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

July 13, 2020

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

Abstract

We develop a framework for prediction or forecasting of multivariate data that follow some known linear constraints, such as the example where some variables are aggregates of others. For point prediction, an increasingly popular technique is reconciliation, whereby predictions or forecasts are made for all series (so called 'base' forecasts) and subsequently adjusted to ensure coherence with the constraints. This paper extends reconciliation from the setting of point prediction to probabilistic prediction. A novel definition of reconciliation is developed and used to construct densities and draw samples from a reconciled probabilistic prediction. In the elliptical case, it is proven that the true predictive distribution can be recovered from reconciliation even when the location and scale matrix of the base prediction are chosen arbitrarily. To find reconciliation weights, an objective function based on scoring rules is optimised. The energy score is chosen since the log score is improper in the context of comparing unreconciled to reconciled forecasts, a result also proved in this paper. To account for the stochastic nature of the energy score, optimisation is achieved using stochastic gradient descent. ADD SOMETHING ABOUT SIMULATION AND APPLICATION.

Keywords: Scoring Rules, Probabilistic Forecasting, Hierarchical Time Series, Stochastic Gradient Descent.

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 Hyndman & Athanasopoulos 2018, for an overview). 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) and 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 the techniques discussed throughout the paper generalise 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 prediction intervals. This idea was extended to constructing a full probabilistic forecast by Jeon et al. (2019) who propose a number of algorithms, one of which is equivalent reconciling a large number of forecast quantiles. Ben Taieb et al. (2020) also propose an algorithm to obtain probabilistic forecasts that cohere to linear constraints. 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 is in the mean, which is adjusted to match that obtained using the MinT reconciliation method (Wickramasuriya et al. 2019).

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 independent base probabilistic forecasts before reconciliation is effective, this may only be true due to the highly dependent time series considered in their application. A limitation of Ben Taieb et al. (2020) is that to ensure their sample from the base probabilistic forecast 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, possibly missing out on valuable information in the probabilistic forecasts of aggregate data.

This paper seeks to address a number of open issues in probabilistic forecast reconciliation. First, we develop in a formal way, definitions and a framework that generalise reconciliation from the point setting to the probabilistic setting. This is achieved by extending the geometric framework proposed by Panagiotelis et al. (2019) for the point forecasting. Second, we utilise these definitions to show how a reconciled forecast can be constructed from an arbitrary base forecast. Solutions are provided in the case where a density of the base probabilistic forecast is available and in the case where it is only possible to draw a sample from the base forecasting distribution. Third, we show that in the elliptical case, the correct predictive distribution can be recovered via linear reconciliation irrespective of the location and scale parameters of the base forecasts. We also derive conditions for when this also holds for the special case of reconciliation via projection. Fourth, we derive theoretical results on the evaluation of reconciled probabilistic forecasts using multivariate scoring rules, including a particularly important result on the impropriety of using the log score to compare reconciled to unreconciled forecasts. Fifth, we propose an algorithm for choosing reconciliation weights by optimising a scoring rule. This algorithm takes advantages of advances in stochastic gradient descent and is thus suited to scoring rules that are themselves often only known up to an approximation.

The remainder of the paper is structured as follows. In Section 2, after brief review of point forecast reconciliation, novel definitions are provided for coherent forecasts and reconciliation in the probabilistic setting. In Section 3, we outline how reconciliation can be achieved in the both the case where the density of the base probabilistic forecast is available, and in the case where a sample has been generated from the base probabilistic forecast. In Section 4, we consider the evaluation of probabilistic hierarchical forecasts via scoring rules, including theoretical results on the impropriety of the log score in the context of forecast reconciliation. The use of scoring rules motivates our algorithm for finding optimal reconciliation weights using stochastic gradient descent, which is described in Section 5 and evaluated in an extensive simulation study in Section 6. An empirical

application on tourism forecasting is contained in Section 7. Finally Section 8 concludes with some discussion and thoughts on future research.

2 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.

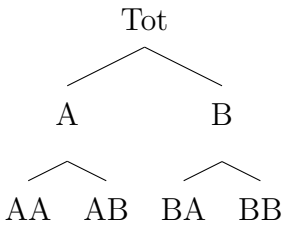

Figure 1: An example of a two level hierarchical structure.

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 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

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 $\tau\text{diag}(\hat{\boldsymbol{\Sigma}}_{sam}) + (1-\tau)\hat{\boldsymbol{\Sigma}}_{sam}$ where $\tau = \frac{\sum_{i\neq j}\hat{\text{Var}}(\hat{\sigma}_{ij})}{\sum_{i\neq j}\hat{\sigma}_{ij}^2}$ and σ_{ij} denotes the (i,j)th element of $\hat{\boldsymbol{\Sigma}}^{\text{sam}}$.

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)	$(\boldsymbol{S}'\hat{\boldsymbol{\Sigma}}_{shr}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\hat{\boldsymbol{\Sigma}}_{shr}^{-1}$

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

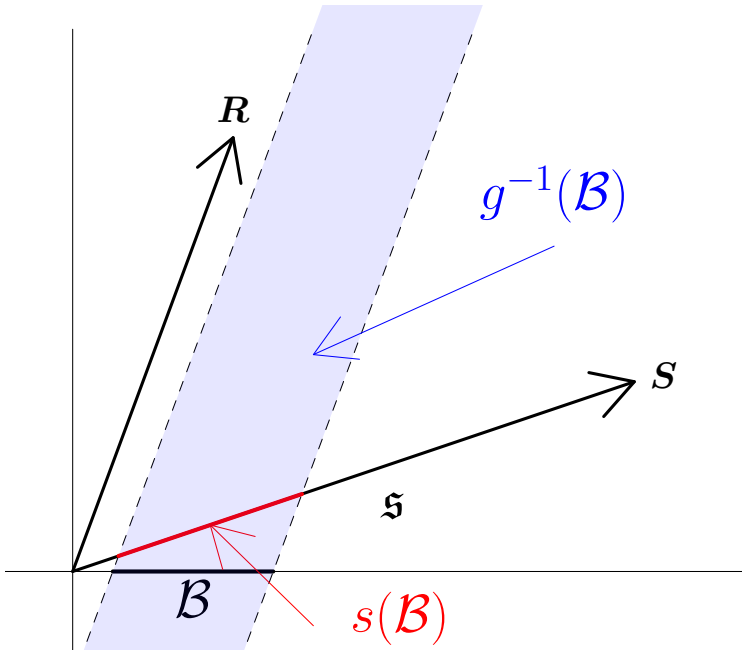

Figure 2: Summary of probabilistic forecast reconciliation. The probability that $\boldsymbol{y}_{t+h}$ lies in the red line segment under the reconciled probabilistic forecast is defined to be equal to the probability that $\boldsymbol{y}_{t+h}$ lies in the shaded blue area under the unreconciled probabilistic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

3 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

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

Noting that

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

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

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

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

3.2 Elliptical distributions

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,

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

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 $\boldsymbol{\mu} = \boldsymbol{S}\boldsymbol{\beta}$. □

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 $\boldsymbol{SG}_{opt}$ 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 $\boldsymbol{SG}_{opt}\hat{\boldsymbol{\mu}} \neq \boldsymbol{\mu}$.

