to H step ahead forecasts are required, we randomly

select H consecutive columns from $\boldsymbol{\mathcal{E}}$ to form $\boldsymbol{\mathcal{E}}^b$ and repeat for $b = 1, \ldots, B$. This preserves both serial correlation of errors as well as dependence across different series.

The $\boldsymbol{\mathcal{E}}^b$ can then be used together with forecasts $\hat{\boldsymbol{y}}_{T+h}$ for $h = 1, \ldots, H$ to recursively form B sample paths of the full hierarchy. As a simple example consider the case where each forecast comes from (the same) AR(1) model and $H = 2$. Then $\hat{y}^b_{i,T+1} = \phi y_{i,T} + e^b_{i,1}$ and $\hat{y}^b_{i,T+2} = \phi \hat{y}^b_{i,T+1} + e^b_{i,2}$ for each $i = 1, \ldots, n$ and $b = 1, \ldots, B$ and where $e^b_{i,h}$ is the element in row i and column h of $\boldsymbol{\mathcal{E}}^b$. Each future path can be stacked into a vector $\hat{\boldsymbol{y}}^b_{T+h} := (\hat{y}^b_{1,T+h}, \ldots, \hat{y}^b_{n,T+h})'$ which in turn are stacked into a $n \times B$ matrix $\hat{\boldsymbol{\Upsilon}}_{T+h} = (\hat{\boldsymbol{y}}^1_{T+h}, \ldots, \hat{\boldsymbol{y}}^B_{T+h})$.

As discussed previously, reconciling each draw from an incoherent distribution yields a sample from the reconciled distribution. Therefore setting $\tilde{\boldsymbol{\Upsilon}}_{T+h} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\Upsilon}}_{T+h}$ yields a matrix whose columns $\tilde{\boldsymbol{y}}^b_{T+h} := \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}^b_{T+h}$ are a sample from the reconciled distribution. In principle $\boldsymbol{G}$ matrices from popular point forecasting reconciliation methods can be used. In Section 6.3.1, we propose a way to find an optimal $\boldsymbol{G}$ for reconciling future paths by minimising a proper multivariate scoring rule.

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. In general, a scoring rule $K(.,.)$, is a function taking a probability measure as the first argument and a realisation as the second argument (although for ease of notation we will at times use the replace the probability measure with its associated density in the first argument). A scoring rule is *proper* if $\mathrm{E}_Q[K(Q, \boldsymbol{\omega})] \leq \mathrm{E}_Q[K(P, \boldsymbol{\omega})]$ for all P, where P is any member of some class of probability measures (densities), Q is the true predictive and ω is a realisation. When this inequality is strict for all $P \neq Q$, the scoring rule is said to be *strictly proper*.

Since hierarchical forecasting is by its very nature a multivariate problem (the linear

constraints affect all variables), our focus is on multivariate scoring rules. We consider the log score (LS), energy score (ES) and the variogram score (VS). However, in our simulations we evaluate individual margins of interest using the univariate counterparts of the log score and energy score (where the latter is the continuous ranked probability score, CRPS).

The log score simply involves evaluating the negative log density at the value of the realisation, $\text{LS}(P, \boldsymbol{\omega}) = -\log f(\boldsymbol{\omega})$, where f 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 this density can also be approximated from a sample of values drawn from the probabilistic forecast (see Jordan et al. 2017). The energy score is most commonly calculated using the following representation

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

where $\boldsymbol{y}$ and $\boldsymbol{y}^*$ are independent copies drawn from the distribution P, and we follow common convention by setting $\alpha = 0.5$. These expectations can be easily approximated via Monte Carlo with samples drawn from the probabilistic forecast.

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

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

where y_i and y_j are the i^{th} and j^{th} elements of $\boldsymbol{y} \sim P$ and we follow common convention by setting $p = 0.5$.

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 is to compare two reconciled probabilistic forecasts to one another, to evaluate which choice of $\boldsymbol{G}$ performs best in practice.

5.1 Scoring Rules for Hierarchical Time Series

When a expression for the density of an incoherent base forecast is available, Section 3 describes how the density of a reconciled forecast can be recovered. With both densities available, the log score is natural and straightforward scoring rule to use. However, the following theorem shows that the log score is improper in the setting of comparing incoherent to coherent forecasts.

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 C. □

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

If a probabilistic forecast is 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 discuss how this depends on the specific scoring rule used.

Consider a coherent probabilistic forecast with density $\tilde{f}_{\boldsymbol{y}}$ for the full hierarchy and $\tilde{f}_{\boldsymbol{b}}$ for the bottom-level series. By Theorem 3.2, $\tilde{f}_{\boldsymbol{y}}(\boldsymbol{y}) = |\boldsymbol{S}^*|\tilde{f}_{\boldsymbol{b}}(\boldsymbol{S}^-\boldsymbol{y})\mathbb{1}\boldsymbol{y} \in \mathfrak{s}$. 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

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

For coherent densities, the log score for the full hierarchy differs from the log score for the bottom-level series only by $-\log|\boldsymbol{S}^*|$. This term is independent from the choices of $\boldsymbol{G}$. As such, rankings of different reconciliation methods using the log score for the full hierarchy will not change if only the bottom-level series is used.

Table 1: Properties of scoring rules for reconciled probabilistic forecasts.

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

The same property does not hold for all scores. For example, the energy score is invariant under orthogonal transformations (Székely & Rizzo 2013, Gneiting & Raftery 2007) but not true under linear transformations in general. Therefore it is possible for one method to outperform another when energy score is calculated using the full hierarchy, but for these ranking to reverse if only bottom-level series are considered. We therefore recommend computing the energy score using the full hierarchy. The same reasoning holds for the variogram score. The properties of multivariate scoring rules in the context of evaluating reconciled probabilistic forecasts are summarised in Table 1.

6 Simulations

6.1 Data Generating Processes and Setup

The aim of the simulations that follow is to assess the performance of both the analytical and sample based reconciliation approaches using a number of different choices of $\boldsymbol{G}$. The data generating process we consider corresponds to the 3-level hierarchical structure presented in Figure 1. Bottom-level series are first generated from ARIMA(p, d, q) processes, which are in-turn aggregated to form the middle and top-level series. We consider a multivariate Gaussian and a non-Gaussian setting for the errors driving the ARIMA processes. More specifically the non-Gaussian errors are drawn from a meta-distribution of a Gumbel copula with Beta margins. Parameter values are chosen so that bottom level series have a lower signal-to-noise ratio that top level series with specific details provided in Appendix D. For each series we generate 2002 observations from which the first 500 are ignored to avoid the impact of initial values. To investigate the impact of model mispecification, both

Gaussian and non-Gaussian probabilistic forecasts will be evaluated for each DGP.

