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

May 18, 2020

Abstract

TBC

1 Introduction

Many multivariate prediction problems involve data that follow some linear constraints. For instance, in retail or tourism it is important to forecast demand in individual regions as well as aggregate demand of a whole country. In recent years reconciliation has become an increasingly popular method for handling such problems [see][for an overview]Hyndman & Athanasopoulos (2018). Reconciliation involves producing predictions for all variables and making a subsequent adjustment to ensure these adhere to known linear constraints. While this methodology has been extensively developed for point prediction there is a paucity of literature dealing with probabilistic predictions. As such, this paper develops of a formal framework for probabilistic reconciliation, derives theoretical results that allow reconciled probabilistic forecasts to be constructed and evaluated and proposes an algorithm for optimally reconciling probabilistic forecasts with respect to a proper scoring rule.

Before describing the need for probabilistic reconciliation we briefly review the literature on point forecast[1] reconciliation. Prior to the development of forecast reconciliation, the focus was on finding a subset of variables that could be subsequently aggregated or disaggregated to find forecasts for all series (see Dunn et al. 1976, Gross & Sohl 1990, and references therin). An alternative approach emerged with Athanasopoulos et al. (2009), Hyndman et al. (2011) who recommended producing forecasts of all series and then adjusting, or 'reconciling', these forecasts to be 'coherent', i.e. adhere to the aggregation constraints. These papers formulated reconciliation as a regression model, however subsequent work has formulated reconciliation as an optimisation problem where weights are chosen to minimise a loss such as a weighted squared error (Van Erven & Cugliari 2015, Nystrup et al. 2020), a penalised version thereof (Ben Taieb & Koo 2019) or the trace of the forecast error covariance (Wickramasuriya et al. 2019).

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 &

---

[1]Such has been the predominance of forecasting in the literature on reconciliation, that we will refer to forecasting throughout the remainder of the paper. However, we note that most of the techniques discussed generalised to prediction problems in general and are not limited to time series.

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 et al. 2017) and retail forecasting (Böse et al. 2017). An early attempt towards probabilistic forecast reconciliation came from Shang & Hyndman (2017) who applied reconciliation to forecast quantiles, rather than to the point forecasts, in order to construct confidence intervals. This idea was extended to constructing a full probabilistic forecast by Jeon et al. (2019) who simply reconcile a large number of forecast quantiles. In addition Jeon et al. (2019) also produce samples from univariate probabilistic forecasts and reconcile these but provide no formal justification for this procedure. Also in the field of energy, Ben Taieb et al. (2020) propose their own reconciliation algorithm. In particular, Ben Taieb et al. (2020) draw a sample from the probabilistic forecasts of univariate models for the bottom level data, reorder these to match the empirical copula of residuals, and aggregate these in a bottom up fashion. The only sense in which top level forecasts are used in the mean which is adjusted to match that obtained using the MinT method of reconciliation.

There are a number of shortcomings to Jeon et al. (2019) and Ben Taieb et al. (2020) which to the best of our knowledge represent the only attempts to develop algorithms for probabilistic forecast reconciliation. First little formal justification is provided for the algorithms or for the sense is which they generalise forecast reconciliation to the probabilistic domain. As such, both algorithms are based on sampling and neither can be used to obtain a reconciled density analytically. Both algorithms are tailored towards specific applications and conflate reconciliation with steps that involve reordering the base forecasts. For example while Jeon et al. (2019) show that ranking draws from base probabilistic forecasts is effective, this may only be true due to the highly dependent temporally aggregated times series they consider. A limitation of Ben Taieb et al. (2020) is that to ensure their sample from the base probabilistic forecasts has the same empirical copula as the data, it must be of the same size as the training data, which could be problematic in applications with fewer observations than the smart meter data they consider. Furthermore, Ben Taieb et al.

(2020) only incorporate information from the forecast mean of aggregate variables, possible missing out on valuable information in the probabilistic forecasts of top level data.

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 4 we consider the evaluation of probabilistic hierarchical forecasts via scoring rules. Parametric forecast reconciliation and some theoretical results related to elliptical distributions are discussed in Section 3 while the non-parametric approach is introduced in Section 3.3. 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.
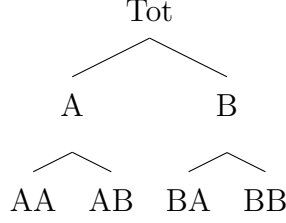
Figure 1: An example of a two level hierarchical structure.

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. Many existing point forecasting approaches including the bottom-up (Dunn et al. 1976), OLS (Hyndman et al. 2011), WLS (Hyndman

Table 1: Summary of reconciliation methods for which $\boldsymbol{SG}$ is a projection matrix. Here $\boldsymbol{W}$ some diagonal matrix, $\hat{\boldsymbol{\Sigma}}_{sam}$ is a sample estimate of the one-step ahead forecast error covariance matrix and $\hat{\boldsymbol{\Sigma}}_{shr}$ is a shrinkage estimator proposed by Schäfer & Strimmer (2005), given by $\tau\mathrm{diag}(\hat{\boldsymbol{\Sigma}}_{sam}) + (1-\tau)\hat{\boldsymbol{\Sigma}}_{sam}$ where $\tau = \frac{\sum_{i\neq j}\hat{\mathrm{Var}}(\hat{\sigma}_{ij})}{\sum_{i\neq j}\hat{\sigma}_{ij}^2}$ and σ_{ij} denotes the (i,j)th element of $\hat{\boldsymbol{\Sigma}}^{\mathrm{sam}}$.

Reconciliation method	$\boldsymbol{G}$
OLS	$(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$
WLS	$(\boldsymbol{S}'\boldsymbol{W}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}$
MinT(Sample)	$(\boldsymbol{S}'\hat{\boldsymbol{\Sigma}}_{sam}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\hat{\boldsymbol{\Sigma}}_{sam}^{-1}$
MinT(Shrink)	$(\boldsymbol{S}'\hat{\boldsymbol{\Sigma}}_{shr}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\hat{\boldsymbol{\Sigma}}_{shr}^{-1}$

et al. 2016, Athanasopoulos et al. 2017) and MinT (Wickramasuriya et al. 2019) methods, are special cases where g involves premultiplication by a matrix $\boldsymbol{G}$ and where $\boldsymbol{SG}$ is a projection matrix. These are summarised in Table 1.