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

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

□

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

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

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.

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

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

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

For coherent densities, the log score for the full hierarchy differs from the log score for the bottom-level series only by $-\log|\boldsymbol{S}^*|$. This term is independent from the 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 an objective function based on scores. For clarity of exposition, we consider the special case of the energy score. However, the algorithm can be generalised to any score that is computed by sampling

from the probabilistic forecast, and in Section 6 and Section 7 we consider optimising with respect to both the energy and variogram score. We consider linear reconciliation of the form $\tilde{\boldsymbol{y}} = \psi_{\boldsymbol{\gamma}}(\hat{\boldsymbol{y}}) = \boldsymbol{S}\left(\boldsymbol{d} + \boldsymbol{G}\hat{\boldsymbol{y}}\right)$, where $\boldsymbol{\gamma} := (\boldsymbol{d}, vec(\boldsymbol{G}))$. This allows for more flexibility than a projection, which would imply the constraints $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{GS} = \boldsymbol{I}$. 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:

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

where $\tilde{f}^{\boldsymbol{\gamma}}_{t+h|t}$ is probabilistic forecast for $\boldsymbol{y}_{t+h}$ made at time t and reconciled with respect to $\psi_{\boldsymbol{\gamma}}(.)$ and R is the number of score evaluations used in forming the objective function.

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

$$\hat{\mathcal{E}}\left(\boldsymbol{\gamma}\right) = \sum_{t=1}^{T-h}\left[\frac{1}{Q}\left(\sum_{q=1}^{Q}||\tilde{\boldsymbol{y}}^{[q]}_{t+h|t} - \boldsymbol{y}_{t+h}|| - \frac{1}{2}||\tilde{\boldsymbol{y}}^{[q]}_{t+h|t} - \tilde{\boldsymbol{y}}^{*[q]}_{t+h|t}||\right)\right]\,, \tag{3}$$

where $\tilde{\boldsymbol{y}}^{[q]}_{t+h|t} = \boldsymbol{S}\left(\boldsymbol{d} + \boldsymbol{G}\hat{\boldsymbol{y}}^{[q]}_{t+h|t}\right)$, $\tilde{\boldsymbol{y}}^{*[q]}_{t+h|t} = \boldsymbol{S}\left(\boldsymbol{d} + \boldsymbol{G}\hat{\boldsymbol{y}}^{*[q]}_{t+h|t}\right)$ and $\hat{\boldsymbol{y}}^{[q]}_{t+h|t}, \hat{\boldsymbol{y}}^{*[q]}_{t+h|t} \overset{iid}{\sim} \hat{f}_{t+h|t}$ for $q = 1, \ldots, Q$, and $t = 1, \ldots, T-h$.

The objective function is 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 3 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). Pseudocode for the full procedure in the case where $h = 1$ is provided in Algorithm 1 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 and addition in step 14 are element-wise operations.

1: **procedure** ScoreOpt($\hat{f}_{2|1}, \ldots, \hat{f}_{T|T-1|1}$, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, β_1,β_2,ϵ,η).
2: **for** t=1:T-1 **do**
3: Find base forecasts $\hat{f}_{t+1|t}$ using $t-R+1, t-R+2, \ldots t$ 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 $\hat{\boldsymbol{y}}^{[q]}_{t+1|t}, \hat{\boldsymbol{y}}^{*[q]}_{t+1|t} \sim \hat{f}_{t+1|t}$ for $q = 1, \ldots, Q$, $t = 1, \ldots T-1$.
8: Compute $\tilde{\boldsymbol{y}}^{[q]}_{t+1|t}$ and $\tilde{\boldsymbol{y}}^{*[q]}_{t+1|t}$ 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 $\tilde{f}^{\boldsymbol{\gamma}_{opt}}_{T+1|T}$ 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 $\boldsymbol{\gamma}$ 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. For all simulations, 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. Also the number of sample periods used to construct the objective function is $R = 250$, while the number of sample used to estimate each score is $Q = 250$. All estimation of base models uses a sample size of $T = 500$ and all evaluations is carried out using a rolling window, also of size $W = 500$.

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 ARIMA(p, d, q) processes, which are in-turn aggregated to form the middle and top-level series. The orders p, and q are randomly selected from $\{1, 2\}$ for each bottom level series. The AR and MA parameters randomly and uniformly generated from $[0.3, 0.5]$ and $[0.3, 0.7]$ respectively, and only accepted if they belong to the stationary and invertible region. In addition a

non-stationary case where d is randomly chosen for each bottom level from $\{0, 1\}$ was considered, these results are omitted for brevity.

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. For each series the first 500 observations are ignored to avoid the impact of initial values.

6.2 Modelling and Base Forecasts

We fit univariate ARIMA models to each series using the `ARIMA()` function 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. Also considered were ETS models which were fit using the `ETS()` function in the `fable` package. These are omitted for brevity. For each series and model, base probabilistic forecasts are constructed in four different ways. Letting $\hat{\boldsymbol{y}}_{t+h|h} = \left(\hat{y}_{1,t+h|h}, \ldots, \hat{y}_{n,t+h|h}\right)'$ 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 $\hat{\sigma}^2_{i,t+h|h}$, where $\hat{\sigma}^2_{i,t+h|h}$ is the sample variance of the residuals in the i^{th} row of $\boldsymbol{E}$,

(b) The base probabilistic forecast is a multivariate Gaussian distribution 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 the residuals,

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

(d) A draw from the joint probabilistic forecast is made as $\hat{\boldsymbol{y}}_{t+h|h} + \boldsymbol{e}_\tau$ where $\boldsymbol{e}_\tau$ is the τ^{th} column of $\boldsymbol{E}$, where τ 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 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 $\boldsymbol{S}\left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1}\boldsymbol{S}'$,

- **MinTShr:** Reconciliation via premultiplication using the shinkage estimator of the covariance matrix used by Wickramasuriya et al. (2019) but applied to probabilistic rather than point forecasting,
- **ScoreOptE:** The algorithm described in Section 5 used to optimise energy score.
- **ScoreOptV:** The algorithm described in Section 5 used to optimise variogram score.

