Stochastic Coherency in Forecast Reconciliation

Kandrika F. Pritularga[a,*], Ivan Svetunkov[a], Nikolaos Kourentzes[b]

[a]*Centre for Marketing Analytics and Forecasting, Department of Management Science, Lancaster University Management School, UK*

[b]*Skövde Artificial Intelligence Lab, School of Informatics, University of Skövde, Sweden.*

Abstract

Hierarchical forecasting has been receiving increasing attention in the literature. The notion of coherency is central to this, which implies that the hierarchical time series follows some linear aggregation constraints. This notion, however, does not take several modelling uncertainties into account. We propose to redefine coherency as stochastic. This allows to accommodate overlooked uncertainties in forecast reconciliation. We show analytically that there are two potential sources of uncertainty in forecast reconciliation. We use simulated data to demonstrate how these uncertainties propagate to the covariance matrix estimation, introducing uncertainty in the reconciliation weights matrix. This then increases the uncertainty of the reconciled forecasts. We apply our understanding to modelling accident and emergency admissions in a UK hospital. Our analysis confirms the insights from stochastic coherency in forecast reconciliation. It shows that we gain accuracy improvement from forecast reconciliation, on average, at the cost of the variability of the forecast error distribution. Users may opt to prefer less volatile error distributions to assist decision making.

Keywords: forecasting, coherency, model uncertainty, forecast combination, covariance estimation

1. Introduction

Forecasting is an essential activity for decision making in organisations. Often forecasts and supported decisions are organised in hierarchies. These hierarchies can be constructed from market segments, products, or other demarcations (Athanasopoulos et al., 2009). Beyond cross-sectional hierarchies, there are temporal hierarchies where different functions in an organisation require forecasts at different sampling frequencies and planning horizons (Athanasopoulos et al., 2017). Combining both is also possible, which aims to provide a coherent view of the future across both dimensions (Kourentzes and Athanasopoulos, 2019).

All the hierarchical forecasting methods are based on the property of coherency (Wickramasuriya et al., 2019; Jeon et al., 2019; Taieb et al., 2020; Athanasopoulos et al., 2020). It implies

*Correspondance: K Pritularga, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK. Tel.: +44 1524 592911

Email addresses: `k.pritularga@lancaster.ac.uk` (Kandrika F. Pritularga), `i.svetunkov@lancaster.ac.uk` (Ivan Svetunkov), `nikolaos@kourentzes.com` (Nikolaos Kourentzes)

that the lower level forecasts add up to the forecasts of the higher levels. For example, sales of individual products in a hierarchy sum up to product category sales at higher levels, in observations and forecasts. When forecasts are produced independently, they are typically not coherent, and this has been one of the motivations for developing hierarchical forecasting methods. This enables aligned planning and actions throughout organisations and stake-holders (Kourentzes and Athanasopoulos, 2019), which is the main motivation for hierarchical forecasting in an organisational context. The concept of coherency has been central in temporal dissagregation, establishing a link between high and low frequency time series. For example, Chow and Lin (1971) uses a highly restrictive generalised linear regression model to this purpose.

In the past, studies attempted to tackle this problem by employing a bottom-up or top-down approach (Fliedner, 2001). The main issue with these methods is that they ignore information either at higher levels or lower levels (Athanasopoulos et al., 2009; Ord et al., 2017), thus leading to less accurate forecasts. Furthermore, implicitly we accept increased modelling risk, as all forecasts in the hierarchy are based on a single (top-down) or a few (bottom-up) forecasting models, which may be misspecified. This misspecification can have adverse effects on the uncertainty of the forecasts across the hierarchy, resulting in increased costs of any supported decisions, such as unmet demand due to poor forecasts.

Nowadays, hierarchical forecasting is seen as a reconciliation problem, where forecasts are generated at all levels and then are reconciled to a common view of the future (Hyndman et al., 2011, 2016; Wickramasuriya et al., 2019). Several studies have shown significant improvements in forecast accuracy in different contexts, when hierarchical reconciliation techniques are used (Yang et al., 2016; Oliveira and Ramos, 2019; Kourentzes and Athanasopoulos, 2019, 2021). In brief, forecast reconciliation is achieved by linearly combining all forecasts from the hierarchy to a set of adjusted bottom-level forecasts, which by construction make use of all available information, and then aggregating these to reconciled forecasts for the complete hierarchy. Note that since we no longer rely on forecasts at any specific level, we mitigate uncertainties stemming from the specification of the forecasting methods that both top-down and bottom-up methods suffer from. Apart from providing coherent forecasts, the reconciliation methods also often improve upon the accuracy of the base independent forecasts in cross-sectional, temporal, and cross-temporal hierarchies (Hyndman et al., 2011; Athanasopoulos et al., 2017; Wickramasuriya et al., 2019; Kourentzes and Athanasopoulos, 2019; Kourentzes et al., 2021).

Nonetheless, in the literature there is empirical evidence that hierarchical forecasting does not universally result in reduced forecast uncertainty and better forecast accuracy. In temporal hierarchies, Athanasopoulos et al. (2017) demonstrate that model selection uncertainty affects the efficacy of forecast reconciliation. Base forecasts from well-specified forecasting models gain little benefit from forecast reconciliation, whereas forecasts from mis-specified models are improved significantly. Furthermore, Kourentzes and Athanasopoulos (2019) find that combining cross-sectional and temporal hierarchies offers 'small yet significant' improvement upon the accuracy of the base forecasts, as the first dimension, the temporal, already mitigates much of the uncertainty in base forecasts. However, in order to reconcile forecasts, a reconciliation weights matrix is needed, and defining this matrix in the cross-temporal case can be challenging (see also Di Fonzo and Girolimetto, 2020). In cross-temporal hierarchies, Kourentzes and Athanasopoulos (2019) average across multiple estimates of the reconciliation weights matrix to avoid unnecessary estimation uncertainty, while retaining coherency. Results in Panagiotelis et al. (2021) show substantially different forecast error variances depending on how the reconciliation weights matrix is calculated. Empirical results from the retail and tourism sectors further demonstrate this variability of performance that appears to depend on the calculation of the reconciliation weights (Wickramasuriya et al., 2019; Oliveira and Ramos, 2019). We argue that there are inherent uncertainties in forecast reconciliation that have not been explored in the literature, which we investigate here.

Thus, it appears that recent studies have overlooked the effect of uncertainties in forecast reconciliation. Panagiotelis et al. (2020, Theorem 3.1) demonstrate that the only source of uncertainty is originating from the base forecasts, and the reconciliation weights matrix is assumed to have no uncertainty and effectively treated as known. Nevertheless, since we estimate the reconciliation weights, we face uncertainty in their estimation. Furthermore, as there are different approximations for the covariance matrix (Hyndman et al., 2011, 2016; Athanasopoulos et al., 2017; Wickramasuriya et al., 2019; Nystrup et al., 2020), this leads to a selection question. Thus, the conventional reconciled forecast variance is potentially underestimated. Note that we consider parameter estimation and forecasting method selection uncertainties as two aspects of the same modelling issue.

Apart from that, there is another complication with hierarchical time series. There is a gap between how the hierarchical time series are collected in practice and how we use the data for forecasting. Suppose that we see the original information coming from the bottom-level of the hierarchy. For example, in macroeconomic variables the data is collected either by surveys, estimates, or a combination of them. As the data are collected for different nodes or levels of the hierarchy, the bottom level does not always add up to the higher levels of the hierarchy. Statistics bureaux use the account 'statistical discrepancy' to fill the gap. Athanasopoulos et al. (2020) treat the discrepancy as another time series in forecast reconciliation. This affects how we perceive coherency in hierarchical time series, both in the observational and population levels, as well as how we understand modelling uncertainty in forecast reconciliation.

In order to address all these issues, we propose the notion of "stochastic coherency". Stochastic coherency is easy to understand if we use the geometric interpretation of forecast reconciliation (Panagiotelis et al., 2021). Incoherent base forecasts (the initial forecasts for each node of the hierarchy) are projected to a coherent subspace. Conventionally, this projection has no uncertainty, while with stochastic coherency, the projection becomes stochastic. Equivalently, if we see forecast reconciliation from a forecast combination interpretation (e.g., Kourentzes and Athanasopoulos, 2019), the combination weights are stochastic.

Suppose we have a set of forecasts from a sample of time series and a hierarchy is given. When we collect additional samples and re-estimate the reconciliation weights matrix, that is bound to change due to the estimation of covariance matrix approximations. However, as long as the estimated reconciliation weights matrix meets the coherency constraint (Wickramasuriya et al., 2019), the forecasts are coherent, but they will change as the weight matrix changes. The key here is to acknowledge that the uncertainty of coherent forecasts comes from the incoherent base forecasts and propagates to the estimation of the reconciliation weights matrix. This increases the uncertainty of coherent forecasts.

Another differentiating characteristic of stochastic coherency is how the error terms in the data generating process are treated. We realise that the error term in the hierarchical time series itself may contain not only the innovations but also potential errors coming from data collection, such as sampling and measurement errors. On top of that, modelling uncertainty is introduced when we produce forecasts. Hence, it allows us to decompose the variance of coherent forecasts. As we show later in the paper, this has important implications for the construction of the estimated covariance matrix and coherent forecasts.

Stochastic coherency affects not only point forecasts but also probabilistic forecasts. In order to understand the effect of stochastic coherency on probabilistic coherent forecasts better, we refer to its definition by Taieb and Koo (2019) and Panagiotelis et al. (2020). The former defines the probabilistic coherent forecasts as convolutions of linear constraints, while the latter defines them in a more flexible manner as to extend to non-linear constraints (Panagiotelis et al., 2020, p. 8). However, both definitions are rooted from the idea of forecast reconciliation where the base point forecasts are projected onto the coherent space by the reconciliation weights matrix, and they assume that the weights matrix is known. Our stochastic coherency highlights the uncertainty of these weights, or the projection, and it will affect both the point and the probabilistic forecasts. This will increase the coherent probabilistic forecasts uncertainty. We argue that these are important characteristics in the application of hierarchical forecasting in organisations, where understanding and controlling the sources of uncertainty is important for mitigating risks associated with decision making, beyond any accuracy improvements.

We explore stochastic coherency in detail in Section 2 and 3. We show when forecast reconciliation becomes beneficial, and we explore uncertainties in forecast reconciliation further, attributing them to their sources. We find that the more complete the covariance matrix approximation is, the better the resulting point forecast accuracy can be but at the cost of the increased variance of the reconciled forecast errors. In Section 4, we conduct a simulation experiment to validate our understanding. In Section 5 we apply this to modelling accident and emergency admissions at a UK hospital, demonstrating the effect of stochastic coherency on a real complex problem. Based on these findings, we discuss and conclude our work in Section 6.

2. Classical and Stochastic Coherency

The notion of coherency in hierarchical forecasting has been proposed and elaborated by a series of hierarchical forecasting works (Athanasopoulos et al., 2009; Hyndman et al., 2011, 2016; Wickramasuriya et al., 2019; Panagiotelis et al., 2021; Athanasopoulos et al., 2020). The literature defines forecasts as coherent forecasts if they adhere to a linear constraint, e.g. they add up according to the hierarchy, often simplified as the bottom level forecasts aggregating to the higher level forecasts. In a similar manner, Taieb et al. (2020) define mean coherent forecasts when the errors between aggregated bottom-level forecasts and the independent forecasts at the upper-level are zero.

