

Probabilistic Forecast Reconciliation: Properties, Evaluation and Score Optimisation

Anastasios Panagiotelis*

Discipline of Business Analytics,

University of Sydney, NSW 2006, Australia.

Email: anastasios.panagiotelis@sydney.edu.au

and

Puwasala Gamakumara

Department of Econometrics and Business Statistics,

Monash University, VIC 3800, Australia.

Email: puwasala.gamakumara@monash.edu

and

George Athanasopoulos

Department of Econometrics and Business Statistics,

Monash University, VIC 3145, 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

March 25, 2021

*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 of multivariate data that follow known linear constraints. This is particularly common in forecasting where some variables are aggregates of others, but also arises in other prediction settings. For point prediction, an increasingly popular technique is reconciliation, whereby predictions are made for all series (so-called base predictions) and subsequently adjusted to cohere with the constraints. We extend reconciliation from 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, we prove that true predictive distributions can be recovered using reconciliation even when the location and scale of base predictions are chosen arbitrarily. Reconciliation weights are estimated to optimise energy or variogram score. The log score is not considered since it is improper when comparing unreconciled to reconciled predictions, a result also proved in this paper. Due to randomness in the objective function, optimisation uses stochastic gradient descent. This method improves upon base predictions in simulated and empirical data, particularly when the base prediction models are severely misspecified. For milder misspecification, extending popular reconciliation methods for point prediction results in similar performance to score optimisation.

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

1 Introduction

Many multivariate prediction problems involve data that follow some linear constraints. For instance, in energy, 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 and Athanasopoulos, 2021, 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 prediction. This paper develops a formal framework for probabilistic reconciliation, derives theoretical results that allow reconciled probabilistic predictions to be constructed and evaluated, and proposes an algorithm for optimal probabilistic reconciliation with respect to a proper scoring rule.

Before describing the need for probabilistic reconciliation we briefly review the literature on point forecast¹ 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 and Sohl, 1990, and references therein). An alternative approach emerged with Athanasopoulos et al. (2009) and Hyndman et al. (2011) who recommended producing forecasts of all series (referred to as ‘base’ forecasts) 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 and Cugliari, 2015; Nystrup et al., 2020), a penalised version thereof (Ben Taieb and Koo, 2019), or the trace of the forecast error covariance (Wickramasuriya et al., 2019).

There is now substantial theoretical and empirical evidence that reconciliation improves forecast accuracy relative to base forecasts, bottom-up and top-down methods. The improvement relative to bottom-up, may at first seem counter-intuitive since aggregates contain no additional information relative to the disaggregate series. However, aggregation can filter out noise leading to smoother signals that can be detected by certain statistical models. In contrast the disaggregate data can be noisy, leading to higher parameter uncertainty and risk model of model mis-specification. Reconciliation thus provides a mechanism for combining information between different forecasting approaches at different levels of aggregation.

In contrast to point forecasts, the entire probability distribution of future values provides a full description of the uncertainty associated with the predictions (Gneiting and Katzfuss, 2014). The importance of probabilistic forecasts can be seen in decision making settings in risk management, when it is critical to quantify the probability of extreme events. Therefore

¹Such has been the dominance 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.

probabilistic forecasting has become of great interest in many disciplines such as, economics (Rossi, 2014), meteorological studies (McLean Sloughter et al., 2013), energy forecasting (Ben Taieb et al., 2017) and retail forecasting (Böse et al., 2017). An early attempt towards probabilistic forecast reconciliation came from Shang and 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 to 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 bottom-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 our best 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 that 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 reorder 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. This will be problematic in applications with fewer observations

than the smart meter data they consider. Further, Ben Taieb et al. (2020) only incorporate information from the forecast mean of aggregate variables, missing out on potentially valuable information in the probabilistic forecasts of aggregate data.

In this paper we seek 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. (2021) for point forecast reconciliation. 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 showing that the log score is improper when used to compare reconciled to unreconciled forecasts. Fifth, we propose an algorithm for choosing reconciliation weights by optimising a scoring rule. This algorithm exploits advances in stochastic gradient descent and is thus suited to scoring rules which are often only known up to an approximation. The algorithm and other methodological contributions described in this paper are implemented in the *ProbReco* package (Panagiotelis, 2020).

The remainder of the paper is structured as follows. In Section 2, after a 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 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 forecasting electricity generation from different sources is contained in Section 7. Finally Section 8 concludes with some discussion and thoughts on future research.

2 Hierarchical probabilistic forecasts

Before extending coherence and reconciliation to the probabilistic setting, we briefly refresh these concepts for point forecasts. We follow the geometric interpretation introduced by Panagiotelis et al. (2021), since this formulation naturally generalises to probabilistic forecasting.

2.1 Point Forecasting

A *hierarchical time series* is a collection of time series adhering to linear constraints. Stacking the value of each series at time t into an n -vector \mathbf{y}_t , the constraints imply that \mathbf{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 $\mathbf{b}_t \in \mathbb{R}^m$ 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 $\mathbf{y}_t = \mathbf{S}\mathbf{b}_t$, where \mathbf{S} is an $n \times m$ constant matrix for a given hierarchical structure.

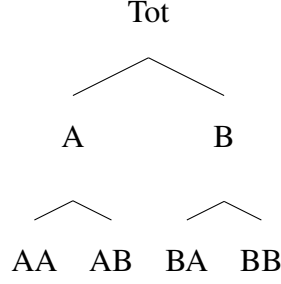


Figure 1: An example of a two-level hierarchical structure.

An example of a hierarchy is shown in Figure 1. There are $n = 7$ series of which $m = 4$ are bottom-level series. Also, $\mathbf{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $\mathbf{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, \mathbf{b}_t']'$, and

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

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

The connection between this characterisation and the coherent subspace is that the columns of \mathbf{S} span \mathfrak{s} . Below, the notation $s : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is used when premultiplication by \mathbf{S} is thought of as a mapping. Finally, while \mathbf{S} is defined in terms of m bottom-level series here, in general any m series can be chosen with the \mathbf{S} matrix redefined accordingly. The columns of all appropriately defined \mathbf{S} matrices span the same coherent subspace \mathfrak{s} .

When forecasts of all n series are produced, they may not adhere to constraints. Such forecasts are called *incoherent base* forecasts and are denoted $\hat{\mathbf{y}}$. To exploit the fact that the target of the forecast adheres to known linear constraints, base forecasts can be adjusted in a process known as *forecast reconciliation*. This involves selecting a mapping $\psi : \mathbb{R}^n \rightarrow \mathfrak{s}$ and then setting $\tilde{\mathbf{y}} = \psi(\hat{\mathbf{y}})$, where $\tilde{\mathbf{y}} \in \mathfrak{s}$ is called the *reconciled* forecast. The mapping ψ may be considered as the composition of two mappings $\psi = s \circ g$. Here, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 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 (Athanasopoulos et al., 2017) and MinT (Wickramasuriya et al., 2019) methods, are special cases where g is premultiplication by a matrix \mathbf{G} and where $\mathbf{S}\mathbf{G}$ is a projection matrix. These are summarised in Table 1.