All forecasts are evaluated using a rolling window with a training sample of 500. This yields 1,000 probabilistic forecasts for evaluation. The predictive accuracy of the alternative methods is evaluated using the scoring rules presented in Section 5. Results are reported in terms of skill scores, i.e., the percentage improvement of a probabilistic forecast over a reference method. A positive (negative) value indicates a percentage gain (loss) in forecast accuracy over the reference method.

6.2 Base Forecasts

We fit univariate ARIMA and exponential smoothing (ETS) models to each series using the `auto.arima()` and `ets()` functions in the `forecast` package (Hyndman et al. 2019) in R (R Core Team 2018). Since the conclusions were similar for both modelling approaches, the results for ETS are omitted here but are available in an online supplement. For each series, (incoherent) point forecasts are obtained for $h = 1, 2$ and 3-steps ahead. We denote these as $\hat{\boldsymbol{\mu}}_{T+h}$. For $h = 2$ and $h = 3$ these are obtained using the recursive method (Hyndman & Athanasopoulos 2018). A Gaussian base probabilistic forecast is obtained as $N(\hat{\boldsymbol{\mu}}_{T+h}, \hat{\boldsymbol{\Sigma}})$ with $\hat{\boldsymbol{\Sigma}}$ estimated using in-sample one-step ahead forecast errors and the shrinkage estimator of Schäfer & Strimmer (2005) (as presented in Table 2). A sample from the non-Gaussian probabilistic forecast is obtained using the bootstrap method described in Section 4.1.

6.3 Reconciliation

For Gaussian probabilistic forecasts, the reconciled forecast will be of the form $N(\boldsymbol{SG}\hat{\boldsymbol{\mu}}, \boldsymbol{SG}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\boldsymbol{S}')$, as described in Section 3. For non-Gaussian probabilistic forecasts all draws are premultiplied by $\boldsymbol{SG}$ as described in Section 4. A number of alternative projections were considered for reconciliation and are summarised in Table 2.

Table 2: Summary of reconciliation methods for which $\boldsymbol{SG}$ is a projection matrix, with $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}^{-1}$. The rows present alternative choices of $\boldsymbol{W}$. $\hat{\boldsymbol{W}}^{\text{sam}}$ is the sample variance-covariance matrix estimated from one-step ahead in-sample forecast errors from the base incoherent forecasts. $\hat{\boldsymbol{W}}^{\text{wls}}$ is a diagonal matrix with elements from $\hat{\boldsymbol{W}}^{\text{sam}}$. $\hat{\boldsymbol{W}}^{\text{shr}}$ is a shrinkage estimator proposed by Schäfer & Strimmer (2005), where $\tau = \frac{\sum_{i\neq j}\hat{\text{Var}}(\hat{w}_{ij})}{\sum_{i\neq j}\hat{w}_{ij}^2}$ and w_{ij} denotes the (i,j)th element of $\hat{\boldsymbol{W}}^{\text{sam}}$. Relevant references are presented at the end of Section 2.1.

Reconciliation method	$\boldsymbol{W}$
OLS	$\boldsymbol{I}$
WLS	$\text{diag}(\hat{\boldsymbol{W}}^{\text{sam}})$
MinT(Sample)	$\hat{\boldsymbol{W}}^{\text{sam}}$
MinT(Shrink)	$\tau\text{diag}(\hat{\boldsymbol{W}}^{\text{sam}}) + (1-\tau)\hat{\boldsymbol{W}}^{\text{sam}}$

6.3.1 Score Optimal Reconciliation

In general, the $\boldsymbol{G}$ matrix that minimises expected score may depend on both the base forecast, true predictive and specific scoring rule used. In the following section we also propose an approach to find a $\boldsymbol{G}$ by optimising the average score over a rolling window constructed within the training data. The respective objective function can be written as,

$$\underset{\boldsymbol{G}_h}{\text{argmin}} \quad \text{E}_Q[K(\tilde{\nu}_{\boldsymbol{G}_h}, \boldsymbol{y}_{T+h})], \tag{1}$$

where $\tilde{\nu}_{\boldsymbol{G}_h}$ is the measure reconciled with respect to $\boldsymbol{G}_h$ (the subscript h emphasises distinct matrices for different forecast horizons), Q is the true predictive distribution so that $\boldsymbol{y}_{T+h} \sim Q$ and $K(.,.)$ is a proper scoring rule as described in Section 5.

To approximate the expectation in Equation (1) we break up the training sample into J rolling windows each of length $T-J+1$. For the j^{th} rolling window we use $\boldsymbol{y}_j, \ldots, \boldsymbol{y}_{T-J+j}$ to evaluate a h-step ahead base probabilistic forecast $\hat{\nu}$. Reconciling with respect to $\boldsymbol{G}_h$ yields the distribution $\tilde{\nu}_{\boldsymbol{G}_h}$. The expectation in Equation (1) can then be approximated by

$$\mathrm{E}_Q[K(\tilde{\nu}_{\boldsymbol{G}_h}, \boldsymbol{y}_{T+h})] \approx \frac{1}{J}\sum_{j=1}^{J} K(\tilde{\nu}_{\boldsymbol{G}_h}, \boldsymbol{y}_{T-J+j+h}).$$

Where this approach is used in simulations, we set $J = 100$ and use a slightly larger training window of $T = 600$ for selecting $\boldsymbol{G}$ only. We implement the algorithm in the Gaussian setting (see Section 6.4) using log score and in the non-Gaussian setting (see Section 6.5) using energy score. We also note that different ways of parameterising $\boldsymbol{G}_h$ were used in the numerical optimisation but all led to identical results. This is discussed in detail in Section F in Appendix.

6.4 Results for Gaussian probabilistic forecasts

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. Hence, we calculate the skill scores with reference to the bottom-up forecasts in all cases leaving blank the cells for log score of the incoherent forecasts. Further, all log scores are evaluated on the basis of bottom-level series only, as these differ from the log scores for the full hierarchy by a fixed constant. Please refer to Table 1 for further explanations.

All reconciliation approaches that use information from all levels of the hierarchy improve on the incoherent base forecasts. The bottom-up that uses information only from the bottom-level does not improve on the base incoherent forecasts when considering the Energy Score. Overall, the MinT methods provide the most improvement over the incoherent base forecasts irrespective of the scoring rule or the nature of the DGP.

Tables 4 (and 6 and 7 in Appendix E break down the forecasting performance of the different methods by considering univariate scores on each individual margin. 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: Forecast evaluation of hierarchical probabilistic forecasts for $h = 1, 2$ and 3 steps ahead. Entries show the percentage (%) skill score with reference to the bottom-up method. A positive (negative) entry shows a gain (loss) in forecast accuracy over the reference method. The top panel shows the results from the Gaussian DGP while the bottom panel shows the results from the non-Gaussian DGP. The Log Score entries are based on the joint forecast distribution for the bottom-level (see Table 1 for a further explanation), while the Energy and Variogram skill scores are based on the joint forecast distribution across the entire hierarchy.