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

In addition to these methods, two further reconciliation methods were considerd; WLS, which reconciles via premultiplication by the same matrix used in Jeon et al. (2019) but without any reordering of the draws, and MinTSam which uses a sample estimate of the covariance matrix rather than a shrinkage estimator. These methods were mostly dominated by OLS and MinTShr respectively and are therefore omitted for brevity. However, a complete set of results for the simulation study, including non-stationary DGPs, ETS base forecasting and all reconciliation methods can be found at the github repository `https://github.com/PuwasalaG/Probabilistic-Forecast-Reconciliation/Simulation/Result_Reports/Reports/`.

6.4 Results for Gaussian probabilistic forecasts

Table 3 shows the mean energy score for different reconciliation methods and different methods of generating base forecasts. When base probabilistic forecasts are generated independently, score optimisation with the energy score (ScoreOptE) performs best, while when base forecasts are generated jointly, using the shrinkage estimator from the MinT method for reconciliation (MinTShr) yields the most accurate forecasts. The bottom up method as well as BTTH and JPP fail to even improve upon base forecasts in all cases. As expected score optimisation using the variogram score does not perform as well as score optimisation using energy score, when evaluation is carried out with respect to the latter.

However, the results are quite close suggesting that score optimisation is fairly robust to using an alternative proper score.

Table 3: Mean Energy score for ARIMA modelling with a Gaussian Stationary DGP. Values in bold indicate the best reconciliation method for a given base forecasting method.

Method	Ind. Bootstrap	Ind. Gaussian	Joint Bootstrap	Joint Gaussian
Base	11.3256	11.3247	11.0662	11.0556
BottomUp	11.9759	11.9705	11.6429	11.6203
BTTH	21.7508	21.8051	21.8542	21.8826
JPP	22.8651	22.8986	22.8573	22.8599
MinTShr	10.8865	10.8905	**10.7475**	**10.7428**
OLS	11.1284	11.1264	10.8396	10.8292
ScoreOptE	**10.8053**	**10.8297**	10.8433	10.8406
ScoreOptV	11.2066	11.2074	11.0413	11.0262

To assess significant differences between the reported results, we use post-hoc Nemenyi tests (Hollander et al. 2013). The Nemenyi test is a non-parametric test that identifies groups of forecasts which cannot be significantly distinguished from one another. We use the implementation of the tests available in the `tsutils` R package (Kourentzes 2019). Figure 3 reports the results which should be looked at column-wise. A blue square indicates that the method in the corresponding row, is statistically indistinguishable from the method in that column. For all four methods of generating base forecasts, MinTShr, ScoreOptE and OLS significantly outperfom base forecasts, bottom up forecasts as well as BTTH and JPP.