Table 1: *Summary of reconciliation methods for which $\mathbf{S}\mathbf{G}$ is a projection matrix. Here \mathbf{W} is some diagonal matrix, $\hat{\Sigma}_{sam}$ is a sample estimate of the residual covariance matrix and $\hat{\Sigma}_{shr}$ is a shrinkage estimator proposed by Schäfer and Strimmer (2005), given by $\tau \text{diag}(\hat{\Sigma}_{sam}) + (1 - \tau)\hat{\Sigma}_{sam}$ where $\tau = \frac{\sum_{i \neq j} \hat{V}ar(\hat{\sigma}_{ij})}{\sum_{i \neq j} \hat{\sigma}_{ij}^2}$ and σ_{ij} denotes the (i, j) th element of $\hat{\Sigma}_{sam}$.*

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

2.2 Coherent probabilistic forecasts

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

Definition 2.1 (Coherent Probabilistic Forecasts). Given the triple, $(\mathbb{R}^m, \mathcal{F}_{\mathbb{R}^m}, \nu)$, a coherent probability triple $(s, \mathcal{F}_s, \check{\nu})$, is given by s , the σ -algebra \mathcal{F}_s and a measure $\check{\nu}$, such that

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

To our best 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.

2.3 Probabilistic forecast reconciliation

Let $(\mathbb{R}^n, \mathcal{F}_{\mathbb{R}^n}, \hat{\nu})$ be a probability triple characterising a probabilistic forecast for all n series. The hat is used for $\hat{\nu}$ analogously with \hat{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 $\mathcal{F}_{\mathfrak{s}}$.

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

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

where $\psi^{-1}(\mathcal{A}) := \{\mathbf{y} \in \mathbb{R}^n : \psi(\mathbf{y}) \in \mathcal{A}\}$ is the pre-image of \mathcal{A} , that is the set of all points in \mathbb{R}^n that $\psi(\cdot)$ 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 transforming an incoherent forecast. For probabilistic forecasts, sets of points are transformed to sets of points, with the same probabilities assigned to these sets under the base and reconciled measures respectively. Also, since ψ can be expressed as a composition $s \circ g$, a reconciled probabilistic distribution ν can be obtained for m series such that $\nu(\mathcal{B}) = \hat{\nu}(g^{-1}(\mathcal{B}))$ for all $\mathcal{B} \in \mathcal{F}_{\mathbb{R}^m}$. 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 $\mathcal{F}_{\mathbb{R}^n}$ 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 most applications are high-dimensional. The arrow labelled \mathbf{S} spans an m -dimensional coherent subspace \mathfrak{s} , while the arrow labelled \mathbf{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.

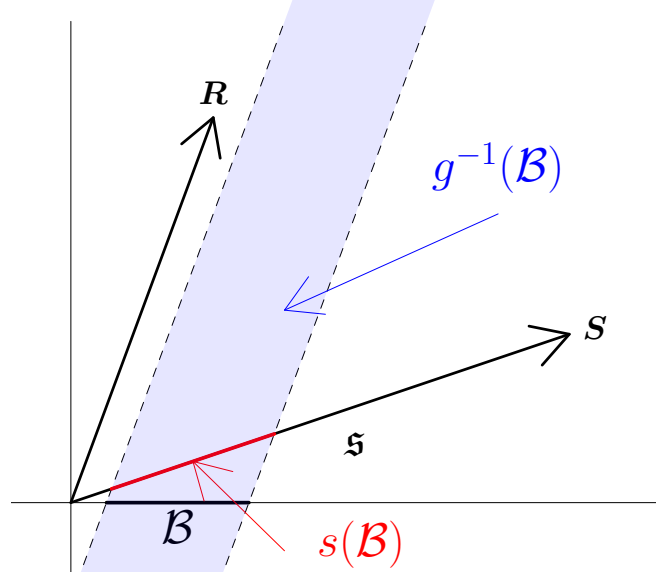


Figure 2: Summary of probabilistic forecast reconciliation. The probability that \mathbf{y}_{t+h} lies in the red segment under the reconciled probabilistic forecast equals the probability that \mathbf{y}_{t+h} lies in the shaded blue area under the unreconciled probabilistic forecast. Since most applications are high-dimensional, 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*

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

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

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

Proof. See Appendix ??.

□

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

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

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

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

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

Proof. See Appendix ??.

□

Applying this result in the Gaussian case is shown in Appendix ?? in the online supplement.

3.2 Elliptical distributions

More generally, consider linear reconciliation of the form $\psi(\hat{y}) = S(d + G\hat{y})$. For an elliptical base probabilistic forecast, with location $\hat{\mu}$ and scale $\hat{\Sigma}$, the reconciled probabilistic forecast will also be elliptical with location $\tilde{\mu} = S(d + G\hat{\mu})$ and scale $\tilde{\Sigma} = SG\hat{\Sigma}G'S'$. 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 (Recovering the true density through reconciliation). *Assume the true predictive distribution is elliptical with location μ and scale Σ . Then for an elliptical base probabilistic forecast with arbitrary location $\hat{\mu}$ and scale $\hat{\Sigma}$, there exists d_{opt} and G_{opt} such that the true predictive distribution is recovered by reconciliation.*

Proof. First consider finding a G_{opt} for which the following holds,

$$\Sigma = SG_{\text{opt}}\hat{\Sigma}G'_{\text{opt}}S'. \quad (1)$$

Note that since the true data must be coherent Σ will not be full rank, while in contrast $\hat{\Sigma}$ usually will be (but need not be) full rank. Equation (1) can be solved as $G_{\text{opt}} = \Omega_0^{1/2}\hat{\Sigma}^{-1/2}$, where $\hat{\Sigma}^{1/2}$ is any matrix such that $\hat{\Sigma} = \hat{\Sigma}^{1/2}(\hat{\Sigma}^{1/2})'$.² Here, $\Omega_0^{1/2}(\Omega_0^{1/2})' = \Omega$ and Ω is the true scale matrix for the bottom-level series. To ensure conformability of matrix multiplication, $\Omega^{1/2}$ must be an $m \times n$ matrix; so it can be set to the Cholesky factor of Ω augmented with an

²For example a Cholesky factor which will be unique if $\hat{\Sigma}$ is full rank. If $\hat{\Sigma}$ is rank deficient, then although the Cholesky factor is no longer unique, any $\hat{\Sigma}^{1/2}$ such that $\hat{\Sigma} = \hat{\Sigma}^{1/2}(\hat{\Sigma}^{1/2})'$ can be used

additional $n - m$ columns of zeros. To reconcile the location, solve the following for \mathbf{d}_{opt}

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

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

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

Although $\mathbf{S}\mathbf{G}_{\text{opt}}$ 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 distribution is also elliptical, then this can be recovered via reconciliation using a projection if $\text{rank}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \leq n - m$.*

Proof. See Appendix ??.

