Forecast reconciliation: A geometric view with new insights on bias correction

Journal:	*International Journal of Forecasting*
Manuscript ID	INTFOR_1909455.R1
Manuscript Type:	Original Article
Keyword:	Aggregation, Coherence, Hierarchical times series, Linear constraints, Projections
Abstract:	A geometric interpretation is developed for so-called *reconciliation* methodologies used to forecast time series that adhere to known linear constraints. In particular, a general framework is established nesting many existing popular reconciliation methods within the class of *projections*. This interpretation facilitates the derivation of novel theoretical results. First, reconciliation via projection is guaranteed to improve forecast accuracy with respect to a class of loss functions based on a generalised distance metric. Second, the MinT method minimises expected loss for this same class of loss functions. Third, the geometric interpretation provides a new proof that forecast reconciliation using projections results in unbiased forecasts provided the initial base forecasts are also unbiased. Approaches for dealing with biased base forecasts are proposed. An extensive empirical study on Australian tourism flows demonstrates the theoretical results of the paper and shows that bias correction prior to reconciliation outperforms alternatives that only bias-correct or only reconcile forecasts

Forecast reconciliation: A geometric view with new insights on bias correction

Abstract

A geometric interpretation is developed for so-called *reconciliation* methodologies used to forecast time series that adhere to known linear constraints. In particular, a general framework is established nesting many existing popular reconciliation methods within the class of *projections*. This interpretation facilitates the derivation of novel theoretical results. First, reconciliation via projection is guaranteed to improve forecast accuracy with respect to a class of loss functions based on a generalised distance metric. Second, the MinT method minimises expected loss for this same class of loss functions. Third, the geometric interpretation provides a new proof that forecast reconciliation using projections results in unbiased forecasts provided the initial base forecasts are also unbiased. Approaches for dealing with biased base forecasts are proposed. An extensive empirical study on Australian tourism flows demonstrates the theoretical results of the paper and shows that bias correction prior to reconciliation outperforms alternatives that only bias-correct or only reconcile forecasts.

1 Introduction

The past decade has seen rapid development in methodologies for forecasting time series that follow a hierarchical aggregation structure. Of particular prominence have been *forecast reconciliation* methods involving two steps: first separate forecasts are produced for all series, then these are adjusted to ensure coherence with aggregation constraints. Forecast reconciliation has mostly been formulated using a regression model that admits a generalised least squares (GLS) solution, see Hyndman et al. (2011) and Wickramasuriya, Athanasopoulos & Hyndman (2019) for examples. Alternatively, Van Erven & Cugliari (2015) and Nystrup et al. (2020) arrive at a GLS solution by formulating reconciliation as an optimisation problem. The regression setup can be counter-intuitive since a vector comprised of forecasts from different time series models is also assumed to be the dependent variable in a regression model. In this paper, we eschew a regression interpretation in favour of a novel, geometric understanding of forecast reconciliation. This allows us to develop novel proofs and a clearer understanding of the interplay between objective functions, loss functions, forecast bias and reconciliation methods.

Multivariate time series following an aggregation structure arise in many sectors such as retail, energy, insurance, health and welfare and economics (see for example Karmy & Maldonado 2019, Ben Taieb et al. 2017, Nystrup et al. 2020, Almeida et al. 2016, Jeon et al. 2019, Mahkya et al. 2017, Li & Tang 2019, Shang & Hyndman 2017, Athanasopoulos et al. 2019). Forecasts of these series should adhere to aggregation constraints to ensure aligned decision making. Earlier studies achieved this by only forecasting a single level of the hierarchy and then either aggregating in a bottom-up fashion (Dunn et al. 1976) or disaggregating in a top-down fashion (Gross & Sohl 1990, Athanasopoulos et al. 2009). For reviews of these approaches, including a discussion of their advantages and disadvantages, see Schwarzkopf et al. (1988), Kahn (1998), Lapide (1998), Fliedner (2001), Athanasopoulos et al. (2009).

In contrast to these methods, Hyndman et al. (2011) proposed forecasting all series in the hierarchy, referring to these as *base* forecasts. Since base forecasts were produced independently they were not guaranteed to adhere to aggregation constraints and could thus be improved via further adjustment. A framework was proposed whereby the aggregation constraints were expressed in a regression model for the base forecasts. The predicted values from this model were guaranteed to adhere to the linear constraints by construction and could thus be used as a new set of forecasts. This approach and later modifications have subsequently been shown to outperform bottom-up and top-down approaches in a variety of empirical settings (see for example Athanasopoulos et al. 2009, 2017, Wickramasuriya, Athanasopoulos & Hyndman 2019, among others). Some theoretical insight into the performance of forecast reconciliation methods has been provided by Van Erven & Cugliari (2015) and Wickramasuriya, Athanasopoulos & Hyndman (2019). Both papers provide a proof that reconciliation is guaranteed to improve base forecasts. The latter paper also proposes a particular version of reconciliation known as the Minimum Trace (MinT) method. This is optimal in a sense of minimising the trace of the reconciled forecast error covariance matrix under the assumption that the base forecasts are unbiased.

Our main contribution is to propose a geometric interpretation of the entire hierarchical forecasting problem. In this setting, we show that reconciled forecasts have a number of attractive properties when they are obtained via projections. We believe that this is clearer and more intuitive than explanations based on regression modelling. In addition to casting existing results in a new light, the geometric interpretation also allows us to derive four new important results.

First, our approach makes it clear that the defining characteristic of so-called *hierarchical time series* is not aggregation but linear constraints. As a result forecast reconciliation can be applied in contexts where there are no clear candidates of *bottom-level* series, an insight that is not apparent when the problem is viewed through the lens of regression modelling. Second, we provide a new proof that reconciled forecasts dominate base forecasts, for a certain class of loss functions. The projection that achieves this depends on the weights used in the loss function but not on the dependence in forecast errors. Furthermore, unlike proofs of similar results by Van Erven & Cugliari (2015) and Wickramasuriya, Athanasopoulos & Hyndman (2019), our proof does not require an assumption about convexity that may not hold in general. Third, we show that when it comes to a different objective, namely minimising the expected loss, the optimal projection depends only on the covariance of the forecast errors, and not the weights used in the loss function. In this case of equal weights, this property is a direct consequence of the trace minimising property already established by Wickramasuriya, Athanasopoulos & Hyndman (2019). We now prove that this result also applies to a more general class of loss functions. Fourth, we provide a new proof that reconciliation using certain projection matrices guarantees unbiased reconciled forecasts provided the base forecasts are also unbiased. A natural question that arises is what to do in the case of biased reconciled forecasts. Rather than addressing this issue by considering matrices that are not projections, we propose to bias-correct before reconciliation. This is evaluated in an extensive empirical study where we find that even when bias correction fails, the extent of the problem is mitigated by reconciling forecasts.

The remainder of this paper is structured as follows. Section 2 deals with the concept of coherence and defines hierarchical time series in a way that does not depend on any notion of bottom-level series. Section 3 defines forecast reconciliation in terms of projections and includes proofs that make the optimality properties of different reconciliation methods clear. In Section 4 we prove the unbiasedness preserving property of reconciliation via certain projection matrices and propose methods for bias correction. In Section 5 we conduct an extensive empirical application to domestic tourism flows in Australia with two objectives; first to demonstrate the theorems discussed in Section 3, and second to evaluate the methods for bias correction discussed in Section 4. Section 6 concludes with some discussion, practical recommendations and thoughts on the future research directions for forecast reconciliation.

2 Coherent forecasts

2.1 Notation and preliminaries

We briefly define the concept of a *hierarchical time series* in a fashion similar to Athanasopoulos et al. (2019), Hyndman & Athanasopoulos (2018) and others, before elaborating on some of the limitations of this understanding. A *hierarchical time series* is a collection of

n variables indexed by time, where some variables are aggregates of other variables. We let $\boldsymbol{y}_t \in \mathbb{R}^n$ be a vector comprising observations of all variables in the hierarchy at time t. The *bottom-level series* are defined as those m variables that cannot be formed as aggregates of other variables; we let $\boldsymbol{b}_t \in \mathbb{R}^m$ be a vector comprised of observations of all bottom-level series at time t. The hierarchical structure of the data implies that the following holds for all t:

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

where $\boldsymbol{S}$ is an $n \times m$ constant matrix that encodes the aggregation constraints.

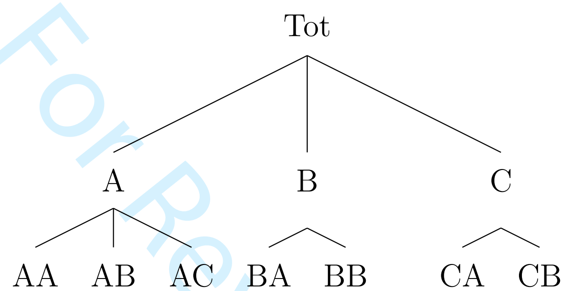

Figure 1: An example of a two level hierarchical structure.

To clarify these concepts, consider the example of the hierarchy in Figure 1. For this hierarchy, $n = 11$, $\boldsymbol{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, y_{C,t}, y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t}, y_{CA,t}, y_{CB,t}]'$, $m = 7$, $\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t}, y_{CA,t}, y_{CB,t}]'$ and

$$\boldsymbol{S} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ & & & \boldsymbol{I}_7 & & & \end{pmatrix},$$

where $\boldsymbol{I}_7$ is the 7×7 identity matrix.

While such a definition is completely serviceable, it obscures the full generality of the literature on so-called hierarchical time series. In fact, concepts such as coherence and reconciliation, defined in full below, require the data to have only two important characteristics: the first is that they are multivariate, the second is that they adhere to linear constraints.

2.2 Coherence

The property that data adhere to some linear constraints is referred to as *coherence*. We now provide definitions aimed at providing geometric intuition for hierarchical time series.

Definition 2.1 (Coherent subspace)**.** The m-dimensional linear subspace $\mathfrak{s} \subset \mathbb{R}^n$ for which some linear constraints hold for all $\boldsymbol{y} \in \mathfrak{s}$ is defined as the *coherent subspace*.

To further illustrate, Figure 2 depicts the simplest three variable hierarchy where $y_{Tot,t} = y_{A,t} + y_{B,t}$. The coherent subspace is depicted as a grey 2-dimensional plane within 3-dimensional space; i.e. $m = 2$ and $n = 3$. It is worth noting that the coherent subspace is spanned by the columns of $\boldsymbol{S}$; i.e. $\mathfrak{s} = \text{span}(\boldsymbol{S})$. In Figure 2, these columns are $\vec{s}_1 = (1, 1, 0)'$ and $\vec{s}_2 = (1, 0, 1)'$. It is equally important to recognise that the hierarchy could also have been defined in terms of $y_{Tot,t}$ and $y_{A,t}$ rather than the bottom-level series, $y_{A,t}$ and $y_{B,t}$. In this case the corresponding '$\boldsymbol{S}$ matrix' would have columns $(1, 0, 1)'$ and $(0, 1, -1)'$. However, while there are multiple ways to define an $\boldsymbol{S}$ matrix, in all cases the columns will span the same coherent subspace, which is unique.

Definition 2.2 (Hierarchical Time Series)**.** A hierarchical time series is an n-dimensional multivariate time series such that all observed values $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ and all future values $\boldsymbol{y}_{T+1}, \boldsymbol{y}_{T+2}, \ldots$ lie in the coherent subspace, i.e., $\boldsymbol{y}_t \in \mathfrak{s} \quad \forall t$.