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 σ-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 et al. (2020) 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, sets of points is mapped to sets of points by a transformation. The same probabilities are assigned to these sets under the base and reconciled measures 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
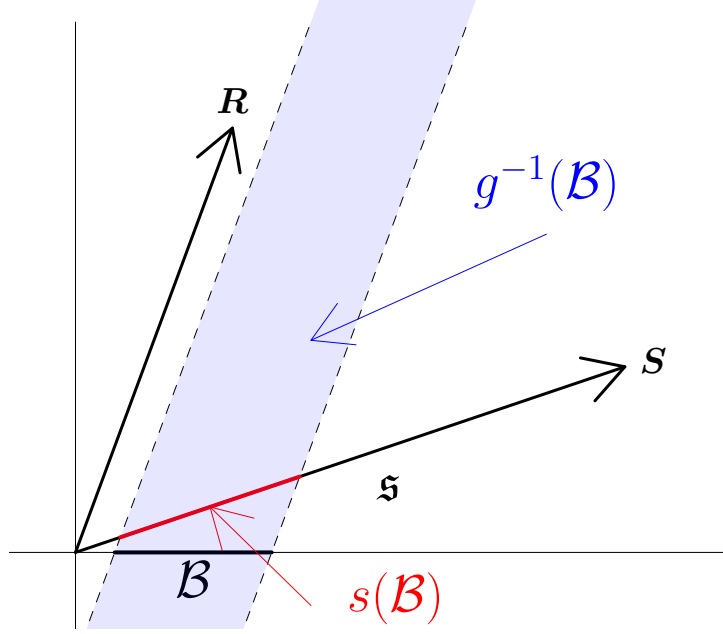
Figure 2: Summary of probabilistic forecast reconciliation. The probability that $\boldsymbol{y}_{t+h}$ lies in the red line segment under the reconciled probabilistic forecast is defined to be equal to the probability that $\boldsymbol{y}_{t+h}$ lies in the shaded blue area under the unreconciled probabilistic forecast. Note that since the smallest possible hierarchy involves three dimensions, this figure is only a schematic.

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.

3   Construction of Reconciled Distribution

In this section we derive theoretical results on how distributions on $\mathbb{R}^n$ can be reconciled to a distribution on $\mathfrak{s}$. In section 3.1 we show how this can be achieved analytically by a change of coordinates and marginalisation when the density is available. In section 3.2 we

explore this result further in the specific case of elliptical distributions. In section 3.3 we consider reconciliation in the case where the density may be unavailable but it is possible to draw a sample from the base probabilistic forecast distribution. Throughout we restrict our attention to linear reconciliation.

3.1   Analytical derivation of reconciled densities

The following theorem shows how a reconciled density can be derived from any base probabilistic forecast on $\mathbb{R}^n$.

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 $\boldsymbol{G}\boldsymbol{G}_\perp = \boldsymbol{0}$, $\boldsymbol{G}^* = \left(\boldsymbol{G}^- \vdots \boldsymbol{G}_\perp \right)$, 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

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

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

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

More generally, consider linear reconciliation of the form $\psi(\hat{\boldsymbol{y}}) = \boldsymbol{S}(\boldsymbol{d} + \boldsymbol{G}\hat{\boldsymbol{y}})$. For an elliptical base probabilistic forecast, with location $\hat{\boldsymbol{\mu}}$ and scale $\hat{\boldsymbol{\Sigma}}$, the reconciled probabilistic forecast will also be elliptical with location $\tilde{\boldsymbol{\mu}} = \boldsymbol{S}(\boldsymbol{d} + \boldsymbol{G}\hat{\boldsymbol{\mu}})$ and scale $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\boldsymbol{S}'$. This is a consequence of the fact that elliptical distributions are closed under linear transformations and marginalisation. While the base and reconciled distribution may be of a different form they will both belong to the elliptical family. This leads to the following result

Theorem 3.3. *Assume the true predictive distribution is elliptical with location $\boldsymbol{\mu}$ and scale $\boldsymbol{\Sigma}$. Then for an elliptical base probabilistic forecast with arbitrary location $\hat{\boldsymbol{\mu}}$ and scale $\hat{\boldsymbol{\Sigma}}$, there exists $\boldsymbol{d}_{opt}$ and $\boldsymbol{G}_{opt}$ such that the true predictive distribution is recovered by reconciliation.*

Proof. First consider finding a $\boldsymbol{G}_{opt}$ for which the following holds,

$$\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{G}_{opt}\boldsymbol{\Sigma}\boldsymbol{G}'_{opt}\boldsymbol{S}'\,.$$

This can be solved as $\boldsymbol{G}_{opt} = \boldsymbol{\Omega}_0^{1/2}\hat{\boldsymbol{\Sigma}}^{-1/2}$, 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), $\boldsymbol{\Omega}_0^{1/2}(\boldsymbol{\Omega}_0^{1/2})' = \boldsymbol{\Omega}$ and $\boldsymbol{\Omega}$ is the true scale matrix for the bottom-level series. 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. To reconcile the location solve the following for $\boldsymbol{d}_{opt}$

$$\boldsymbol{\mu} = \boldsymbol{S}(\boldsymbol{d}_{opt} + \boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}})$$

which is given by $\boldsymbol{d}_{opt} = \boldsymbol{\beta} - \boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}}$, where $\boldsymbol{\beta}$ is defined so that $\boldsymbol{\mu} = \boldsymbol{S}\boldsymbol{\beta}$. $\qquad\square$

While the above theorem is not feasible in practice (exploiting the result requires knowledge of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$), it does nonetheless have important consequences for the algorithm that we introduce in Section 5. In particular, note that $\boldsymbol{S}\boldsymbol{G}_{opt}$ is not a projection matrix in general. This implies that in the probabilistic forecasting setting, it is advised to include a translation $\boldsymbol{d}$ in the reconciliation procedure. This holds even if the base forecasts are unbiased (i.e. $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$) since in general $\boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} \neq \boldsymbol{\mu}$.