\square

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 $(\hat{\mathbf{y}}^{[1]}, \dots, \hat{\mathbf{y}}^{[L]})$ is a sample drawn from an incoherent probability measure $\hat{\nu}$. Then $(\tilde{\mathbf{y}}^{[1]}, \dots, \tilde{\mathbf{y}}^{[L]})$ where $\tilde{\mathbf{y}}^{[\ell]} := \psi(\hat{\mathbf{y}}^{[\ell]})$ for $\ell = 1, \dots, L$, is a sample drawn from the reconciled probability measure $\tilde{\nu}$ as defined in Definition 2.2.*

Proof. For any $\mathcal{A} \in \mathcal{F}_s$

$$\begin{aligned}
\Pr(\hat{\mathbf{y}} \in \psi^{-1}(\mathcal{A})) &= \lim_{L \rightarrow \infty} \sum_{\ell=1}^L \mathbb{1}\{\hat{\mathbf{y}}^{[\ell]} \in \psi^{-1}(\mathcal{A})\} \\
&= \lim_{L \rightarrow \infty} \sum_{\ell=1}^L \mathbb{1}\{\psi(\hat{\mathbf{y}}^{[\ell]}) \in (\mathcal{A})\} \\
&= \Pr(\tilde{\mathbf{y}} \in (\mathcal{A}))
\end{aligned}$$

□

This result implies that reconciling each member of a sample drawn from an incoherent distribution provides a sample from the reconciled distribution, a strategy already 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 and Raftery, 2007; Gneiting and 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 $E_Q[K(Q, \omega)] \leq E_Q[K(P, \omega)]$ 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 inherently a multivariate problem (the linear constraints affect all variables), our focus is on multivariate scoring rules. Arguably the simplest multivariate scoring rule is the log score. The log score simply evaluates the negative log density at the value of the realisation, $LS(P, \omega) = -\log f(\omega)$, 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.

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 Gneiting and Raftery, 2007, for details) which is a multivariate generalisation of the popular Cumulative Rank Probability Score (CRPS). The energy score is given by

$$ES(P, \omega) = E_P ||\mathbf{y} - \omega||^\alpha - \frac{1}{2} E_P ||\mathbf{y} - \mathbf{y}^*||^\alpha, \quad \alpha \in (0, 2], \quad (2)$$

where \mathbf{y} and \mathbf{y}^* are independent copies drawn from the distribution P . In the empirical results described later, we follow common convention by setting $\alpha = 1$. While the expectations in Equation (2) may have no closed form, they can be easily approximated via simulations using a sample drawn from the probabilistic forecast. Other scores with similar behaviour are kernel-based scores (Gneiting and Raftery, 2007) and the variogram score (Scheuerer and Hamill, 2015).

4.1 The Log Score for Hierarchical Time Series

When an 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 a 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 ??.

□

As a result of Theorem 4.1 we recommend avoiding the log score when comparing reconciled and unreconciled probabilistic forecasts.

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. For example, in the simulations and the empirical application of Sections 6 and 7 we consider optimising with respect to both the energy and variogram scores. Motivated by Theorem 3.3, which shows that the true predictive density can be recovered (albeit only infeasibly) by linear reconciliation we consider reconciliation of the form $\tilde{\mathbf{y}} = \psi_{\boldsymbol{\gamma}}(\hat{\mathbf{y}}) = \mathbf{S}(\mathbf{d} + \mathbf{G}\hat{\mathbf{y}})$, where $\boldsymbol{\gamma} := (\mathbf{d}, \text{vec}(\mathbf{G}))$. This allows for more flexibility than a projection, which would imply the constraints $\mathbf{d} = \mathbf{0}$ and $\mathbf{GS} = \mathbf{I}$. This added flexibility is also 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 an 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

$$\mathcal{E}(\boldsymbol{\gamma}) = \sum_{t=T}^{T+R-1} ES(\tilde{f}_{t+h|t}^{\boldsymbol{\gamma}}, \mathbf{y}_{t+h}), \quad (3)$$

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

One of the challenges in 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

simulation as

$$\hat{\mathcal{E}}(\boldsymbol{\gamma}) = \sum_{t=T}^{T+R-1} \left[\frac{1}{Q} \left(\sum_{q=1}^Q \|\tilde{\mathbf{y}}_{t+h|t}^{[q]} - \mathbf{y}_{t+h}\| - \frac{1}{2} \|\tilde{\mathbf{y}}_{t+h|t}^{[q]} - \tilde{\mathbf{y}}_{t+h|t}^{*[q]}\| \right) \right], \quad (4)$$

where $\tilde{\mathbf{y}}_{t+h|t}^{[q]} = \mathbf{S}(\mathbf{d} + \mathbf{G}\hat{\mathbf{y}}_{t+h|t}^{[q]})$, $\tilde{\mathbf{y}}_{t+h|t}^{*[q]} = \mathbf{S}(\mathbf{d} + \mathbf{G}\hat{\mathbf{y}}_{t+h|t}^{*[q]})$ and $\hat{\mathbf{y}}_{t+h|t}^{[q]}, \hat{\mathbf{y}}_{t+h|t}^{*[q]} \stackrel{iid}{\sim} \hat{f}_{t+h|t}$ for $q = 1, \dots, Q$.

The objective function is optimised by Stochastic Gradient Descent (SGD). The SGD method 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 and Welling, 2013). There is also a recent but growing literature on using SGD to optimise scoring rules (see Gasthaus et al., 2019; Janke and Steinke, 2020; Hofert et al., 2020, and references therein for examples). These are typically high dimensional problems, deep neural networks handle millions of parameters, so this tool is well suited to our problem. It requires an estimate of the gradient $\partial \hat{\mathcal{E}} / \partial \boldsymbol{\gamma}$ which is computed by automatically differentiating Equation (4) 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 and Ba, 2014, for details). Pseudo-code for the full procedure in the case where $h = 1$ is provided in Algorithm 1 and is implemented in the R package *ProbReco* (Panagiotelis, 2020).