Despite the common use of the term *hierarchical time series*, it should be clear from the definition that the data need not necessarily follow a hierarchy. Also notable by its absence in the above definition is any reference to *aggregation*. In some ways, terms such as *hierarchical* and *aggregation* can be misleading since the literature has covered instances that cannot be depicted in a similar fashion to Figure 1 and/or do not involve aggregation. Examples include, temporal hierarchies which involve grouped structures (see Athanasopoulos et al. 2017), overlapping temporal hierarchies (see Jeon et al. 2019), applications for which

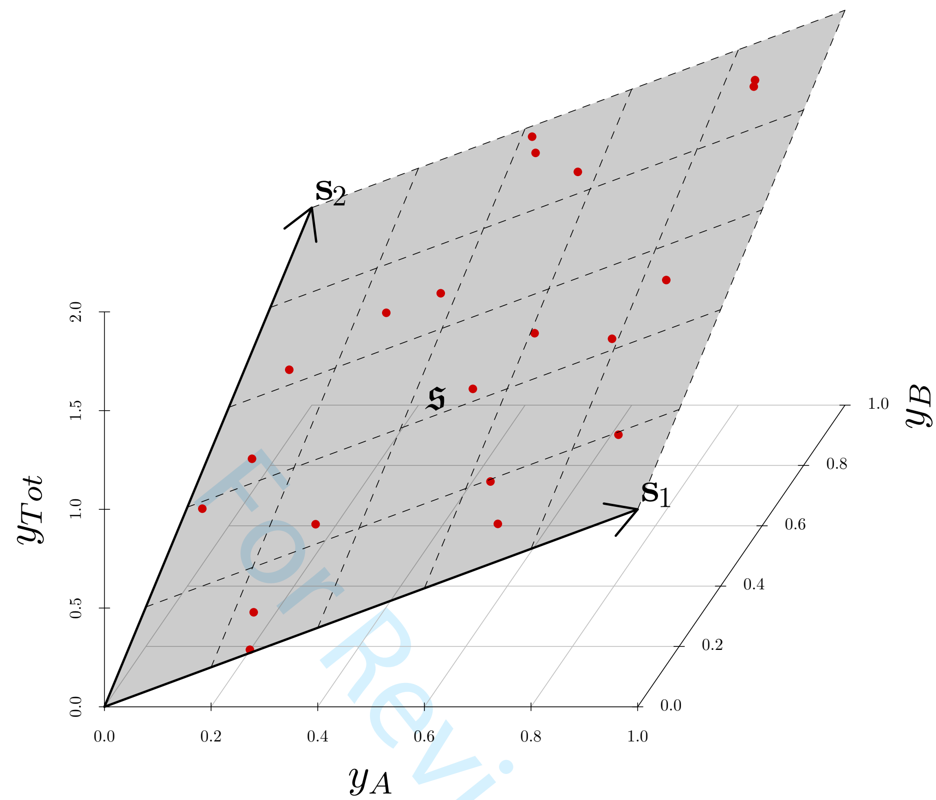

Figure 2: Depiction of a three dimensional hierarchy with $y_{\text{Tot}} = y_{\text{A}} + y_{\text{B}}$. The gray coloured two dimensional plane depicts the coherent subspace $\mathfrak{s}$ where $\vec{s}_1 = (1, 1, 0)'$ and $\vec{s}_2 = (1, 0, 1)'$ are basis vectors that span $\mathfrak{s}$. The red points in $\mathfrak{s}$ represent realisations or coherent forecasts.

the difference rather than the aggregate is of interest (see Li & Tang 2019), or structures that involve both cross-sectional and temporal dimensions referred to as cross-temporal structures (see Kourentzes & Athanasopoulos 2019*a*). Finally, although Definition 2.2 makes reference to time series, this definition can be easily generalised to any vector-valued data for which some linear constraints are known to hold for all realisations.

Definition 2.3 (Coherent Point Forecasts)**.** Let $\breve{\boldsymbol{y}}_{t+h|t} \in \mathbb{R}^n$ be a vector of point forecasts of all series in the hierarchy where the subscript $t + h|h$ implies that the forecast is made as time t for a period h steps into the future. Then $\breve{\boldsymbol{y}}_{t+h|t}$ is *coherent* if $\breve{\boldsymbol{y}}_{t+h|t} \in \mathfrak{s}$.

Without any loss of generality, the above definition could also be applied to prediction for multivariate data in general, rather than just forecasting of time series.

Much of the early literature that dealt with the problem of forecasting hierarchical time series (see Gross & Sohl 1990, and references therein) produced forecasts at a single level of the hierarchy in the first stage. Subsequently forecasts for all series were recovered through either aggregation, disaggregation according to historical or forecast proportions, or some combination of both. Consequently, incoherent forecasts were not a challenge in these earlier papers.

Forecasting a single level of the hierarchy did not echo common practice within many industries. In many organisations different departments or 'silos' each produced their own forecasts, often with their own information sets and judgemental adjustments.[1] This approach does have several advantages over only forecasting a single level. First, there is no loss of information since all levels and series are modelled. Second, modelling higher level series often identifies features such as trend and seasonality that cannot be detected in noisy disaggregate data. However, when forecasts are produced independently at all levels, forecasts are likely to be incoherent.[2] While the problem of incoherent forecasts can be addressed by some multivariate approaches, including state space models, these can not always be generalised to models with more complicated features or scaled up to high-dimensional problems. Instead, the solution is to make an adjustment that ensures coherence, a process known as *forecast reconciliation*

3 Forecast reconciliation

The concept of forecast reconciliation is predicated on there being an n-vector of forecasts that are incoherent. We call these *base forecasts* and denote them as $\hat{\boldsymbol{y}}_{t+h|t}$. We will sometimes drop this subscript for ease of exposition. In the most general terms, reconciliation can be defined as follows.

[1] Chase (2013) discusses silos and the importance of information and data sharing across an organisation.

[2] There are some special cases of using simple approaches such as naïve, which extrapolate the coherent nature of the data to the forecasts.

Definition 3.1 (Reconciled forecasts)**.** Let ψ be a mapping, $\psi : \mathbb{R}^n \to \mathfrak{s}$. The point forecast $\tilde{\boldsymbol{y}}_{t+h|t} = \psi(\hat{\boldsymbol{y}}_{t+h|t})$ "reconciles" a base forecast $\hat{\boldsymbol{y}}_{t+h|t}$ with respect to the mapping $\psi(.)$

All reconciliation methods that we are aware of consider a linear mapping for ψ, which involves pre-multiplying base forecasts by an $n \times n$ matrix that has $\mathfrak{s}$ as its image. One way to achieve this is with a matrix $\boldsymbol{SG}$, where $\boldsymbol{G}$ is an $m \times n$ matrix (in some papers, $\boldsymbol{P}$ is used in place of $\boldsymbol{G}$). This facilitates an interpretation of reconciliation as a two-step process. In the first step, base forecasts $\hat{\boldsymbol{y}}_{t+h|t}$ are combined to form a new set of bottom-level forecasts. In the second step, these are mapped to a full vector of coherent forecasts via pre-multiplication by $\boldsymbol{S}$.

Although pre-multiplying base forecasts by $\boldsymbol{SG}$ will result in coherent forecasts, a number of desirable properties arise when $\boldsymbol{SG}$ has the specific structure of a *projection* matrix onto $\mathfrak{s}$. In general a projection matrix is defined via its idempotence property, i.e. $(\boldsymbol{SG})^2 = \boldsymbol{SG}$. We also rely on another property of projection matrices, namely that any vector lying in the image of a projection is mapped to itself by that projection (see Lemma 2.4 in Rao 1974, for a proof). In our context this implies that for any $\boldsymbol{v} \in \mathfrak{s}$, $\boldsymbol{SGv} = \boldsymbol{v}$.

We begin by considering the special case of an orthogonal projection whereby $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$. This is equivalent to so-called OLS reconciliation as introduced by Hyndman et al. (2011). We refrain from any discussion of regression models focusing instead on geometric interpretations. Nonetheless, the connection between OLS and orthogonal projection should be clear; in the context of regression modelling predicted values from OLS are obtained via an orthogonal projection onto the span of the regressors.

3.1 Orthogonal projection

In this section we discuss two sensible properties that can be achieved by reconciliation via orthogonal projection.

1. The first is that reconciliation should adjust the base forecasts as little as possible; i.e., the base and reconciled forecasts should be 'close'.

2. The second is that reconciliation in some sense should improve forecast accuracy, or more loosely, that the reconciled forecast should be 'closer' to the realised value targeted by the forecast.

To address the first of these properties we make the concept of closeness more concrete by considering the Euclidean distance between the base forecast $\hat{\boldsymbol{y}}_{t+h|t}$ and the reconciled forecast $\tilde{\boldsymbol{y}}_{t+h|t}$. A property of an orthogonal projection is that the distance between $\hat{\boldsymbol{y}}_{t+h|t}$ and $\tilde{\boldsymbol{y}}_{t+h|t}$ is minimal over any possible $\tilde{\boldsymbol{y}}_{t+h|t} \in \mathfrak{s}$. In this sense reconciliation via orthogonal projection $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$ leads to the smallest possible adjustments of the base forecasts. Alternatively, Euclidean distance can be interpreted as a loss function $L(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|$ where $\|.\|$ denotes the L2 norm, in which case an orthogonal projection solves the optimisation problem $\min\limits_{\boldsymbol{G}} L(\hat{\boldsymbol{y}}_{t+h|t}, \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t})$. This is a special case of the optimisation problem considered by Nystrup et al. (2020), the more general case will be covered in the next section.

The aim of the second property is to guarantee that reconciled forecasts should always be closer to the target than base forecasts and is related to the difference in loss functions $L(\boldsymbol{y}_{t+h}, \hat{\boldsymbol{y}}_{t+h|t}) - L(\boldsymbol{y}_{t+h}, \tilde{\boldsymbol{y}}_{t+h|t})$. This is expressed as a minimax optimisation by Van Erven & Cugliari (2015) and was also touched upon in Section 2.3 of Wickramasuriya, Athanasopoulos & Hyndman (2019), albeit in both cases for a slightly different loss function. Here we provide a new explicit proof of this distance reducing property. We do so first in the case of the loss function defined here in terms of Euclidean distance where the geometric intuition of the proof is clear and then generalise the result to more general loss functions Section 3.2

Consider the Euclidean distance between the target and a forecast. This is equivalent to the square root of the sum of squared forecast errors over the entire hierarchy. Let $\boldsymbol{y}_{t+h}$ be the realisation of the data generating process at time $t+h$. The following theorem shows that reconciliation never increases the sum of squared errors of point forecasts.

Theorem 3.1 (Distance reducing property)**.** *If $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t}$, where $\boldsymbol{G}$ is such that $\boldsymbol{S}\boldsymbol{G}$ is an orthogonal projection (in the Euclidean sense) onto $\mathfrak{s}$ then:*

$$\|(\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t})\| \leq \|(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t})\|.$$

Proof. Since $\boldsymbol{y}_{t+h|t}, \tilde{\boldsymbol{y}}_{t+h|t} \in \mathfrak{s}$ and since the projection is orthogonal, by Pythagoras' theorem

$$\|(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t})\|^2 = \|(\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t})\|^2 + \|(\tilde{\boldsymbol{y}}_{t+h|t} - \hat{\boldsymbol{y}}_{t+h|t})\|^2.$$

Since $\|(\tilde{\boldsymbol{y}}_{t+h|t} - \hat{\boldsymbol{y}}_{t+h|t})\|^2 \geq 0$ this implies,

$$\|(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t})\|^2 \geq \|(\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t})\|^2,$$

with equality only holding when $\tilde{\boldsymbol{y}}_{t+h|t} = \hat{\boldsymbol{y}}_{t+h|t}$. Taking the square root of both sides proves the desired result. □

The simple geometric intuition behind the proof is demonstrated in Figure 3. In this schematic, the coherent subspace is depicted as a black arrow, and the base forecast $\hat{\boldsymbol{y}}$ is shown as a blue dot. Since $\hat{\boldsymbol{y}}$ is incoherent, $\hat{\boldsymbol{y}} \notin \mathfrak{s}$ and in this case the inequality is strict. Reconciliation is an orthogonal projection from $\hat{\boldsymbol{y}}$ to the coherent subspace yielding the reconciled forecast $\tilde{\boldsymbol{y}}$ shown in red. Finally, the target of the forecast $\boldsymbol{y}$ is displayed as a black point, and although its exact location is unknown to the forecaster, it is known that it will lie somewhere along the coherent subspace.

Figure 3 clearly shows that $\hat{\boldsymbol{y}}$, $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}$ form a right angled triangle with $\tilde{\boldsymbol{y}}$ at the right-angled vertex. In this triangle, the line between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ is the hypotenuse and therefore must be longer than the distance between $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$. Therefore reconciliation is guaranteed to reduce a loss function based on distance, or indeed any monotonic function of distance.

Theorem 3.1 is in some ways more powerful than perhaps previously understood. Crucially, the result is not a result that requires taking expectations. This distance reducing property will hold for any realisation and any forecast and not just on average. Nothing needs to be assumed about the statistical properties of the data generating process or the process by which forecasts are made.