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 3.4 (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. □

3.3   Simulation from a Reconciled Distribution

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 3.5 (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 $\mathcal{A} \in \mathscr{F}_\mathfrak{s}$

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

$$= \lim_{L \to \infty} \sum_{l=1}^{L} \mathbb{1}\left\{\psi(\hat{\boldsymbol{y}}^{[l]}) \in (\mathcal{A})\right\}$$

$$= \Pr(\tilde{\boldsymbol{y}} \in (\mathcal{A}))$$

□

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. This result allows coherent

forecasts to be built in a general and modular fashion, the mechanism for simulating base forecasts is separated from the question of reconciliation. This will become clear in the Simulation Study covered in Section 6.

4   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 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. Arguably the simplest and most common multivariate scoring rule is the log score. The log score simply involves evaluating the negative log density at the value of the realisation, $\mathrm{LS}(P, \boldsymbol{\omega}) = -\log f(\boldsymbol{\omega})$, 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).

Alternatively there are a number of other multivariate scoring rules that are difficult to compute using the probabilistic forecast density alone, but can be approximated using a sample drawn from that density. An example is the energy score (ES) (see Székely 2003, Gneiting & Raftery 2007, for details) which is multivariate generalisation of the popular

Cumulative Rank Probability Score (CRPS). The energy score is given by

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

where $\boldsymbol{y}$ and $\boldsymbol{y^*}$ are independent copies drawn from the distribution P, and we follow common convention by setting $\alpha = 0.5$. While the expectations in Equation 1 may have no closed form, they can be easily approximated via Monte Carlo using a sample drawn from the probabilistic forecast. Other scores with similar behaviour are kernel-based scores (Dawid 2007, Gneiting & Raftery 2007) and the variogram score (Scheuerer & Hamill 2015).

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

Proof. See Appendix C. □

As a result of Theorem 4.1 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. This is true for the log score.

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

Table 2: 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 Score	Proper	Full hierarchy should be used

score is therefore

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

$$= -\log|\boldsymbol{S}^*| - \log\tilde{f}_{\boldsymbol{b}}(\boldsymbol{b}^*).$$

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 choice 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.

The same property does not hold for all scores. For example, the energy score is invariant under orthogonal transformations (Székely & Rizzo 2013) but not 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 properties of multivariate scoring rules in the context of evaluating reconciled probabilistic forecasts are summarised in Table 2.

5   Score Optimal Reconciliation

We now propose an algorithm for finding reconciliation weights by optimising the total score. We will consider the special case of the energy score, however the algorithm can be generalised to any score that is computed by sample from the probabilistic forecast, including the variogram and kernel score. We consider linear reconciliation of the form $\tilde{\boldsymbol{y}} = \psi_{\boldsymbol{\gamma}}(\hat{\boldsymbol{y}}) = \boldsymbol{S}(\boldsymbol{d} + \boldsymbol{G}\hat{\boldsymbol{y}})$, where $\boldsymbol{\gamma} := (\boldsymbol{d}, vec(\boldsymbol{G}))$. This allows for more flexibility than a projection, which implies the constraints $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$. This added flexibility is

motivated by Theorem 3.3 which shows that projections in general are not guaranteed to recover the true predictive distribution even in the elliptical case. When making a h-step ahead forecast at time T the objective used to determine an optimal value of $\boldsymbol{\gamma}$ is the total energy score based on in-sample information given by:

$$\mathcal{E}\left(\boldsymbol{\gamma}\right) = \sum_{t=1}^{T-h} ES(\tilde{f}_{t+h|t}^{\gamma}, \boldsymbol{y}_{t+1}),\tag{2}$$

where $\tilde{f}_{t+h|t}^{\gamma}$ is probabilistic forecast for $\boldsymbol{y}_{t+h}$ made at time t and reconciled with respect to $\psi_{\boldsymbol{\gamma}}(.)$.

One of the challenges to optimising this objective function is that there is, in general, no closed form expression for the energy score. However, it can be easily approximated by Monte Carlo as

$$\hat{\mathcal{E}}\left(\boldsymbol{\gamma}\right) = \sum_{t=1}^{T-h}\left[\frac{1}{Q}\left(\sum_{q=1}^{Q}||\tilde{\boldsymbol{y}}_{t+h|t}^{[q]} - \boldsymbol{y}_{t+h}|| - \frac{1}{2}||\tilde{\boldsymbol{y}}_{t+h|t}^{[q]} - \tilde{\boldsymbol{y}}_{t+h|t}^{*[q]}||\right)\right],$$

where $\tilde{\boldsymbol{y}}_{t+h|t}^{[q]} = \boldsymbol{S}\left(\boldsymbol{d} + \boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t}^{[q]}\right)$, $\tilde{\boldsymbol{y}}_{t+h|t}^{*[q]} = \boldsymbol{S}\left(\boldsymbol{d} + \boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t}^{*[q]}\right)$ and $\hat{\boldsymbol{y}}_{t+h|t}^{[q]}, \hat{\boldsymbol{y}}_{t+h|t}^{*[q]} \overset{iid}{\sim} \hat{f}_{t+h|t}$ for $q = 1, \ldots, Q$, and $t = 1, \ldots, T - h$. Therefore the objective function can be optimised by Stochastic Gradient Descent (SGD). The SGD technique has become increasingly popular in machine learning and statistics over the past decade having been applied to training neural networks (Bottou 2010) and Variational Bayes (Kingma & Welling 2013). The method requires an estimate of the gradient $\partial\hat{\mathcal{E}}/\partial\gamma$ which is computed by automatically differentiating Equation 5 using the header only C++ library of the Stan project (Carpenter et al. 2015). The learning rates used for SGD are those of the Adam method (see Kingma & Ba 2014, for details). The tuning parameters are set as $\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$, which are the values recommended by Kingma & Ba (2014) and used in popular software packages such as TensorFlow, Keras and Torch amongst others. Convergence is achieved when the change in all gradients is less than 10% of the step size η. The full procedure is described in Algorithm 1 for one step ahead forecasts and is implemented in the R package `ProbReco` publicly available at the GitHub repository `anastasiospanagiotelis/ProbReco`.

Algorithm 1 SGD with Adam for score optimal reconciliation (one-step ahead forecasts). The initial value of $\boldsymbol{\gamma}$ is given by OLS reconciliation. Steps 9-14 are the standard steps for SGD with Adam. Squaring $\boldsymbol{g}_j$ in step 11 and division in step 14 are element-wise operations.