Let us explore the hierarchical approach in detail. Suppose that $\boldsymbol{y}_t$ is a $n \times 1$ vector of hierarchical time series across the hierarchy, at period t, where $\boldsymbol{y}_t$ is constructed from $\boldsymbol{b}_t$, a $m \times 1$ vector of the bottom-level time series, and a summation matrix, $\boldsymbol{S}$. In this case, $\boldsymbol{S}$ maps the bottom-level onto the upper-level of the hierarchy. The coherent hierarchical time series is denoted as,

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t. \tag{1}$$

We argue that Eq (1) is not general. Let us consider how time series data is collected in different organisations. In any retailer which records demand of every stock keeping unit at the bottom-level in real-time, they can update new information in the middle and the top of hierarchy at time t, across the hierarchy, instantaneously. This means that the hierarchical time series are coherent, even as new information becomes available. On the other extreme, we may need to estimate the data, even though by nature it is a part of a hierarchy, for example the gross domestic product (GDP). For instance, the Office of National Statistics United Kingdom measures national accounts through surveys, forecasts, and estimates from models, which are subject to errors (Office for National Statistics, 2011). Once an account is measured, they need to reconcile the number from different methods and sources. Hence, in the case of GDP, the values in the hierarchy from aggregating the bottom-level data and collecting data from each level will be different. To accommodate the potential gap, the statistical bureaux create an account called statistical discrepancy (Australian Bureau of Statistics, 2015, p. 471). This discrepancy captures any potential error coming from measurement and sampling errors. Athanasopoulos et al. (2020) treat the discrepancy as a time series. It is easy to identify scenarios where such measurement issues violate the classical coherency, from individual companies, to national statistics. Therefore, due to the measurement errors, we redefine hierarchical time series $\boldsymbol{y}_t$ as,

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

where $\boldsymbol{\delta}_t$ is the statistical discrepancy at time t. By nature, $\boldsymbol{\delta}_t$ is zero when data collection is done perfectly and able to measure the variables of interest accurately.

First, we discuss the time series in population. Suppose that we know the true data generating process of $\boldsymbol{b}_t$, which has an additive state-space structure. We use this framework illustratively and we are not restricted to it. Nonetheless, the state-space modelling framework is very flexible and encompasses many popular forecasting model families. Let:

$$\boldsymbol{b}_t = \boldsymbol{\mu}_{b,t} + \boldsymbol{\eta}_{b,t}, \tag{2}$$

where $\boldsymbol{\mu}_{b,t}$ denotes the structure of the time series and $\boldsymbol{\eta}_{b,t}$ is the innovation term at period

t, which for simplicity follows a multivariate normal distribution with zero mean and has a covariance matrix of $\boldsymbol{\Sigma}_b$. We aggregate $\boldsymbol{b}_t$ by multiplying with $\boldsymbol{S}$ and from Eq (2), and we get $\boldsymbol{y}_t$ as a vector time series,

$$\begin{aligned}\boldsymbol{y}_t &= \boldsymbol{S}\boldsymbol{b}_t + \boldsymbol{\delta}_t \\ &= \boldsymbol{S}\boldsymbol{\mu}_{b,t} + \boldsymbol{S}\boldsymbol{\eta}_{b,t} + \boldsymbol{\delta}_t \\ &= \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t \end{aligned} \tag{3}$$

where $\boldsymbol{\varepsilon}_t$ is the total residual of the process, which consists of the aggregated innovations and the statistical discrepancy, denoted as $\boldsymbol{\varepsilon}_t = \boldsymbol{S}\boldsymbol{\eta}_{b,t} + \boldsymbol{\delta}_t$, and $\boldsymbol{\mu}_t = \boldsymbol{S}\boldsymbol{\mu}_{b,t}$. In this case, we assume that $\mathrm{E}(\boldsymbol{\delta}_t|\boldsymbol{\mathcal{I}}_t) = \mathbf{0}$, and from definition $\mathrm{E}(\boldsymbol{\eta}_{b,t}|\boldsymbol{\mathcal{I}}_t) = \mathbf{0}$, thus $\mathrm{E}(\boldsymbol{S}\boldsymbol{\eta}_{b,t}|\boldsymbol{\mathcal{I}}_t) = \mathbf{0}$. In expectation, Eq (3) becomes,

$$\begin{aligned}\mathrm{E}(\boldsymbol{y}_t|\boldsymbol{\mathcal{I}}_t) &= \mathrm{E}(\boldsymbol{S}\boldsymbol{b}_t + \boldsymbol{\delta}_t|\boldsymbol{\mathcal{I}}_t) \\ &= \mathrm{E}(\boldsymbol{S}\boldsymbol{\mu}_{b,t}|\boldsymbol{\mathcal{I}}_t) + \mathrm{E}(\boldsymbol{S}\boldsymbol{\eta}_{b,t}|\boldsymbol{\mathcal{I}}_t) + \mathrm{E}(\boldsymbol{\delta}_t|\boldsymbol{\mathcal{I}}_t) \\ &= \mathrm{E}(\boldsymbol{\mu}_t|\boldsymbol{\mathcal{I}}_t)\end{aligned}$$

where $\boldsymbol{\mathcal{I}}_t$ is the available information at t. We can also infer that $\boldsymbol{\mu}_t = \boldsymbol{S}\boldsymbol{\mu}_{b,t}$ and this also holds at period $t + h$. This shows that the time series is coherent in expectations, meaning that the linear hierarchical structure, $\boldsymbol{S}$, guarantees coherency in the structures of the time series, but does not necessarily guarantee coherency in the residuals.

In observations, we exploit $\boldsymbol{\mathcal{I}}_t$ by differentiating between the type of the information, namely $\boldsymbol{\theta}$ as a set of forecasting models in the hierarchy, and $\boldsymbol{Y}_t$ as the available hierarchical time series, where $\boldsymbol{Y}_t = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t\}$. Note that $\boldsymbol{\theta}$ is not restricted to a single family of forecasting models and can be different forecasting models or methods for each series across the hierarchy.

Using forecasting models $\boldsymbol{\theta}$, we produce h-step ahead base forecasts. The forecasts, typically, adhere to the classical coherency, but are inaccurate. Following the hierarchical forecasting literature, we can reconcile base forecast as:

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t}, \tag{4}$$

where $\tilde{\boldsymbol{y}}_{t+h|t}$ is h-step ahead reconciled forecast, and $\boldsymbol{G}$ is a reconciliation weights matrix, which combines all forecasts across the hierarchy to create adjusted bottom-level forecasts. As $\boldsymbol{S}$

and $\hat{\boldsymbol{y}}_{t+h|t}$ are available prior to the reconciliation, $\boldsymbol{G}$ is estimated. Wickramasuriya et al. (2019) propose the MinT Reconciliation to obtain $\boldsymbol{G}$, by minimising the trace of covariance matrix of the reconciled forecast error ($\tilde{\boldsymbol{e}}_{t+h|t} = \boldsymbol{y}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t}$), instead of reconciliation error ($\boldsymbol{\epsilon}_{t+h|t} = \hat{\boldsymbol{y}}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t}$):

$$\min \mathrm{Tr}\left(\boldsymbol{S}\boldsymbol{G}\boldsymbol{W}_{t+h|t}\boldsymbol{G}^\top\boldsymbol{S}^\top\right)$$

subject to $\boldsymbol{S}\boldsymbol{G}\boldsymbol{S} = \boldsymbol{S}$, or alternatively $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$, where $\boldsymbol{W}_{t+h|t} = \mathrm{E}(\hat{\boldsymbol{e}}_{t+h|t}\hat{\boldsymbol{e}}_{t+h|t}^\top|\boldsymbol{\mathcal{I}}_t)$ and $\hat{\boldsymbol{e}}_{t+h|t}$ is the h-step ahead base forecast error, $\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}$. They show that forecasts are unbiasedly coherent when the unbiasedness constraint, or $\boldsymbol{S}\boldsymbol{G}\boldsymbol{S} = \boldsymbol{S}$, holds and also implies that $\boldsymbol{S}\boldsymbol{G}$ is a projection matrix. Thus, the optimal reconciliation weights matrix is formulated as:

$$\boldsymbol{G} = (\boldsymbol{S}^\top\boldsymbol{W}_{t+h|t}^{-1}\boldsymbol{S})^\top\boldsymbol{S}^\top\boldsymbol{W}_{t+h|t}^{-1}. \tag{5}$$

Eq (5) shows that $\boldsymbol{G}$ is valid under a set of forecasting models $\boldsymbol{\theta}$ and depends on the expected value of the h-step ahead base forecast error covariance matrix, which contains the uncertainties from the corresponding forecasting models. In a limited sample, $\hat{\boldsymbol{W}}_{t+h|t}$, is constructed from the estimated parameters of the forecasting models $\boldsymbol{\theta}$ and the one-step ahead base forecast error covariance matrix, $\hat{\boldsymbol{W}}_{t+1|t}$. Being estimated, $\hat{\boldsymbol{W}}_{t+h|t}$ is uncertain due to modelling uncertainty, and this influences the uncertainty of $\boldsymbol{G}$. In the observational level where the sample size is limited, we denote the estimated reconciliation weights matrix as $\hat{\boldsymbol{G}}$, thus $\boldsymbol{S}\hat{\boldsymbol{G}}\boldsymbol{S} = \boldsymbol{S}$ is subject to uncertainty and the coherency constraint depends on how we utilise the available information, given a limited sample. To avoid confusion we clarify the notion here: $\boldsymbol{G}$ refers to the weights matrix of the conventional coherency from the literature. Here, we use $\hat{\boldsymbol{G}}$ to highlight that $\hat{\boldsymbol{G}}$ is estimated. In our stochastic coherency framework $\boldsymbol{G}$ and $\hat{\boldsymbol{G}}$ are coincide. We also introduce $\boldsymbol{\Gamma}$ that is the weights matrix in population.

The expectation of the reconciled forecasts conditional to $\boldsymbol{\mathcal{I}}_t$ is:

$$\begin{aligned}
\mathrm{E}(\tilde{\boldsymbol{y}}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) &= \mathrm{E}(\boldsymbol{S}\hat{\boldsymbol{G}}\hat{\boldsymbol{y}}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) \\
&= \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{\mu}_{t+h|t} \\
&= \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S}\boldsymbol{\mu}_{b,t+h|t},
\end{aligned}$$

where $\mathrm{E}(\hat{\boldsymbol{y}}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) = \boldsymbol{\mu}_{t+h|t} = \boldsymbol{S}\boldsymbol{\mu}_{b,t+h|t}$, and $\boldsymbol{\Gamma} = \mathrm{E}(\hat{\boldsymbol{G}}|\boldsymbol{\mathcal{I}}_t)$. Coherency needs the unbiasedness property to ensure that the forecasts are coherent via distributing information across the hierarchy through linear combination. The multiplication between $\boldsymbol{S}$ and $\boldsymbol{\Gamma}$ results in a projection which maintains coherency with regard to any overlooked errors from forecasting models, such as the estimation errors or any statistical discrepancies.

As $\boldsymbol{S\Gamma}$ and $\boldsymbol{S\hat{G}}$ are both projection matrices, this property should be maintained. For example, we maintain the projection matrix to be idempotent. This should hold in both the population and the estimation level, $\boldsymbol{S\Gamma S\Gamma} = \boldsymbol{S\Gamma}$ and $\boldsymbol{S\hat{G}S\hat{G}} = \boldsymbol{S\hat{G}}$. The linear projection, which maintains unbiasedness and coherency, basically ensures the projected forecasts lie on the coherent subspace (Panagiotelis et al., 2021). The issue now is how uncertain the estimated projection matrix is. In the case of $\boldsymbol{S\Gamma}$, it projects to $\boldsymbol{\mu}_{t+h|t}$, whereas $\boldsymbol{S\hat{G}}$ may project the forecasts a bit further from $\boldsymbol{\mu}_{t+h|t}$. Thus, the uncertainty in the projection highlights the importance of modelling uncertainty, since the former originates from the latter. Therefore, we redefine coherency by treating the coherent projection matrix, $\boldsymbol{S\Gamma S} = \boldsymbol{S}$ to mitigate overlooked errors from the forecasting models in forecast reconciliation. We call it **stochastic coherency**.

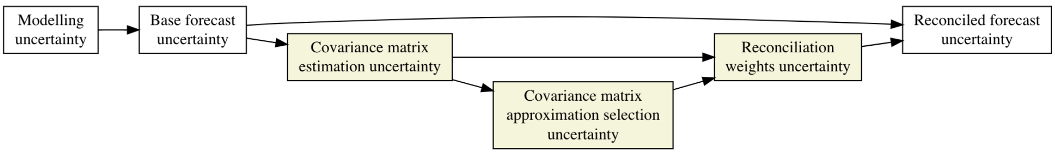


Figure 1: Forecast reconciliation uncertainty framework. Boxes in beige depict our novel understanding in forecast reconciliation uncertainty framework.