In other ways, Theorem 3.1 is weaker than perhaps often understood. First, when improvements in forecast accuracy are discussed in the context of the theorem, this refers to a very specific measure of forecast accuracy. In particular, this measure is the sum of squared forecast errors of *all* variables in the hierarchy (or any monotonic function thereof). Second, although an orthogonal projection is guaranteed to improve on base forecasts, it is

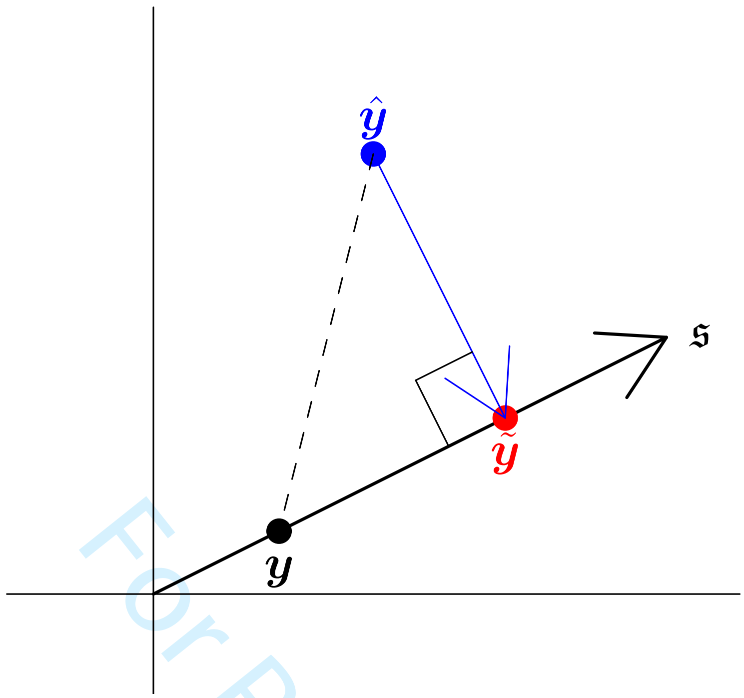

Figure 3: Orthogonal projection of $\hat{\boldsymbol{y}}$ onto $\mathfrak{s}$ yielding the reconciled forecast $\tilde{\boldsymbol{y}}$.

not necessarily the projection that leads to the greatest improvement in forecast accuracy in expectation. Therefore, referring to OLS reconciliation as 'optimal' is somewhat misleading since it does not have the optimality properties of some oblique projections. We now turn our attention to the way oblique projections can address both of these shortcomings.

3.2 Oblique Projections

One justification for using an orthogonal projection is that it leads to improved forecast accuracy in terms of a loss function based on Euclidean distance that involves *all* variables in the hierarchy. A clear shortcoming of this measure of forecast accuracy is that forecast errors in all series should not necessarily be treated equally. For example, in hierarchies, top-level series tend to have a much larger scale than bottom-level series. Alternatively, the context of the forecast problem itself may suggest a loss function that weights series differently. For example in our tourism application in Section 5 we will consider weighting forecast errors by average tourism expenditure in each region.[3] An even more sophisticated understanding may take linear combinations of series into account. All of these considerations lead towards a loss function $L_{\boldsymbol{W}}(\boldsymbol{y}_{t+h}, \tilde{\boldsymbol{y}}_{t+h|t}) = ||\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t}||_{\boldsymbol{W}}$ where $||\boldsymbol{v}||_{\boldsymbol{W}} = \boldsymbol{v}'\boldsymbol{W}\boldsymbol{v}$, and $\boldsymbol{W}$ is a symmetric matrix assumed to be invertible. The geometry defined by the norm $||.||_{\boldsymbol{W}}$, will be referred to as the generalised Euclidean geometry with respect to $\boldsymbol{W}$.

One way to understand the generalised Euclidean geometry is that it is the same as Euclidean geometry when all vectors are first transformed by pre-multiplying by $\boldsymbol{W}^{1/2}$, where $\boldsymbol{W} = (\boldsymbol{W}^{1/2})'\boldsymbol{W}^{1/2}$. This leads to a transformed $\boldsymbol{S}$ matrix $\boldsymbol{S}^* = \boldsymbol{W}^{1/2}\boldsymbol{S}$ and transformed $\hat{\boldsymbol{y}}$ and $\tilde{\boldsymbol{y}}$ vectors $\hat{\boldsymbol{y}}^* = \boldsymbol{W}^{1/2}\hat{\boldsymbol{y}}$ and $\tilde{\boldsymbol{y}}^* = \boldsymbol{W}^{1/2}\tilde{\boldsymbol{y}}$. A projection of the form $\tilde{\boldsymbol{y}} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}\hat{\boldsymbol{y}}$, is an orthogonal projection in the transformed space since

$$\begin{aligned}
\tilde{\boldsymbol{y}}^* &= \boldsymbol{W}^{1/2}\tilde{\boldsymbol{y}} \\
&= \boldsymbol{W}^{1/2}\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}\hat{\boldsymbol{y}} \\
&= \boldsymbol{W}^{1/2}\boldsymbol{S}((\boldsymbol{W}^{1/2}\boldsymbol{S})'\boldsymbol{W}^{1/2}\boldsymbol{S})^{-1}(\boldsymbol{W}^{1/2}\boldsymbol{S})'\boldsymbol{W}^{1/2}\hat{\boldsymbol{y}} \\
&= \boldsymbol{S}^*(\boldsymbol{S}^{*'}\boldsymbol{S}^*)^{-1}\boldsymbol{S}^{*'}\hat{\boldsymbol{y}}^*.
\end{aligned}$$

Thinking of the generalised Euclidean space as a transformed version of Euclidean space also allows the distance reducing property of Theorem 3.1 to be generalised to any loss function $L_{\boldsymbol{W}}$

Theorem 3.2 (General distance reducing property). *If $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t}$, where $\boldsymbol{G}$ is such that $\boldsymbol{S}\boldsymbol{G}$ is an orthogonal projection (in the generalised Euclidean sense) onto $\mathfrak{s}$ then:*

$$\|(\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t})\|_{\boldsymbol{W}} \leq \|(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t})\|_{\boldsymbol{W}}.$$

Proof. The proof is identical to the proof for Theorem 3.1 but relies on the Generalised Pythagorean Theorem (applicable to Generalised Euclidean space) rather than the Pythagorean Theorem. □

[3]Similar consideration are made in the Kourentzes & Athanasopoulos (2019*b*) discussion paper, and also are taken into account for the loss functions used in the M5 forecasting competition (see https://mofc.unic.ac.cy/m5-competition).

The implication of Theorem 3.2 is that if the loss function is a monotonic function of $L_{\boldsymbol{W}}$ with some $\boldsymbol{W}$ given a priori then the projection matrix $\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}$ is guaranteed to improve forecast accuracy over base forecasts. This result does not necessarily involve the covariance of forecast errors unless $\boldsymbol{W}$ is explicitly chosen to depend on these covariances.

Wickramasuriya, Athanasopoulos & Hyndman (2019) and Van Erven & Cugliari (2015) both prove special cases of this result, the former in the case where $\boldsymbol{W}$ is the inverse of the forecast error covariance and the latter in the case where $\boldsymbol{W}$ is diagonal. Note that we rely here on the Generalised Pythagorean Theorem (which involves an equality). In contrast, Wickramasuriya, Athanasopoulos & Hyndman (2019) follow Van Erven & Cugliari (2015) in stating their result in terms of the Generalised Pythagorean Inequality. These proofs require an assumption of convexity so that the angle between the base forecast and coherent subspace must be greater than 90 degrees. The proof we have provided here requires no such assumption, since this may not hold for an arbitrary $\boldsymbol{W}$. As such the statement from Wickramasuriya, Athanasopoulos & Hyndman (2019) that *"MinT reconciled forecasts are at least as good as the incoherent forecasts"* should be qualified — this is true only for loss that is a monotonic function of $L_{\boldsymbol{\Sigma}^{-1}}$, where $\boldsymbol{\Sigma} = E(\boldsymbol{y} - \hat{\boldsymbol{y}})(\boldsymbol{y} - \hat{\boldsymbol{y}})'$. If Euclidean distance (or mean squared error) is used as a loss function, there will be realisations where the MinT estimator does not improve forecast accuracy relative to base forecasts. This will be demonstrated using a real data set in the empirical study in Section 5.2.

3.3 MinT

The discussion so far provides a roadmap, such that for a given choice of $\boldsymbol{W}$, a projection with distance-reducing properties can be derived. While the MinT method of Wickramasuriya, Athanasopoulos & Hyndman (2019) is a special case of such a projection, it was in fact motivated by a different optimality property. This was to minimise the trace of the forecast error covariance matrix of reconciled forecasts, i.e,

$$\min_{\boldsymbol{G}} \operatorname{tr}(E\left[(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}})(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}})'\right]). \tag{1}$$

The solution is the oblique projection $\boldsymbol{S}\left(\boldsymbol{S}'\Sigma^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\Sigma^{-1}$. While MinT is used here to refer to the case where Σ is known, in practice it is unknown. It can be estimated using in-sample errors, with some specific estimators found in Wickramasuriya, Athanasopoulos & Hyndman (2019) and also Nystrup et al. (2020).

Figure 4 provides geometrical intuition into the MinT method. Suppose that the orange points in panel (a) represent in-sample forecast errors. These provide information on the most likely direction of large deviations from the coherent subspace $\mathfrak{s}$. This direction is denoted by $\boldsymbol{R}$. Panel (b) shows a target value of $\boldsymbol{y}$, while the grey points indicate possible values for the base forecasts (the base forecasts are of course stochastic). One possible value of the forecast is depicted in blue as $\hat{\boldsymbol{y}}$. An oblique projection of the blue point back along the direction of $\boldsymbol{R}$, yields a reconciled forecast closer to the target, especially compared to an orthogonal projection. Panel (c) shows the orthogonal projection of every potential base forecast onto the coherent subspace. Panel (d) depicts an oblique projection along $\boldsymbol{R}$ for all the gray points. The oblique projection yields reconciled forecasts tightly packed near the target $\boldsymbol{y}$. In this sense, the oblique MinT projection minimises the forecast error variance of the reconciled forecasts. In contrast to the result in Theorem 3.2, this property is a statistical property in the sense that MinT is optimal in expectation.

3.4 Expected loss minimisation and MinT

We now make explicit the connection between MinT and a loss function based on the squared Euclidean distance $L^2(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|^2$ before generalising to $L^2_{\boldsymbol{W}}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|^2_{\boldsymbol{W}}$. [4] By the properties of the trace operator, the objective function in Eq. (1) can be rearranged as

$$\begin{aligned}
\mathrm{tr}(E\left[(\boldsymbol{y}-\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}})(\boldsymbol{y}-\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}})'\right]) &= \mathrm{tr}(E\left[(\boldsymbol{y}-\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}})'(\boldsymbol{y}-\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}})\right]) \\
&= E\left[\|\boldsymbol{y}-\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}\|^2\right] \\
&= E\left[L^2(\boldsymbol{y},\tilde{\boldsymbol{y}})\right]
\end{aligned}$$

[4] Since taking the square is monotonic over the range of L and $L_{\boldsymbol{W}}$, the properties in Theorem 3.1 and Theorem 3.2 also apply to L^2 and $L^2_{\boldsymbol{W}}$.