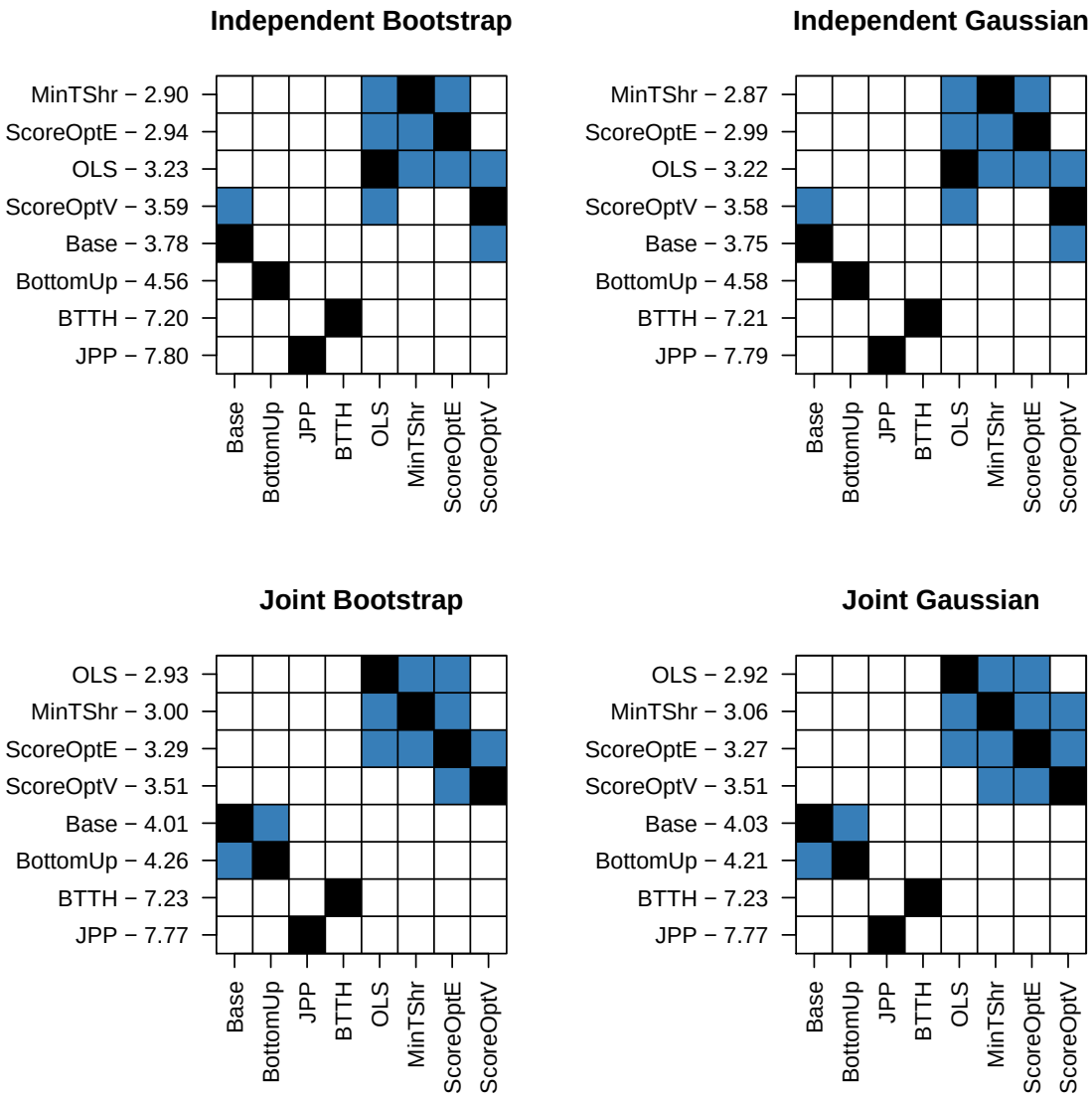

Figure 3: Nemenyi matrix for Variogram score for ARIMA modelling with a Gaussian Stationary DGP.

Table 4 and Figure 4 report the same output but using the variogram score for evaluation. For this specific DGP and for this specific the BTTH method significantly outperforms all other methods. However, this result was not observed when using BTTH for any other simulation scenario, including those reported only in the online supplement. Excluding this result, score optimisation with respect to the variogram score is the best performing

method with MinTShr and OLS also performing well. Score optimisation, OLS, MinTShr and BTTH all lead to significantly improvements relative to base forecasts, bottom up methods and the method of JPP.

Table 4: Mean Variogram score for ARIMA modelling with a Gaussian Stationary DGP. Values in bold indicate the best reconciliation method for a given base forecasting method.

Method	Ind. Bootstrap	Ind Gaussian	Joint Bootstrap	Joint Gaussian
Base	694.5012	694.6362	694.7861	694.5443
BottomUp	888.7264	891.1163	697.9468	697.7384
BTTH	**630.8518**	**632.3063**	**632.7646**	**631.8727**
JPP	1032.9742	1039.0791	1033.3112	1038.1928
MinTShr	683.3065	683.0490	677.8987	677.4923
OLS	713.4531	713.5246	680.3692	680.1407
ScoreOptE	686.0447	686.8067	683.5021	684.2645
ScoreOptV	675.1864	675.2327	673.5918	672.3905

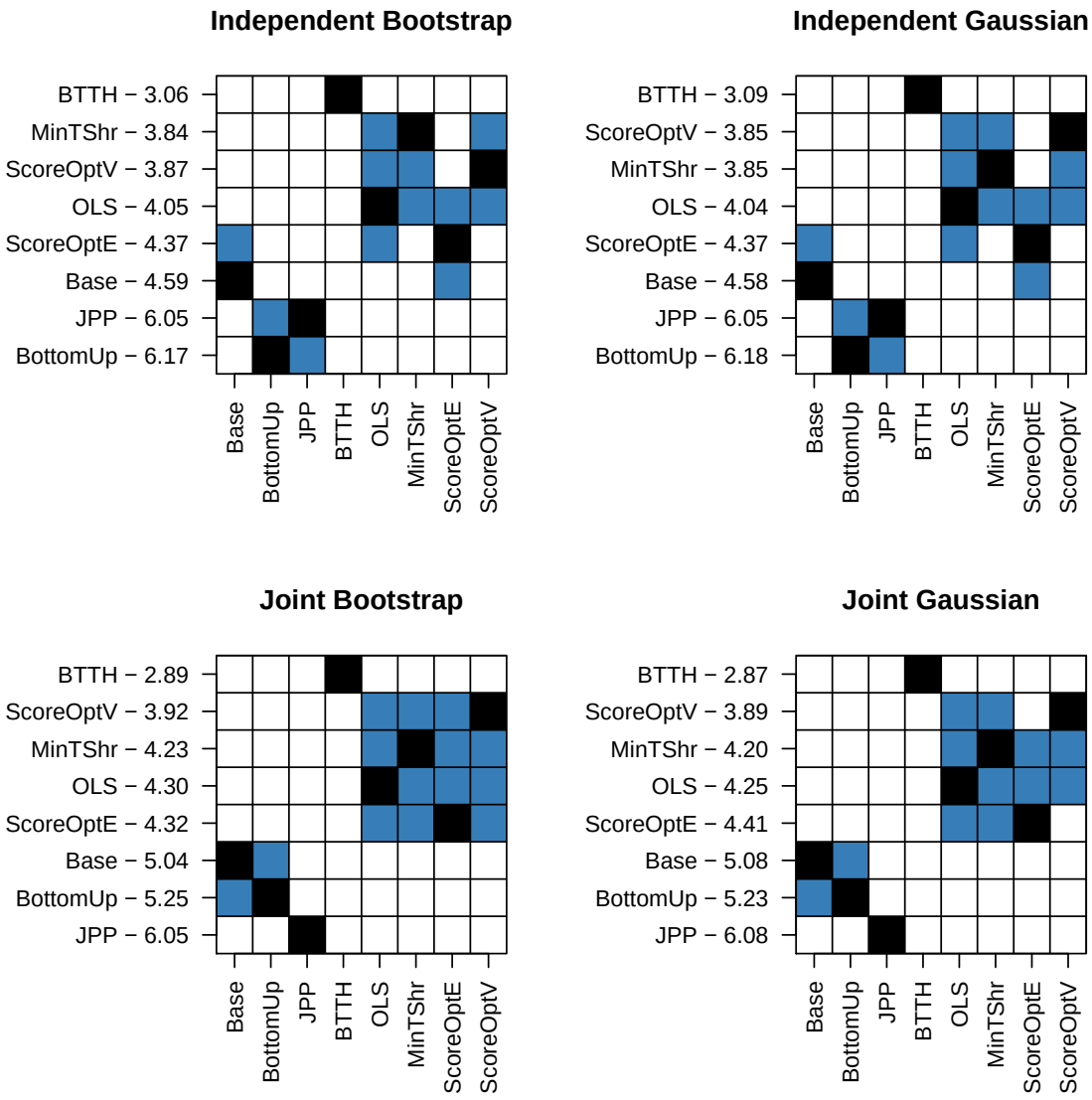

Figure 4: Nemenyi matrix for Variogram score for ARIMA modelling with a Gaussian Stationary DGP.