Figure 1 summarises the view of the uncertainties in forecast reconciliation we gain from stochastic coherency. Modelling uncertainty leads to the uncertainty in the base forecasts, as in conventional forecasting (Chatfield, 1995). This contributes to the uncertainty of reconciled forecast, which is well understood in the hierarchical forecasting literature (for example, Athanasopoulos et al., 2017). With stochastic coherency we demonstrated that there are additional sources of uncertainty, that can help explain the observations in the literature (Panagiotelis et al., 2021). There is uncertainty in the covariance matrix approximation, which is naturally connected to the uncertainty of the base forecasts. This additional uncertainty is both due the estimation and selection of an appropriate covariance matrix approximation method. Both contribute to the uncertainty of the reconciliation weights, which adds to total uncertainty of the reconciled forecasts.

As modelling uncertainty plays an important role in forecast reconciliation, i.e. how we exploit $\boldsymbol{\mathcal{I}}_t$, we discuss the effect of model specification on the reconciliation. We illustrate the effects by focusing on the case when the forecasts are unbiased. Then, we move to two special cases, namely reconciling biased forecasts and reconciling forecasts from perfectly specified forecasting models. We demonstrate how modelling uncertainty, as in the structure of the models and the parameter estimation, affects forecast reconciliation.

3. Reconciling Unbiased Forecasts with Stochastic Coherency

In this scenario, we consider well-specified forecasting models. The forecasting models are able to capture the structure of the data generating process well, but suffer from parameter estimation uncertainty. Given the limited sample size in $\boldsymbol{Y}_t$, we produce h-step ahead base forecasts, $\hat{\boldsymbol{y}}_{t+h|t}$, and we expect that $\mathrm{E}(\hat{\boldsymbol{y}}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) = \boldsymbol{\mu}_{t+h|t}$. Thus, the uncertainty due to parameter estimation is $\hat{\boldsymbol{y}}_{t+h|t} - \mathrm{E}(\hat{\boldsymbol{y}}_{t+h|t}) = \boldsymbol{v}_{t+h|t}$, where $\mathrm{E}(\boldsymbol{v}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) = \boldsymbol{0}$. Note that the irreducible forecast error at period $t+h$ is defined as $\boldsymbol{\zeta}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h|t}$. This differs from $\boldsymbol{\varepsilon}_{t+h}$ as the latter is unconditional.

As the base forecasts are $\hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{\mu}_{t+h|t} + \boldsymbol{v}_{t+h|t}$, the base forecast errors become:

$$\begin{aligned}\hat{\boldsymbol{e}}_{t+h|t} &= \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t} \\ &= \boldsymbol{\mu}_{t+h|t} + \boldsymbol{\zeta}_{t+h|t} - \boldsymbol{\mu}_{t+h|t} - \boldsymbol{v}_{t+h|t} \\ &= \boldsymbol{\zeta}_{t+h|t} - \boldsymbol{v}_{t+h|t}. \end{aligned} \tag{6}$$

From Eq (6), we can see that the base forecast errors consist of the irreducible error and the error due to parameter estimation. The latter is affected by the sample sizes.

We aim to reconcile the base forecasts with regard to the hierarchical structure, using $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{G}}\hat{\boldsymbol{y}}_{t+h|t}$. To estimate $\hat{\boldsymbol{G}}$, we need to estimate the h-step ahead base forecast error covariance matrix,

$$\hat{\boldsymbol{W}}_{t+h|t} = \boldsymbol{Z}_{t+h|t} + \boldsymbol{V}_{t+h|t} + \boldsymbol{C}_{t+h|t},$$

where $\boldsymbol{Z}_{t+h|t}$ is the covariance matrix of $\boldsymbol{\zeta}_{t+h|t}$ and $\boldsymbol{V}_{t+h|t}$ is the covariance matrix of $\boldsymbol{v}_{t+h|t}$, where $\mathrm{E}(\boldsymbol{Z}_{t+h|t}) = \boldsymbol{\Sigma}$ and $\mathrm{E}(\boldsymbol{V}_{t+h|t}) = \boldsymbol{V}$, and $\boldsymbol{C}_{t+h|t}$ is the covariance matrix between $\boldsymbol{\zeta}_{t+h|t}$ and $\boldsymbol{v}_{t+h|t}$. Therefore,

$$\hat{\boldsymbol{G}} = (\boldsymbol{S}^\top(\hat{\boldsymbol{W}}_{t+h|t})^{-1}\boldsymbol{S})^\top\boldsymbol{S}^\top(\hat{\boldsymbol{W}}_{t+h|t})^{-1}. \tag{7}$$

Looking at Eq (7), $\hat{\boldsymbol{G}}$ is uncertain, because the variances and the covariances of $\hat{\boldsymbol{W}}_{t+h|t}$

depend on the parameter estimation uncertainty, given a limited sample size. We can see that the uncertainty in forecasting models is transferred to $\hat{\boldsymbol{G}}$, which will affect the reconciled forecast errors. The reconciliation weights matrix may not be able to improve the base forecast accuracy due to this uncertainty.

From Eq (7), we can produce the reconciled forecasts, $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{G}}\hat{\boldsymbol{y}}_{t+h|t}$, and decompose the reconciled forecast error:

$$
\begin{aligned}
\tilde{\boldsymbol{e}}_{t+h|t} &= \boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t} \\
&= \boldsymbol{y}_{t+h} - \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{\mu}_{t+h|t} + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{\mu}_{t+h|t} - \boldsymbol{S}\hat{\boldsymbol{G}}\hat{\boldsymbol{y}}_{t+h|t} \\
&= \boldsymbol{y}_{t+h} - \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{\mu}_{t+h|t} + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{\mu}_{t+h|t} - \boldsymbol{S}\hat{\boldsymbol{G}}(\boldsymbol{\mu}_{t+h|t} + \boldsymbol{v}_{t+h|t}) \\
&= \underbrace{\boldsymbol{\zeta}_{t+h|t}}_{\text{irreducible error}} + \underbrace{(\boldsymbol{S}\boldsymbol{\Gamma} - \boldsymbol{S}\hat{\boldsymbol{G}})\boldsymbol{\mu}_{t+h|t}}_{\substack{\text{reconciliation matrix} \\ \text{estimation error}}} + \underbrace{(-\boldsymbol{S}\hat{\boldsymbol{G}}\boldsymbol{v}_{t+h|t})}_{\substack{\text{reconciled} \\ \text{estimation error}}}
\end{aligned} \tag{8}
$$

where $\boldsymbol{\zeta}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h|t}$, and $\boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{\mu}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$ as $\boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S} = \boldsymbol{S}$ and $\boldsymbol{\mu}_{t+h|t} = \boldsymbol{S}\boldsymbol{\mu}_{b,t+h|t}$. Eq (8) shows that the reconciled forecast error consists of the irreducible error, the reconciliation matrix estimation error, and the reconciled estimation error, which will affect the uncertainty of reconciled forecast error variance.

Looking at the relations between different forecast errors in forecast reconciliation, Panagiotelis et al. (2020) and Panagiotelis et al. (2021) use generalised Pythagoras theorem to establish their relationships. We argue that it needs a relaxation to accommodate the uncertainty by using triangular inequality, where the relationship is shown as,

$$
\begin{aligned}
||\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}||^2 &\leq ||\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t}||^2 + ||\hat{\boldsymbol{y}}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t}||^2, \\
\text{SSE}_{base} &\leq \text{SSE}_{recon} + \text{SSE}_{\boldsymbol{\epsilon}},
\end{aligned} \tag{9}
$$

where $\text{SSE}_{\boldsymbol{\epsilon}}$ is the sum squared reconciliation error and $\text{SSE}_{\boldsymbol{\epsilon}} = ||\hat{\boldsymbol{y}}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t}||^2 \geq 0$. If the left hand side of Eq (9) is equal to the right hand side, then $||\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}||^2 \geq ||\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t}||^2$. However, Eq (9) demonstrates that SSE_{recon} may exceed SSE_{base} as a result of the overall uncertainty in forecast reconciliation. Given the case of unbiased forecasts, we discuss two special cases when the forecasting models are mis-specified and perfectly specified.

3.1. Special Case I: Mis-specified Forecasting Models

Due to unknown data generating processes, it is possible to obtain mis-specified models, denoted by †, i.e. adding a redundant variable, wrong transformation, or omitted variables. In the case of mis-specified forecasting models, we produce biased h-step ahead base forecasts, where $\hat{\boldsymbol{y}}^{\dagger}_{t+h|t} = \hat{\boldsymbol{y}}_{t+h|t} + \boldsymbol{o}^{\dagger}_{t+h|t}$ and $\mathrm{E}(\boldsymbol{o}^{\dagger}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) = \boldsymbol{o}^{\dagger}$, which may be nonzero. The base forecast error is shown as,

$$\begin{aligned}
\hat{\boldsymbol{e}}^{\dagger}_{t+h|t} &= \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}^{\dagger}_{t+h|t} \\
&= \boldsymbol{\mu}_{t+h|t} + \boldsymbol{\zeta}_{t+h|t} - \boldsymbol{\mu}_{t+h|t} - \boldsymbol{v}_{t+h|t} - \boldsymbol{o}^{\dagger}_{t+h|t} \\
&= \boldsymbol{\zeta}_{t+h|t} - \boldsymbol{v}_{t+h|t} - \boldsymbol{o}^{\dagger}_{t+h|t}. \qquad (10)
\end{aligned}$$

Consequently, the h-step ahead biased base forecast error covariance matrix can be constructed as:

$$\hat{\boldsymbol{W}}^{\dagger}_{t+h|t} = \boldsymbol{Z}_{t+h|t} + \boldsymbol{V}_{t+h|t} + \boldsymbol{O}^{\dagger}_{t+h|t} + \boldsymbol{C}^{\dagger}_{t+h|t},$$

where $\boldsymbol{O}^{\dagger}_{t+h|t}$ is the estimated covariance matrix of $\boldsymbol{o}^{\dagger}_{t+h|t}$ and $\mathrm{E}(\boldsymbol{O}^{\dagger}_{t+h|t}) = \boldsymbol{O}^{\dagger}$. In this case, $\boldsymbol{C}^{\dagger}_{t+h|t}$ collects all covariances between $\boldsymbol{\zeta}_{t+h|t}$, $\boldsymbol{v}_{t+h|t}$, and $\boldsymbol{o}^{\dagger}_{t+h|t}$. Hence, we can calculate $\hat{\boldsymbol{G}}$ in the case of mis-specified models, or $\hat{\boldsymbol{G}}^{\dagger}$, such as,

$$\hat{\boldsymbol{G}}^{\dagger} = (\boldsymbol{S}^{\top}(\hat{\boldsymbol{W}}^{\dagger}_{t+h|t})^{-1}\boldsymbol{S})^{\top}\boldsymbol{S}^{\top}(\hat{\boldsymbol{W}}^{\dagger}_{t+h|t})^{-1}. \qquad (11)$$

Note that Eq (5) is obtained from the assumption of unbiased base forecasts. However, we aim to show that we are still able to reconcile the forecasts, even if the forecasts are biased, but it will come at a cost of more variability due to an additional element in the modelling uncertainty.