While Algorithm 1 is not the first instance of calibrating parameters by optimising scoring rules (see Gneiting et al., 2005, for an earlier example), to the best of our knowledge it is the first instance of doing so in the context of forecast reconciliation. It is amenable to parallel computing architectures: the loop beginning at line 2 of the pseudo-code of Algorithm 1 can be done in parallel as can the computation of the gradient. Finally, the total score in Equation (3) 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.

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

```

1: procedure SCOREOPT( $\mathbf{y}_1, \dots, \mathbf{y}_{T+R}, \beta_1, \beta_2, \epsilon, \eta$ ).
2:   for  $t = T : T + R - 1$  do
3:     Find base forecasts  $\hat{f}_{t+1|t}$  using  $t - T + 1, t - T + 2, \dots, t$  as training data.
4:   end for
5:   Initialise  $\mathbf{m}_0 = \mathbf{0}, \mathbf{v}_0 = \mathbf{0}$  and  $\gamma_0 = (\mathbf{0}, \text{vec}((\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'))$ 
6:   for  $j = 1, 2, 3, \dots$  up to convergence do
7:     Draw  $\hat{\mathbf{y}}_{t+1|t}^{[q]}, \hat{\mathbf{y}}_{t+1|t}^{*[q]} \sim \hat{f}_{t+1|t}$  for  $q = 1, \dots, Q, t = T, \dots, T + R - 1$ .
8:     Compute  $\tilde{\mathbf{y}}_{t+1|t}^{[q]}$  and  $\tilde{\mathbf{y}}_{t+1|t}^{*[q]}$  for  $q = 1, \dots, Q, t = T, \dots, T + R - 1$  using  $\gamma_{j-1}$ .
9:      $\mathbf{g}_j \leftarrow \partial \hat{\mathcal{E}} / \partial \gamma|_{\gamma=\gamma_{j-1}}$  ▷ Compute gradient
10:     $\mathbf{m}_j \leftarrow \beta_1 \mathbf{m}_{j-1} + (1 - \beta_1) \mathbf{g}_j$  ▷ Moving average of gradient
11:     $\mathbf{v}_j \leftarrow \beta_2 \mathbf{v}_{j-1} + (1 - \beta_2) \mathbf{g}_j^2$  ▷ Moving average of squared gradient
12:     $\hat{\mathbf{m}}_j \leftarrow \mathbf{m}_j / (1 - \beta_1^j)$  ▷ Bias correct
13:     $\hat{\mathbf{v}}_j \leftarrow \mathbf{v}_j / (1 - \beta_2^j)$  ▷ Bias correct
14:     $\gamma_j \leftarrow \gamma_{j-1} + \eta \frac{\hat{\mathbf{m}}_j}{(\hat{\mathbf{v}}_j + \epsilon)}$  ▷ Update weights
15:   end for
16:   Set the reconciled forecast as  $\tilde{f}_{T+R+1|T+R}^{\gamma_{\text{opt}}}$  where  $\gamma_{\text{opt}}$  is the converged value of  $\gamma$ .
17: end procedure

```

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 for the SGD 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 and Ba (2014) and used in popular software packages such

as TensorFlow, Keras and Torch amongst others. Convergence is achieved when the change in all gradients is less than 10% of the step size η . The number of sample periods used to construct the objective function is $R = 250$, while the number of draws used to estimate each score is $Q = 250$. All estimation of base models uses a sample size of $T = 500$. All forecast evaluations are 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 $\text{ARIMA}(p, d, q)$ 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 are 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. A complete set of results are available at the github repository <https://git.io/JJwQB>.

We consider a multivariate Gaussian and a non-Gaussian setting for the errors driving the ARIMA processes. Specifically, the non-Gaussian errors are drawn from a meta-distribution of a Gumbel copula with $\text{Beta}(1, 3)$ margins. After simulating from the ARIMA models, additional noise is added to ensure bottom-level series have a lower signal-to-noise ratio than bottom-level series with details provided in Appendix ?? of the online supplement. 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 and

Khandakar (2008), 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. We also considered exponential smoothing (ETS) models using the `ETS()` function in the `fable` package. These are omitted for brevity; please refer to <https://git.io/JJwQB> for a full set of results.

Let $\hat{\mathbf{y}}_{t+h|t} = (\hat{y}_{1,t+h|t}, \dots, \hat{y}_{n,t+h|t})'$, where $\hat{y}_{i,t+h|t}$ is the h -step-ahead point forecast for series i , and $\mathbf{E} := \{e_{i,t}\}_{i=1,\dots,n;t=1,\dots,T}$ is an $(n \times T)$ matrix of stacked residuals $e_{i,t}$. For each series and model, base probabilistic forecasts for $h = 1$ are constructed in the following four ways:

- **Independent Gaussian:** The base probabilistic forecast is made up of independent Gaussian distributions with the forecast mean and variance of variable i given by $\hat{y}_{i,t+h|t}$ and $\hat{\sigma}_{i,t+h|t}^2$, where $\hat{\sigma}_{i,t+h|t}^2$ is the sample variance of the residuals in the i th row of \mathbf{E} .
- **Joint Gaussian:** The base probabilistic forecast is a multivariate Gaussian distribution with the forecast mean $\hat{\mathbf{y}}_{t+h|t}$ and variance covariance matrix $\hat{\Sigma}$, where $\hat{\Sigma}$ is the variance covariance matrix of the residuals.
- **Independent Bootstrap:** Draw from the base probabilistic forecast independently for each variable as $\hat{y}_{i,t+h|t} + e_{i,\tau}$ with τ is drawn randomly (with replacement) from $1, 2, \dots, T$.
- **Joint Bootstrap:** Draw from the joint probabilistic forecast with $\hat{\mathbf{y}}_{t+h|t} + \mathbf{e}_\tau$ where \mathbf{e}_τ is the τ th column of \mathbf{E} , where τ is drawn randomly (with replacement) from $1, 2, \dots, T$.

We restrict our attention to $h = 1$ although these methods can be generalised to larger h using the recursive method (Hyndman and Athanasopoulos, 2021). For multi-step-ahead forecasts, bootstraps should be block-wise 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 reconciling 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-multiplied by $S (S'WS)^{-1} S'W$ where 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 of Wickramasuriya et al. (2019) described in Table 1.
- **BottomUp:** Reconciliation via premultiplication by SG where $G = (\mathbf{0}_{m \times (n-m)}, \mathbf{I}_{m \times m})$.
- **OLS:** Reconciliation via pre-multiplication by $S (S'S)^{-1} S'$.
- **MinTShr:** Reconciliation via pre-multiplication using the shrinkage 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 previously exist in the literature. The methods BottomUp, OLS, and MinTShr have been used extensively for point forecasting but their application to probabilistic forecasting for general base forecasts is, to our best knowledge, novel.