	Log Score (%)			Energy Score (%)			Variogram Score (%)		
h	1	2	3	1	2	3	1	2	3
Gaussian DGP									
Incoherent base				12.46	9.58	7.19	4.91	5.89	6.02
Bottom-up	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OLS	5.56	-1.55	-10.74	16.83	13.43	10.44	9.17	9.63	9.02
WLS	6.66	-2.79	-15.79	19.06	15.34	11.88	11.52	12.24	11.62
MinT(Sample)	**7.73**	-2.59	-17.07	**21.55**	**17.36**	13.60	**15.74**	15.75	15.51
MinT(Shrink)	7.70	-2.63	-17.12	21.44	17.33	13.60	15.65	**15.98**	**15.58**
Optimal	5.31	**2.86**	**4.28**	8.08	13.75	**15.10**	-14.12	3.60	6.57
Non-Gaussian DGP									
Incoherent base				7.18	7.37	7.13	-0.44	-0.35	-0.22
Bottom-up	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OLS	12.11	7.48	-1.61	9.73	10.28	10.18	0.50	0.52	0.87
WLS	19.80	1.56	-39.14	11.20	12.00	11.52	0.49	0.46	0.62
MinT(Sample)	**22.89**	4.69	-36.42	**12.94**	14.07	**13.44**	**1.78**	2.16	2.07
MinT(Shrink)	22.67	4.43	-36.76	12.92	**14.13**	13.39	1.63	**2.18**	**2.20**
Optimal	19.15	**9.26**	-4.11	-3.31	3.48	2.41	-27.79	-19.37	-22.70

Table 4: Forecast evaluation of univariate hierarchical probabilistic forecasts for $h = 1$-step ahead. Entries show the percentage (%) skill score base on the Log Score and CRPS with reference to the incoherent base forecasts. A positive (negative) entry shows a gain (loss) in forecast accuracy over the incoherent base forecasts. The left panel shows the results from the Gaussian DGP while the right panel shows the results from the non-Gaussian DGP.

	Gaussian							Non-Gaussian						
Series	Tot	A	B	AA	AB	BA	BB	Tot	A	B	AA	AB	BA	BB
						Log Score (%)								
Incoherent 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
Bottom up	-29.59	-2.99	-4.24	0.00	0.00	-5.94	8.11	-263.36	-0.72	-3.54	0.00	0.00	-0.81	-12.22
OLS	12.26	2.51	0.43	1.80	1.06	-3.41	7.59	49.53	0.38	2.58	-1.40	3.51	0.55	-11.33
WLS	**16.10**	-3.73	-14.04	2.34	1.27	-3.08	0.45	75.34	4.86	-269.27	-1.38	3.36	0.91	-5.38
MinT(Sample)	-0.07	**4.80**	**0.54**	**3.86**	**2.55**	-2.34	**8.20**	0.24	0.44	**5.18**	**0.61**	3.73	**2.46**	-10.21
MinT(Shrink)	-0.05	**4.80**	**0.54**	3.65	2.48	-2.36	**8.20**	0.27	**0.45**	**5.18**	-1.42	**3.75**	2.30	-10.32
Optimal	13.86	1.37	-18.41	-0.28	-1.75	-6.75	4.13	**70.23**	-5.37	-287.34	-6.99	-2.02	-4.07	-17.38
						CRPS (%)								
Incoherent 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
Bottom up	-134.06	-10.99	-15.42	-0.05	-0.14	-21.39	25.07	-485.01	-1.41	-9.35	-0.01	0.08	-1.85	-36.37
OLS	35.17	8.45	1.17	6.49	3.07	-11.45	23.77	74.22	1.30	6.77	-3.85	9.21	1.82	-33.37
WLS	**43.24**	-13.62	-50.06	8.16	3.82	-10.32	1.12	87.17	**12.20**	-503.54	-3.90	8.85	2.85	-16.01
MinT(Sample)	-0.20	**15.84**	**1.75**	**12.48**	**7.98**	-7.53	**25.38**	0.24	1.34	12.79	**1.39**	9.67	**6.70**	-29.83
MinT(Shrink)	-0.19	15.69	1.59	12.06	7.76	-7.61	25.35	0.26	1.52	**12.88**	-3.82	**9.84**	6.38	-30.00
Optimal	37.22	3.37	-74.40	-2.85	-8.50	-27.27	11.28	**85.14**	-16.83	-599.74	-21.16	-7.42	-13.08	-55.62

6.5 Results for non-Gaussian probabilistic forecast

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 5 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

Table 5: 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.

	Energy Score						Variogram Score					
	Gaussian DGP			Non-Gaussian DGP			Gaussian DGP			Non-Gaussian DGP		
h	1	2	3	1	2	3	1	2	3	1	2	3
Base	11.7	14.6	17.8	5.71	5.94	6.27	5.56	6.61	7.87	1.28	1.37	1.49
OLS	11.1	13.7	16.7	5.51	5.70	5.98	4.86	5.35	6.05	1.23	1.30	1.40
WLS	10.7	13.2	16.0	5.43	5.60	5.89	4.86	5.35	5.86	1.23	1.30	1.40
MinT(Shrink)*	10.5	12.8	15.7	5.33	5.50	5.77	4.77	5.24	5.86	1.19	1.26	1.34
Optimal*	10.6	12.9	15.7	5.36	5.51	5.83	4.85	5.30	5.86	1.21	1.27	1.40

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

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.

7 Forecasting Australian domestic tourism flows

Previous studies have shown that point forecast reconciliation can generate more accurate forecasts compared to incoherent base forecasts or traditional methods such as bottom-up for forecasting Australian tourism flows (see for example, Athanasopoulos et al. 2009, Hyndman et al. 2011, Wickramasuriya et al. 2019). This study is the first to apply reconciliation methods for forecasting Australian tourism in a probabilistic framework. We use "overnight trips" to different destinations across the country as a measure of tourism flows. These naturally disaggregate through a geographic hierarchy consisting of 7 states, 27 zones and 76 regions. Hence, this 3-level hierarchy comprises 111 series in total. More details about the data and the individual series are provided in Appendix H.

We consider monthly data for all series spanning the period January 1998 to December 2018. This gives 252 observations per series. Using a rolling window of 100 observations as the training sample we generate incoherent base probabilistic forecasts for $h = 1$ to 12-steps ahead from univariate ARIMA and ETS models for each series using `auto.arima()` and `ets()` from the `forecast` package (Hyndman et al. 2019) in R software (R Core Team 2018).