Using Eq (11), we construct the reconciled forecasts as $\tilde{\boldsymbol{y}}^{\dagger}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{G}}^{\dagger}\hat{\boldsymbol{y}}^{\dagger}_{t+h|t}$ and the reconciled forecast error is shown as,

$$\begin{aligned}
\tilde{\boldsymbol{e}}^{\dagger}_{t+h|t} &= \boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}^{\dagger}_{t+h|t} \\
&= \boldsymbol{y}_{t+h} - \boldsymbol{S\Gamma}(\boldsymbol{\mu}_{t+h|t} + \boldsymbol{o}^{\dagger}_{t+h|t}) + \boldsymbol{S\Gamma}(\boldsymbol{\mu}_{t+h|t} + \boldsymbol{o}^{\dagger}_{t+h|t}) - \boldsymbol{S\hat{G}}^{\dagger}\hat{\boldsymbol{y}}^{\dagger}_{t+h|t} \\
&= \boldsymbol{y}_{t+h} - \boldsymbol{S\Gamma}\boldsymbol{\mu}_{t+h|t} + \boldsymbol{S\Gamma}\boldsymbol{\mu}_{t+h|t} - \boldsymbol{S\hat{G}}^{\dagger}(\boldsymbol{\mu}_{t+h|t} + \boldsymbol{v}_{t+h|t} + \boldsymbol{o}^{\dagger}_{t+h|t}) \\
&= \underbrace{\boldsymbol{\zeta}_{t+h|t}}_{\substack{\text{irreducible}\\ \text{error}}} + \underbrace{(\boldsymbol{S\Gamma} - \boldsymbol{S\hat{G}})\boldsymbol{\mu}_{t+h|t}}_{\substack{\text{reconciliation matrix}\\ \text{estimation error}}} + \underbrace{(-\boldsymbol{S\hat{G}}^{\dagger}\boldsymbol{v}_{t+h|t})}_{\substack{\text{reconciled}\\ \text{estimation error}}} + \underbrace{(-\boldsymbol{S\hat{G}}^{\dagger}\boldsymbol{o}^{\dagger}_{t+h|t})}_{\substack{\text{reconciled}\\ \text{bias error}}},
\end{aligned} \tag{12}$$

where $\boldsymbol{\zeta}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h|t}$, and $\boldsymbol{S\Gamma o}^{\dagger}_{t+h|t}$ cancels out and similar to Eq (8) $\boldsymbol{S\Gamma\mu}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$ as $\boldsymbol{S\Gamma S} = \boldsymbol{S}$ and $\boldsymbol{\mu}_{t+h|t} = \boldsymbol{S\mu}_{b,t+h|t}$. Hence, Eq (12) shows that the reconciled forecast error from biased unreconciled forecast consists of the irreducible error, the reconciliation matrix estimation error, the reconciled estimation error, and the reconciled bias error. This additional error affects the uncertainty of the sum squared reconciled forecast error.

3.2. Special Case II: Perfectly-Specified Models

Suppose we were able to produce forecasts from perfectly-specified forecasting models, where the parameters and the data generating process are known. The h-step ahead base forecasts will match with the structure of the hierarchical time series in expectations and in the observational level, shown as $\hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$. Hence, the h-step ahead base forecast error is the irreducible error, shown as $\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h|t} = \boldsymbol{\zeta}_{t+h|t}$.

Suppose we aim to reconcile the base forecasts, the reconciled forecasts are shown as,

$$\begin{aligned}
\tilde{\boldsymbol{y}}_{t+h|t} &= \boldsymbol{S\hat{G}}\boldsymbol{\mu}_{t+h|t} \\
&= \boldsymbol{S\hat{G}S}\boldsymbol{\mu}_{b,t+h|t} \\
&= \boldsymbol{\mu}_{t+h|t}
\end{aligned} \tag{13}$$

where $\boldsymbol{\mu}_{t+h|t} = \boldsymbol{S\mu}_{b,t+h|t}$ and $\boldsymbol{S\Gamma S} = \boldsymbol{S}$. In this case, if the structure and the parameters are known, $\boldsymbol{S\hat{G}} = \boldsymbol{S\Gamma}$, and $\hat{\boldsymbol{G}}$ becomes irrelevant because the forecasts are coherent already. Following Eq (9), since the forecast errors between both forecasts are the same, then $||\hat{\boldsymbol{y}}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t}||^2 = 0$. Consequently, $||\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}||^2 = ||\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t}||^2$, as $\tilde{\boldsymbol{y}}_{t+h|t} = \hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$. In perfetly specified models, MinT Reconciliation does not improve or worsen the forecast accuracy as the models are able to produce coherent structures of the time series, $\boldsymbol{\mu}_{t+h|t}$. This is in agreement with Athanasopoulos et al. (2017).

3.3. Uncertainty in $\boldsymbol{G}$

The previous discussion shows that the accuracy improvement due to forecast reconciliation depends on the quality of the estimated projection, $\boldsymbol{S}\hat{\boldsymbol{G}}$. Since $\hat{\boldsymbol{G}}$ is a function of $\hat{\boldsymbol{W}}_{t+h|t}$ and $\hat{\boldsymbol{W}}_{t+h|t}$ depends on the model specification, $\hat{\boldsymbol{G}}$ is stochastic. With regard to Eq (4), we can say that the reconciled forecasts are the result of a linear combination of all base forecasts, in which the weights are stochastic.

In order to deal with uncertain weights in $\hat{\boldsymbol{G}}$, we draw on the arguments from linear forecast combination literature by Smith and Wallis (2009) and Claeskens et al. (2016). As $\hat{\boldsymbol{G}}$ contains the estimated weights, Smith and Wallis (2009) and Claeskens et al. (2016) note that estimated combination increases the variance of the combined forecasts. Furthermore, in forecast pooling, for any forecast added in the combination to be beneficial there are conditions on the forecast variance (Kourentzes et al., 2019). In order to manage the uncertainty in forecast reconciliation, it could be possible that not all parts of $\boldsymbol{S}$ are equally informative, i.e. these may increase the uncertainty of $\hat{\boldsymbol{G}}$. This may explain the marginal improvements observed with cross-temporal hierarchies, but more importantly it suggests that $\hat{\boldsymbol{G}}$ could be restricted further.

A restricted $\hat{\boldsymbol{G}}$ can be achieved by controlling the information which enters the forecast reconciliation via $\hat{\boldsymbol{W}}_{t+h|t}$. Suppose that $\hat{\boldsymbol{W}}_{t+h|t}$ is assumed to be a fixed covariance matrix, e.g. an identity matrix, then the weights in $\hat{\boldsymbol{G}}$ are fixed and constructed from $\boldsymbol{S}$ only. Alternatively, we can include the sample variances and the covariances of the base forecast errors, but the level of the randomness on the weights in $\hat{\boldsymbol{G}}$ are subject to the uncertainty from the forecasting models. A balance between these is to use the sample variances and manage the off-diagonal elements, for example by shrinking the covariances or restricting them to zero. This may enable us to balance the trade-off between more information and reducing uncertainty of the weights in $\hat{\boldsymbol{G}}$. We note here that stochastic coherency is a general concept which can be applied to any covariance matrix in forecast reconciliation. Fixed weights in $\hat{\boldsymbol{G}}$ due to the identity covariance matrix, or OLS reconciliation, is seen as a means to limit the uncertainty of $\hat{\boldsymbol{G}}$ to zero. This way, we can restrict the uncertainty propagation from the forecasting models to the reconciled forecast uncertainty.

Stochastic coherency acknowledges two potential sources of uncertainty in forecast reconciliation, originating from the modelling or the collection of data. We demonstrate that the main source of uncertainty originates from the forecasting models. The uncertainty in the forecasting model propagates to the estimation of the reconciliation weights matrix via the covariance matrix of the h-step ahead base forecast error and this affects the uncertainty of the reconciled forecasts.

The difference between stochastic and deterministic coherency is not in the reconciled point forecasts, but rather in the variability of the error distribution. Our stochastic interpretation demonstrates that there is an increased uncertainty in the error distribution. In the following section we use simulated and real data to show that our theoretical discussion of stochastic coherency is observable in practice using the widely used MinT Reconciliation and can help explain results in the literature.

4. Simulation Study

4.1. Experimental Design

In this section we perform two simulations: first, with a small hierarchy, controlling for the model uncertainty, so as to validate the theoretical discussion above; second, with a large hierarchy, as to see the effect of the hierarchy size.

We specify the data generating process of each bottom-level time series as an AR(1) process for the small hierarchy:

$$b_{q,t} = 0.4y_{q,t-1} + \varepsilon_{q,t},$$

where q is an index from 1 to 4, denoting the bottom level time series in the hierarchy. The innovation term $\varepsilon_{q,t} = \{\varepsilon_{1,t}, \varepsilon_{2,t}, \varepsilon_{3,t}, \varepsilon_{4,t}\}$ and $\boldsymbol{\varepsilon}_{b,t} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon_b})$, where $\boldsymbol{\varepsilon}_{b,t} = \begin{bmatrix} \varepsilon_{1,t} & \varepsilon_{2,t} & \varepsilon_{3,t} & \varepsilon_{4,t} \end{bmatrix}^\top$,

$$\boldsymbol{\Sigma}_{\varepsilon_b} = \begin{bmatrix} 3 & 2 & 1 & 1 \\ 2 & 3 & 1 & 1 \\ 1 & 1 & 3 & 2 \\ 1 & 1 & 2 & 3 \end{bmatrix}, \text{ and } \quad \boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \boldsymbol{I}_4 & & \end{bmatrix}.$$

The top and middle level series result from the aggregation of the bottom-level, as presented by $\boldsymbol{S}$, where $\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t$ and $\boldsymbol{b}_t = \{b_{q,t}\}$. For example, $y_{Top,t} = 0.4(b_{1,t-1} + b_{2,t-1} + b_{3,t-1} + b_{4,t-1}) + \varepsilon_{1,t} + \varepsilon_{2,t} + \varepsilon_{3,t} + \varepsilon_{4,t}$. We simulate this setting with sample sizes of 24, 120, and 240 and a burn-in period of 200, to eliminate any initialisation issues.

For the large hierarchy we use 50 bottom-level series, two levels in the middle-level and a top-level time series. All bottom-level time series are generated from ARIMA with $\boldsymbol{\varepsilon}_{b,50} \sim$

$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon_{b,50}})$ and $\mathbf{\Sigma}_{\varepsilon_{b,50}}$ is generated randomly at each iteration of the simulation. In both simulations, we assume that $\boldsymbol{\delta}_t = \mathbf{0}$. For ARIMA we allow randomness in the data generating process, i.e., the AR and MA orders are sampled from 0 to 3 and the integration is from 0 to 1. We simulate the same sample sizes as for the small hierarchy with the same burn-in setting.

4.1.1. Forecasting Models

For the small hierarchy, we generate individual base forecasts using different model specification settings, summarised in Table 1. The first option, referred to as *DGP*, assumes that we know the process fully. The second option assumes the model structure is known, but the model is subject to parameter uncertainty. We call this *AR(1)*. The third option employs ARIMA with automatic model selection, named as *AutoARIMA*, which has potentially reduced model uncertainty and parameter uncertainty, as the data generating process is encompassed. The fourth option uses exponential smoothing and represents a mis-specified model by using *ETS(AAN)* that is equivalent to ARIMA(0,2,2), introducing superfluous terms. For the large hierarchy, we use ARIMA and exponential smoothing with automatic selection. These match the latter two options in Table 1. We produce 1- to 6-step ahead base forecasts for both hierarchies. For each combination of sample sizes, forecast horizon, and model specification scenario, we repeat the simulations 1000 times. ARIMA and ETS models are implemented using the auto.arima() in the forecast package (Hyndman et al., 2019) and the es() in the smooth package (Svetunkov, 2019) for R (R Core Team, 2018), and we rely on Akaike Information Criterion for selecting the appropriate model form.

Models	DGP	AR(1)	AutoARIMA	ETS
Specification	Known	Known	Approximated	Wrong
Parameter	Known	Estimated	Estimated	Estimated

Table 1: Model specification for each scenario in the experimental design

4.1.2. Forecast Reconciliation

We reconcile the base forecasts using the MinT Reconciliation methodology. We use several approximation methods for $\hat{\boldsymbol{W}}_{t+h|t}$ from the literature, summarised in Table 2. Hyndman et al. (2011) use a diagonal covariance matrix with equivariant variances, they call this method *OLS*. Athanasopoulos et al. (2017) propose Structural Scaling (*SCL*), where they set equal variances to the bottom-level, and then calculate the covariance matrix as $\boldsymbol{S}\boldsymbol{\sigma}_b$. In this case, $\boldsymbol{\sigma}_b = c\boldsymbol{I}_m$, where c is a scalar and m is the number of the bottom-level time series. Hyndman et al. (2016) propose

WLS, which uses a diagonal covariance matrix allowing for heterogeneity. Wickramasuriya et al. (2019) propose MinT-Sample, a fully unrestricted estimated covariance matrix of one-step ahead in-sample base forecast errors. This method is denoted here as *EMP*. However, as it is difficult to estimate the off-diagonals, they implement shrinkage on the off-diagonals towards zero by Schäfer and Strimmer (2005), called MinT-Shrink method. This is denoted *SHR* in Table 2.