In addition to these, two further reconciliation methods were considered; WLS, which reconciles via pre-multiplication by the matrix used in Jeon et al. (2019) but with no 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 omitted for brevity; please refer to <https://git.io/JJwQB> for a full set of results.

6.4 Energy score results for Gaussian probabilistic forecasts

The left panel of Figure 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, the MinT method for reconciliation using the shrinkage estimator (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.

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 4 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 outperform base forecasts, bottom-up forecasts, BTTH and JPP.

The right panel of Figure 3 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; please refer to <https://git.io/JJwQB> for a full set of results. However, these lead to similar conclusion

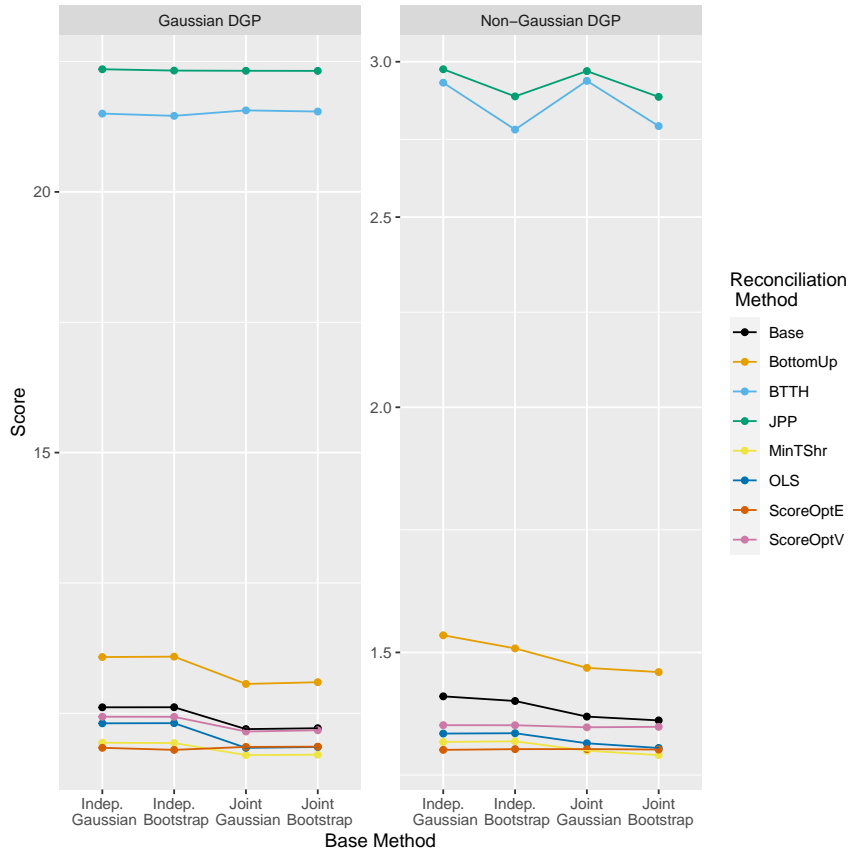


Figure 3: Mean energy scores using different base forecast and reconciliation methods. Left panel is the Gaussian data, right panel is the non-Gaussian data.

to Figure 4. The methods ScoreOptE, MinTShr and OLS are statistically indistinguishable from one another but are significantly better than 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.

Similar conclusions can be drawn based on the results for both Gaussian and non-Gaussian probabilistic forecasts considering the mean variogram score. These are presented in Appendix ??.

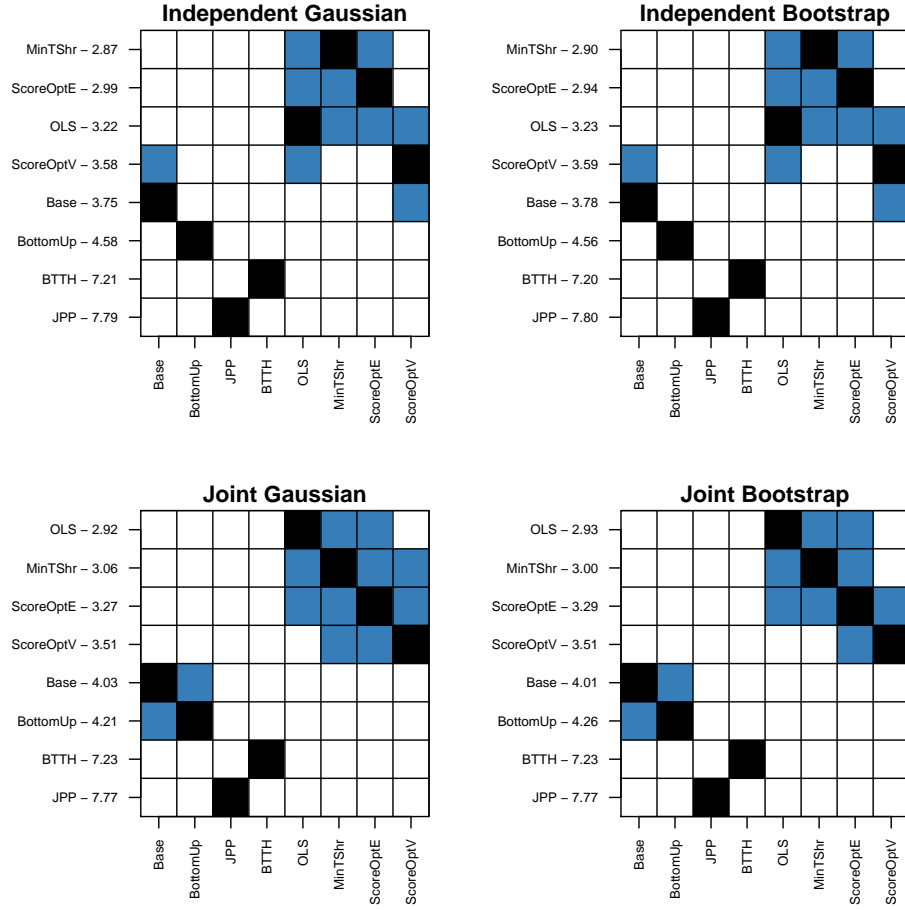


Figure 4: Nemenyi matrix for Energy Score for Gaussian DGP.

7 Forecasting Australian Electricity Generation

7.1 Data Description and Base Forecasts

To demonstrate the potential of the proposed methods, we consider an application to forecasting electricity generation from different sources of energy. Daily time series were obtained from opennem.org.au, a website that compiles publicly available data from the Australian Energy Market Operator (AEMO). Probabilistic day-ahead forecasts are crucial inputs into operational and planning decisions that ensure efficiency and stability of the power network. This has become a more challenging problem with growth in intermittent sources of generation such as wind and solar. The hierarchy comprises three levels of aggregation.