6.5 Results for non-Gaussian probabilistic forecast

Table 5 reports the mean energy score for the non-Gaussian DGP. Overall, the results are quite similar to the Gaussian DGP. The best performing reconciliation method is ScoreOptE when base probabilistic forecasts are independent, and MinTShr when base forecasts are

dependent. The Nemenyi matrix is omitted for brevity, but leads to similar conclusion to Figure 3. The methods ScoreOptE, MintShr and OLS are statistically indistinguishable from one another but are significantly better that base forecasts and the bottom up method. The methods BTTH and JPP lead to a statistically significant deterioration in forecast quality relative to base forecasts.

Table 5: Mean Energy score for ARIMA modelling with a Gaussian Stationary DGP. Values in bold indicate the best reconciliation method for a given base forecasting method.

Method	Ind. Bootstrap	Ind. Gaussian	Joint Bootstrap	Joint Gaussian
Base	1.4169	1.4246	1.3850	1.3912
BottomUp	1.5071	1.5305	1.4658	1.4731
BTTH	2.7704	2.9271	2.7817	2.9335
JPP	2.8805	2.9737	2.8788	2.9672
MinTShr	1.3514	1.3504	**1.3297**	**1.3367**
OLS	1.3643	1.3638	1.3409	1.3483
ScoreOptE	**1.3391**	**1.3379**	1.3384	1.3391
ScoreOptV	1.3772	1.3772	1.3746	1.3738

Finally, Table 6 and Figure 6.5 report results for the non-Gaussian DGP using the variogram score to evaluate forecasts. In this case, score optimisation with respect to the variogram score yields the best performance when base forecasts are dependent, while MinTShr yields the best performance when base forecasts are independent. In contrast to the Gaussian DGP, the JPP method leads to significant improvements over base forecasts, while the BTTH method leads to a significantly worse performance than base forecasts.

Table 6: Mean Variogram score for ARIMA modelling with a Gaussian Stationary DGP. Values in bold indicate the best reconciliation method for a given base forecasting method.

Method	Ind. Bootstrap	Ind. Gaussian	Joint Bootstrap	Joint Gaussian
Base	26.2480	26.2638	26.2339	26.2643
BottomUp	28.5977	28.7956	27.6925	27.7625
BTTH	29.2544	29.1507	29.0767	29.0854
JPP	26.3968	26.3993	26.3833	26.4030
MinTShr	**25.4412**	**25.4215**	25.4799	25.5248
OLS	25.6498	25.6362	25.6860	25.7348
ScoreOptE	25.6992	25.6954	25.5763	25.6889
ScoreOptV	25.5433	25.5325	**25.4541**	**25.4798**

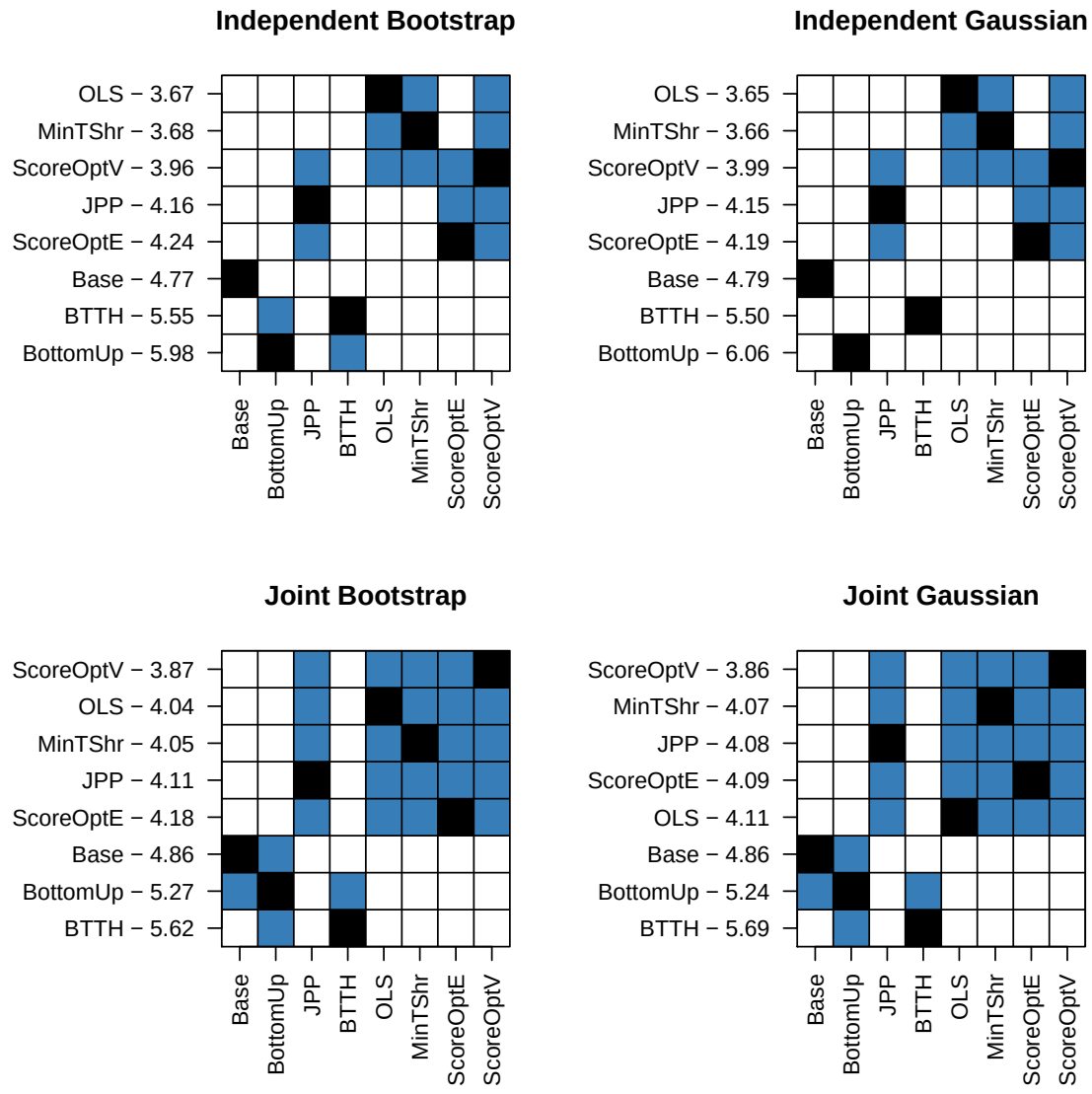

Figure 5: Nemenyi matrix for Variogram score for ARIMA modelling with a non-Gaussian Stationary DGP.