We apply both the analytic, by assuming Gaussian incoherent base forecasts, and the non-parametric sampling reconciliation approaches discussed in Sections 3 and 4 respectively. We do not implement the MinT(Sample) approach as the sample size of the training data is smaller than the dimension of the hierarchy. The process is repeated by rolling the training window forward one month at a time 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. In what follows we only present the results for ARIMA. The results for ETS lead to similar conclusions and are available upon request.

Figure 3 shows energy and variogram scores across the entire hierarchy for the different reconciliation methods, calculated over the rolling windows. Results from the analytic approach are presented on the left. Results from the non-parametric sampling approach are presented on the right. Recall that we follow negatively-oriented scoring rules so that a lower (higher) score implies a more (less) accurate forecast. For both scoring rules all

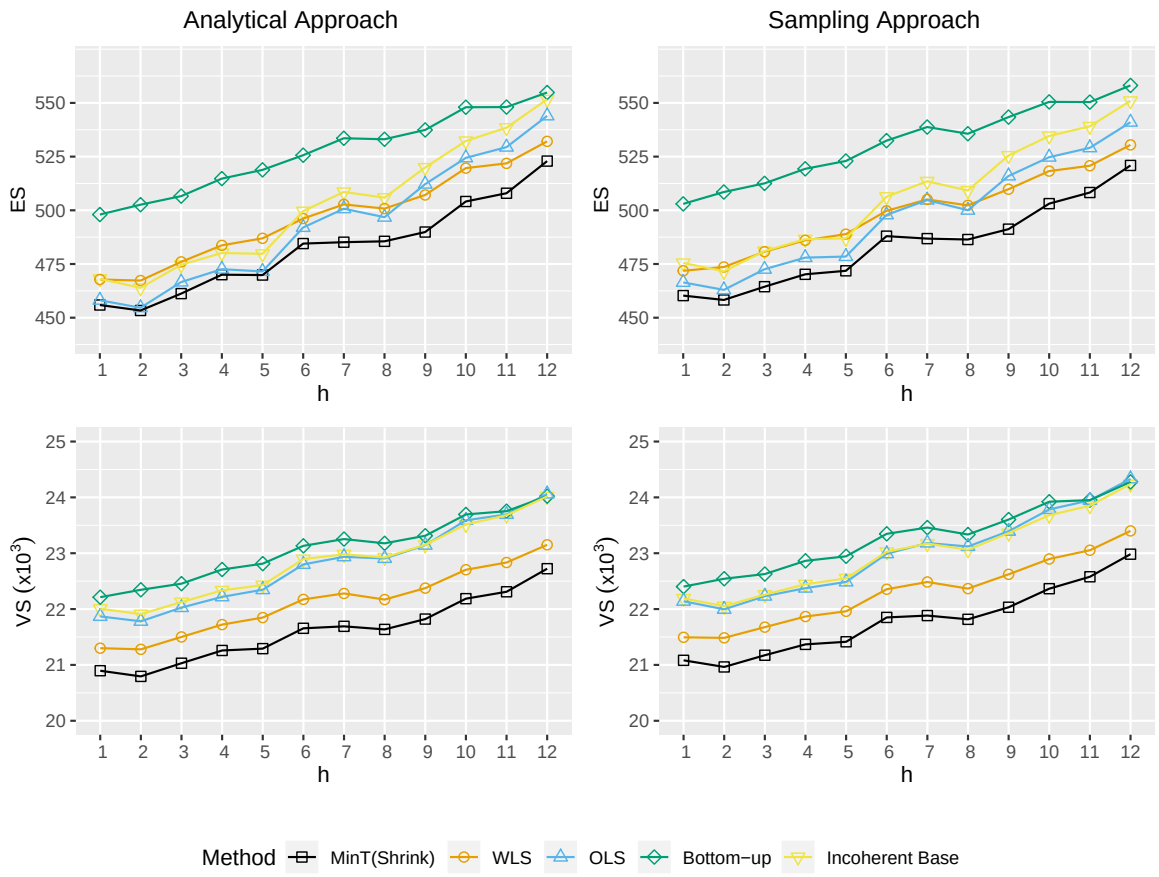

Figure 3: Energy and variogram scores for multivariate predictive distributions across the entire hierarchy. A lower (higher) score indicates a more (less) accurate forecast. Results from the analytic approach assuming Gaussian incoherent base forecasts are presented on the left while results from the non-parametric approach are presented on the right.

forecasts from the analytic Gaussian approach are more accurate (although marginally) than the forecasts from the non-parametric sampling approach. A result also verified when comparing scores across each level of the hierarchy. This indicates that assuming Gaussianity for the incoherent base forecasts and using the analytic approach is adequate for this data set. Hence, in what follows we concentrate only on the analytic Gaussian reconciliation results. All other results are available upon request.

As shown in Figure 3, applying MinT(Shrink) for reconciliation of incoherent base forecasts generates the most accurate forecasts in all cases. As expected accuracy for all

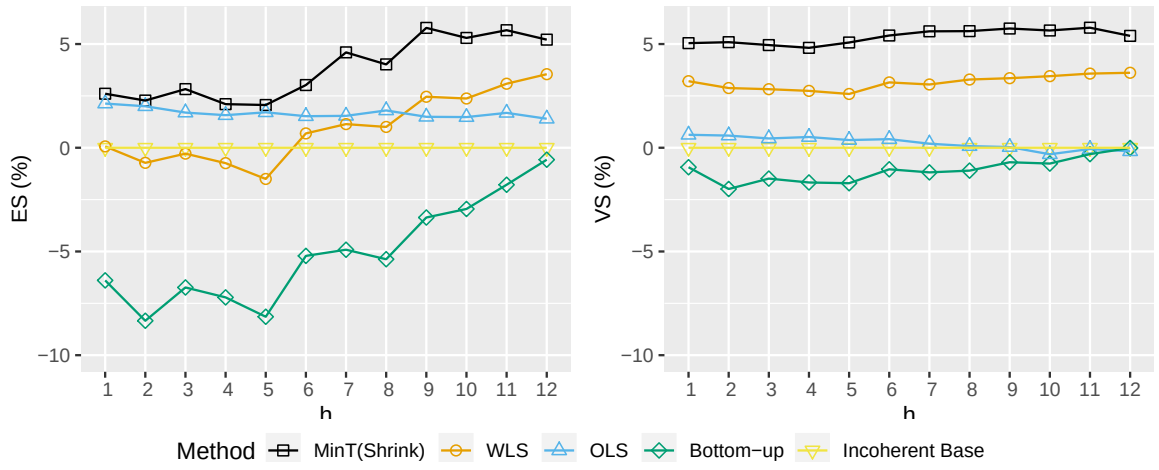

Figure 4: Skill scores (%) relative to incoherent base forecasts, across the entire Australian tourism hierarchy based on energy score (on the left) and variogram score (on the right). A higher (lower) score indicates a gain (loss) in forecast accuracy relative to the incoherent base forecasts. The results are for the analytic solution assuming Gaussian incoherent base forecasts.