Estimation	OLS	SS	WLS	SHR	EMP
Approximation	$c\boldsymbol{I}$	$\boldsymbol{S}\boldsymbol{\sigma}_b$	$\hat{\boldsymbol{W}}_{d,t+1\|t}$	$\hat{\boldsymbol{W}}^{\text{SHR}}_{t+1\|t}$	$\hat{\boldsymbol{W}}_{t+1\|t}$
$\hat{\boldsymbol{G}}$	$\hat{\boldsymbol{G}}_{\text{OLS}}$	$\hat{\boldsymbol{G}}_{\text{SCL}}$	$\hat{\boldsymbol{G}}_{\text{WLS}}$	$\hat{\boldsymbol{G}}_{\text{SHR}}$	$\hat{\boldsymbol{G}}_{\text{EMP}}$

Table 2: Different approximations of $\boldsymbol{W}_{t+h|t}$

Apart from the established covariance matrix approximations, we explore three alternative covariance matrices, motivated by our theoretical discussion. Our motivation is to either construct them from the bottom-level or by ignoring some of the off-diagonals in order to mitigate the uncertainty, instead of estimating the whole covariance matrix. A similar study was done by Nystrup et al. (2020) who exploited autocorrelations between time series in temporal hierarchies. Furthermore, we provide a covariance matrix approximation continuum.

Figure 2 illustrates the covariance matrix approximations. In the first method we collect a vector of the bottom-level variances from one-step ahead in-sample forecast errors, $\hat{\boldsymbol{\sigma}}_b$, and construct the covariance matrix with the variances of $\boldsymbol{S}\hat{\boldsymbol{\sigma}}_b$. We call it *cWLS*. Second, we estimate the bottom level covariance matrix and construct it according to the hierarchy. We force a block diagonal structure, making other elements zero and shrinking the remaining, named *bShrink*. Third, we estimate MinT-Shrink and retain the bottom-level covariance matrix. Then we aggregate it to the hierarchy and force covariances between the bottom-level and the upper-levels to be zero. We call this *pShrink*. We sacrifice information utilisation on bShrink and pShrink by forcing hierarchically block diagonals to mitigate the variability of the forecast error.

Figure 3 depicts the covariance matrix approximation continuum. It represents the utilisation of information with regards to the forecast error variability. On the left side of the continuum, OLS provides the least information as $\hat{\boldsymbol{G}}$ is constructed from the hierarchy only. However, it produces the least variable forecast error. Beside OLS, there are SCL and WLS. They provide more information than OLS by allowing heteroscedasticity. Consequently, they produce more variable forecast error than OLS.

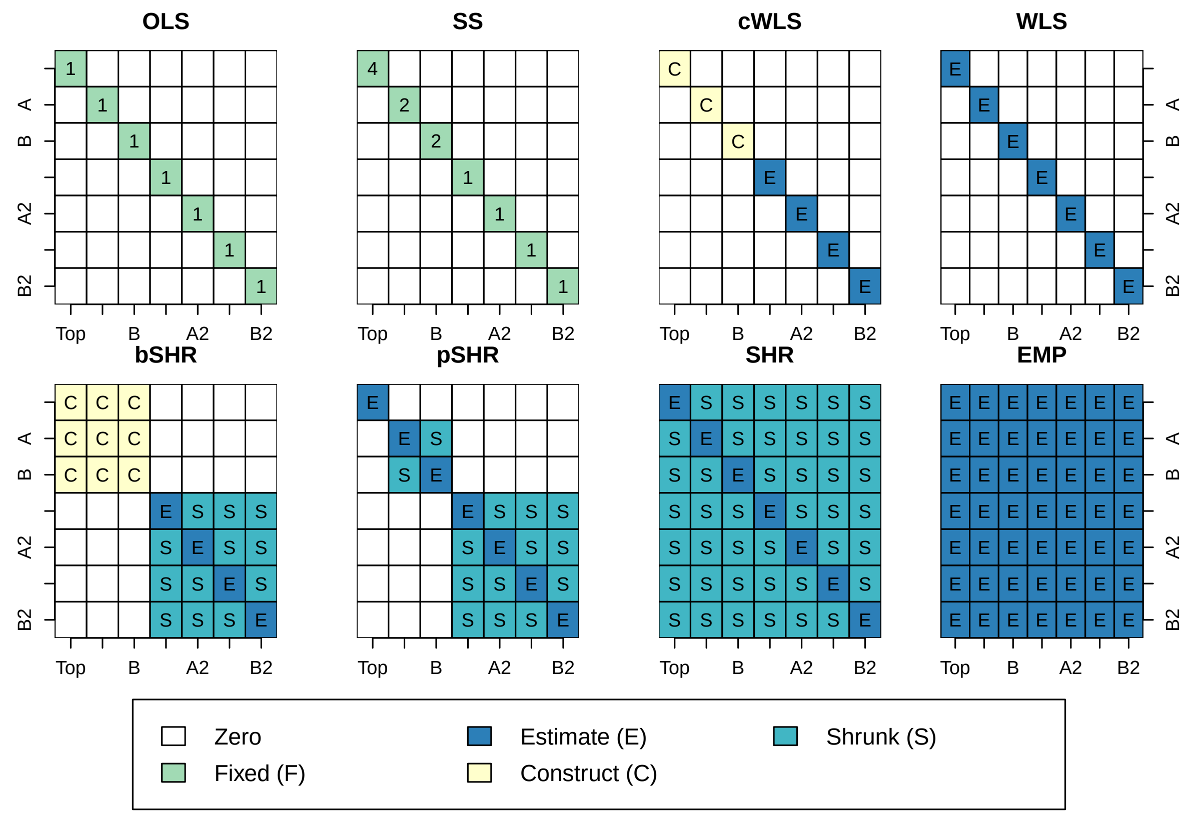


Figure 2: Illustration of covariance matrix approximations for a hierarchy of seven time series. It consists of four bottom-level, two middle-level, and a top-level time series. S on the upper-level of pShrink is constructed, then shrunk. All covariance matrix approximations are positive definite, except bSHR and EMP. bSHR is a positive semi-definite covariance matrix and the positive definiteness of EMP here depends on the sample (Hyndman et al., 2011)

On the right hand side of the continuum, EMP provides full information as we estimate the unrestricted covariance matrix. Consequently, EMP will produce the most variable forecast error. Next to EMP, SHR provides full information with some restrictions, thus produces less variable forecast error than EMP.

Our alternative covariance matrices fill the gap between WLS and SHR. We sacrifice some correlations to manage the variability. We retain the correlations between the parent nodes and the children nodes, but we dismiss the correlations between the parent and the children from different parent nodes, and vice versa. By constructing the covariance matrix from the bottom-level information, it is expected to produce less variable forecast errors, but have a similar performance with SHR in terms of forecast accuracy.

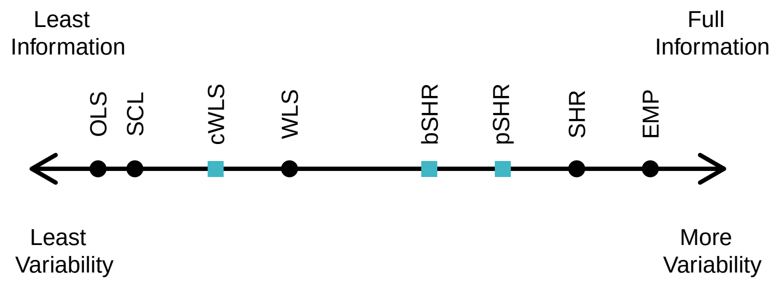


Figure 3: Covariance matrix approximation continuum againts information used and forecast error variability for the whole hierarchy. Square points denote the alternative covariance matrix approximations

4.1.3. Error Metrics

We consider two different measures in hierarchical forecasting: (a) a measure which aligns to the objective function (Wickramasuriya et al., 2019; Panagiotelis et al., 2020, 2021); (b) a measure which is more relevant to decision makers (Kourentzes and Athanasopoulos, 2019; Athanasopoulos and Kourentzes, 2020). The former deals with measuring the average accuracy of base and reconciled forecasts across the complete hierarchy. The latter measures performance of individual time series and then summarises them across the complete hierarchy. A relevant discussion about the evaluation of hierarchical forecasts is given by Athanasopoulos and Kourentzes (2020).

We focus on the mean squared error (MSE) for each time series i, as

$$\text{MSE}_{i,h} = \frac{1}{J}\sum_{j=1}^{J}(y_{ij,t+h} - \hat{y}_{ij,t+h|t})^2,$$

where J is the simulation run. Then, we measure the performances across the hierarchy from the loss function perspective, using Relative Total Squared Error:

$$\text{RelTotSE}_h = \frac{\sum_{i=1}^{n} MSE_{ih,recon}}{\sum_{i=1}^{n} MSE_{ih,base}},$$

where n is the number of time series in the hierarchy. Essentially, RelTotSE_h measures the relative accuracy between SSE_{recon} and SSE_{base}. From the decision-focused perspective, we use

Average Relative MSE, inspired by Davydenko and Fildes (2013):

$$\text{AvgRelMSE}_h = \left(\prod_{i=1}^{n} \frac{MSE_{ih,recon}}{MSE_{ih,base}} \right)^{\frac{1}{n}}.$$

4.2. Findings: Small Hierarchy

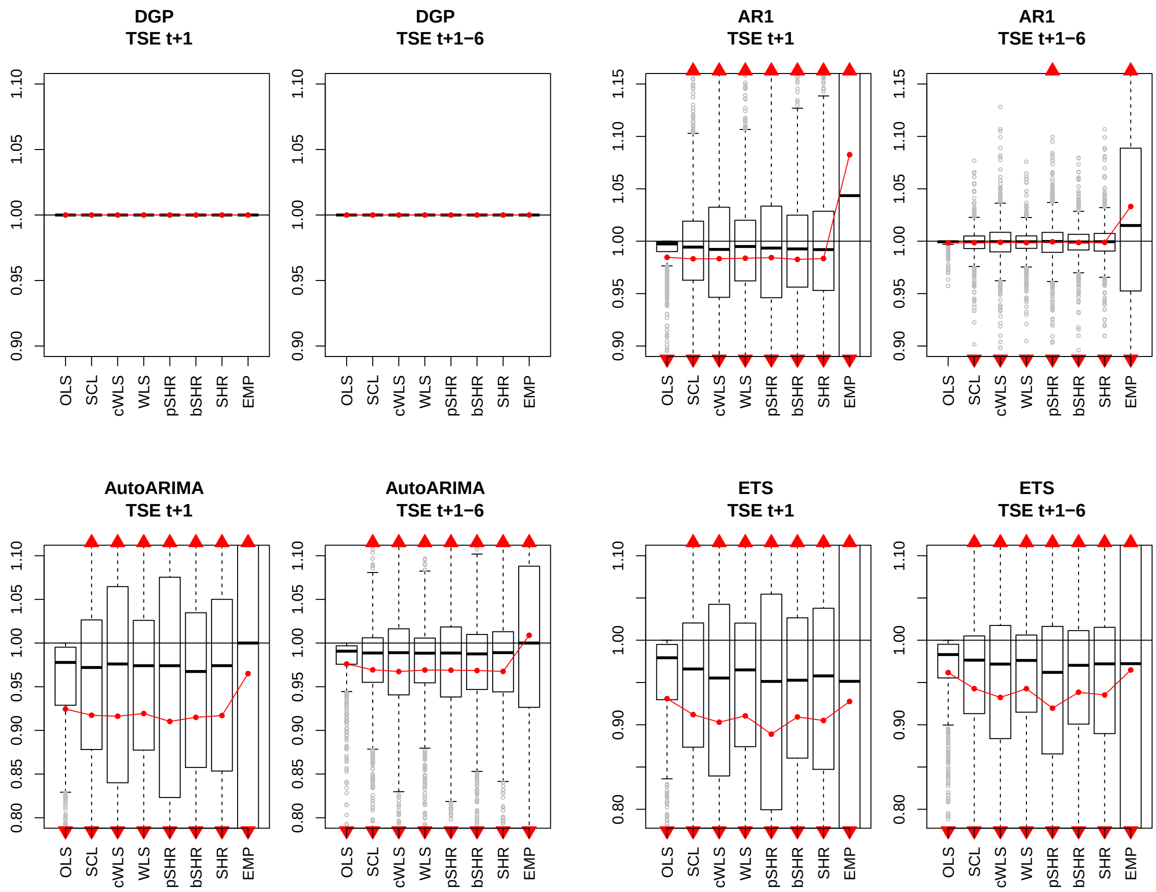


Figure 4: Distributions of RelTotSE (TSE) for the small hierarchy and the sample size of 24. The red dotted lines denotes the geometric average of each error distribution. The red arrows show that some parts of the distribution are not plotted. Outliers are denoted by grey dots.