Overall, the main conclusion from the simulation study is that score optimisation leads to significant improvements in forecast performance over base forecasts irrespective of whether the DGP is Gaussian or non-Gaussian and irrespective of whether the energy or variogram score is used for evaluation. For the DGPs considered in the simulation study, MinTShr and to a lesser extent OLS also provided significant improvements over base and

bottom up forecasts. While the existing probabilistic forecast reconciliation methods considered in the literature (BTTH and JPP) performed well in some scenarios (particularly BTTH for the Gaussian DGP evaluated by variogram score), overall they lacked consistency and even led to a statistically significant deterioration in forecast quality relative to base forecasts in some settings.

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 F.

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 6 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 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 6, applying MinT(Shrink) for reconciliation of incoherent base forecasts generates the most accurate forecasts in all cases. As expected accuracy for all forecasts deteriorates as the forecast horizon increases. The skill scores presented in Figure 7 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 8 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 12 in Appendix F 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

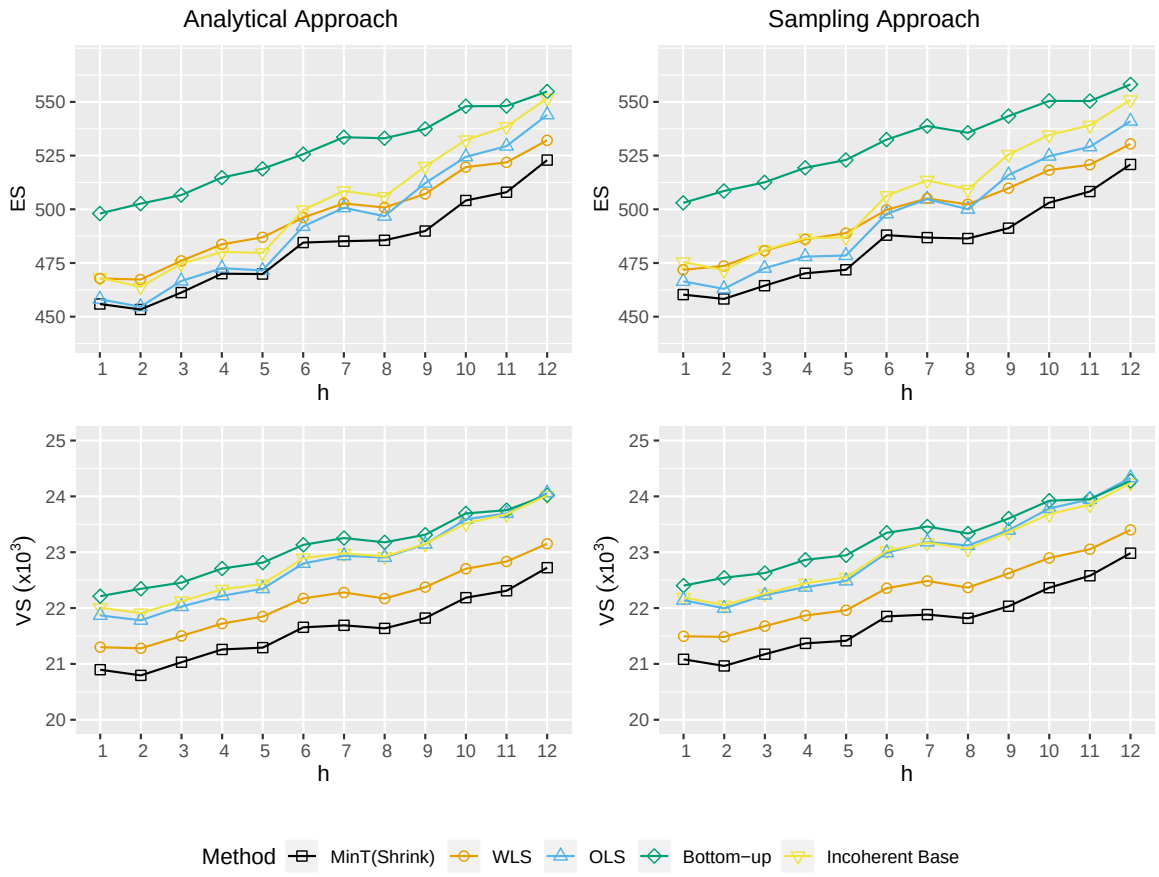

Figure 6: Energy and variogram scores for multivariate predictive distributions across the entire hierarchy. A lower (higher) score indicates a more (less) accurate forecast. Results from the analytic approach assuming Gaussian incoherent base forecasts are presented on the left while results from the non-parametric approach are presented on the right.

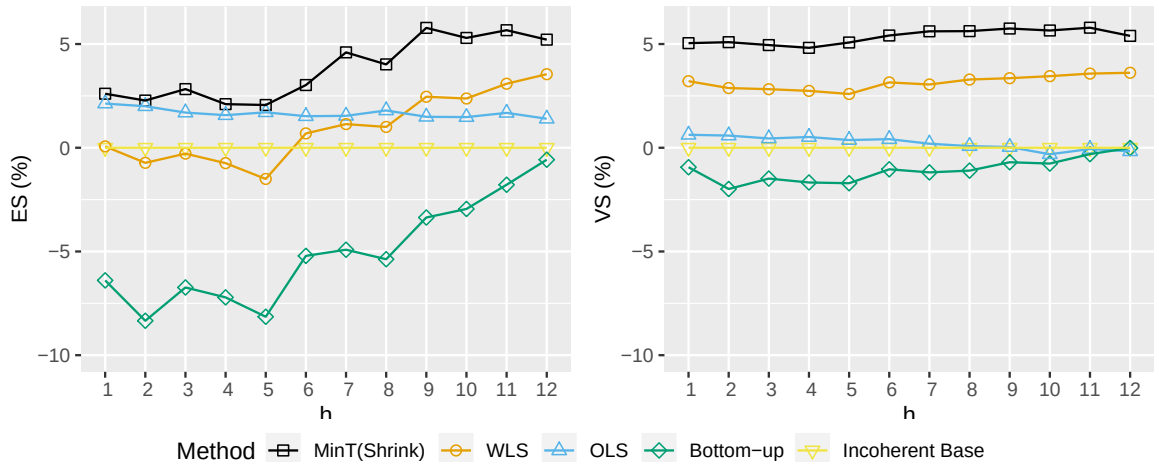

Figure 7: Skill scores (%) relative to incoherent base forecasts, across the entire Australian tourism hierarchy based on energy score (on the left) and variogram score (on the right). A higher (lower) score indicates a gain (loss) in forecast accuracy relative to the incoherent base forecasts. The results are for the analytic solution assuming Gaussian incoherent base forecasts.