1: **procedure** SCOREOPT($\hat{f}_{2|1}, \ldots, \hat{f}_{T|T-1|1}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T, \beta_1, \beta_2, \epsilon, \eta$).

2:      **for** t=1:T-1 **do**

3:          Find base forecasts $\hat{f}_{t+1|t}$ using $t - S, t - S + 1, \ldots t$ as training data.

4:      **end for**

5:      Initialise $\boldsymbol{m}_0 = \boldsymbol{0}, \boldsymbol{v}_0 = \boldsymbol{0}$ and $\boldsymbol{\gamma}_0 = (\boldsymbol{0}, vec((\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'))$

6:      **for** j = 1,2,3,... up to convergence **do**

7:          Draw $\hat{\boldsymbol{y}}_{t+1|t}^{[q]}, \hat{\boldsymbol{y}}_{t+1|t}^{*[q]} \sim \hat{f}_{t+1|t}$ for $q = 1, \ldots, Q, t = 1, \ldots T - 1$.

8:          Compute $\tilde{\boldsymbol{y}}_{t+1|t}^{[q]}$ and $\tilde{\boldsymbol{y}}_{t+1|t}^{*[q]}$ for $q = 1, \ldots, Q, t = 1, \ldots T - 1$ using $\boldsymbol{\gamma}_{j-1}$.

9:          $\boldsymbol{g}_j \leftarrow \partial\hat{\mathcal{E}}/\partial\boldsymbol{\gamma}\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_{j-1}}$ ▷ Compute gradient

10:         $\boldsymbol{m}_j \leftarrow \beta_1\boldsymbol{m}_{j-1} + (1 - \beta_1)\boldsymbol{g}_j$ ▷ Moving average of gradient

11:         $\boldsymbol{v}_j \leftarrow \beta_2\boldsymbol{v}_{j-1} + (1 - \beta_2)\boldsymbol{g}_j^2$ ▷ Moving average of squared gradient

12:         $\hat{\boldsymbol{m}}_j \leftarrow \boldsymbol{m}_j/(1 - \beta_1^j)$ ▷ Bias correct

13:         $\hat{\boldsymbol{v}}_j \leftarrow \boldsymbol{v}_j/(1 - \beta_2^j)$ ▷ Bias correct

14:         $\boldsymbol{\gamma}_j \leftarrow \boldsymbol{\gamma}_{j-1} + \eta\frac{\hat{\boldsymbol{m}}_j}{(\hat{\boldsymbol{v}}_j+\epsilon)}$ ▷ Update weights

15:      **end for**

16:      Set the reconciled forecast as $\tilde{f}_{T+1|T}^{\boldsymbol{\gamma}_{opt}}$ where $\boldsymbol{\gamma}_{opt}$ is the converged value of $\boldsymbol{\gamma}$.

17: **end procedure**

While Algorithm 1 is not the first instance of calibrating parameters by optimising scoring rules (see Gneiting & Raftery 2005, for an example), to the best of our knowledge it is the first instance of doing so using SGD and the first application to forecast reconciliation. It is amenable to parallel computing architectures, the loop beginning at line 2 of the pseudocode of Algorithm 1 can be done in parallel as can the computation of the gradient. Finally, the total score in Equation 2 can be replaced with a weighted sum where appropriate, for instance weights that decay for scores computed further in the past will favour choices of γ that produced better forecasting performance for more recent forecast windows.

6 Simulations

The aim of the simulations that follow is to demonstrate probabilistic forecast reconciliation including the algorithm discussed in Section 5. A number of different data generating processes, models and approaches to obtain base forecasts are considered.

- describe all DGPS stationary/non-stationary gaussian/nongaussian and models ARIMA/ETS

- Explain all 4 ways of doing base forecasts independent/joint gaussian/non-gaussian

- Reconiciliation Methods: Base, Ben Taieb et al (?), Jeon et all. Projections (BU, OLS, WLS, MinT) Score Optimal.

- Completely rewrite results

6.1 Data Generating Processes

The data generating process we consider corresponds to the 3-level hierarchical structure presented in Figure 1. Bottom-level series are first generated from $\text{ARIMA}(p, d, q)$ processes, which are in-turn aggregated to form the middle and top-level series. The orders p, and d are randomly selected from $\{1, 2\}$ for each bottom level series. Both a non-stationary case where d is randomly chosen for each bottom level $\{0, 1\}$ and a stationary case where

d is set to zero for all bottom level series. The parameters for the AR and MA components are randomly and uniformly generated from $[0.3, 0.5]$ and $[0.3, 0.7]$ respectively where in the case of the stationary series these are only accepted if they belong to the stationary and invertible region.

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$(1, 3)$ margins. After simulating from the ARIMA models, additional noise is added to ensure bottom level series have a lower signal-to-noise ratio that top level series with specific details provided in Appendix D. This yields a total of four data generating processes (DGPs) for each combination of stationary/non-stationary and Gaussian/non-Gaussian. For each series we generate 2000 observations from which the first 500 are ignored to avoid the impact of initial values.

6.2 Modelling and Base Forecasts

We fit univariate ARIMA and exponential smoothing (ETS) models to each series using the `auto.arima()` and `ets()` functions in the `fable` package (O'Hara-Wild et al. 2020) in R (R Core Team 2018). Note that the order of the ARIMA models is not set to the true order but is chosen using the algorithm of Hyndman et al. (2007), allowing for the possibility of misspecification. Indeed, an advantage of forecast reconciliation is the ability to down-weight the forecasts of series within the hierarchy that come from misspecified models. For each series and model, base probabilistic forecasts are constructed in four different ways. Letting $\hat{\boldsymbol{y}}_{t+h|h} L = \hat{y}_{1,t+h|h}, \ldots, \hat{y}_{n,t+h|h}$ where $\hat{y}_{i,t+h|h}$ is the point forecast for series i and $e_{i,t}$ be residuals stacked in a $(n \times T)$ matrix $\boldsymbol{E} := \{e_{i,t}\}_{i=1,\ldots,n,t=1,\ldots,T}$, these are:

(a) The base probabilistic forecast is made up of independent Gaussian distributions with the forecast mean and variance of variable i given by $\hat{y}_{i,t+h|h}$ and $\hat{\sigma}^2_{i,t+h|h}$,

(b) The base probabilistic forecast is multivariate Gaussian distributions with the forecast mean $\hat{\boldsymbol{y}}_{t+h|h}$ and variance covariance matrix $\hat{\boldsymbol{\Sigma}}$, where $\hat{\boldsymbol{\Sigma}}$ is the variance covariance matrix of residuals,

(c) A draw from the base probabilistic forecast is made independently for each variable as $\hat{y}_{i,t+h|h} + e_{i,s}$ where s is drawn randomly (with replacement) from $1, 2, \ldots, T$, and

(d) A draw from the joint probabilistic is made as $\hat{\boldsymbol{y}}_{t+h|h} + \boldsymbol{e}_s$ where $\boldsymbol{e}_s$ is the sth column of $\boldsymbol{E}$, where s is drawn randomly (with replacement) from $1, 2, \ldots, T$.

We restrict our attention to the case of $h = 1$ although these methods can be generalised to larger h using the recursive method (Hyndman & Athanasopoulos 2018). For multi-step ahead forecasts, methods (c) and (d) should sample in blocks to preserve serial dependence in the residuals.

6.3   Reconciliation

For each DGP, model and method for obtaining base forecasts, reconciled probabilistic forecasts are obtained using each of the following techniques.

- **Base:** The base forecasts with no reconciliation.

- **JPP:** The best method of Jeon et al. (2019). This is equivalent to reconcile quantiles. A sample is drawn from the base forecast, these are ranked, one variable at a time (so that the smallest value drawn from each variable are put together, etc.). These are then pre-multipled by $\boldsymbol{S}\left(\boldsymbol{S}'\boldsymbol{W}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{W}$ where $\boldsymbol{W}$ is a diagonal matrix with elements $(1/4^2, 1/2^2, 1/2^2, 1, 1, 1, 1)$. These are the squared reciprocals of the number of bottom level series used to form an aggregate.

- **BTTH** The method of Ben Taieb et al. (2020). This is a method whereby draws from the probabilistic forecasts of the bottom level series are permuted so that they have the same empirical copula as the residuals. These are then aggregated to form a sample from the distribution of all series. The mean is adjusted to be equivalent to the mean that would be obtained using the MinT shrink method Wickramasuriya et al. (2019) described in Table 1.

- **BottomUp:** Reconciliation via premultiplication by $\boldsymbol{S}\boldsymbol{G}$ where $\boldsymbol{G} = \left(\boldsymbol{0}_{m\times(n-m)}, \boldsymbol{I}_{m\times m}\right)$

- **OLS:** Reconciliation via premultiplication by $\boldsymbol{S}\left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1}\boldsymbol{S}'$,

- **WLS:** Reconciliation via premultiplication by the same matrix used in Jeon et al. (2019) but without any reordering of the draws,

- **MinT:** Reconciliation via premultiplication by the same matrix used in Wickrama-suriya et al. (2019) but applied to probabilistic rather than point forecasting,

- **ScoreOpt:** The algorithm described in Section 5.

Note JPP and BTTH are two methods previously existing in the literature. The methods BottomUp, OLS, WLS and MinT have been used extensively in the point forecasting literature but their application to probabilistic forecasting for general base forecasts is novel.

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 2 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 2 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.

Series	Gaussian							Non-Gaussian						
	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 ?? in Appendix. From these we note that the different parameterisation of 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 G matrix. Thus we suggest using the MinT 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 G.