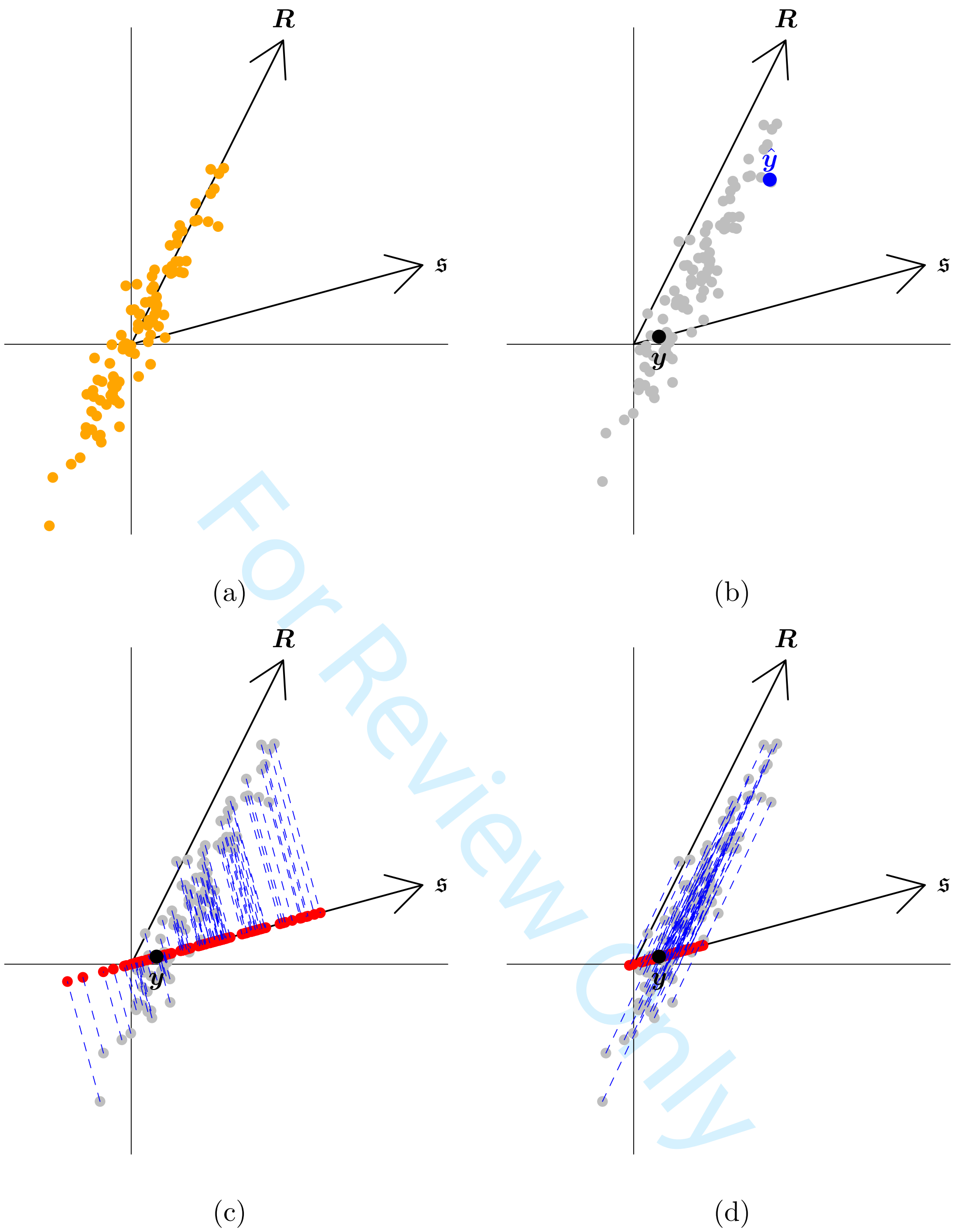

Figure 4: A schematic representation of orthogonal and oblique reconciliations. The orange points in (a) represent in-sample errors and $\boldsymbol{R}$ shows the most likely direction of deviations from the coherent subspace $\mathfrak{s}$. Grey points in (b) indicate potential base forecasts while the blue dot $\hat{\boldsymbol{y}}$ represents one such realisation. The black dot $\boldsymbol{y}$ denotes the (unknown) target of the forecast. (c) shows the orthogonal projection of all potential base forecasts onto the coherent subspace while (d) shows an oblique projection.

Note that trace minimisation implies a different optimality property to the distance reducing property described in Section 3.1. Theorem 3.1 implies optimality in the sense that reconciled forecasts always improve on base forecasts. For MinT, optimality refers to minimising the loss function *in expectation.*

Suppose the optimisation problem is generalised to a loss function based on some $\boldsymbol{W}$:

$$\min_{\boldsymbol{G}} \left(E\left[(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}})' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}) \right] \right). \tag{2}$$

The following theorem proves that the solution to this optimisation problem does not depend on the choice of $\boldsymbol{W}$ used in the loss function.

Theorem 3.3 (Expected loss minimisation and MinT)**.** *The usual MinT reconciliation method* $\tilde{\boldsymbol{y}} = \boldsymbol{S}\left(\boldsymbol{S}'\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{y}}$ *solves the optimisation problem in Eq.* (2) *for any choice of* $\boldsymbol{W}$.

Proof. The loss function in Eq. (2) is equivalent to Euclidean distance in the transformed space and can therefore be minimised by using the MinT method in this space. The MinT method in the transformed space is given by

$$\tilde{\boldsymbol{y}}^* = \boldsymbol{S}^* \left(\boldsymbol{S}^{*\prime} \boldsymbol{\Sigma}^{*-1} \boldsymbol{S}^* \right)^{-1} \boldsymbol{S}^{*\prime} \boldsymbol{\Sigma}^{*-1} \hat{\boldsymbol{y}}^*, \tag{3}$$

where $\boldsymbol{y}^* = \boldsymbol{W}^{1/2}\boldsymbol{y}$, $\boldsymbol{S}^* = \boldsymbol{W}^{1/2}\boldsymbol{S}$, $\hat{\boldsymbol{y}}^* = \boldsymbol{W}^{1/2}\hat{\boldsymbol{y}}$ and

$$\begin{aligned}
\boldsymbol{\Sigma}^* &= E\left[(\boldsymbol{y}^* - \hat{\boldsymbol{y}}^*)(\boldsymbol{y}^* - \hat{\boldsymbol{y}}^*)' \right] \\
&= E\left[\boldsymbol{W}^{1/2} (\boldsymbol{y} - \hat{\boldsymbol{y}})(\boldsymbol{y} - \hat{\boldsymbol{y}})' (\boldsymbol{W}^{1/2})' \right] \\
&= \boldsymbol{W}^{1/2} E\left[(\boldsymbol{y} - \hat{\boldsymbol{y}})(\boldsymbol{y} - \hat{\boldsymbol{y}})' \right] (\boldsymbol{W}^{1/2})' \\
&= \boldsymbol{W}^{1/2} \boldsymbol{\Sigma} (\boldsymbol{W}^{1/2})'
\end{aligned}$$

Noting that

$$\begin{aligned}
(\boldsymbol{\Sigma}^*)^{-1} &= \left(\boldsymbol{W}^{1/2} \boldsymbol{\Sigma} (\boldsymbol{W}^{1/2})' \right)^{-1} \\
&= (\boldsymbol{W}^{1/2\prime})^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{W}^{-1/2}
\end{aligned}$$

and substituting the expressions for $\boldsymbol{\Sigma}^*$, $\boldsymbol{S}^*$, $\boldsymbol{y}^*$ and $\hat{\boldsymbol{y}}^*$ into Eq. (3) yields,

$$\boldsymbol{W}^{1/2}\tilde{\boldsymbol{y}} = \boldsymbol{W}^{1/2}\boldsymbol{S}\left((\boldsymbol{W}^{1/2}\boldsymbol{S})'(\boldsymbol{W}^{1/2'})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{-1/2}\boldsymbol{W}^{1/2}\boldsymbol{S}\right)^{-1} (\boldsymbol{W}^{1/2}\boldsymbol{S})'(\boldsymbol{W}^{1/2'})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{-1/2}\boldsymbol{W}^{1/2}\hat{\boldsymbol{y}}$$

Rearranging and cancelling gives

$$\begin{aligned}\tilde{\boldsymbol{y}} &= \boldsymbol{S}\left(\boldsymbol{S}'(\boldsymbol{W}^{1/2})'(\boldsymbol{W}^{1/2'})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{-1/2}\boldsymbol{W}^{1/2}\boldsymbol{S}\right)^{-1} \boldsymbol{S}'(\boldsymbol{W}^{1/2})'(\boldsymbol{W}^{1/2'})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{-1/2}\boldsymbol{W}^{1/2}\hat{\boldsymbol{y}} \\ &= \boldsymbol{S}\left(\boldsymbol{S}'\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{y}}\end{aligned}$$

which corresponds to the usual MinT method. □

The implication of this result is that irrespective of the $\boldsymbol{W}$ used in the loss function, an oblique projection based on the forecast error covariance will always minimise *expected* loss (where loss is based on squared generalised Euclidean distance). From the point of view of minimising expected loss, for loss defined in Eq. (2), considerations about sensible weights for an error metric are not relevant. This is an important caveat to the statement by Van Erven & Cugliari (2015) that *"one should not assume that the choice $\boldsymbol{\Sigma}^{-1} = \boldsymbol{W}$ will adequately take care of sharing information between hierarchical levels!"*. While this statement is correct in the context of Theorem 3.2, which is the case considered by Van Erven & Cugliari (2015), it is not true for the objective described in Eq. (2). This will be empirically demonstrated in Section 5.

4 Bias in forecast reconciliation

Before turning our attention to the issue of bias itself it is important to state a desirable property that any reconciliation method should have. That is if base forecasts are already coherent then reconciliation should not change the forecasts. As stated in Section 3, this property holds only when $\boldsymbol{SG}$ is a projection matrix. As a corollary, reconciling using an arbitrary $\boldsymbol{G}$, may in fact change an already coherent forecast.

The property that projections map all vectors in the coherent subspace onto themselves is also useful in proving the unbiasedness preserving property of Wickramasuriya, Athanasopoulos & Hyndman (2019). Before restating this proof using a clear geometric interpretation, we discuss in a precise fashion what is meant by unbiasedness.

Suppose that the target of a point forecast is $\boldsymbol{\mu}_{t+h|t} := \mathrm{E}(\boldsymbol{y}_{t+h} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$ where the expectation is taken over the predictive density. Our point forecast can be thought of as an estimate of this quantity. The forecast is random due to uncertainty in the training sample and it is with respect to this uncertainty that unbiasedness is defined. Specifically, the point forecast will be unbiased if $\mathrm{E}_{1:t}(\hat{\boldsymbol{y}}_{t+h|t}) = \boldsymbol{\mu}_{t+h|t}$, where the subscript $1:t$ denotes an expectation taken over the training sample.

Theorem 4.1 (Unbiasedness preserving property)**.** *For unbiased $\hat{\boldsymbol{y}}_{t+h|t}$, the reconciled point forecast is also an unbiased prediction as long as $\boldsymbol{SG}$ is a projection onto $\mathfrak{s}$.*

Proof. The expected value of the reconciled forecast is given by

$$\mathrm{E}_{1:t}(\tilde{\boldsymbol{y}}_{t+h|t}) = \mathrm{E}_{1:t}(\boldsymbol{SG}\hat{\boldsymbol{y}}_{t+h|t}) = \boldsymbol{SG}\mathrm{E}_{1:t}(\hat{\boldsymbol{y}}_{t+h|t}) = \boldsymbol{SG}\boldsymbol{\mu}_{t+h|t}.$$

Since $\boldsymbol{\mu}_{t+h|t}$ is an expectation taken with respect to the degenerate predictive density it must lie in $\mathfrak{s}$. We have already established that when $\boldsymbol{SG}$ is a projection onto $\mathfrak{s}$ then it maps all vectors in $\mathfrak{s}$ onto themselves. As such $\boldsymbol{SG}\boldsymbol{\mu}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$ when $\boldsymbol{SG}$ is a projection matrix. □

The above result holds when the projection $\boldsymbol{SG}$ has the coherent subspace $\mathfrak{s}$ as its image and not for all projection matrices in general. To describe this more explicitly suppose $\boldsymbol{SG}$ has as its image $\mathfrak{L}$ which is itself a lower dimensional linear subspace of $\mathfrak{s}$, i.e., $\mathfrak{L} \subset \mathfrak{s}$. Then for $\left\{\boldsymbol{\mu}_{t+h|t} : \boldsymbol{\mu}_{t+h|t} \in \mathfrak{s}, \boldsymbol{\mu}_{t+h|t} \notin \mathfrak{L}\right\}$, $\boldsymbol{SG}\boldsymbol{\mu}_{t+h|t} \neq \boldsymbol{\mu}_{t+h|t}$. This is depicted in Figure 5 where $\boldsymbol{\mu}$ is projected to a point $\boldsymbol{\mu}^*$ in $\mathfrak{L}$. In this case, the expectation of reconciled forecast will be $\boldsymbol{\mu}^*$ rather than $\boldsymbol{\mu}$ and hence biased.

This result has implications in practice. The top-down method (Gross & Sohl 1990) has

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{p} & \boldsymbol{0}_{(m \times n-1)} \end{pmatrix},$$

where $\boldsymbol{p} = (p_1, \ldots, p_m)'$ is an m-dimensional vector consisting a set of proportions used to disaggregate the top-level forecast. In this case it can be verified that $\boldsymbol{SG}$ is idempotent, i.e., $\boldsymbol{SGSG} = \boldsymbol{SG}$ and therefore $\boldsymbol{SG}$ is a projection matrix. However, the image of this projection is not an m-dimensional subspace but a 1-dimensional subspace. As such, top-down reconciliation produces biased forecasts even when the base forecasts are unbiased.