forecasts deteriorates as the forecast horizon increases. The skill scores presented in Figure 4 for the analytic solution show improvements upon the incoherent Gaussian base forecasts across the hierarchy as a whole. The improvements start at 2.5% for $h = 1$ and increase to above 5% for $h \geq 9$ for the energy score and are consistently above 5% for the variogram score. Note that in all cases the bottom-up forecasts are always inferior to the incoherent base forecasts. This comes to no surprise reflecting upon the fact that the bottom-level series are the noisiest and most challenging to forecast and information is lost when levels above are not considered as with a reconciliation approach.

Figure 5 shows skill scores (%) for the Gaussian analytic approach, relative to the incoherent base forecasts across each level of the Australia tourism hierarchy (please see Figure 9 in Appendix H for the raw scores). The top-panel presents the results for the aggregate level based on the CRPS, the univariate equivalent to the Energy score. For the levels below skill scores based on both the energy and variogram scores are presented.

Based on both scoring rules MinT(Shrink) improves upon the incoherent base forecasts at all levels and all forecast horizons (the only exception being forecast horizons 4 and 5 at

the top-level for which a marginal loss is shown). The improvements for the top-level seem to be higher for the longer forecast horizons, increasing to 5% or more for $h \geq 7$. For the levels below the improvements seem to be more homogenous across the forecast horizons. Based on the energy score improvements are around 3%, 4% and 2% for States, Zones and Regions respectively. These are considerably higher based on the variogram score for which gains around 10%, 7% and 3% are shown. We could comment on the bottom-up and the OLS results but I am not sure it is worth it.

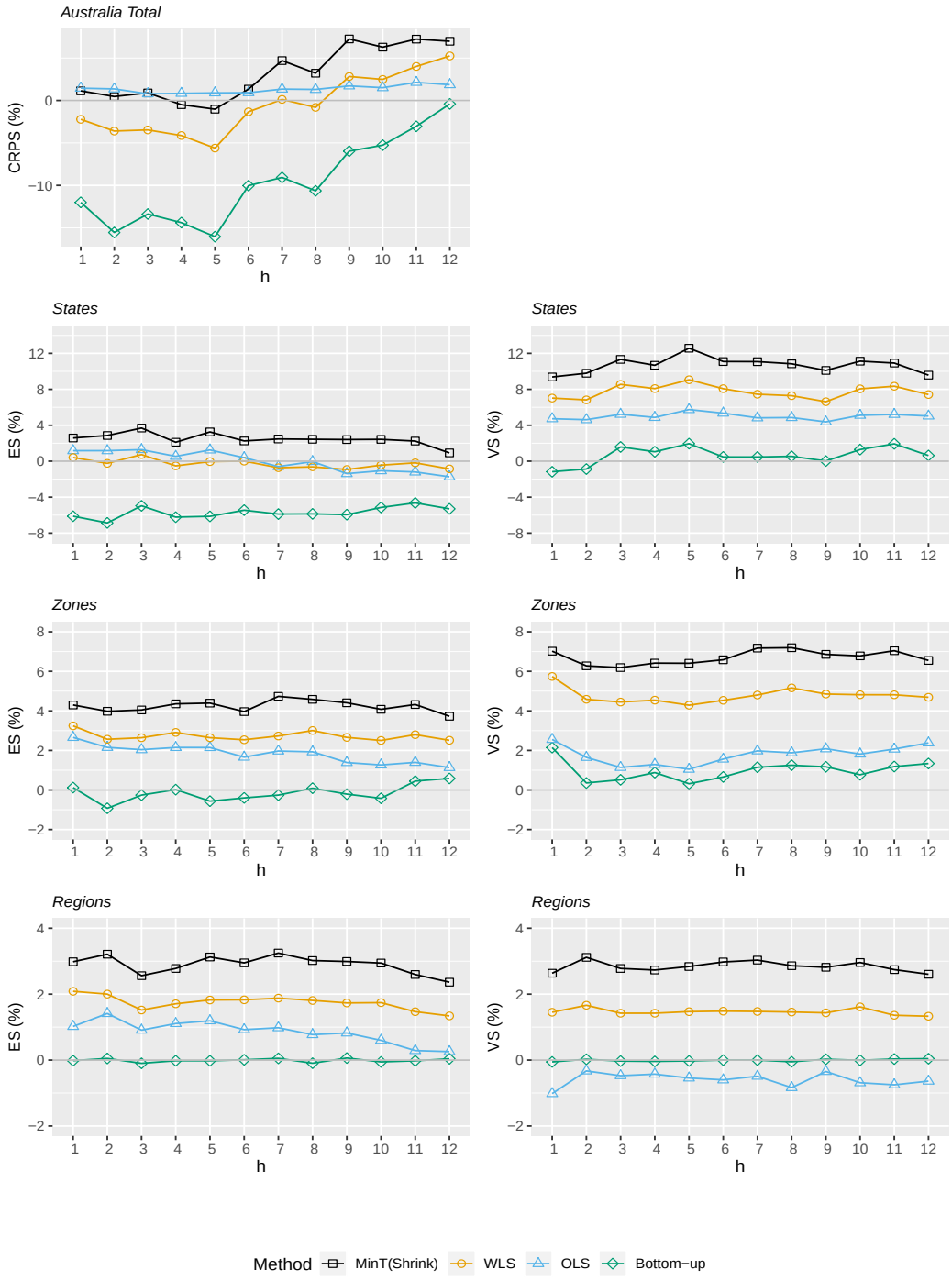

Figure 5: Skill scores (%) relative to incoherent base forecasts, for the CRPS for the top-level and energy and variogram scores for the levels below for the Australia tourism hierarchy. A higher (lower) score indicates a gain (loss) in forecast accuracy relative to the incoherent base forecasts. All results are for the analytic solution assuming Gaussian incoherent base forecasts.