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 3.3 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
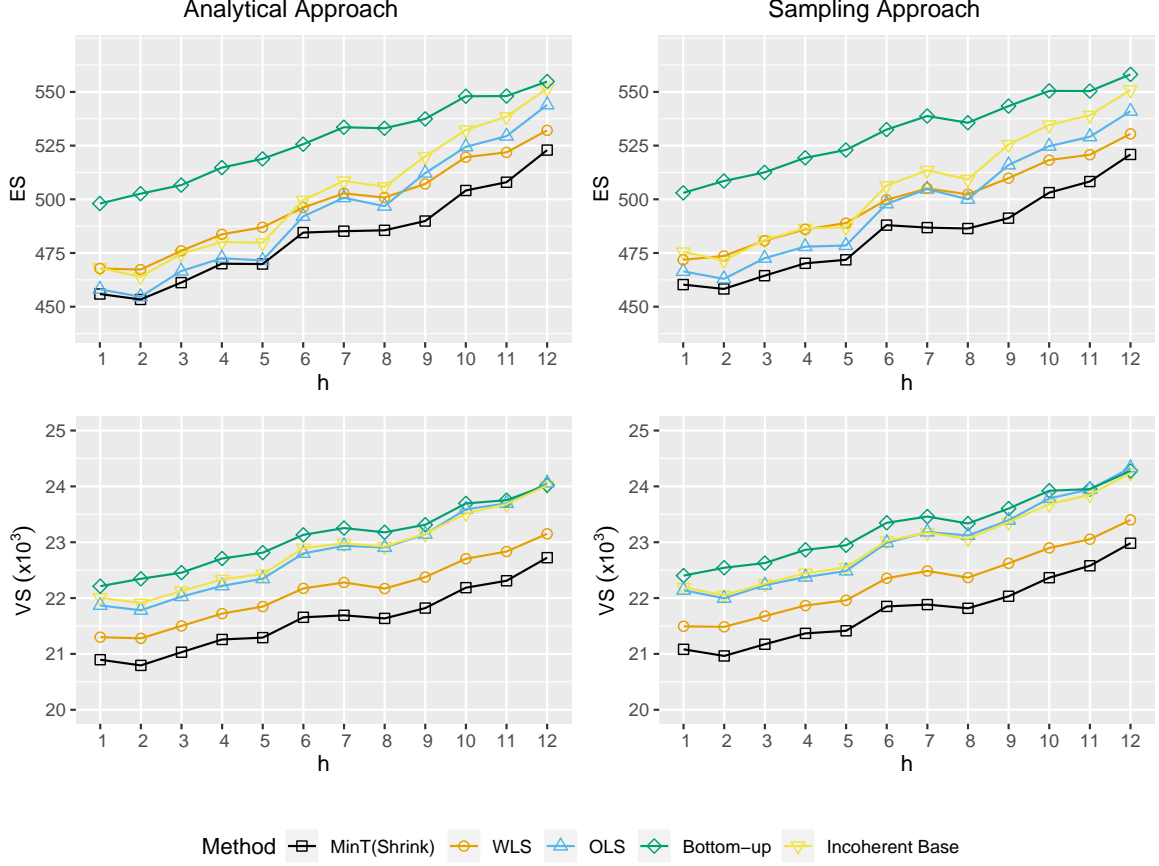
Figure 3: Energy and variogram scores for multivariate predictive distributions across the entire hierarchy. A lower (higher) score indicates a more (less) accurate forecast. Results from the analytic approach assuming Gaussian incoherent base forecasts are presented on the left while results from the 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
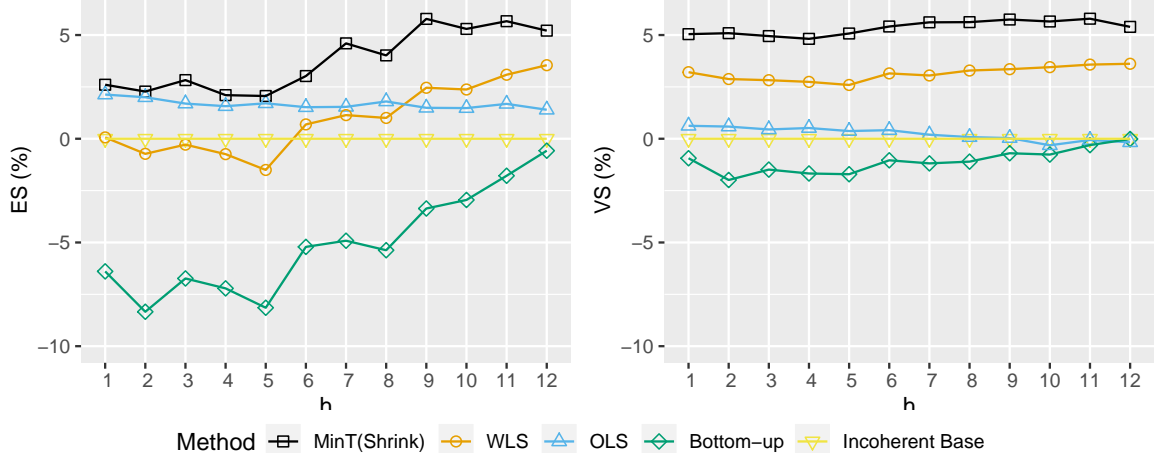
Figure 4: Skill scores (%) relative to incoherent base forecasts, across the entire Australian tourism hierarchy based on energy score (on the left) and variogram score (on the right). A higher (lower) score indicates a gain (loss) in forecast accuracy relative to the incoherent base forecasts. The results 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 G 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.
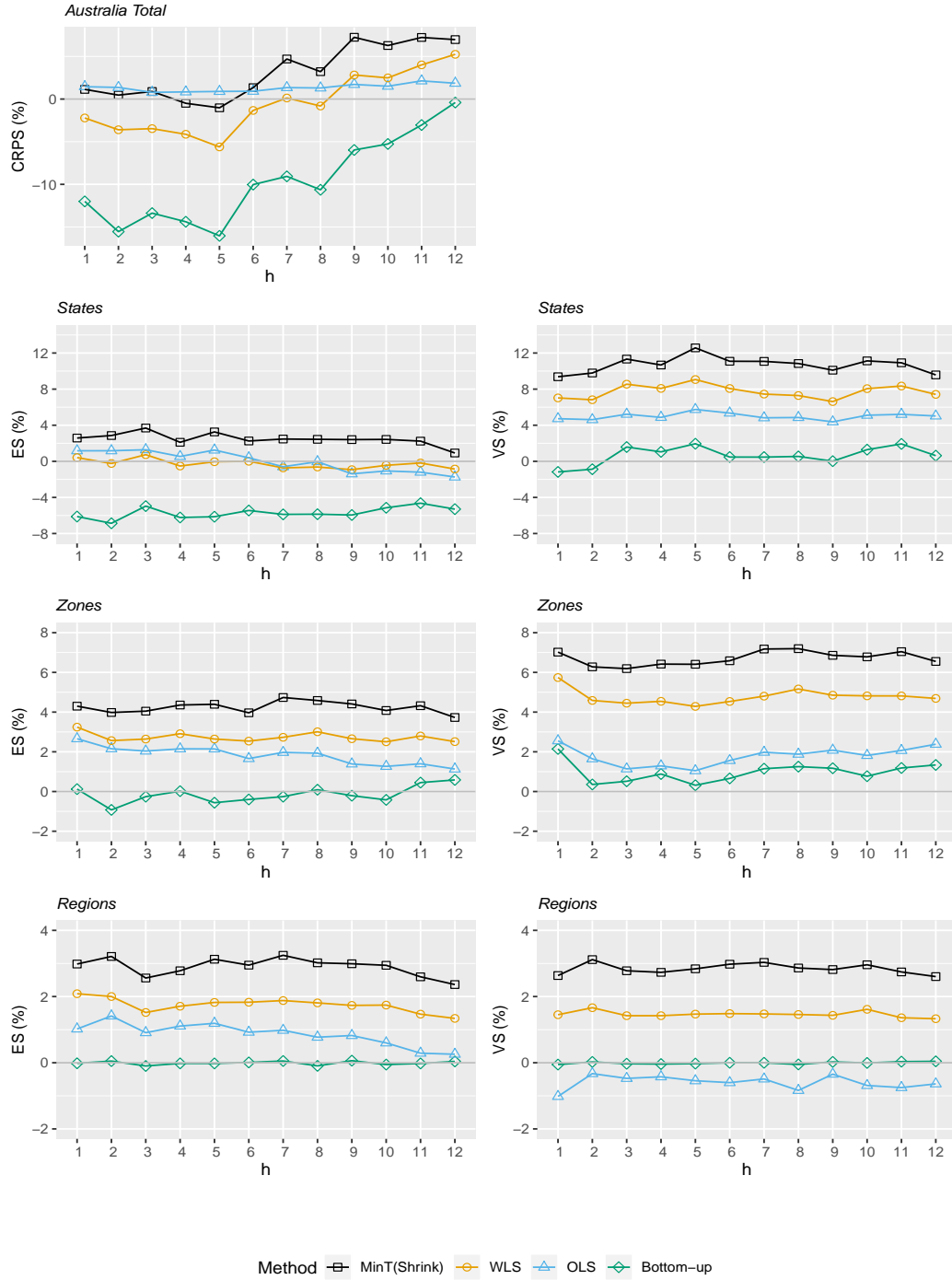
Figure 5: Skill scores (%) relative to incoherent base forecasts, for the CRPS for the top-level and energy and variogram scores for the levels below for the Australia tourism hierarchy. A higher (lower) score indicates a gain (loss) in forecast accuracy relative to the incoherent base forecasts. All results are for the analytic solution assuming Gaussian incoherent base forecasts.