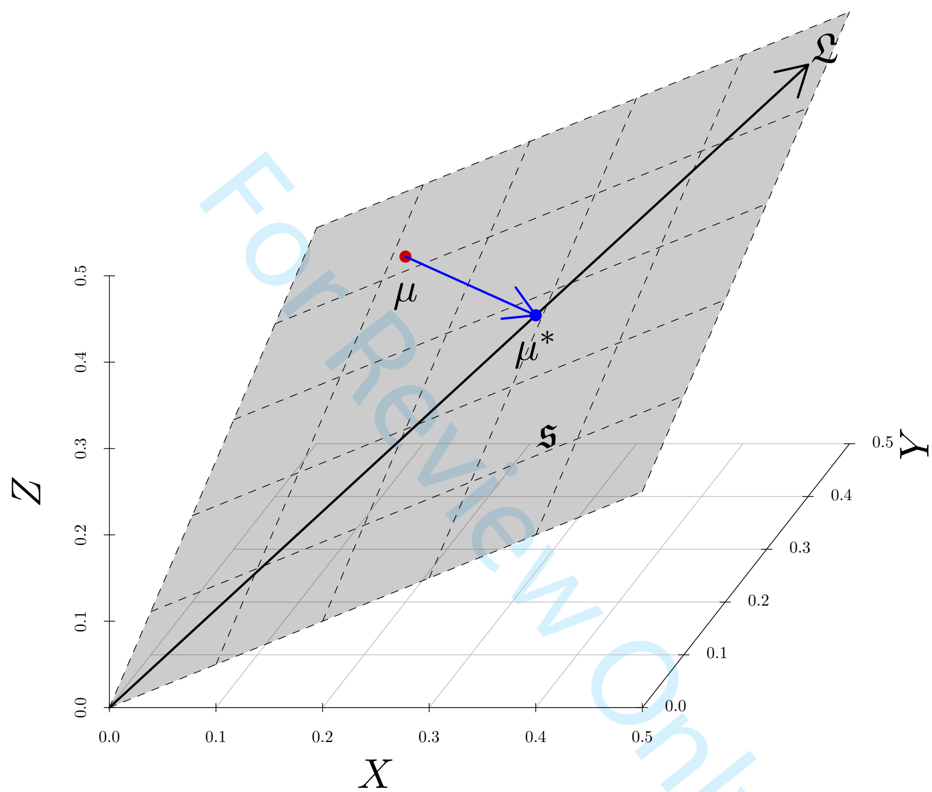

Figure 5: $\mathfrak{L}$ is a linear subspace of the coherent subspace $\mathfrak{s}$. If a projection is onto $\mathfrak{L}$ instead of $\mathfrak{s}$, then $\boldsymbol{\mu} \in \mathfrak{s}$ will be moved to $\boldsymbol{\mu}^* \in \mathfrak{L}$.

Finally, it is often stated that an assumption required to prove the unbiasedness preserving property is that $\boldsymbol{SGS} = \boldsymbol{S}$ or alternatively that $\boldsymbol{GS} = \boldsymbol{I}$. Both of these conditions are equivalent to assuming that $\boldsymbol{SG}$ is a projection matrix (see Section A.1 in Appendix A for a proof). Despite this connection, problems arise when viewing the preservation of unbiasedness through the prism of imposing the constraint $\boldsymbol{GS} = \boldsymbol{I}$. This thinking suggests that a way to deal with biased forecasts is to select $\boldsymbol{G}$ in an unconstrained manner. Equipped with a geometric understanding of the problem, we would advise against this approach. The constraint $\boldsymbol{GS} = \boldsymbol{I}$ is not just about bias. Dropping the constraint compromises all of the attractive properties of projections. It also opens the door to reconciliation methods that change already coherent base forecasts, which suggests an increase in the variability of the forecasts. This seems particularly perverse when the motivation for using a biased method in the first place is to reduce variance.

4.1 Bias correction

Our own solution to dealing with biased forecasts is to bias correct *before* reconciliation. In many cases the method for bias correction will be context specific. For instance, in our empirical study in Section 5 we consider a scenario where data are modelled after taking either a log transformation or a Box-Cox transformation. Since linear constraints that hold on the original scale do not hold for the transformed series, back-transforming to the original scale is necessary. Since this step induces a bias we propose to bias correct after this back-transformation step, but before reconciliation. In the well-known case of the Box-Cox transformation a number of bias correction methods exist based on Taylor expansions.

Alternatively, a more general purpose approach to bias correction is to simply estimate the bias by taking the sample mean of $\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}$ for all $t+h$ in the training sample. This can then be subtracted from future forecasts. As stated in the discussion of MinT, in-sample errors are already used to estimate the optimal direction of projection. As such it may be possible to use the same errors to bias correct. Geometrically, the intuition is simple. In panel (a) of Figure 4, the orange points are centered around the origin as would be expected from an unbiased forecast. If forecasts are biased, then errors should simply be translated until they are centered at the origin. Nonetheless there are also a number of pitfalls to such an approach. First, for the very construction we consider, where bias is induced by taking a log or Box Cox transformation, bias should be corrected by a multiplicative rather than an additive factor. Second, if in-sample errors are non-stationary due to model misspecification or structural breaks, then the proposed method for bias correction may break down.

5 Empirical study

Using an empirical application to forecast Australian domestic tourism flows, we illustrate the usefulness of projection-based reconciliation in practice. Previous studies have found that reconciliation improves point forecast accuracy in domestic tourism flows for Australia (see for example Athanasopoulos et al. 2009, Hyndman et al. 2011, Wickramasuriya, Athanasopoulos & Hyndman 2019). Our motivation in this study is twofold. First, we demonstrate the implications of Theorems 3.1, 3.2 and 3.3 by comparing reconciled and base forecasts. In contrast to previous studies, we consider individual periods as well as compute averages over a rolling window. Second, we demonstrate how the bias correction methods discussed in the previous section along with the projection-based reconciliation help to improve forecast accuracy.

5.1 Data

We consider "overnight trips" across Australia as a measure of domestic tourism flows. The data are provided by the National Visitor Survey and are collected through telephone interviews from an annual sample of $120,000$ Australian residents aged 15 years or more. We disaggregate tourism flows into 7 states, 27 zones and 75 regions forming a natural geographical hierarchy that is of interest to tourism operators and policy makers amongst others. Hence, there are 110 series across the hierarchy with 75 bottom-level series. More information about the series and the geographical hierarchy is presented in Table 3 in Appendix B. The data span the period January 1998 to December 2017, which gives a total of 240 observations per series.

Figure 6 shows time, sub-series and seasonal plots of the aggregate overnight trips. As is usual with tourism data, these show a strong seasonal pattern with peaks observed every January corresponding to the summer vacation season in Australia. There are also some lower peaks observed in April, July and October corresponding to school term breaks. On the other hand, the month with the least overnight trips is February indicating that people travel least for the month following their summer vacation. The time plot also shows a pronounced upward trend starting from around 2010 to the end of the sample, with flows being fairly flat from the beginning of the sample and a slight downward trend during 2004–2010.

The top panel of Figure 7 shows time plots for the six states and Northern Territory, hence the first level of the geographical hierarchy. The panels below show some selected series from the second-level zones and the bottom-level regions. The plots display the diversity of time series features, within but also between levels. For example, noticeable at the first level is the asynchronous seasonal pattern between the Northern Territory and the states. For the Northern Territory the high tourist season occurs during June-August with July being the peak, while the low season is during December-February. This reflects the tropical climate of the Northern Territory, with Australians mostly visiting the north during its dry winter-season rather than the wet summer season. Noticeable as we move to the lower levels is the variation in the signal-to-noise ratio, with the regional bottom-level series being much noisier compared to the series from levels above. This of course highlights the importance of modelling series at all levels without any loss of valuable information. We should note here that we observed an anomalous (extremely high) observation for 'Adelaide Hills' for December 2002. We replaced this observation with the average overnight trips on December 2001 and December 2003 for the same destination.

5.2 Comparison to Base Forecasts

To demonstrate the implications of Theorem 3.1 we consider the improvement of different reconciliation methods over base forecasts using different loss functions. For each series the ARIMA model minimising AICc is chosen using the `auto.arima()` function in the `forecast` package. Using these fitted models, base forecasts are produced for $h = 1$ to 6-steps ahead for each series in the hierarchy. This is first carried out with a training sample of 100 observations, i.e., Jan-1998 to Apr-2006. The training window is then rolled forward one observation at a time until the end of the sample. This generates 140 1-step-ahead,

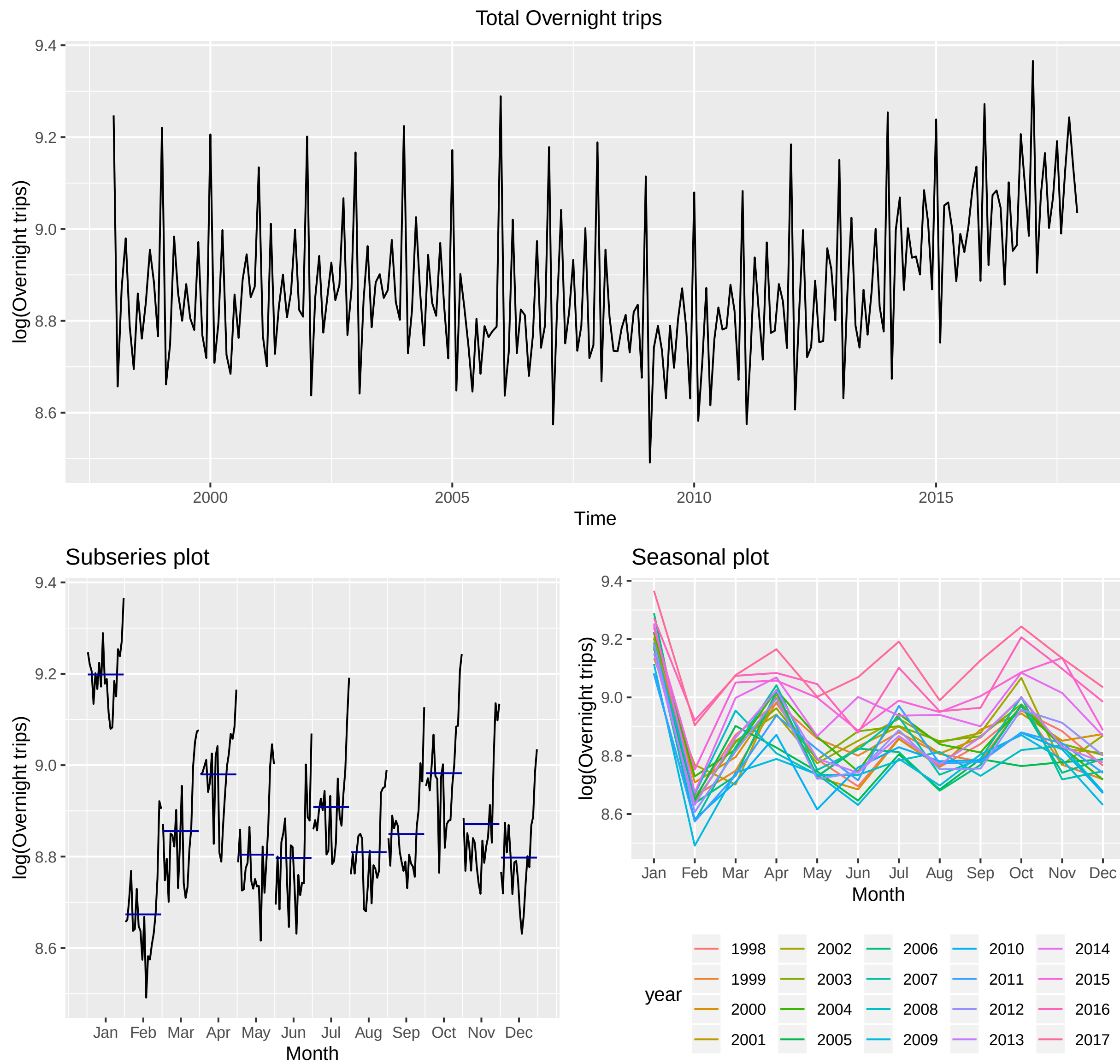

Figure 6: Total domestic overnight trips (in logs) for Australia from January 1998 to December 2017. The top-panel shows a time plot; the bottom-left panel a sub-series plot for each month; the bottom-right panel shows a seasonal plot coloured by year.