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 region $\mathcal{I}$ given by the Cartesian product of intervals $(l_1, u_1), (l_2, u_2), \ldots (l_m, u_m)$. We derive the probability, under the reconciled measure, that the bottom-level series lie in $\mathcal{I}$, i.e. $\Pr(\boldsymbol{l} \succ \boldsymbol{b} \succ \boldsymbol{u})$, 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 $\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{y} \succ \boldsymbol{u}$. Using 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 $\boldsymbol{y} = \boldsymbol{G}^*\boldsymbol{z}$. Recall, $\boldsymbol{G}^* = \left(\boldsymbol{G}^- \vdots \boldsymbol{G}_\perp\right)$, $\boldsymbol{G}^-$ is a generalised 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

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

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

$$\tilde{f}_{\boldsymbol{b}}(\boldsymbol{b}) = \int \hat{f}(\boldsymbol{G}^-\boldsymbol{b} + \boldsymbol{G}_\perp\boldsymbol{a}) |\boldsymbol{G}^*| d\boldsymbol{a}\,, \tag{2}$$

is the density of $\boldsymbol{b}$. To obtain the density of the full hierarchy we first augment the density in Equation (2) by $n - m$ variables denoted $\boldsymbol{u}$

$$f(\boldsymbol{b}, \boldsymbol{u}) = \tilde{f}_b(\boldsymbol{b}) \mathbb{1}\left\{\boldsymbol{u} = 0\right\} , \tag{3}$$

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

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

where $\boldsymbol{S}_{\perp}^{-}$ is a generalised inverse such that $\boldsymbol{S}_{\perp}' \boldsymbol{S}_{\perp}^{-} = \boldsymbol{I}$ and noting the inverse of $\left(\boldsymbol{S} \vdots \boldsymbol{S}_{\perp}\right)$ is given by

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

it can be seen that $\boldsymbol{b} = \boldsymbol{S}^{-}\boldsymbol{y}$ and $\boldsymbol{u} = \boldsymbol{S}_{\perp}'\boldsymbol{y}$. Applying this change of variables yields the density

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

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

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

B Proof of Theorem 3.3

Let

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \boldsymbol{D} = \boldsymbol{S}\boldsymbol{\Omega}\boldsymbol{S}' + \boldsymbol{D}\,.$$

If reconciliation is carried out via a projection onto $\mathfrak{s}$, then $\boldsymbol{S}\boldsymbol{G}\boldsymbol{S} = \boldsymbol{S}$ and

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}} &= \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\boldsymbol{S}' \\
&\boldsymbol{S}\boldsymbol{G}\boldsymbol{S}\boldsymbol{\Omega}\boldsymbol{S}'\boldsymbol{G}'\boldsymbol{S}' + \boldsymbol{S}\boldsymbol{G}\boldsymbol{D}\boldsymbol{G}'\boldsymbol{S}' \\
&\boldsymbol{S}\boldsymbol{\Omega}\boldsymbol{S}' + \boldsymbol{S}\boldsymbol{G}\boldsymbol{D}\boldsymbol{G}'\boldsymbol{S}' \\
&\boldsymbol{\Sigma} + \boldsymbol{S}\boldsymbol{G}\boldsymbol{D}\boldsymbol{G}'\boldsymbol{S}'\,.
\end{aligned}$$

Therefore to recover the true predictive using a projection, some $\boldsymbol{G}_0$ must be found such that $\boldsymbol{G}_0\boldsymbol{D} = \boldsymbol{0}$. Let the eigenvalue decomposition of $\boldsymbol{D}$ be given by $\boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}'$, where $\boldsymbol{R}$ is an $n \times q$ matrix with $q = \text{rank}(\boldsymbol{D})$ and $\boldsymbol{\Lambda}$ is an $q \times q$ diagonal matrix containing non-zero eigenvalues of $\boldsymbol{D}$. By the rank nullity theorem, $\boldsymbol{R}$ will have an orthogonal complement $\boldsymbol{R}_\perp$ of dimension $n \times (n-q)$. If $q = n - m$ then the number of columns of $\boldsymbol{R}_\perp$ is m and $\boldsymbol{G}_0$ can be formed as the $m \times n$ matrix $(\boldsymbol{R}_\perp'\boldsymbol{S})^{-1}\boldsymbol{R}_\perp'$. If $q < n - m$ the number of columns of $\boldsymbol{R}_\perp$ is greater than m, and any m columns of $\boldsymbol{R}_\perp$ can be used to form $\boldsymbol{G}_0$ in a similar fashion. However when $q > n - m$, the number of columns of $\boldsymbol{R}_\perp$ is less than m and no such $m \times n$ matrix $\boldsymbol{G}_0$ can be formed. Therefore the true predictive can only be recovered via a projection when $\text{rank}(\boldsymbol{D}) \leq n - m$.

With respect to the location, if $\boldsymbol{S}\boldsymbol{G}$ is a projection then reconciled forecasts will be unbiased as long as the base forecasts are also unbiased. When base forecasts are biased they can be bias corrected before reconciliation as described by Panagiotelis et al. (2019) in the point forecasting setting.

C Proof of Theorem 5.1

The proof relies on the following change of variables,

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

Also recall from the proof of Theorem 3.2 that $\boldsymbol{S}^* = \left(\boldsymbol{S} \vdots \boldsymbol{S}_{\perp}\right)^{-1}$

Let the density of the true predictive $f(\boldsymbol{y})$ after a change of variables, be given by $|\boldsymbol{S^*}|^{-1} f_{\boldsymbol{b}}(\boldsymbol{b}) \mathbb{1}\{\boldsymbol{u} = \boldsymbol{0}\}$. To prove that the log score is improper we construct an incoherent base density $\hat{f}$ such that $E_f\left[LS\left(\hat{f}, \boldsymbol{y}\right)\right] < E_f\left[LS\left(f, \boldsymbol{y}\right)\right]$. This incoherent density is constructed, so that after the same change of variables it can be written as $|\boldsymbol{S^*}|^{-1} \hat{f}_{\boldsymbol{b}}(\boldsymbol{b}) \hat{f}_{\boldsymbol{u}}(\boldsymbol{u})$. We require $\hat{f}_{\boldsymbol{u}}(\boldsymbol{0}) > 1$, i.e., $\boldsymbol{u}$ is highly concentrated around $\boldsymbol{0}$ but still non-degenerate. An example is an independent normal with mean 0 and variances less than $(2\pi)^{-1}$. Now, let $\boldsymbol{y}^*$ be a realisation from f. Let the first m elements of $\boldsymbol{S}^* \boldsymbol{y}^*$ be $\boldsymbol{b^*}$, and the remaining elements be $\boldsymbol{u}^*$. The log score for f is thus,

$$\begin{aligned} LS\left(f, \boldsymbol{y}^*\right) &= -\log f(\boldsymbol{y}^*) \\ &= -\log |\boldsymbol{S^*}| - \log f_{\boldsymbol{b}}\left(\boldsymbol{b}^*\right) - \log\left(\mathbb{1}\left\{\boldsymbol{u}^* = \boldsymbol{0}\right\}\right) \\ &= -\log |\boldsymbol{S^*}| - \log f_{\boldsymbol{b}}\left(\boldsymbol{b}^*\right), \end{aligned} \tag{4}$$

where the third term in Equation 4 is equal to zero since the fact that $\boldsymbol{y}^* \in \mathfrak{s}$ implies that $\boldsymbol{u}^* = \boldsymbol{0}$. The log score for $\hat{f}$ 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, it will also hold after taking expectations (by the monotonicity of expectations). Thus $\hat{f}$ violates the condition for a proper scoring rule.