Figure 4 and 5 present the distributions of RelTotSE and AvgRelMSE for different forecasting models and covariance matrix approximations for the sample size of 24. Each pair of subplots corresponds to a modelling case from Table 1, where the first subplot provides $t+1$ error distributions, while the second provides the average across $t+1$ to $t+6$. The red lines indicate the geometric mean. Furthermore, in all plots, arrows at the top and at the bottom indicate that there are outliers and a part of the distribution is not plotted. Note that the covariance matrix approximations are ordered by its completeness of the information, i.e. from an identity matrix to utilising variances and covariances fully to estimate the reconciliation weights matrix.

From Figure 4, for RelTotSE, we can see that when we have perfectly-specified forecasting models, there is no gain from reconciliation. That is because the models are able to produce coherent forecasts. As the base forecasts are coherent already, the reconciled forecasts are the same as the base forecasts.

However, once we introduce modelling uncertainty, we gain some benefit from forecast reconciliation. Imposing parameter uncertainty only, i.e. employing estimated AR(1), induces relatively small gains from reconciliation. This shows that as the modelling uncertainty increases reconciliation provides gains but again at an increased variability, meaning that the mean of the relative errors decreases, but the variance of error measure distributions increases. The relative accuracy gain is more noticeable when we use AutoARIMA compared to AR(1), at the cost of higher variability of RelTotSE. Using ETS we benefit the most from forecast reconciliation, but also at the cost of the highest error variability among other modelling options. These gains are less pronounced when the multi-step base forecast errors are introduced, even though the variances are non-zero.

Looking at the covariance matrix approximations, the results verify Theorem 3.1 by Panagiotelis et al. (2021) that OLS reconciliation improves or matches the accuracy of base forecasts regardless the model specification. However, when we approximate the covariance matrix, it is possible to get less accurate reconciled forecasts on some observations, as uncertainties are introduced. We can see that some distributions go well beyond the accuracy of base forecasts. As expected the simpler the approximation of the covariance matrix is, the less the variability is, and vice versa. Our proposed covariance matrix approximations, e.g. cWLS, pSHR, and bSHR, are able to reduce the variability of RelTotSE yet provide relative accuracy, on average, similar to WLS and SHR.

In Figure 5 the model specification does not affect the accuracy improvement much, but affect the variability of the relative measure, for AvgRelMSE. We can see from the figure that the variability increases as the forecasting models become increasingly mis-specified.

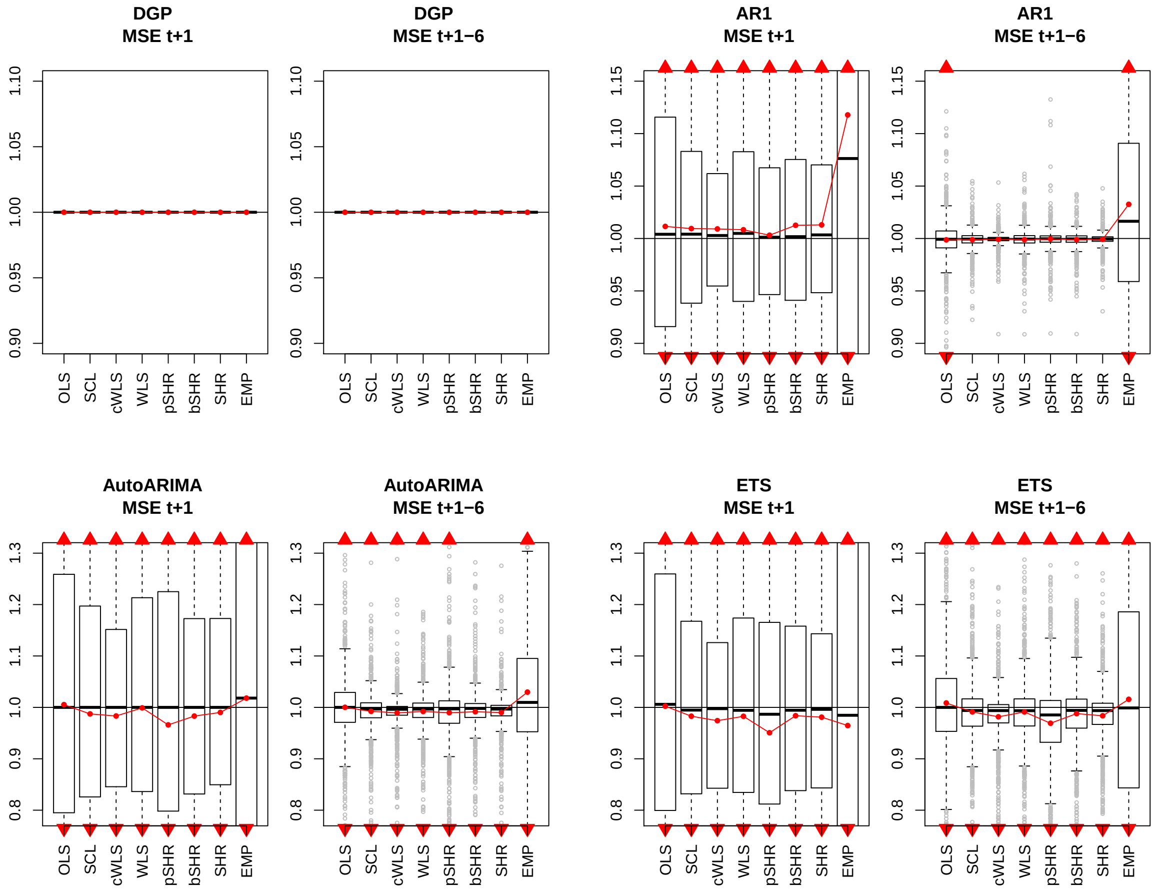


Figure 5: Distributions of AvgRelMSE (MSE) for the small hierarchy and the sample size of 24. The red dotted lines denotes the geometric average of each error distribution. The red arrows show that some parts of the distribution are not plotted. Outliers are denoted by grey dots.

Here, the effect of the covariance matrix approximations differ from RelTotSE. For AvgRelMSE there is no clear increase in error variability as more complete covariance matrix approximations are used. However, the simplest covariance matrix approximation results in very variable performance. This can be explained by considering that another role of the covariance matrix in forecast reconciliation is to scale the reconciled forecast errors. At more aggregate levels of the hierarchy the scale of errors increases. Conversely SHR is able to scale the forecast errors better than OLS. The same is true for the other approximations. This argument aligns to the discussion on temporal hierarchies where SCL performs well (Athanasopoulos et al., 2017).

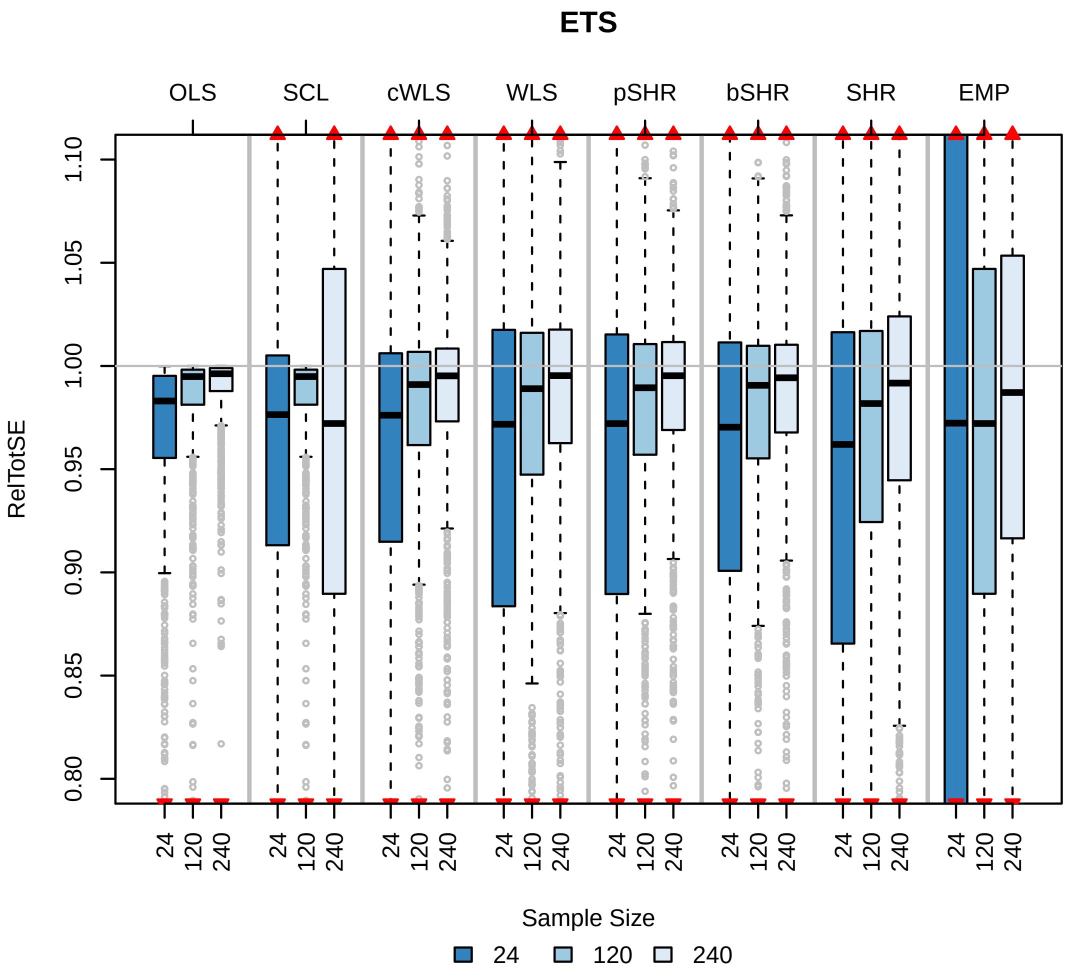


Figure 6: ETS, h=6, different sample sizes, RelTotSE. The red dotted lines denotes the geometric average of each error distribution. The red arrows show that some parts of the distribution are not plotted. Outliers are denoted by grey dots.

Figure 6 presents the effect of sample sizes using mis-specified models with RelTotSE. For RelTotSE, the benefits of forecast reconciliation reduce as the sample sizes increase, together with the decrease of the variability. As the estimation of the parameters improves, the uncertainty reduces, and therefore this result is expected. We can see this effect by looking at the lower error bars. Nevertheless, we still observe some variability in longer sample sizes.

Regardless of what error metrics is used, we observe variability in the performance of forecast reconciliation. For example, we observe a trade-off between accuracy and variability of RelTotSE, i.e. the more complete covariance matrix is, the more accurate the reconciled forecasts are. This, however, comes at a cost, which is introducing more error variability.