32

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 $\boldsymbol{l} \succ \boldsymbol{Gy} \succ \boldsymbol{u}$. Using Definition 2.2,

$$\Pr(\boldsymbol{l} \succ \boldsymbol{b} \succ \boldsymbol{u}) = \int_{\boldsymbol{l} \succ \boldsymbol{Gy} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{y}) d\boldsymbol{y},$$

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

$$\Pr(\boldsymbol{l} \succ \boldsymbol{b} \succ \boldsymbol{u}) = \int_{\boldsymbol{l} \succ \boldsymbol{Gy} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{y}) d\boldsymbol{y}$$

$$= \int_{\boldsymbol{l} \succ \boldsymbol{GG}^* \boldsymbol{z} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{G}^* \boldsymbol{z}) |\boldsymbol{G}^*| d\boldsymbol{z}$$

$$= \int_{\boldsymbol{l} \succ \boldsymbol{z}_1 \succ \boldsymbol{u}} \hat{f}(\boldsymbol{G}^* \boldsymbol{z}) |\boldsymbol{G}^*| d\boldsymbol{z},$$

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

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

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

$$\tilde{f}_{\boldsymbol{b}}(\boldsymbol{b}) = \int \hat{f}(\boldsymbol{G}^- \boldsymbol{b} + \boldsymbol{G}_\perp \boldsymbol{a}) |\boldsymbol{G}^*| d\boldsymbol{a}, \tag{3}$$

is the density of $\boldsymbol{b}$. To obtain the density of the full hierarchy we first augment the density in Equation (3) by $n - m$ variables denoted $\boldsymbol{u}$

$$f(\boldsymbol{b}, \boldsymbol{u}) = \tilde{f}_b(\boldsymbol{b})\mathbb{1}\{\boldsymbol{u} = 0\} , \tag{4}$$

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

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

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

$$\tilde{f}_y(\boldsymbol{y}) = |\boldsymbol{S}^*|\tilde{f}_b(\boldsymbol{S}^- \boldsymbol{y})\mathbb{1}\{\boldsymbol{y} \in \mathfrak{s}\} .$$

B   Proof of Theorem 3.4

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 $\boldsymbol{SGS} = \boldsymbol{S}$ and

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{SG}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\boldsymbol{S}'$$

$$\boldsymbol{SGS}\boldsymbol{\Omega}\boldsymbol{S}'\boldsymbol{G}'\boldsymbol{S}' + \boldsymbol{SGD}\boldsymbol{G}'\boldsymbol{S}'$$

$$\boldsymbol{S}\boldsymbol{\Omega}\boldsymbol{S}' + \boldsymbol{SGD}\boldsymbol{G}'\boldsymbol{S}'$$

$$\boldsymbol{\Sigma} + \boldsymbol{SGD}\boldsymbol{G}'\boldsymbol{S}'\,.$$

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 $q = \operatorname{rank}(\boldsymbol{D})$ 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 $\operatorname{rank}(\boldsymbol{D}) \le n - m$.

With respect to the location, if $\boldsymbol{SG}$ 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 4.1

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,

$$
\begin{aligned}
LS\left(f, \boldsymbol{y}^* \right) &= -\log f(\boldsymbol{y}^*) \\
&= -\log |\boldsymbol{S^*}| - \log f_{\boldsymbol{b}}\left(\boldsymbol{b}^* \right) - \log \left(\mathbb{1}\{\boldsymbol{u}^* = \boldsymbol{0}\} \right) \qquad (5) \\
&= -\log |\boldsymbol{S^*}| - \log f_{\boldsymbol{b}}\left(\boldsymbol{b}^* \right),
\end{aligned}
$$

where the third term in Equation 5 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

To ensure that bottom level series are noisier than top level series (a feature often observed empirically), noise is added to the bottom level series in the following manner

$$y_{AA,t} = w_{AA,t} + u_t - 0.5v_t,$$

$$y_{AB,t} = w_{AB,t} - u_t - 0.5v_t,$$

$$y_{BA,t} = w_{BA,t} + u_t + 0.5v_t,$$

$$y_{BB,t} = w_{BB,t} - u_t + 0.5v_t,$$

where $w_{AA,t}, w_{AB,t}, w_{BA,t}, w_{BB,t}$ generated from ARIMA processes as described in Section 6.1 with innovations $\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}$. Aggregating the bottom-level series gives the following for the aggregated levels,

$$y_{A,t} = w_{AA,t} + w_{AB,t} - v_t,$$

$$y_{B,t} = w_{BA,t} + w_{BB,t} + v_t,$$

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

For the Gaussian DGP, $u_t \sim \mathcal{N}(0, \sigma_u^2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$ and the errors driving the bottom-level ARIMA processes are jointly generated from a Normal distribution while . More specifically, $\{\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}\} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \; \forall t$. We follow Wickramasuriya et al. (2019) and set

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

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

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

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

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

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

For the non-Gaussian case, errors are generated from a Gumbel copula with Beta margins as described in Section 6.1. Rather than add Gaussian noise, we simulate u_t and v_t from skew t distributions using the package **?** The scale, skew and degrees of freedom parameters are chosen as 0.5,1.5 and 4 and 0.9,2 and 8 for u_t and v_t respectively. Monte Carlo simulations show that these values satisfy the inequalities described above.

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

Series	Gaussian							Non-Gaussian						
	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 Application

F.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.