1. *Total* generation is the sum of generation from *Renewable* and *non-Renewable* sources.

2. *Renewable* generation is the sum of *Batteries*, *Hydro (inc. Pumps)*, *Solar*, *Wind* and *Biomass*. *Non-Renewable* is the sum of *Coal*, *Gas* and *Distillate*
3. *Battery* generation is given by *Battery (Discharging)* minus *Battery (Charging)*, *Hydro (inc. Pumps)* is *Hydro* generation minus *Pumps* (energy used to pump water upstream), while *Solar* generation is the sum of *Solar (Rooftop)* and *Solar (Utility)*. *Coal* generation is the sum of *Black Coal* and *Brown Coal*, while *Gas* is the sum of *Gas (OCGT)*, *Gas (CCGT)*, *Gas (Steam)*, *Gas (Reciprocating)*.

In total, there are $n = 23$ series of which $m = 15$ are bottom-level series.

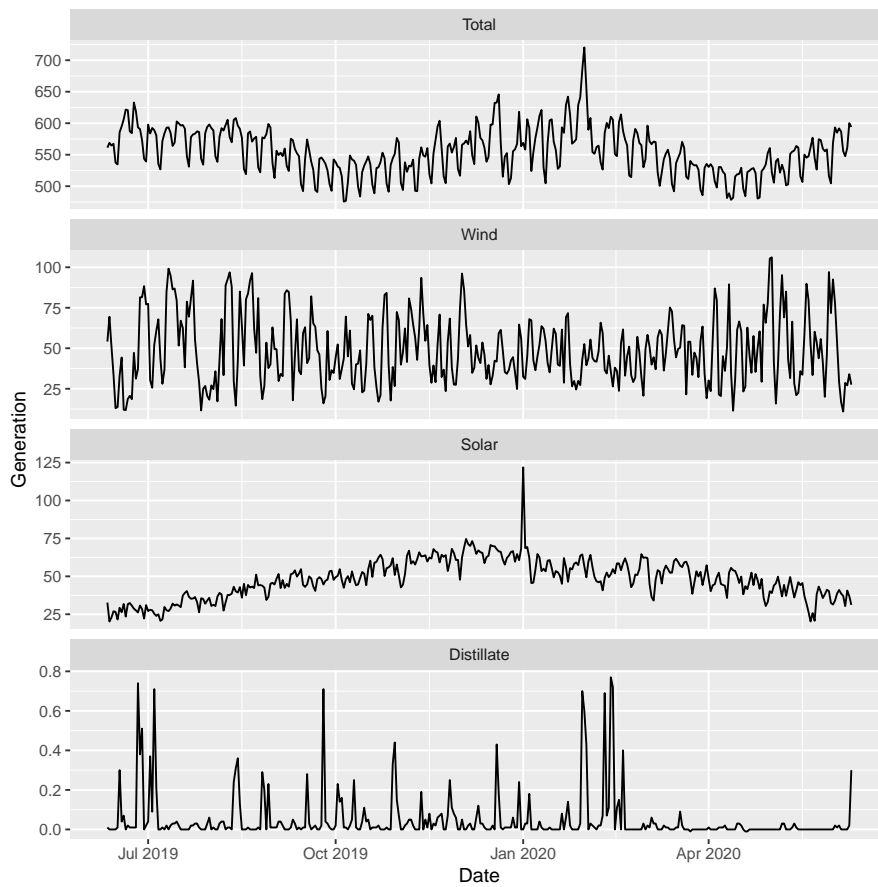


Figure 5: Time series plots for selected series from 11 June 2019 to 10 June 2020.

Figure 5 shows time plots for some selected series³. The series exhibit some interesting and unique features. At the aggregate level, *Total* generation shows strong weekly seasonality, with troughs corresponding to weekends. An annual seasonal pattern is also displayed with

³Time plots for the remaining series area available from <https://git.io/JJwd0>.

peaks occurring during the months of June–August as well as December–February. These periods correspond to the winter and summer months in Australia for which electricity demand peaks for heating and cooling purposes respectively. As expected, generation from *Solar* peaks during the summer months of December–February. There are also some unusually large spikes observed in both the *Total* and *Solar* series during February and January 2020 respectively. *Wind* displays higher volatility (especially outside the summer months), while generation from *Distillate* exhibits aperiodic spikes. The diversity and prominence of the features in each series and each level of aggregation highlights the importance of modelling and forecasting each series on its own merits and then applying a reconciliation approach.

The forecast evaluation is based on a rolling window. Each training window consists of 140 days (20 weeks) of data. One-step-ahead forecasts were generated leading to 170 daily forecasts for evaluation. Each series was independently modelled using a one-layer feed-forward neural network with up to 28 lags of the target variable as inputs. This was implemented using the `NNETAR` function in the `fable` package. Neural networks are used to highlight the versatility of reconciliation to different forecasting approaches. While including more layers or meteorological variables as predictors will probably lead to improved base forecasts, the primary objective is to assess the effectiveness of different forecast reconciliation methods. For base forecasts, we also considered a multivariate model, namely a Vector Autoregression (VAR) with shrinkage, implemented using the R package `BigVAR` package (Nicholson et al., 2019). Since, for both base forecasts and reconciled forecasts the neural network outperformed the VAR, we focus only on the former. However, we note that even for the VAR results (summarised in Appendix ?? of the online supplement) reconciliation methods improve upon base forecasts. This highlights that forecast reconciliation should not be thought of as an alternative for forecasting from multivariate models, but rather as an option for improving both univariate and multivariate approaches.

Four situations were considered where base forecasts are assumed to be either Gaussian or bootstrapped from residuals, and either dependent or independent. Figure 6 demonstrates departures from normality in the residuals of base forecasting models, while the correlation heatmap of these residuals in Figure 7 demonstrates departures from independence. Therefore, independent Gaussian probabilistic forecasts are likely to represent severe misspecification.