D Data generating process

The hierarchy considered in the simulations is the 2-level structure shown in Figure 1. Data are first generated for the bottom-level series from

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

where $w_{AA,t}, w_{AB,t}, w_{BA,t}, w_{BB,t}$ are ARIMA(p,d,q) processes with error terms $\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}$. (p,q) and (d) take integer values from $\{1,2\}$ and $\{0,1\}$ respectively with equal probability and the parameters for the AR and MA components are randomly and uniformly generated from $[0.3, 0.5]$ and $[0.3, 0.7]$ respectively. $u_t \sim \mathcal{N}(0, \sigma_u^2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$. Aggregating the bottom-level series gives the series for the aggregated levels, such that,

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

Gaussian errors

The errors driving the bottom-level ARIMA processes are jointly generated from a Normal distribution. More specifically, $\{\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}\} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})\ \forall t$. A commonly observed feature of hierarchical time series in practice, is that upper-level aggregated series are relatively less noisy than their corresponding lower-level components. This is due to the smoothing effect of aggregation, eliminating some random variation. Similarly to

Wickramasuriya et al. (2019), setting

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

and $\sigma_u^2 = 28$ and $\sigma_v^2 = 22$ ensures that the following inequalities are satisfied,

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

Non-Gaussian errors

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

$$C_\theta(e_1, e_2) = exp\{-[(-ln(e_1))^\theta + (-ln(e_2))^\theta]^{1/\theta}\}.$$

We generate random variates $\{e_{AA}, e_{AB}\}$ from $C_{\theta=10}(.)$ and $\{e_{BA}, e_{BB}\}$ from $C_{\theta=8}(.)$ for series $\{AA, AB\}$ and $\{BA, BB\}$ respectively. The ARIMA errors $\{\varepsilon_{AA}, \varepsilon_{AB}, \varepsilon_{BA}, \varepsilon_{BB}\}$ are generated as the quantiles from beta distributions with shape parameters $\alpha = 1$ and $\beta = 3$ that correspond to $\{e_{AA}, e_{AB}, e_{BA}, e_{BB}\}$. We then choose $\sigma_u^2 = 10$ and $\sigma_v^2 = 7$ such that they satisfy the inequalities explained above. Tas to double check. Do the inequalities hold for nonlinear dependence. Do we need to drop the last statement?

E Simulation results from parametric solution for marginal forecast distributions for $h = 2$ and $h = 3$

Table 6: 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 $h = 2$ 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

	Gaussian							Non-Gaussian						
Series	Tot	A	B	AA	AB	BA	BB	Tot	A	B	AA	AB	BA	BB
							Log Score (%)							
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
Bottom up	**9.87**	-0.72	-4.28	0.00	0.00	-10.02	**18.99**	-8.05	-0.30	-3.12	0.00	0.00	-1.46	-11.80
OLS	-22.52	1.31	0.50	1.81	0.25	-6.72	18.12	30.16	0.20	3.28	-1.53	4.33	0.18	-10.70
WLS	-27.45	-13.81	26.50	2.41	0.57	-6.17	5.10	12.65	**7.09**	-6.06	-1.55	4.11	0.71	-1.67
MinT(Sample)	-0.28	2.65	0.57	**3.51**	**1.80**	-5.36	18.86	-0.02	0.43	**6.80**	**0.64**	4.66	**2.44**	-9.36
MinT(Shrink)	-0.29	2.60	0.56	3.37	1.72	-5.34	18.90	-0.06	0.42	6.79	-1.50	**4.67**	2.26	-9.46
Optimal	-12.70	**5.42**	**26.71**	2.86	0.52	-6.42	17.11	**34.04**	-2.98	-7.45	-6.28	-0.19	-2.31	-14.48
							CRPS (%)							
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
Bottom up	-45.40	-5.11	-16.82	0.02	-0.05	-37.16	46.65	-185.36	-0.54	-8.18	0.08	-0.03	-3.54	-35.55
OLS	-14.62	6.88	1.64	7.06	0.78	-23.24	45.12	61.90	0.64	8.48	-4.32	11.30	1.10	-31.86
WLS	-10.18	-41.19	23.92	8.75	1.83	-21.19	14.60	70.13	**17.33**	-171.66	-4.30	10.76	2.52	-6.06
MinT(Sample)	-0.13	**12.14**	2.13	11.19	**6.18**	-17.67	46.49	0.26	1.29	16.54	**1.60**	11.95	**6.76**	-27.69
MinT(Shrink)	-0.08	12.03	2.05	**11.62**	5.74	-17.91	**46.64**	0.32	1.44	**16.71**	-4.15	**12.14**	6.54	-27.79
Optimal	-10.80	10.78	**22.61**	7.57	-0.96	-25.69	41.91	**71.52**	-8.89	-184.63	-18.29	-1.88	-7.13	-44.68

Table 7: 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 $h = 3$ 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

	Gaussian							Non-Gaussian						
Series	Tot	A	B	AA	AB	BA	BB	Tot	A	B	AA	AB	BA	BB
						Log Score (%)								
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
Bottom up	34.32	1.21	-3.97	0.00	0.00	-14.82	29.16	56.40	-0.15	-2.89	0.00	0.00	-1.73	-11.81
OLS	-67.52	-0.68	0.48	0.27	-0.27	-10.97	28.06	-5.91	0.47	2.87	-1.78	5.14	0.07	-10.68
WLS	-85.70	-26.29	50.39	0.68	0.12	-10.26	10.40	-129.42	10.02	57.99	-1.71	4.84	0.75	1.18
MinT(Sample)	-0.40	-1.26	0.45	2.29	1.61	-9.34	28.94	-0.09	0.65	5.79	0.75	5.58	2.48	-9.97
MinT(Shrink)	-0.42	-1.32	0.46	1.39	1.50	-9.27	28.99	-0.16	0.63	5.77	-1.74	5.60	2.34	-9.93
Optimal	-24.58	10.50	51.73	2.73	3.03	-8.64	27.28	-12.25	-2.65	57.70	-7.44	0.17	-2.67	-15.72
						CRPS (%)								
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
Bottom up	-22.10	-3.52	-16.10	-0.04	-0.09	-53.20	58.23	-80.43	-0.35	-8.48	-0.17	-0.05	-4.20	-36.03
OLS	-47.20	4.67	2.04	2.52	0.48	-36.36	56.33	44.08	1.46	7.70	-4.75	13.40	0.90	-31.65
WLS	-44.96	-60.44	42.85	3.92	1.71	-33.62	25.60	46.73	22.77	-66.37	-4.90	12.58	2.58	-0.12
MinT(Sample)	0.02	7.99	2.21	9.03	6.23	-29.83	57.85	0.21	1.88	14.24	1.88	14.20	6.81	-29.36
MinT(Shrink)	-0.03	7.89	2.34	6.61	6.05	-29.59	57.94	0.15	1.75	14.18	-4.92	14.34	6.61	-29.31
Optimal	-30.12	13.11	42.74	4.33	5.29	-34.52	54.09	54.69	-8.71	-75.04	-22.55	-0.69	-8.13	-50.60