4.3. Findings: Large Hierarchy

Next we discuss the findings from the large hierarchy. Table 3 presents a comparison between the small and the large hierarchy, with RelTotSE and AvgRelMSE for one-step ahead forecast and different sample sizes. We present the geometric mean and the logarithm of geometric standard deviation of the relative error distribution from ETS only. A negative (positive) number denotes an improvement (deterioration) on the error measure. The bold highlights the most accurate reconciliation approach, for the geometric mean, and the least volatile reconciliation approach, for the standard deviation. All numbers in the geometric mean are in percentages.

Statistics	Geometric Mean (%)				Geometric St. Dev. (log)			
Hierarchy	Small		Large		Small		Large	
Sample	24	240	24	240	24	240	24	240
				RelTotSE				
OLS	-3.8	-1.2	-7.4	-3.4	**6.1**	**2.7**	**10.1**	**4.0**
SCL	-5.7	-1.4	-18.6	-13.5	12.7	4.9	22.9	15.7
CWLS	-5.7	-1.4	-17.4	-13.0	12.8	4.9	21.4	15.3
WLS	-6.8	-1.4	-18.4	-15.3	17.8	7.0	25.4	19.8
pSHR	-6.5	-1.2	-18.4	-14.6	16.5	5.4	25.4	18.2
bSHR	-6.2	-1.4	**-21.8**	-16.8	14.7	5.4	29.8	23.3
SHR	**-8.0**	**-2.6**	-18.3	**-18.4**	20.9	11.8	25.5	29.4
EMP	-3.5	-2.3	293.2	-11.1	44.7	17.5	119.2	51.0
				AvgRelMSE				
OLS	0.9	-0.6	12.6	10.8	13.1	6.1	33.5	29.9
SCL	-0.9	-0.7	2.3	2.4	7.6	3.1	25.8	23.6
cWLS	-0.9	-0.7	-0.4	-0.4	7.8	3.1	19.4	17.9
WLS	-1.8	-0.7	-1.0	-1.5	**6.8**	**2.3**	**18.6**	16.9
pSHR	-1.6	-0.4	-1.0	-1.2	6.9	3.5	**18.6**	17.1
bSHR	-1.2	-0.7	**-1.5**	-1.4	7.6	3.0	19.4	17.8
SHR	**-3.1**	**-1.9**	-0.9	**-2.6**	11.5	8.0	19.3	**15.9**
EMP	1.6	-1.7	1730.7	3.3	34.8	13.1	156.1	26.7

Table 3: A comparison between the small and the large hierarchy with RelTotSE and AvgRelMSE for one-step ahead forecast. A negative (positive) number denotes an improvement (deterioration) of the error measure, on average. Numbers in bold highlight the smallest numbers.

Considering the geometric mean of RelTotSE and AvgRelMSE, SHR outperforms the other alternatives, apart from the case of small sample size for the large hierarchy, where bSHR is the best. We find that pSHR ia also competitive. As expected EMP is very sensitive to estimation uncertainty. The increased size of the hierarchy substantially reduces its performance, while increasing the sample size helps.

In terms of the standard deviation of RelTotSE, we observe similar findings to the small hierarchy, where OLS is the least variable, while EMP and SHR are the most volatile. Overall, as the completeness of the covariance matrix increases, so does the variance of the errors. The larger size of the hierarchy increases it further, requiring more terms to be estimated, while sample size helps. On the other hand, for AvgRelMSE, the least variable methods are between WLS, pSHR, and SHR, which have more complete information than OLS does.

The differences between RelTotSE and AvgRelMSE can be largely explained by the changing scale across the levels of the hierarchy. Improvements in the top-level dominates RelTotSE, which is scale dependent. On the other hand, the scale independent AvgRelMSE balances the gains across all levels, and therefore differences are less pronounced. It is important to consider both views. The RelTotSE matches the operation performed by MinT Reconciliation and directly demonstrates the effects of uncertainty highlighted by the stochastic coherency. Furthermore, although on different scales, both RelTotSE and AvgRelMSE indicate that hierarchical forecasting is beneficial in terms of accuracy. As evidence, as the complexity of the covariance matrix approximation increases, so does the variance of the errors, in agreement with our theoretical discussion.

5. Forecasting A&E hospital admissions

We apply our understanding of stochastic coherency to Accident and Emergency (A&E) admission data in a hospital in the United Kingdom. Hospitals in the UK, as is the case globally, face increased pressure due to the global pandemic, requiring many resources. This has often caused disruptions in their normal operations, such as scheduled surgeries, but also in the operations of their A&E departments. To this end, it is important to have reliable forecasts of demand, across the different groups of interest, so that the hospital can allocate resources best. In normal conditions, A&E forecasting is important in the United Kingdom due to the worrying mismatch between the hospital service quality and financial efficiency (Limb, 2014). Forecasts can be useful for multiple decisions, such as staff scheduling, procurement of drugs and other medical supplies, bed utilisation, etc.

The time series consists of 64 bottom-level time series, which are structured according to age (under 3 years old, between 4-16 years old, between 17-74 years old, and more than 75 years old), gender (male; female), and disposal type (admitted, discharged, referred to clinics, transferred, died, referred to health care professionals, left, and others). Figure 8 provides a plot of representative time series from the A&E hospital admissions dataset. There are multiple ways to aggregate from the bottom-level time series to the total number of admissions, for example aggregating by gender, age, or type first, and then across one of the remaining two characteristics, and so on. This results in a grouped hierarchy of 135 time series with eight distinct groups/levels. A map of the hierarchy is presented in Figure 7. Note that some time series in the bottom-level are sparse and these pose challenges in the modelling.

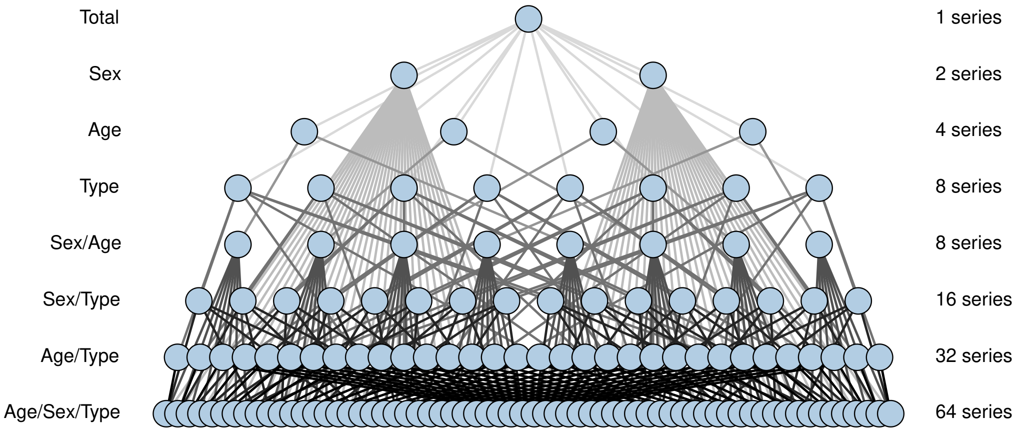
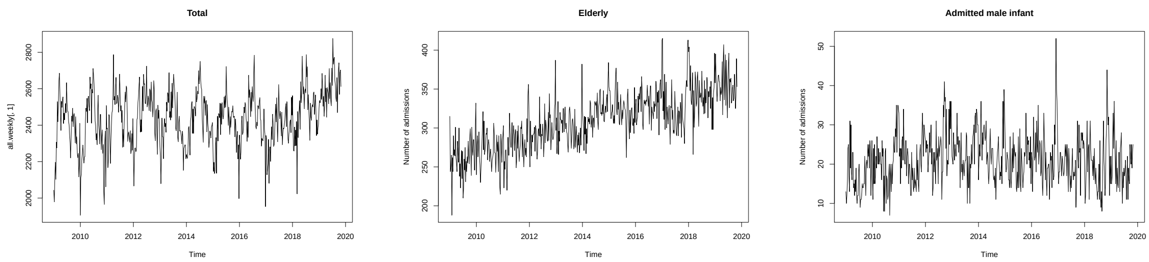


Figure 7: Map of the A&E admission hierarchy. Labels on the left indicate the nature of time series at each level, while on the right provide the number of time series at that level. The lines indicate how the time series are aggregated between the different levels.

We have been provided weekly data from January 2009 to October 2019. We produce from 1- to 4-step ahead base forecasts with two sets of in-sample data. The longer set uses 536 weeks, while a much shorter set has only 100 weeks. The second introduces additional modelling uncertainty as the number of time series is larger than the number of observations making the approximation of the covariance matrix challenging. This helps us validate our findings from stochastic coherency on real complex hierarchical time series. For both cases we use the same test set of 29 weeks, allowing for 25 rolling origin forecasts.



(a) Total (b) Elderly (c) Admitted male infant

Figure 8: Representative time series of A&E hospital admissions dataset for total, elderly, and admitted male infant group. We observe that the time series exhibit seasonal patterns, local trend, and outliers.

Afilal et al. (2016) point that the A&E admission data can be structured in a hierarchy according to the patients' characteristics, which may be correlated. Athanasopoulos et al. (2017) use ARIMA to model UK A&E admission data, but at a country level. Forecasting models can also incorporate exogenous variables, such as special events, holidays, and temperatures, to improve forecast accuracy using regression models, ARIMA, or ETS (Kam et al., 2010; Xu et al., 2016; Rostami-Tabar and Ziel, 2020).

We use ARIMA and ETS with automatic model selection, as setup for the large simulation above. We acknowledge that these can be prone to model misspecification problems. First, we omit important information for A&E forecasting, such as special events. Second, we do not treat differently the any sparse time series at the bottom level of the hierarchy. This potential misspecification is of interest, to explore how the reconciliation approaches impact the forecasts. Thirdly, the automatic ARIMA function on forecast package does not capture seasonality while some time series are seasonal. We reconcile base forecasts using all covariance matrix approximations, as in Figure 2 and evaluate the forecasts using RelTotSE and AvgRelMSE.

Statisics	Geometric Mean (%)				Geometric St. Dev. (log)			
Model	ETS		ARIMA		ETS		ARIMA	
Sample	Short	Long	Short	Long	Short	Long	Short	Long
RelTotSE								
OLS	-0.90	-0.70	-1.40	-3.10	**0.50**	**0.60**	**0.70**	**2.30**
SCL	-0.90	-2.70	-2.20	-7.10	3.70	3.50	7.20	5.90
cWLS	-0.70	-2.60	**-3.10**	-7.40	3.10	2.90	6.90	5.00
WLS	-0.90	-3.10	-2.90	-7.90	4.30	3.70	8.00	6.20
pSHR	**-1.40**	-3.30	-2.90	-5.60	4.90	3.30	8.40	4.50
bSHR	-1.10	-3.10	-1.30	-6.20	5.90	4.80	9.20	6.30
SHR	0.80	-6.60	0.30	**-11.30**	5.80	8.00	9.90	9.20
EMP	55.20	**-6.90**	64.50	-2.60	43.50	41.40	39.90	32.60
AvgRelMSE								
OLS	0.30	1.40	1.20	1.30	3.40	3.50	4.40	7.70
SCL	-2.20	-1.60	-1.50	-1.50	2.70	**3.00**	**3.70**	4.70
cWLS	-2.00	-1.50	-2.30	-2.90	**2.60**	3.40	**3.70**	**4.80**
WLS	-2.70	-2.00	-2.40	-3.10	3.40	3.30	4.20	4.60
pSHR	**-3.50**	-2.30	**-2.70**	-1.60	3.80	3.80	4.50	6.30
bSHR	-3.20	-2.30	-1.70	-1.60	4.00	3.60	4.60	5.70
SHR	-1.30	**-5.00**	-1.20	**-7.00**	4.00	5.80	5.50	7.50
EMP	65.40	2.20	51.40	-0.20	35.30	31.70	27.80	27.10

Table 4: A comparison between AutoARIMA and ETS models for the sample size of 100 and 536, over 1-4 step ahead forecast. A negative (positive) number denotes an improvement (deterioration) of the error measure, on average. Numbers in bold highlight the best performing results.