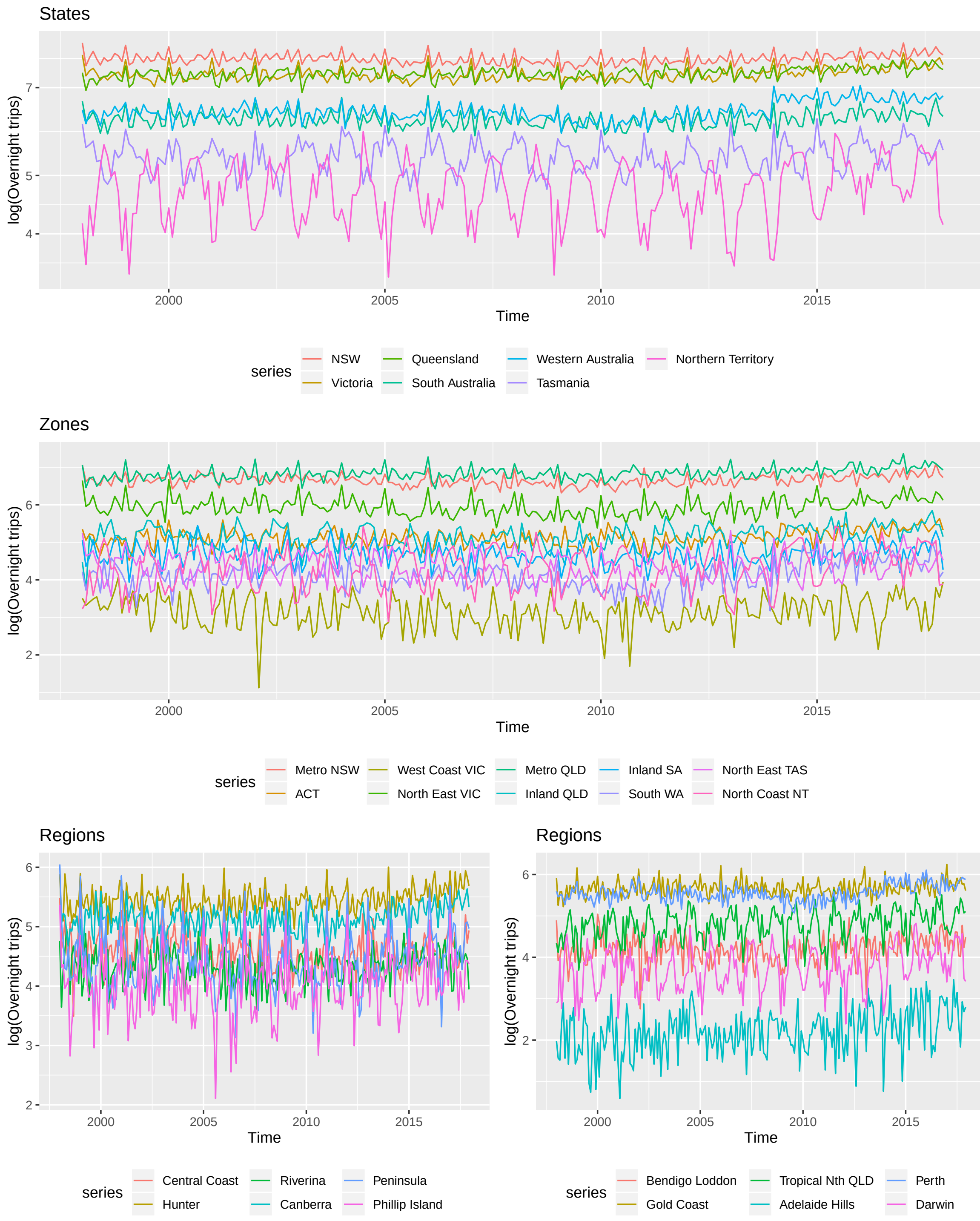

Figure 7: Time plot of overnight trips for some selected series from different disaggregate levels of the hierarchy. All values are presented in log scale. To avoid impact from the zero values we added a constant 1 to all observations

139 2-steps-ahead through to 135 6-steps-ahead forecasts available for forecast evaluation.

After obtaining the base forecasts these are reconciled using various projection methods. In particular: OLS reconciliation, MinT and WLS reconciliation with two different choices of weights that are defined below. For MinT the shrinkage estimator of Schäfer & Strimmer (2005) is used to estimate $\mathbf{\Sigma}$. It is given by $\tau \text{diag}(\hat{\mathbf{\Sigma}}) + (1-\tau)\hat{\mathbf{\Sigma}}$ where $\hat{\mathbf{\Sigma}}$ is the sample estimate of the variance covariance matrix of the in-sample, one-step ahead forecast errors and,

$$\tau = \frac{\sum_{i \neq j} \hat{\text{Var}}(\hat{\sigma}_{ij})}{\sum_{i \neq j} \hat{\sigma}_{ij}^2},$$

where $\hat{\sigma}_{ij}$ denotes the (i,j)th element of $\hat{\mathbf{\Sigma}}$.

For each method we compute three loss functions. The first is the total squared error (TSE), computed as

$$\text{TSE}_t^q = \sum_{i=1}^{n} (y_{i,t} - \tilde{y}_{i,t}^{(q)})^2, \tag{4}$$

where $\tilde{y}^{(q)}$ is the reconciled forecast using method q for series i and replication t. This loss function is L^2, the square of the usual Euclidean distance described in Section 3. We also consider weighted squared error

$$\text{WSE}_t^q = \sum_{i=1}^{n} w_i \left((y_{i,t} - \tilde{y}_{i,t}^{(q)}) \right)^2,$$

which is a loss function based on a squared generalised Euclidean distance $L^2_{\boldsymbol{W}}$, with diagonal $\boldsymbol{W}$. We consider two choices of weights. The first is the squared inverse of the number of bottom-level series included in forming a specific aggregate. For example, for all bottom-level series this weight is 1, whereas for the top-level series this is 1/75. The idea is to ensure top-level series, which are on a much larger scale, do not dominate the forecast evaluation metric. Using these weights in WLS reconciliation is equivalent to what Athanasopoulos et al. (2017) refer to as *structural scaling*. As such we refer to the reconciliation method that uses these weights as 'Stuctural-WLS' and the loss function based on these weights as 'Structural-WSE'.

The second choice of weights is motivated by our empirical example. In addition to visitor numbers per region, we also have access to data on average spend per region. In some settings, it may be desirable to have greater forecast accuracy in regions where tourists spend more. By using average spend per region as weights, the error metric (and transformed space associated with this metric) can be interpreted in terms of revenue, measured in dollars, rather than raw tourist numbers. We refer to the reconciliation method that uses these weights as 'Spend-WLS' and the loss function based on these weights as 'Spend-WSE'.

For each replication we compute the ratio of the loss of each alternative reconciliation method to the loss of base forecasts. A value less than 1 indicates that a reconciliation method has a lower relative error than the base forecast for that replication, while a value greater than 1 indicates the opposite. The boxplots in Figure 8 summarise the distribution of these ratios over each rolling window. We only present the results for $h = 1$, but the results and conclusions that follow are almost identical for the other longer forecast horizons. We do not present these here to save space but they are available in an online supplement[5].

For TSE, relative loss is always less than 1 only for OLS reconciliation, for Structural-WSE the same is only true for Structural-WLS and for Spend-WSE the same is only true for Spend-WLS. Therefore, Figure 8 demonstrates that a reconciliation method is guaranteed to improve upon base forecasts only when the $\boldsymbol{W}$ used in the loss function and reconciliation coincide. This is precisely what Theorem 3.1 and Theorem 3.2 would predict. On the other hand, for every loss function, MinT will perform worse than base for some realisations. For Theorem 3.2 to hold for MinT, the loss function would need to set $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$. Since the estimate of $\boldsymbol{\Sigma}$ will change with each replication we do not believe this is a sensible loss function to use.

The advantage of MinT however is clearly seen when loss functions are averaged (an estimate of expected loss). Table 1 reports the relative total error for each loss function. For example, for TSE the relative mean total squared error (RMTSE) is defined as

$$\text{RMTSE}^q = \frac{\frac{1}{140}\sum_{t=1}^{140}\text{TSE}_t^q}{\frac{1}{140}\sum_{t=1}^{140}\text{TSE}_t^{\text{Base}}} \tag{5}$$

[5]Available at https://anastasiospanagiotelis.netlify.com/papers/GeomRecoSupplement.pdf

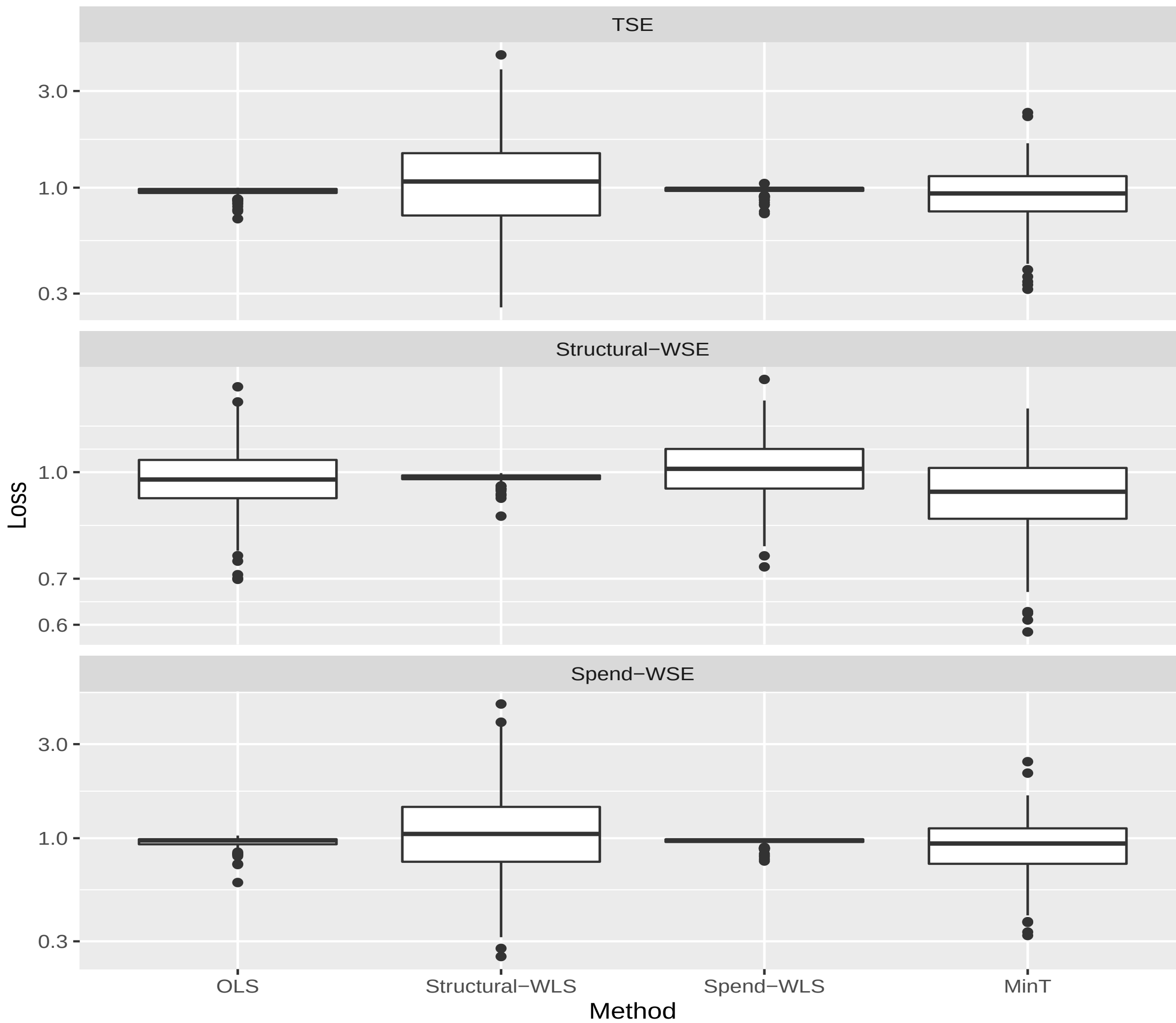

Figure 8: Ratio of loss of reconciled forecast to loss of base forecast for $h = 1$. A value less than 1 indicates that the reconciled forecasts improve upon base forecasts. A log scale is used for the y axis.

where $\text{TSE}_t^{\text{Base}}$ is the total squared error of the base forecasts at replication t. In contrast with what is displayed in the boxplots, here the average is taken over the replications before taking a ratio. Table 1 shows that the average loss for MinT is lower than for all other reconciliation methods, irrespective of the loss function used. This is precisely what would be expected from Theorem 3.3. For Structural-WSE, a Friedman test confirms that the average loss for MinT is significantly lower when tested against every other method[6].

Loss Function	Base	Bottom-up	OLS	Structural-WLS	Spend-WLS	MinT
TSE	1.00	1.22	0.97	1.13	0.98	0.96
Structural-WSE	1.00	1.01	0.96	0.98	1.00	0.93
Spend-WSE	1.00	1.20	0.97	1.12	0.98	0.96

Table 1: Means of different loss functions for 1-step ahead forecasts using different reconciliation methods in the tourism application. All figures are reported relative to base forecasts.

5.3 Transformations and bias adjustment

We first transform each series in the hierarchy using two types of transformations. Namely, we perform a log-transformation and also the more general Box-Cox transformation. A Box-Cox transformation is defined as,

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ \frac{y_t^\lambda - 1}{\lambda} & \text{otherwise.} \end{cases}$$

We first set $\lambda = 0$ and hence consider only a log transformation. For the second more general Box-Cox transformation we select λ using the "Guerrero" method (Guerrero 1993) implemented in the `BoxCox.lambda()` function in the `forecast` package in R (Hyndman et al. 2019). In order to avoid extreme and volatile transformations we restrict $\lambda \in (-0.5, 2)$. As zero observations exist in some of the bottom-level series, before transforming we add a constant (specifically 1) to each series. This overcomes the challenge of undefined transformed values for zero observations when we specifically implement the log transformation or when λ is selected to be zero by the "Guerrero" method. The constant is subtracted from the final forecasts.