F 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,

$$\boldsymbol{G}_h = (\boldsymbol{S}'\boldsymbol{W}_h\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h. \tag{5}$$

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

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

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

	Energy Score						Variogram Score					
	Gaussian DGP			Non-Gaussian DGP			Gaussian DGP			Non-Gaussian DGP		
h	1	2	3	1	2	3	1	2	3	1	2	3
Optimal(Method-1)	10.6	12.9	15.7	5.36	5.51	5.83	4.85	5.30	5.86	1.21	1.27	1.40
Optimal(Method-2)	10.6	13.0	15.8	5.37	5.53	5.83	4.86	5.32	5.88	1.21	1.27	1.37
Optimal(Method-3)	10.6	13.0	15.8	5.37	5.53	5.83	4.86	5.32	5.87	1.21	1.27	1.37
Optimal(Method-2)	10.6	13.0	15.8	5.38	5.54	5.83	4.86	5.32	5.88	1.21	1.27	1.38

G Application

G.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.

H Australian Tourism Hierarchy

Data are collected through the National Visitor Survey 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).

Table 9: Geographic hierarchy of Australian tourism flows

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						

* We consider the Australian Capital Territory as a part of New South Wales and the Northern Territory as a state.

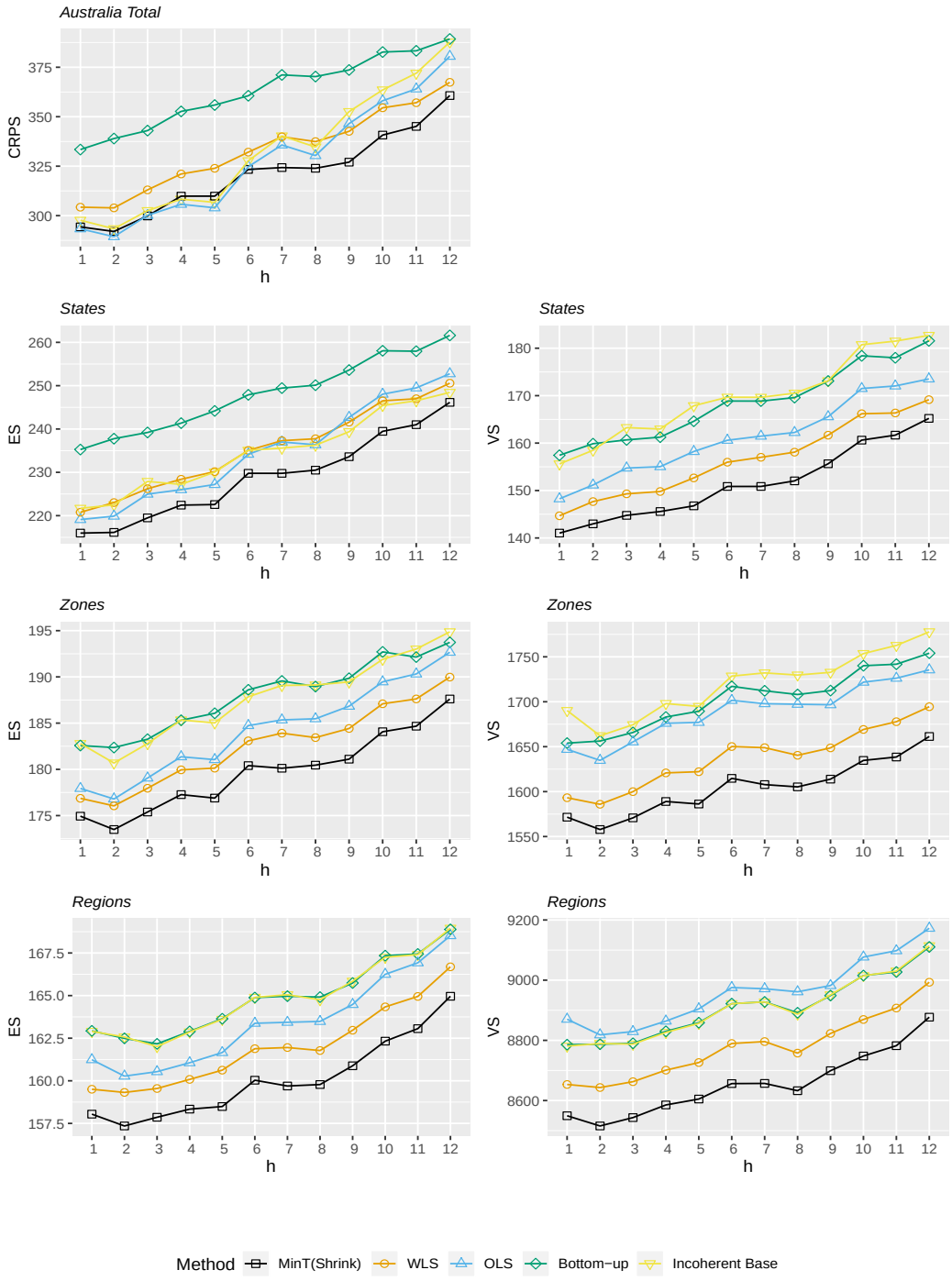

Figure 9: Forecast accuracy results across the different levels of the Australia tourism hierarchy. CRPS results are presented for the top-level and energy and variogram scores for the levels below. A lower (higher) score indicates a more (less) accurate forecast. All results are for the analytic solution assuming Gaussian incoherent base forecasts.

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.

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. (2018), *Forecasting: principles and practice, 2nd Edition*, OTexts.

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*

Hyndman, R. J., Lee, A. J. & Wang, E. (2016), 'Fast computation of reconciled forecasts for hierarchical and grouped time series', *Computational Statistics and Data Analysis* **97**, 16–32.
URL: *http://dx.doi.org/10.1016/j.csda.2015.11.007*

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.