G 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 8: Geographic hierarchy of Australian tourism flows

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

* We consider the Australian Capital Territory as a part of New South Wales and the Northern Territory as a state.
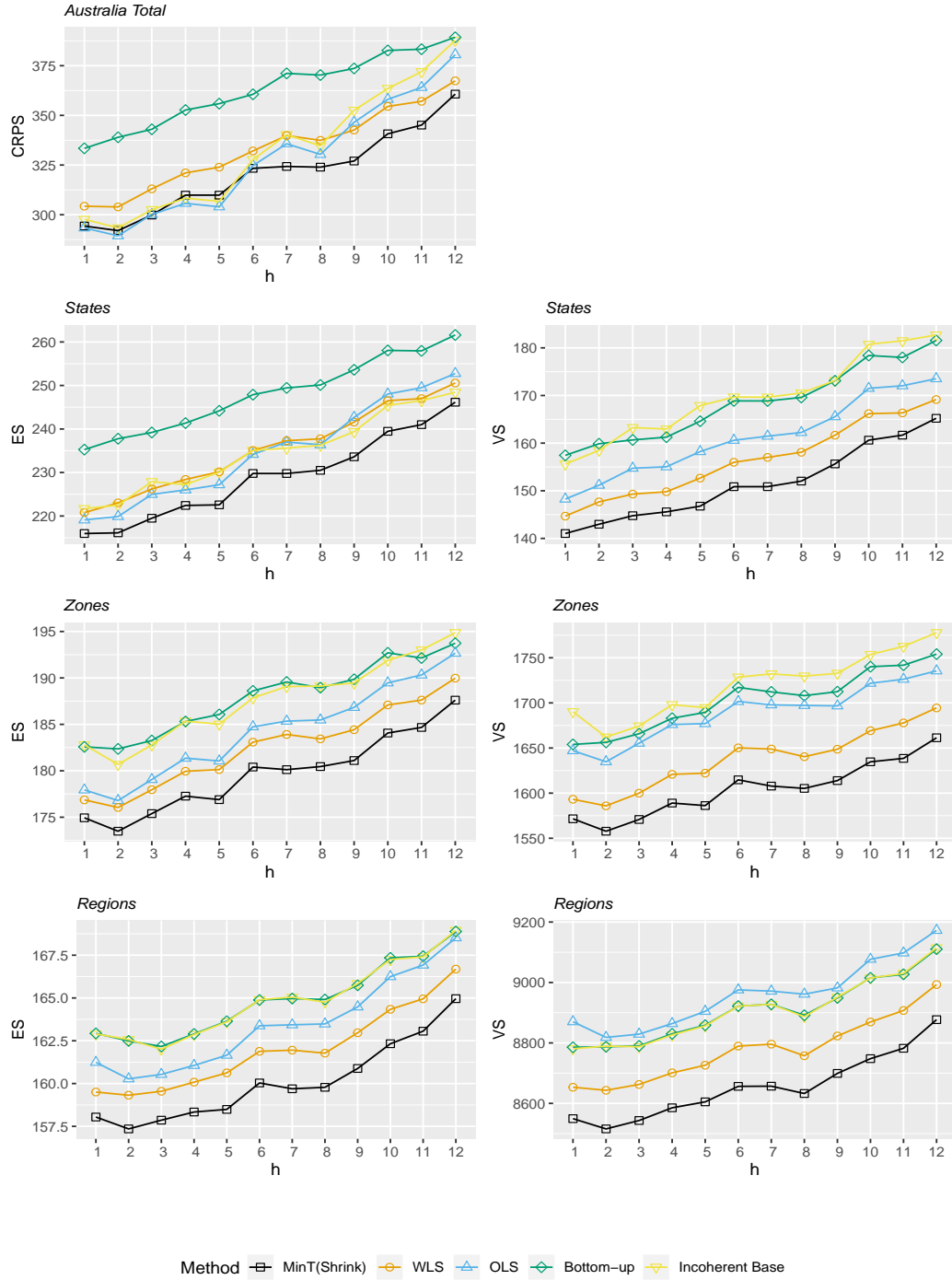
Figure 9: Forecast accuracy results across the different levels of the Australia tourism hierarchy. CRPS results are presented for the top-level and energy and variogram scores for the levels below. A lower (higher) score indicates a more (less) accurate forecast. All results are for the analytic solution 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. & Koo, B. (2019), Regularized regression for hierarchical forecasting without unbiasedness conditions, pp. 1337–1347.

Ben Taieb, S., Taylor, J. W. & Hyndman, R. J. (2020), 'Hierarchical probabilistic forecasting of electricity demand with smart meter data', *Journal of the American Statistical Association* **0**(0), 1–17.

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.

Bottou, L. (2010), Large-scale machine learning with stochastic gradient descent, *in* 'Proceedings of COMPSTAT'2010', Springer, pp. 177–186.

Carpenter, B., Hoffman, M. D., Brubaker, M., Lee, D., Li, P. & Betancourt, M. (2015), 'The stan math library: Reverse-mode automatic differentiation in c++', *arXiv preprint arXiv:1509.07164* .

Dawid, A. P. (2007), 'The geometry of proper scoring rules', *Annals of the Institute of Statistical Mathematics* **59**(1), 77–93.

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. (2005), 'Weather forecasting with ensemble methods', *Science* **310.5746**, 248–249.

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., Khandakar, Y. et al. (2007), *Automatic time series for forecasting: the forecast package for R*, number 6/07, Monash University, Department of Econometrics and Business Statistics .

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*

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kingma, D. P. & Welling, M. (2013), 'Auto-encoding variational bayes', *arXiv preprint arXiv:1312.6114* .

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.

Nystrup, P., Lindstrm, E., Pinson, P. & Madsen, H. (2020), 'Temporal hierarchies with autocorrelation for load forecasting', *European Journal of Operational Research* **280**(3), 876 – 888.

O'Hara-Wild, M., Hyndman, R. & Wang, E. (2020), *fable: Forecasting Models for Tidy Time Series*. R package version 0.2.0.
URL: *https://CRAN.R-project.org/package=fable*

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.

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. (2003), 'E-statistics: The energy of statistical samples', *Bowling Green State University, Department of Mathematics and Statistics Technical Report* **3**(05), 1–18.

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.