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.

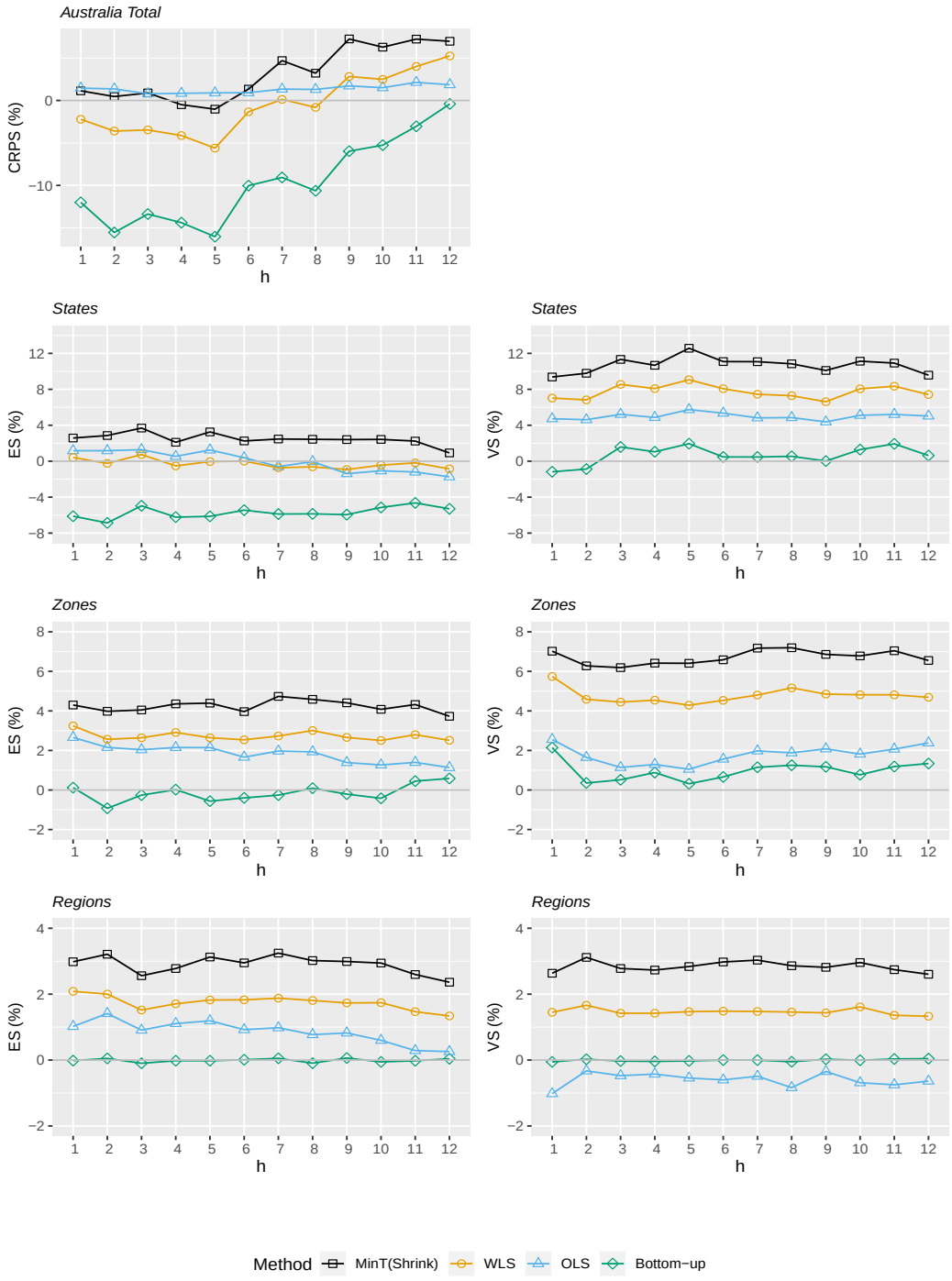

Figure 8: Skill scores (%) relative to incoherent base forecasts, for the CRPS for the top-level and energy and variogram scores for the levels below for the Australia tourism hierarchy. A higher (lower) score indicates a gain (loss) in forecast accuracy relative to the incoherent base forecasts. All results are for the analytic solution assuming Gaussian incoherent base forecasts.

8 Conclusions

Although hierarchical point forecasting is well studied in the literature, there has been a relative lack of attention given to the probabilistic setting. We fill this gap in the literature by providing a mathematically rigorous formulation of coherence and reconciliation for probabilistic forecasts.

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

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

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

In many ways this chapter sets up a substantial future research agenda. For example, having defined what amounts to an entire class of reconciliation methods for probabilistic forecasts it will be worthwhile investigating which specific projections are optimal. This is likely to depend on the specific scoring rule employed as well as the properties of the base forecasts. Another avenue worth investigating is to consider whether it is possible to recover the true predictive distribution for non-elliptical distributions via a non-linear

function $g(.)$.