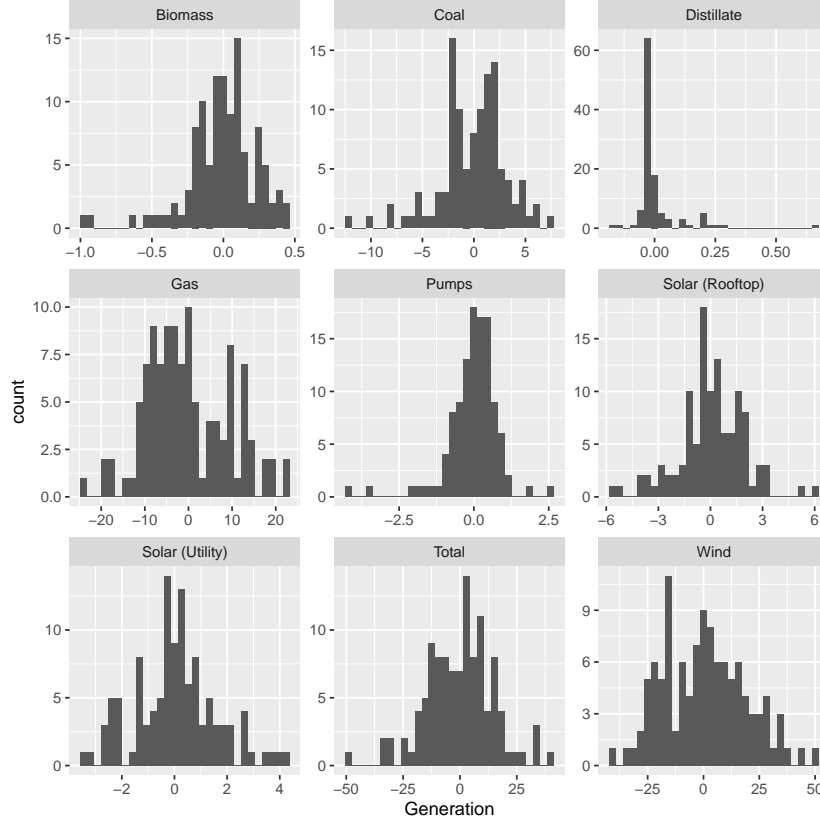


Figure 6: *Densities of residuals for selected series from a typical training window of 2 October 2019 to 21 January 2020.*

7.2 Reconciliation

The same reconciliation methods were used as in the simulation study with score optimisation based on an objective function with 56 days of score evaluations. For brevity, only the energy score results are presented; please refer to <https://git.io/JJMEH> for a full set of results.

The mean energy score for all four base forecasting methods is summarised in Figure 8. When base forecasts are generated assuming both independence and a Gaussian distribution, score optimisation achieves a mean energy score that is considerably smaller than all other

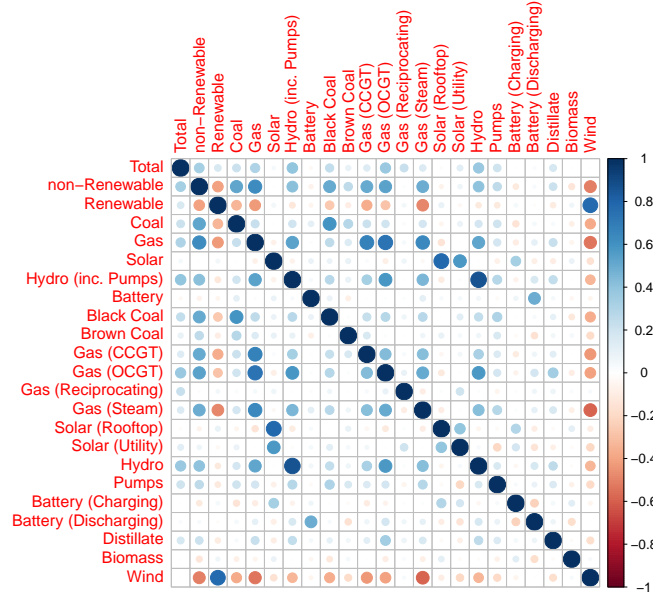


Figure 7: *Correlation heatmap of residuals from a typical training window of 2 October 2019 to 21 January 2020. Blue circles indicate positive correlation, while red circles indicate negative correlation with larger circles indicating stronger correlations.*

competing methods, with MinT providing the second smallest value. The superior forecasting performance of score optimisation is statistically significant, see the Nemenyi matrix in Figure 9 (left). This suggests that score optimisation is best for guarding against severe model misspecification.

For all other methods the best performing method is OLS. This difference is statistically significant.⁴ Although score optimisation does not improve upon base, the differences are not significant. For all base forecasts, both JPP and BTTH are significantly worse than base forecasts.

8 Conclusions

This paper introduces a rigorous formulation of forecast reconciliation in the probabilistic setting. It can be applied when the base forecast is either available as a density or when a sample

⁴See the Nemenyi matrix for jointly bootstrapped base forecasts in Figure 9 (right). The corresponding figures for joint Gaussian and independent bootstrap look mostly similar to the right panel of Figure 9; please refer to <https://git.io/JJMEH> for a full set of results

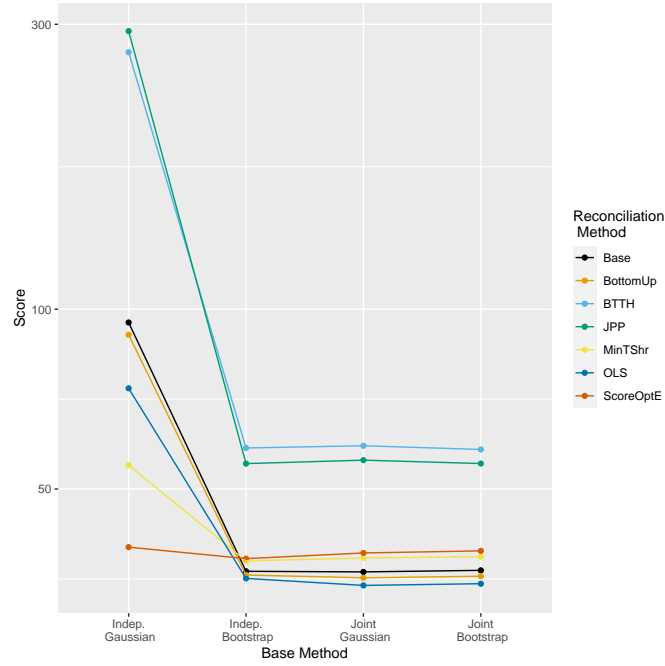


Figure 8: Mean Energy score for the electricity application for different base forecasting methods and different reconciliation methods.