Table 4 presents a comparison between ARIMA and ETS models for the short and long samples, for all covariance matrix approximations. Similar to Table 3, we present summary statistics of the error distribution, with the geometric mean and the logarithm of the geometric standard deviation. The results are ordered in terms of completeness of the covariance matrix approximation.

We note that across all results, the more complete covariance matrices, such as the pSHR, and SHR, offer good forecast accuracy. First, we focus on the cases of the short in-sample set. We find that pSHR performs overall best. For RelTotSE and ARIMA the cWLS is best but closely followed by pSHR. We note that the more complete approximations (SHR and EMP) perform poorly in terms of RelTotSE, while for AvgRelMSE the SHR improves upon the base forecasts, but still performs worse than all simpler approximations.

The results for the large sample are contrasting. The SHR performs best. In the case of RelTotSE and ETS we observe that EMP outperforms all alternatives, although closely followed by SHR. The long sample size allows for reliable estimation. We note that the less complete covariance approximations, although perform worse, all improve upon the base forecasts.

Looking at the standard deviation of the forecast errors, OLS provides the most stable relative accuracy for RelTotSE, and SCL together with cWLS for the AvgRelMSE. Overall, simpler covariance approximations exhibit a low standard deviation of the forecast errors. More complete ones, such as the SHR and EMP, exhibit increased deviations, even for the long in-sample set. The proposed covariance matrix approximation, namely cWLS, pSHR, and bSHR, they enable us to compromise between the accuracy gain and the variability of the error distribution.

The results in Table 4 are relative to the base forecasts and do not permit a direct comparison between ETS and ARIMA, as this is not the aim of the evaluation. If we compare the two, we find that the reconciled forecasts from ARIMA outperform the ones from ETS for small sample sizes, and vice versa.

Therefore, we argue that with complex data generating processes, observed in real data, we again find variability in the performance of forecast reconciliation, especially when we need to estimate the covariance matrix approximation, instead of relying on fixed values. This emphasises the importance of stochastic coherency to be considered in the application of hierarchical forecasting. For the particular case of A&E hospital admissions, we find that the pSHR covariance approximation that was developed with the understanding we gained from stochastic coherency resulted overall in good forecast accuracy, and stability. The performances of cWLS and bSHR were similar. As a group these performed well against approximations from the literature that were either too restrictive or they did not consider the additional uncertainty arising from stochastic coherency.

6. Conclusions

Stochastic coherency shifts our paradigm from deterministic to stochastic forecast reconciliation. We have to deal with the uncertainties in estimating the reconciliation weights matrix, originating modelling uncertainty due to limited sample size. This directly affects the performance of the forecast error, either on average or its variability, due to the approximation of the covariance matrix.

Our findings show that there are two sources of uncertainty in forecast reconciliation originating from modelling, namely the base forecast uncertainty and the reconciliation weight uncertainty. It becomes obvious that the base forecast uncertainty is carried forward to the reconciled forecast uncertainty. Model and parameter uncertainty contaminate the covariance matrix approximation and introduce the second source of error. Naturally, the sample size affects modelling uncertainty. Moreover, a larger hierarchy produces more uncertain reconciled forecasts, as there are more terms to estimate. These become evident with stochastic coherency and the results from both simulated and real data corroborate with this understanding.

Due to these uncertainties, we cannot say that forecast reconciliation improves the accuracy consistently all the time. Our findings show in some cases that the reconciled forecast accuracy can be worse than the base forecast accuracy in some cases, even if on average it ranks better.

In relation to different model specifications, there are some conditions when the degree of specification affects the efficacy of forecast reconciliation. As stated previously, if the forecasting models capture the bottom level data adequately, stochastic coherency indicates that the bottom-up approach is sufficient, and reconciliation will not add value. For instance, our simulation demonstrates that by having perfect information, the models can estimate coherent mean and errors, hence forecast reconciliation does not change anything. When the modelling uncertainty is limited, we obtain limited gains from forecast reconciliation. However, when we have mis-specified models, forecast reconciliation becomes useful, which matches typical cases in reality.

The benefit of forecast reconciliation appears when there are modelling uncertainties. The MinT Reconciliation reduces the forecast error by redistributing the modelling uncertainty, which contains the uncertainty of parameter estimation as well as the unobservable statistical discrepancy, across the hierarchy. As long as the data generating process is unknown and the forecasts are produced from individual forecasting models, the MinT Reconciliation can help. Here we did not explore the effect of statistical discrepancy and it should be explored further in future work.

We also observe a significant accuracy improvement when forecast error covariances are incorporated into the estimated reconciliation weights matrix. Even though it improves the forecast accuracy generally, it comes at the cost of increased variability of the error measure. One of the solutions to deal with the variability is to obtain a good quality reconciliation weights matrix, which reduces the effect of modelling uncertainty. We can obtain this by managing variances and covariances on the estimated h-step ahead covariance matrix and this determines the quality of the combination weights estimation in the estimated reconciliation weights matrix. A weaker argument for this is given by Kourentzes and Athanasopoulos (2019).

We can estimate a useful reconciliation weights matrix from approximating the base forecast error covariance matrix. Simple and fixed approximations of the covariance matrix, namely OLS and SCL, are immune to modelling uncertainty and the fixed estimation of the reconciliation weights matrix is able to limit the variability of the error measure. On the other extreme, the estimation of EMP and SHR relies heavily on the base forecast errors, and is prone to modelling uncertainty. Consequently, the reconciliation weights matrix becomes uncertain. SHR relying on shrinkage remains widely useful, while EMP is useful only for a very large estimation sample size.

Managing the off-diagonals in the covariance matrix construction enables to balance the accuracy gain and the variability of the forecast error. We argue that using bSHR and pSHR are potential solutions, because they introduce restrictions, yet maintain structurally important information. Our findings also show that bSHR and pSHR results in a similar accuracy gain, but less variable to SHR, while being competitive to the simpler WLS and cWLS. Naturally, this is important in applications of hierarchical forecasting, where both aspects of accuracy and reliability over time are important. We find strong evidence of this when we model accident and emergency admissions for the UK hospital of our case study, where the covariance matrices developed with our understanding of stochastic coherency performed very competitively, offering a good balance between accurate and stable forecasts. We argue that these can aid decision making. Naturally, less variable forecasts are beneficial widely for operations. For example, in a production setting less erratic forecasts result in more resilient plans and lower costs (Sagaert et al., 2019). Similar examples can be drawn from inventory management, where maximum accuracy forecasts do not necessarily result in the best inventory performance (Kourentzes et al., 2020).

Our discussion extends to probabilistic hierarchical forecasting. The literature does not take into account modelling uncertainty (Jeon et al., 2019; Taieb et al., 2020). The density of the reconciled forecasts is also affected by the reconciliation weights matrix and so is their performance. Future research on this area will help highlighting the exact influence of modelling uncertainties on probabilistic hierarchical forecasting.

In conclusion, we introduce stochastic coherency to overcome a limitation in the definition of classical coherency in forecast reconciliation and hierarchical forecasting. Using the concept of stochastic coherency, we give more attention to the error term from the data generating process. We are able to demonstrate that stochastic coherency is relevant to forecast reconciliation via simulations and a case study of A&E admissions in a hospital. It allows us to explain observations from the literature, where well performing approximations for the covariance matrix introduce variability in the error distribution, and provides a framework to consider the setup of hierarchical forecasting in applications.

7. Acknowledgement

We would like to thank George Athanasopoulos for his constructive feedback. We would also like to thank the anonymous reviewers for their feedback that helped improve the paper further. This research did not receive support from funding agencies in the public, commercial, or not-for-profit sectors. We declare no conflicts of interest.

References

Afilal, M., Yalaoui, F., Dugardin, F., Amodeo, L., Laplanche, D., and Blua, P. (2016). Forecasting the Emergency Department Patients Flow. *Journal of Medical Systems*, 40(7).

Athanasopoulos, G., Ahmed, R. A., and Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, 25:146–166.

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R. J., and Affan, M. (2020). Hierarchical Forecasting. In Baltagi, B., Hong, Y., Koop, G., Krämer, W., and Matyas, L., editors, *Advanced Studies in Theoretical and Applied Econometrics*, volume 52, pages 689–719. Springer.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., and Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, 262(1):60–74.

Athanasopoulos, G. and Kourentzes, N. (2020). On the evaluation of hierarchical forecasts.

Australian Bureau of Statistics (2015). Australian System of National Accounts: Concepts, Sources and Methods. Technical report.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419–466.

Chow, G. C. and Lin, A.-l. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The review of economics and statistics*, 53(4):372–375.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32:754–762.

Davydenko, A. and Fildes, R. (2013). Measuring Forecasting Accuracy: The Case Of Judgmental Adjustments To SKU-Level Demand Forecasts. *International Journal of Forecasting*, 29:510–522.

Di Fonzo, T. and Girolimetto, D. (2020). Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives.

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

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2019). *forecast: Forecasting functions for time series and linear models*. R package version 8.9.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis*, 55:2579–2589.

Hyndman, R. J., Lee, A. J., and Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics and Data Analysis*, 97:16–32.

Jeon, J., Panagiotelis, A., and Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research*, 279:364–379.

Kam, H. J., Sung, J. O., and Park, R. W. (2010). Prediction of Daily ED Patient Numbers for a Regional Emergency Medical Center using Time Series Analysis. *Healthcare Information Research*, 16(3):158–165.

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

Kourentzes, N. and Athanasopoulos, G. (2021). Elucidate structure in intermittent demand series. *European Journal of Operational Research*, 288:141–152.

Kourentzes, N., Barrow, D., and Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, 209:226–235.

Kourentzes, N., Saayman, A., Jean-Pierre, P., Provenzano, D., Sahli, M., Seetaram, N., and Volo, S. (2021). Visitor arrivals forecasts amid covid-19: A perspective from the africa team. *Annals of Tourism Research*, 88:103197.

Kourentzes, N., Trapero, J. R., and Barrow, D. K. (2020). Optimising forecasting models for inventory planning. *International Journal of Production Economics*, 225:107597.

Limb, M. (2014). Hospitals forecast a rise in emergency admissions, while commissioners forecast a fall. *BMJ (Clinical research ed.)*, 349:3–4.

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

Office for National Statistics (2011). Labour Market Quality and Methodology Information. Technical Report October, Office for National Statistics.

Oliveira, J. M. and Ramos, P. (2019). Assessing the Performance of Hierarchical Forecasting Methods on the Retail Sector. *Entropy*, 21(436):1–22.

Ord, J. K., Fildes, R., and Kourentzes, N. (2017). *Principles of business forecasting.* Wessex Press Inc., New York, second edition.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., and Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, 37(1):343–359.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., and Hyndman, R. J. (2020). Probabilistic Forecast Reconciliation: Properties, Evaluation and Score Optimisation.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rostami-Tabar, B. and Ziel, F. (2020). Anticipating special events in Emergency Department forecasting. Technical report.

Sagaert, Y. R., Aghezzaf, E. H., Kourentzes, N., and Desmet, B. (2019). Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research*, 264(2):558–569.

Schäfer, J. and 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):1–30.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Svetunkov, I. (2019). *smooth: Forecasting Using State Space Models*. R package version 2.5.3.

Taieb, S. B. and Koo, B. (2019). Regularized Regression for Hierarchical Forecasting Without Unbiasedness Conditions. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM.

Taieb, S. B., Taylor, J. W., and Hyndman, R. J. (2020). Hierarchical Probabilistic Forecasting of Electricity Demand With Smart Meter Data. *Journal of the American Statistical Association*, pages 1–17.

Wickramasuriya, S. L., Athanasopoulos & Rob, G., and Hyndman, J. J. (2019). Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association*, 114:804–819.

Xu, Q., Tsui, K. L., Jiang, W., and Guo, H. (2016). A Hybrid Approach for Forecasting Patient Visits in Emergency Department. *Quality and Reliability Engineering International*, 32(8):2751–2759.

Yang, D., Goh, G. S., Jiang, S., and Zhang, A. N. (2016). Forecast UPC-level FMCG demand, Part III: Grouped reconciliation. In *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, pages 3813–3819.