A Proof of Theorem 3.1 and Theorem 3.2

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

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

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

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

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

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

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

$$\tilde{f}_{\boldsymbol{b}}(\boldsymbol{b}) = \int \hat{f}(\boldsymbol{G}^-\boldsymbol{b} + \boldsymbol{G}_\perp \boldsymbol{a}) |\boldsymbol{G}^*| d\boldsymbol{a}\,, \tag{4}$$

is the density of $\boldsymbol{b}$. To obtain the density of the full hierarchy we first augment the density in Equation (4) by $n - m$ variables denoted $\boldsymbol{u}$

$$f(\boldsymbol{b}, \boldsymbol{u}) = \tilde{f}_b(\boldsymbol{b}) \mathbb{1}\left\{\boldsymbol{u} = 0\right\} , \tag{5}$$

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

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

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

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

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

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

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

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

B Proof of Theorem 3.4

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

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

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

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

C Proof of Theorem 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,

$$\begin{aligned} LS\left(f, \boldsymbol{y}^{*}\right) &= -\log f(\boldsymbol{y}^{*}) \\ &= -\log|\boldsymbol{S^*}| - \log f_{\boldsymbol{b}}\left(\boldsymbol{b}^{*}\right) - \log\left(\mathbb{1}\left\{\boldsymbol{u}^{*} = \boldsymbol{0}\right\}\right) \\ &= -\log|\boldsymbol{S^*}| - \log f_{\boldsymbol{b}}\left(\boldsymbol{b}^{*}\right), \end{aligned} \tag{6}$$

where the third term in Equation 6 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

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

where $w_{AA,t}, w_{AB,t}, w_{BA,t}, w_{BB,t}$ are generated from ARIMA processes as described in Section 6.1 with innovations $\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}$.

For the Gaussian DGP, $u_t \sim \mathcal{N}(0, \sigma_u^2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$ and $\{\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}\} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ $\forall t$. 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,

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

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 `sn` package (Azzalini 2020). 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 Application

E.1 Results from ETS base forecasts

Figure 9: 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 10: 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 11: 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.

F 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 7: Geographic hierarchy of Australian tourism flows

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

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

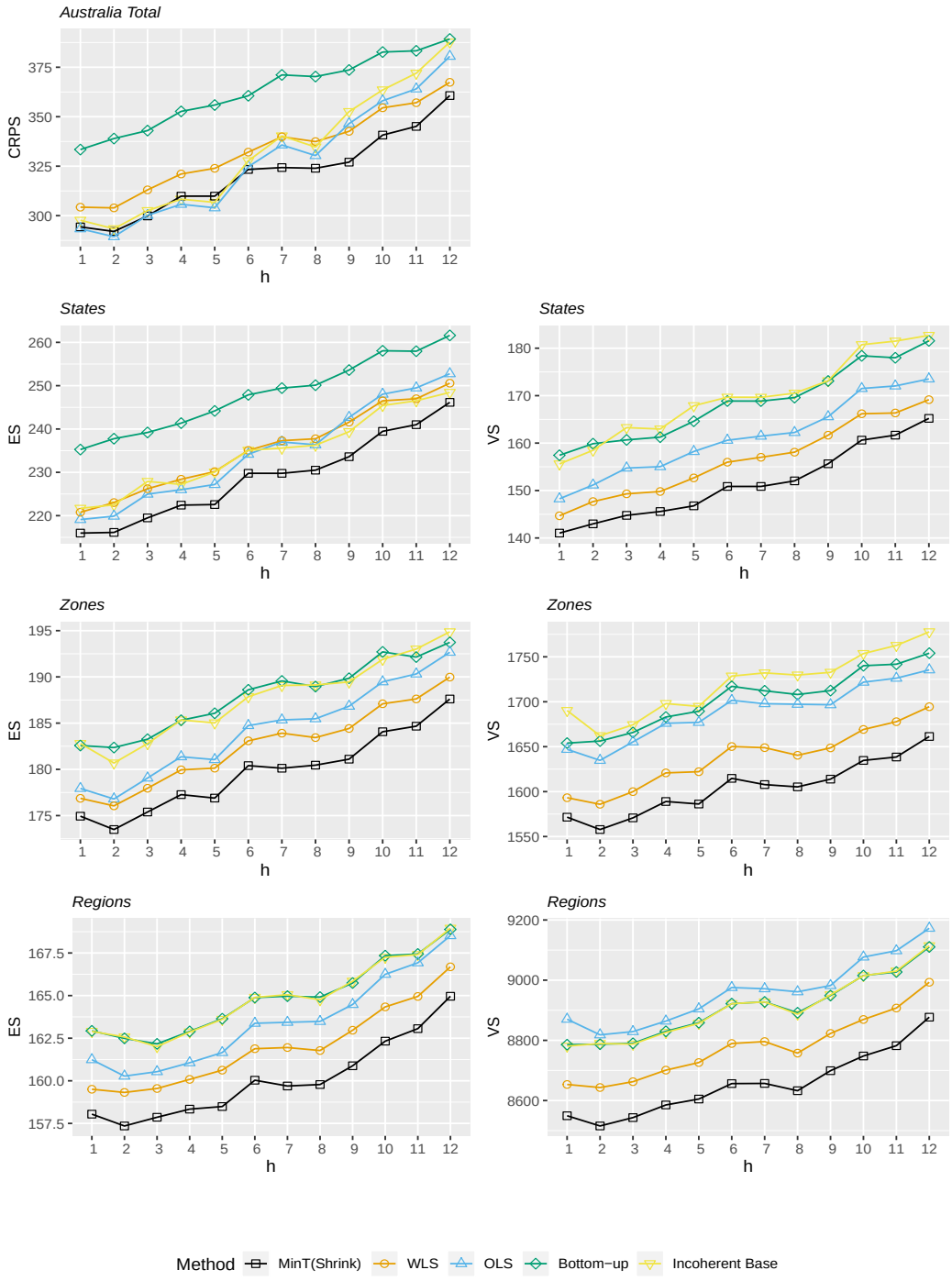

Figure 12: Forecast accuracy results across the different levels of the Australia tourism hierarchy. CRPS results are presented for the top-level and energy and variogram scores for the levels below. A lower (higher) score indicates a more (less) accurate forecast. All results are for the analytic solution 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.

Azzalini, A. (2020), *The R package* `sn`*: The Skew-Normal and Related Distributions such as the Skew-t (version 1.6-1).*, Università di Padova, Italia.

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.

Hollander, M., Wolfe, D. A. & Chicken, E. (2013), *Nonparametric statistical methods*, Vol. 751, John Wiley & Sons.

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* .

Kourentzes, N. (2019), *tsutils: Time Series Exploration, Modelling and Forecasting*. R package version 0.9.0.
URL: *https://CRAN.R-project.org/package=tsutils*

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.