[6]While differences are not significant for the other loss functions, the scale of the aggregate series leads to a large variance in loss, reducing the power of the test. Strutural-WSE stabilises this effect.

After transformation we fit univariate ARIMA models to each transformed series. The `auto.arima()` function in the `forecast` package is used to choose the best model that minimises the AICc. Using the fitted models, forecasts are produced for $h = 1$ to 12-steps ahead for each series in the hierarchy. The same rolling window described in Section 5.2 is used here as well.

The forecasts are then back-transformed by simply reversing the Box-Cox transformation using,

$$\hat{y}_{t+h|t} = \begin{cases} \exp(\hat{w}_{t+h|t}) & \text{if } \lambda = 0, \\ (\lambda \hat{w}_{t+h|t} + 1)^{1/\lambda} & \text{otherwise.} \end{cases} \tag{6}$$

These back-transformed forecasts are potentially biased as they are not the mean of the forecast distribution but the median (assuming that the distribution of the transformed space is symmetric). Hence, the reconciled forecasts that follow from these forecasts will also be biased. We refer to these as "Biased" base forecasts in the results that follow. This is the exact scenario we want to demonstrate in this study and we next move to our proposed solution of bias correcting the base forecasts before reconciling for which we explore two scenarios.

Using a Taylor series expansion (Guerrero 1993), the back-transformed mean of the forecast distribution for a Box-Cox transformation is given by

$$\hat{y}_{t+h|t} = \begin{cases} \exp(\hat{w}_{t+h|t}) \left[1 + \frac{\sigma_h^2}{2}\right] & \text{if } \lambda = 0, \\ (\lambda \hat{w}_{t+h|t} + 1)^{1/\lambda} \left[1 + \frac{\sigma_h^2(1-\lambda)}{2(\lambda \hat{w}_{t+h|t}+1)^2}\right] & \text{if } \lambda \neq 0, \end{cases} \tag{7}$$

where $\hat{w}_{t+h|t}$ is the h-step-ahead forecast from the Box-Cox transformed series and σ_h^2 is the variance of $\hat{w}_{t+h|t}$. Using the mean of the forecast distribution returns bias-adjusted base forecasts compared to the simple back-transformation of Eq. (6). We refer to this as "Method-1" in the results that follow. The second scenario of bias adjustment we explore is using the in-sample forecast error mean of the biased forecasts to adjust the out-of-sample forecasts. We refer to this as "Method-2" in the results that follow.

Using the three sets of base forecasts from each of the two transformations, we generate coherent forecasts implementing OLS and MinT reconciliation projections, and also the bottom-up approach and compare the results for when the base forecasts are biased and bias-adjusted, i.e., unbiased. In addition to the relative mean total squared error (RMTSE) defined in Eq. (5), the relative mean absolute total error (RMATE) is used to measure bias. Total error (TE) is first calculated

$$\mathrm{TE}_i^q = \sum_{t=1}^{140}(y_{i,t} - \tilde{y}_{i,t}^{(q)})$$

reflecting the total bias of method q across the 140 iterations for each series i. Taking the absolute value of each of these, so that positive and negative biases do not cancel across series, and then averaging over the 110 series, RMATE is defined as,

$$\mathrm{RMATE}^q = \frac{\frac{1}{110}\sum\limits_{i=1}^{110}|\mathrm{TE}_i^q|}{\frac{1}{110}\sum\limits_{i=1}^{110}|\mathrm{TE}_i^{\mathrm{Base}}|}.$$

Table 2 reports both RMTSE and RMATE for 1-step-ahead forecasts.[7] An asterisk (*) indicates that forecasts are significantly different from the biased base forecasts. Statistical significance of the differences in the forecast errors is based on the non-parametric Friedman and post-hoc Nemenyi tests, at a 5% level of significance (Hollander et al. 2013). The Friedman test first establishes whether at least one of the forecasts is significantly different from the rest. If this is the case, we use the Nemenyi test to identify groups of forecasts for which there is no evidence of statistically significant differences. This testing approach does not impose any distributional assumptions and does not require multiple pairwise testing between forecasts, which would distort the outcome of the tests. We use the implementation of the tests available in the `tsutils` (Kourentzes 2019) package for R.

Recall that reconciliation approaches via projections preserve unbiasedness in the reconciled forecasts iff the base forecasts are unbiased. Hence, the two columns labelled "Biased" contain results for biased base but also reconciled forecasts. Using Method-1 for first bias adjusting the base forecasts and then reconciling, results in forecast improvements for all

[7]Results and conclusions that follow are almost identical for the other longer forecast horizons. We do not present these here to save space but they are available at `https://anastasiospanagiotelis.netlify.com/papers/GeomRecoSupplement.pdf`.

methods for both RMATE and RMTSE and both the log and the Box-Cox transformations. There are improvements over the biased base forecasts OLS returns the best results for RMATE while MinT returns the best results for RMTSE. In contrast to the results from using Method-1 for bias adjusting before reconciliation, using Method-2 has an adverse effect on the forecast accuracy of the reconciled forecasts. In this case the reconciled unbiased forecasts leads to a significantly worse RMATE and RMTSE compared to base forecasts. This sends the warning that implementing inappropriate bias adjustment, in this case using an additive rather than a multiplicative factor, will hinder forecast accuracy and

Table 2: RMATE and RMTSE of 1-step-ahead forecasts from log and Box-Cox transformed series. Biased denotes forecasts from simply reversing the transformation via Eq. (6). Unbiased(Method-1) performs bias adjustment via a Taylor series expansion as shown in Eq. (7) whereas Unbiased(Method-2) bias adjusts by subtracting the in-sample forecast error mean.

	Log Transformation			Box-Cox Transformation		
Method	Biased	Unbiased (Method-1)	Unbiased (Method-2)	Biased	Unbiased (Method-1)	Unbiased (Method-2)
			RMATE			
Base	1.00	0.58*	1.40*	1.00	0.73*	1.38*
OLS	0.63*	0.54*	0.76	0.65*	0.67*	0.86
MinT	0.77*	0.57*	1.15	0.77*	0.70*	1.09
Bottom-up	1.76*	0.69*	2.72*	1.73	0.84*	2.57*
			RMTSE			
Base	1.00	0.99	1.01	1.00	0.98	1.04
OLS	0.97*	0.96*	0.98*	0.97*	0.96*	1.01
MinT	0.97*	0.93*	1.03	0.93*	0.91*	0.99
Bottom-up	1.42	1.18	1.80	1.35	1.16	1.63

* indicates a statistically significant difference from the biased base forecasts.

extra care must be taken in this bias adjustment procedure.

Also of interest is the fact that reconciliation can to some extent mitigate bias even without bias correction. In particular, using OLS reconciliation without bias correction leads to a statistically significant reduction in bias relative to base forecasts. This is likely to occur since the direction of bias lies in a direction that is close to orthogonal to the coherent subspace. Projection therefore eliminates this bias to some extent.

6 Conclusions

Defining concepts such as coherence and reconciliation in geometric terms provides new insights into hierarchical forecasting methods. We recommend the following steps for practitioners.

1. Choose an objective, either

   a. To guarantee that reconciled forecasts improve upon base forecasts.

   b. To find the reconciliation method that is best on average.

2. Select a $\boldsymbol{W}$ to use in loss function $L^2_{\boldsymbol{W}}$ that is well suited to the empirical problem. For objective (a) our results also apply to any monotonic function of $L^2_{\boldsymbol{W}}$.

3. Select a reconciliation method.

   - For objective (a) this should be $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}$. The dependencies between forecast errors is not relevant.
   - For objective (b) this should be $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{\Sigma}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{\Sigma}^{-1}$. The choice of $\boldsymbol{W}$ is not relevant.

Furthermore, for the second objective, base forecasts must be unbiased. We recommend carrying out bias correction before reconciliation and provide evidence that this improves forecast accuracy compared to approaches that do not bias correct and/or do not use reconciliation.

Our intention in proposing a geometric interpretation is also to provoke research into new areas. We now discuss three such possibilities. First, it should be possible to extend the concept of coherence to examples where the coherent space is not a linear plane in $\mathbb{R}^n$. This includes the case where in addition to aggregation constraints, forecasts are also constrained to be non-negative. We note the work of Wickramasuriya, Turlach & Hyndman (2019) as an attempt to address this issue. Another possibility is non-linear constraints where the coherent space may need to be defined by a manifold. Although much more challenging, it is still possible to define reconciled forecasts in terms of projections onto a manifold. Second, since we have established that the concept of bottom-level series is not crucial in forecast reconciliation, an open question is whether it may be better to construct base forecasts of linear combinations of the time series rather than the time series themselves. Finally, the geometric interpretations of hierarchical forecast reconciliation facilitates an extension into a probabilistic framework. The latter two are issues we investigate in separate papers.

A Appendix

A.1 Proof $\boldsymbol{SGS} = \boldsymbol{S}$ implies $\boldsymbol{SG}$ is a projection

Here we establish that if $\boldsymbol{SG}$ is a projection onto the linear subspace spanned by $\boldsymbol{S}$ then $\boldsymbol{SGS} = \boldsymbol{S}$. We also prove that the converse holds, namely that if the condition $\boldsymbol{SGS} = \boldsymbol{S}$ holds then $\boldsymbol{SG}$ must be a projection onto the linear subspace spanned by $\boldsymbol{S}$.

To establish the first statement, let $\boldsymbol{s}_j$ be the jth column of $\boldsymbol{S}$. Since by definition, $\boldsymbol{s}_j$ lies in $\mathfrak{s}$, it must hold that $\boldsymbol{SGs}_j = \boldsymbol{s}_j$. Stacking these vectors horizontally

$$\begin{aligned}\boldsymbol{SGS} &= \begin{pmatrix}\boldsymbol{SGs}_1, & \boldsymbol{SGs}_2, & \cdots & \boldsymbol{SGs}_m\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{s}_1, & \boldsymbol{s}_2, & \cdots & \boldsymbol{s}_m\end{pmatrix}\\ &= \boldsymbol{S}.\end{aligned}$$

To establish the converse it suffices to postmultiply the condition $\boldsymbol{SGS} = \boldsymbol{S}$ by $\boldsymbol{G}$. This yields $\boldsymbol{SGSG} = \boldsymbol{SG}$, which in turn implies idempotence since $(\boldsymbol{SG})^2 = \boldsymbol{SG}$.

B Australian Tourism Data

Table 3: Geographical hierarchy of Australian tourism flow

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

References

Almeida, V., Ribeiro, R. & Gama, J. (2016), Hierarchical time series forecast in electrical grids, *in* K. J. Kim & N. Joukov, eds, 'Information Science and Applications', Springer, Singapore, pp. 995–1005.

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., Gamakumara, P., Panagiotelis, A., Hyndman, R. J. & Affan, M. (2019), Hierarchical forecasting, *in* P. Fuleky, ed., 'Macroeconomic Forecasting in the Era of Big Data', Springer, Honolulu, chapter 21, pp. 703–733.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N. & Petropoulos, F. (2017), 'Forecasting with temporal hierarchies', *European Journal of Operational Research* **262**, 60–74.

Ben Taieb, S., Taylor, J. W. & Hyndman, R. J. (2017), Coherent probabilistic forecasts for hierarchical time series, *in* 'Proceedings of the 34th International Conference on Machine Learning', Vol. 70, PMLR, pp. 3348–3357.

Chase, C. W. (2013), 'Using big data to enhance demand-driven forecasting and planning', *Journal of Business Forecasting* **32**(2), 27–32.

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.

Fliedner, G. (2001), 'Hierarchical forecasting: issues and use guidelines', *Industrial Management & Data Systems* **101**(1), 5–12.

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

Guerrero, V. M. (1993), 'Time-series analysis supported by power transformations', *Journal of Forecasting* **12**(1), 37–48.

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 edn, OTexts, Melbourne, Australia.
URL: *OTexts.com/fpp2*

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.

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.

Kahn, K. B. (1998), 'Revisiting top-down versus bottom-up forecasting', *The Journal of Business Forecasting* **17**(2), 14.

Karmy, J. P. & Maldonado, S. (2019), 'Hierarchical time series forecasting via support vector regression in the European travel retail industry', *Expert Systems with Applications* **137**, 59–73.