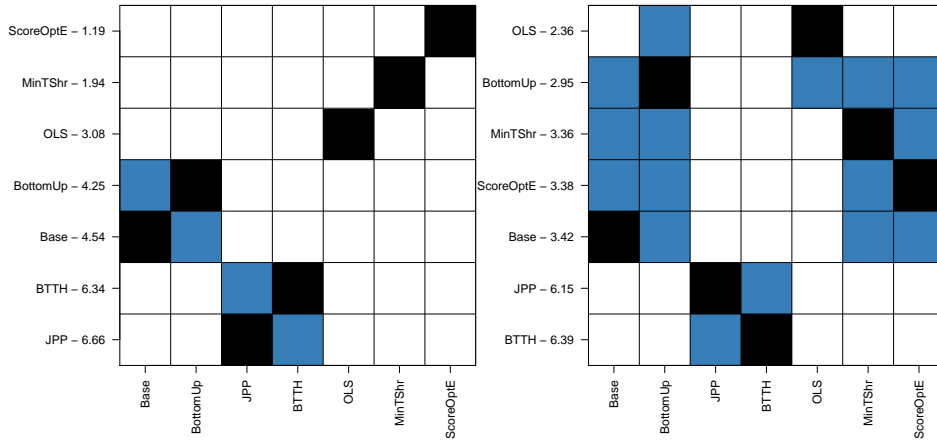


Figure 9: Nemenyi matrices for Energy score for the electricity application. Left: base forecasts are independent and Gaussian. Right: base forecasts are obtained by jointly bootstrapping residuals.

has been drawn from the base forecast. In the elliptical case, we prove that reconciliation can recover the correct probability distribution if the base forecast is of the correct form, irrespective of the scale and location of the base forecast. Probably due to this reason, score

optimisation works well in applications even when the base forecasts are assumed to be independent.

We also prove that the log score is not proper when comparing incoherent and coherent forecasts. Consequently, we introduce a new algorithm that trains reconciliation weights by minimising the energy score or variogram score. Since the scores are approximated by Monte Carlo simulation, stochastic gradient descent is used for optimisation. This method is shown to lead to significant improvements over base forecasts, bottom-up methods and existing probabilistic reconciliation approaches across a wide variety of simulated and empirical examples.

An interesting result is that projection methods with certain optimality properties in the point forecasting setting, also work well when extended to the probabilistic case. In particular, a simple least squares projection is the best performing method in the high-dimensional empirical example, provided the base forecasts are not too badly misspecified. This may arise since projections implicitly provide constrained versions of the reconciliation weights. A promising future research avenue may involve regularised versions of score optimisation that add an L_1 or L_2 penalty to the objective function. Alternatively, early stopping (Bühlmann and Yu, 2003) of the gradient descent may lead to a better bias-variance tradeoff in learning reconciliation weights.

A final important avenue of future research is the development of probabilistic forecast reconciliation for domains other than the real line. These may include domains constrained above zero, discrete domains, or domains that are a mixture of continuous distributions and discrete point masses. While such problems are challenging, the geometric interpretation of probabilistic forecast introduced in this paper, lays the foundation for this research agenda.

References

- Athanasopoulos, G., R. A. Ahmed, and R. J. Hyndman (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting* 25(1), 146 – 166.
- Athanasopoulos, G., R. J. Hyndman, N. Kourentzes, and F. Petropoulos (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research* 262(1), 60–74.
- Azzalini, A. (2020). *sn: The Skew-Normal and Related Distributions such as the Skew-t*. Università di Padova, Italia. R package version 1.6-1.
- Ben Taieb, S., R. Huser, R. J. Hyndman, and M. G. Genton (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. and B. Koo (2019, 07). Regularized regression for hierarchical forecasting without unbiasedness conditions. In *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1337–1347.
- Ben Taieb, S., J. W. Taylor, and R. J. Hyndman (2020). Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*. in press.
- Böse, J.-H., V. Flunkert, J. Gasthaus, T. Januschowski, D. Lange, D. Salinas, S. Schelter, M. Seeger, and Y. Wang (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 Y. Lechevallier and G. Saporta (Eds.), *Proceedings of COMPSTAT'2010*, pp. 177–186. Physica-Verlag HD.

- Bühlmann, P. and B. Yu (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association* 98(462), 324–339.
- Carpenter, B., M. D. Hoffman, M. Brubaker, D. Lee, P. Li, and M. Betancourt (2015). The Stan math library: Reverse-mode automatic differentiation in C++.
- Dunn, D. M., W. H. Williams, and T. L. Dechaine (1976). Aggregate versus subaggregate models in local area forecasting. *Journal of the American Statistical Association* 71(353), 68–71.
- Gasthaus, J., K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski (2019). Probabilistic forecasting with spline quantile function rnns. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1901–1910. PMLR.
- Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application* 1, 125–151.
- Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477), 359–378.
- Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review* 133(5), 1098–1118.
- Gross, C. W. and J. E. Sohl (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting* 9(3), 233–254.
- Hofert, M., A. Prasad, and M. Zhu (2020). Applications of multivariate quasi-random sampling with neural networks.
- Hollander, M., D. A. Wolfe, and E. Chicken (2013). *Nonparametric statistical methods*. John Wiley & Sons.

- Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis* 55(9), 2579–2589.
- Hyndman, R. J. and G. Athanasopoulos (2021). *Forecasting: Principles and Practice* (3rd ed.). Melbourne, Australia: OTexts.
- Hyndman, R. J. and Y. Khandakar (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* 26(3), 1–22.
- Janke, T. and F. Steinke (2020, Aug). Probabilistic multivariate electricity price forecasting using implicit generative ensemble post-processing. *2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*.
- Jeon, J., A. Panagiotelis, and F. Petropoulos (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research* 279(2), 364–379.
- Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization.
- Kingma, D. P. and M. Welling (2013). Auto-encoding variational Bayes.
- Kourentzes, N. (2019). *tsutils: Time Series Exploration, Modelling and Forecasting*. R package version 0.9.0.
- McLean Sloughter, J., T. Gneiting, and A. E. Raftery (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review* 141(6), 2107–2119.
- Nicholson, W., D. Matteson, and J. Bien (2019). *BigVAR: Dimension Reduction Methods for Multivariate Time Series*. R package version 1.0.6.

- Nystrup, P., E. Lindstrom, P. Pinson, and H. Madsen (2020). Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research* 280(3), 876 – 888.
- O’Hara-Wild, M., R. Hyndman, and E. Wang (2020). *fable: Forecasting Models for Tidy Time Series*. R package version 0.2.0.
- Panagiotelis, A. (2020). *ProbReco: Score Optimal Probabilistic Forecast Reconciliation*. R package version 0.1.0.
- Panagiotelis, A., G. Athanasopoulos, P. Gamakumara, and R. J. Hyndman (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting* 37(1), 343–359.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rossi, B. (2014). Density forecasts in economics, forecasting and policymaking. Technical report, Els Opuscles del CREI.
- Schäfer, J. and K. Strimmer (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. and T. M. Hamill (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review* 143(4), 1321–1334.
- Shang, H. L. and R. J. Hyndman (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. and M. L. Rizzo (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* 143(8), 1249–1272.

- Van Erven, T. and J. Cugliari (2015). Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. In *Modeling and Stochastic Learning for Forecasting in High Dimensions*, pp. 297–317. Springer.
- Wickramasuriya, S. L., G. Athanasopoulos, and R. J. Hyndman (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association* 114(526), 804–819.