Kourentzes, N. (2019), *tsutils: Time Series Exploration, Modelling and Forecasting*. R package version 0.9.1.
URL: *https://github.com/trnnick/tsutils/*

Kourentzes, N. & Athanasopoulos, G. (2019*a*), 'Cross-temporal coherent forecasts for Australian tourism', *Annals of Tourism Research* **75**, 393–409.

Kourentzes, N. & Athanasopoulos, G. (2019*b*), Elucidate structure in intermittent demand series, Working Paper 27/19, Department of Econometrics & Business Statistics, Monash University.

Lapide, L. (1998), 'A simple view of top-down vs bottom-up forecasting', *Journal of Business Forecasting Methods & Systems* **17**, 28–31.

Li, H. & Tang, Q. (2019), 'Analyzing mortality bond indexes via hierarchical forecast reconciliation', *ASTIN Bulletin* **24**(3), 823–846.

Mahkya, D., Ulama, B. & Suhartono (2017), 'Hierarchical time series bottom-up approach for forecast the export value in Central Java', *Journal of Physics: Conference Series* **893**(012033).

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

Rao, C. R. (1974), 'Projectors, generalized inverses and the BLUE's', *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(3), 442–448.

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).

Schwarzkopf, A. B., Tersine, R. J. & Morris, J. S. (1988), 'Top-down versus bottom-up forecasting strategies', *International Journal of Production Research* **26**(11), 1833–1843.

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.

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.

Wickramasuriya, S. L., Turlach, B. A. & Hyndman, R. J. (2019), Optimal non-negative forecast reconciliation, Working Paper 15/19, Department of Econometrics & Business Statistics, Monash University.
URL: *robjhyndman.com/publications/nnmint/*

Detailed Response to Referee 1: Summary

The authors provide a geometric interpretation for reconciliation of hierarchical forecasts. They show why and how reconciliation via projection is guaranteed to improve squared forecast errors. They explore a couple of different ways for dealing with biased base forecasts in an application to Australian tourism flows. Overall the paper is well written and the geometric interpretations are an important contribution to the growing literature on forecast reconciliation. The authors do a very good job explaining the geometric aspects, which lead to new insights. That being said, the contribution of the paper in its current form is mainly theoretical, as the empirical evaluation is not much of a contribution. I do not think the paper meets the high standard set by the IJF in terms of empirical evaluation. I recommend that the authors revise their paper with a particular focus on strengthening its empirical contribution to more clearly show the practical value of the geometric insights they derive. I hope that the authors will find my comments useful for improving their paper.

We thank the referee for their comments acknowledging the theoretical contributions of the paper. While we believe that more theory focused papers lie within the scope of the International Journal of Forecasting, we also agree with the referee that the empirical section could be greatly improved. We have endeavoured to do this in ways outlined in our responses below.

Detailed Response to Referee 1: Major comments

1. *Lack of empirical contribution: The authors apply a couple of different transformations to the tourism data before constructing base forecasts. Failing to correctly transform the base forecasts back before they are reconciled introduces a bias. That the highest accuracy is obtained when applying the correct back-transformation is a very small contribution.*

   We agree that if our empirical results only showed that applying the correct back-transformation improves accuracy, that this would be a very marginal contribution. The referee's suggestions have allowed us to make additional insights that we believe have greatly improved the paper. We discuss these in detail at the relevant points below.

   When the authors introduce the log-transformation and in the graphs in Figures 6 and 7, they do not clarify the interplay between the transformation and the reconciliation constraints. If $A + B = C$ then clearly $\log A + \log B \neq \log C$. In other words, a log-transformation affects the reconciliation constraints. This needs to be explained more clearly; e.g., is the reconciliation constraint imposed on the trans-

formed or the raw data?

The observation regarding the constraints not holding on the log scale is correct. When series are transformed to the log scale, forecasts are first produced on the log scale and these are back transformed to the original scale (possibly with bias correction) prior to carrying out reconciliation. Therefore while all modelling is done on the log scale, all reconciliation is carried out on the original scale. The fact that the constraints do not hold on the log scale is therefore not relevant.

On re-reading the paper we do agree with the referee that there is scope for confusion here. Therefore we have added some additional exposition to the beginning of section 4.1. The aim here is to make it clear that after producing forecasting on some tranformed scale, the correct sequence of steps is to 1) back-transform to the original scale, 2) bias correct and 3) reconcile.

When the empirical evaluation is focused on bias, then it would make sense to include an error that measures bias in addition to the squared error.

We now also report a measure of bias, defined in section 5.2. Here we see that bias correction method 1 removes bias, while bias correction method 2 in fact worsens bias. A new insight that we are able to make is to recognise that reconciliation without bias correction can still mitigate bias. For example, even if bias correction is not used at all, OLS reconciliation significantly reduces bias for both the log transformation and Box Cox transformation. This is likely to occur since bias lies in a direction that is almost orthogonal to the coherent subspace.

Moreover, simply showing the MSE without any confidence intervals or measures of significance is not sufficient for the reader to assess the results. It would also be useful to show the MSE relative to the base forecasts or the percentage improvement that is obtained.

We now report all results relative to the base forecasts. We also have added stars to the tables that report our results to indicate when a method significantly upon base forecasts.

The best performing reconciliation method is MinT with shrinkage, but the authors never state the value of the shrinkage parameter or how it was chosen. Similarly, they compare with variance scaling without explaining what they mean by variance scaling.

The shrinkage estimator is that of Schäfer and Strimmer (2005). Details on this method (including the choice of shrinkage parameter) are now provided. Due to other changes in the paper we no longer consider variance scaling, we do however consider two alternative WLS reconciliation methods. The weights for these are explained in Section 5.2.

In addition to the above mentioned shortcomings, I think the authors should reconsider their empirical evaluation. Maybe a second case study or a simulation study is needed to show the value of the geometric insights provided. We already know that MinT is better than OLS and WLS. What is the new and better reconciliation approach that has come from the geometric insights?

The objective of this paper is not to propose a new and better reconciliation approach. Rather it is to establish new results that lead to clearer understanding of the properties of existing methods that can assist practitioners and in turn motivate research into new methods. Our recommendations to practitioners are now summarised in an expanded conclusion. Some insights into how new methods should (or should not) be developed by researchers are also provided in the conclusion and also just before Section 4.1. Furthermore, we note that Spend-WLS is a new reconciliation method (albeit somewhat context specific).

In this paper, the purpose of the empirical evaluation is to provide a demonstration of the theoretical results that we have established. We have now improved the empirical evaluation. We have greatly expanded Section 5.2 to more clearly investigate three different loss functions and demonstrate the different ways in which reconciliation methods can be considered optimal.

2. *Improvement guarantees: The boxplot in Figure 8 shows that OLS always improves MSE, while this is not the case for the other reconciliation approaches. To gain a better understanding of the implications of Theorem 3.2, it would be useful to show that the other approaches always improve accuracy in their transformed spaces.*

We now expand Section 5.2 to report results for three different error metrics. In each case we show that the improvement guarantee only holds if the reconciliation method is aligned with the choice of error metric / loss function (we use error metric and loss function synonymously here).

This result holds in a special case for MinT, when the loss function depends on the error covariance matrix, but since this matrix has to be reestimated for each replication, we do not believe this is a sensible loss function. Instead we now highlight that MinT has a rather different optimality property. While we briefly discussed this in the previous submission, we have now greatly expanded this discussion (mostly in Section 3.4). In particular we now show that MinT minimises expected loss for any loss function that is a squared generalised Euclidean distance. The choice of $\boldsymbol{W}$ is irrelevant in this setting.

What is the interpretation of the transformed spaces and can the authors make the connection between these spaces and the choice of reconciliation approach and error measure more clear?

Reflecting on this issue has led us towards making a number of changes. To reiterate, we may define a loss function based on the notion of Euclidean distance. In this case an orthogonal projection is guaranteed to lead to reconciled forecasts that are better than base forecasts. Alternatively we may define a more general loss function, for instance by taking weights into account. We now consider two such loss functions, one based on the structure of the hierarchy (so-called 'structural scaling') and another based on average spend in each region. To highlight how the transformed space can be interpreted in the latter context, we have added the following sentence to the manuscript: "By using average spend per region as weights, the error metric (and transformed space associated with this metric) can be interpreted in terms of revenue measured in dollars rather than raw tourist numbers."

In cases where the transformed space cannot be easily interpreted in terms of the empirical example, it is nevertheless an important construction in our proofs. This includes the new proof that MinT minimises expected loss for loss functions based on squared generalised Euclidean distance. In the revision we believe that we have more clearly decoupled the notion of optimality, loss functions and reconciliation methods that should be used in each case. The transformed space is critical to this understanding.

For example, Hyndman et al. (2011); van Erven and Cugliari (2015) argued for selecting OLS to increase the importance of forecasting the aggregate. What is the argument for WLS or MinT and what is the corresponding consistent error measure?

We would like to point out that any argument made by Hyndman et al. (2011) or van Erven and Cugliari (2015) that OLS reconciliation somehow targets and improves the aggregate series is based on empirical rather than theoretical evidence. Addressing this misconception, which appears to be common, has in fact been a major motivation behind us writing the paper. We hope that with the revision to the paper these distinctions are clearer.

Detailed Response to Referee 1:Minor Comments

1. *In the first half of the paper it feels like every other sentence includes a however. I suggest reducing the use of however.*

   We have reduced the usage of the word "however" substantially, from twelve instances to just four.

2. *P. 2, l. 11: In several places the authors talk about adjusting forecasts ex-post. Although I understand what is meant, it gives the impression that forecasts are adjusted after observing the realized values.*

   We have either removed all use of the term 'ex post' to avoid the potential for confusion.

3. *P. 2, l. 12: The authors discuss the regression formulation of forecast reconciliation. It would be useful to also make the connection to the optimization formulation considered by, e.g., van Erven and Cugliari (2015); Nystrup et al. (2020). This could also be useful for clarifying the connection between reconciliation approaches and error measures.*

   We now discuss the connection to the optimisation formulation considered by these two papers on page 2 line 12 as requested. We believe that some of the other revisions we have made, particularly the new discussion in 3.4 further clarify the connection between reconciliation approaches and error measures.

4. *P. 4, l. 22: forf*

   This has been corrected.

5. *P. 10, l. 12: the comma should not be there.*

   The comma has been removed.

6. *P. 11, Figure 3: usually a small square is drawn in the corner of the triangle to show orthogonality.*

   We have made this change.

7. *P. 17, l. 18: i,e.*

   We have made this correction.

8. *P. 26, l. 14: the authors mention that the full results are available upon request. I suggest including them in an online supplementary appendix.*

   We now include these an an online supplement available at https://anastasiospanagiotelis.netlify.com/papers/GeomRecoSupplement.pdf

9. *P. 27, Conclusions: The authors should comment on the implications of the non-uniqueness of the S matrix for future work on cross-temporal reconciliation (Kourentzes and Athanasopoulos, 2019).*

   It is not clear to us how the implications for the non-uniqueness of the S matrix for cross-temporal reconciliation are substantial (or at least more substantial than for any other paper in the literature). To reiterate, while S is not unique, the span of the columns of S is unique. While the S matrix used to construct a projection onto this linear subspace is non-unique, the final reconciled forecast will be the same.

Detailed Response to Referee 2

This type of paper makes me regret not investing more time into geometric interpretation because as shown here, it offers an elegant and intuitive way to showcase results related to data integration and reconciliation. The paper is extremely well written, with a great flow and thoughtful considerations. Figure 4 and its description are exemplary successful in their simplicity and effectiveness. The discussion of theorem 3.1 on page 11 is another example of thoroughness and clever insight.

We thank the referee for these kind comments.

I found only one statement in the paper that could be better supported by evidence on page 8 lines 14 when the author(s) refer to multivariate modeling. State-space approaches have also been shown to be theoretically successful is solving these problems although maybe not to the large scale needed for very detailed and complex hierarchical systems. A comment or comparative discussion to the multivariate modeling may be useful.

We have rewritten this sentence and now explicitly include the case of state space models in our discussion.

In the context of real-life application and either for small discussion here or future work, I am wondering if and how the author(s) coherent subspace that would be defined with hard boundaries. For example, the set of Australian tourism flow data and any forecasts that would be considered useful should be non-negative and likely upper bounded (if only by the size of the global population or other more realistic subject matter expert opinion). In many other reconciliation problems, these boundary constraints affect the feasible space. In the context here, could a convoluted case lead to an orthogonal projection be coherent but outside the desired constrained subspace?

It is indeed possible for an orthogonal projection to reconcile a set of positive base forecasts into a vector that includes some negative values. This is a fairly pathological case and does not arise in the application that we consider (and many other applications that we have considered in other work). We do agree that this problem may arise in other contexts so we now include some discussion of this issue in the conclusion and cite some recent work that addresses this problem.

Please correct the minor typo just before section 2 (forf).

We have made this correction.