Enhancing Forecast Reconciliation: A Study of Alternative Covariance Estimators

Vincent Su Shanika Wickramasuriya (supv.)
George Athanasopoulos (supv.)

Abstract

A collection of time series connected via a set of linear constraints is known as hierarchical time series. Forecasting these series without respecting the hierarchical nature of the data can lead to incoherent forecasts across aggregation levels and, in practice, reduced accuracy. Forecast reconciliation corrects this by adjusting base forecasts to satisfy such constraints. Among modern reconciliation methods, Minimum Trace (MinT) is widely used, however, it requires a good estimate of the forecast error covariance matrix. The current practice is to use the linear shrinkage towards a diagonal target. Furthermore, the covariance estimate is based on 1-step-ahead residuals, then proportionally scale it to approximate the h-step-ahead covariance matrix. This leaves a question of whether this method is appropriate for all real-world applications. We study the shortcomings of current practice and propose alternative covariance estimators, including the NOVELIST estimator (shrinkage towards a soft-thresholded target), PC-adjusted shrinkage (which utilises latent factor structures), and horizon-specific estimators that relax proportional scaling. We evaluate MinT using these covariance estimates for both point and probabilistic reconciliation, and demonstrate their effectiveness and improvements over the shrinkage estimator in a complex, large-hierarchy dataset.

1 Introduction

In time series forecasting, aggregation occurs in a variety of settings. For example, Starbucks Corporation operates in many countries, and each country has multiple cities where they have outlets. The sales data is structured *hierarchically*: the top level is the company's total sales, which disaggregates into sales by country, then sales by city within each country, and finally down to sales by individual outlet within each city. As a result, there are over 50,000 sales series across all *aggregation levels*, and decision makers need forecasts at each level to manage inventory and plan marketing strategies effectively. The hierarchy can be even more complex if we consider the sales of different categories of products (e.g., beverage, food, etc.) and the sales of product within each category (e.g., latte, cappuccino, etc.) at each aggregation level. In this case, the structure is called a *grouped structure*, where the aggregation paths are not unique. Such structures also arises in many other decision-making contexts, from supply chains (Angam et al., 2025; Seaman & Bowman, 2022) and energy planning (Di Modica et al., 2021), to macroeconomics (El Gemayel et al., 2022; Li et al., 2019) and tourism analysis (Athanasopoulos et al., 2009). Stakeholders in these settings need forecasts at several aggregation levels to allocate resources and manage risk.

In practice, when forecasts are produced for all series (often called *base forecasts*), they typically violate the aggregation constraints observed in the data (e.g., the sum of all countries' sales forecasts does not equal the total sales forecast). Such forecasts are called *incoherent*. Incoherence undermines downstream decisions that require internal consistency and can degrade forecasting performance. To tackle this problem, forecast reconciliation was introduced. Forecast reconciliation, a post-processing step, utilises the information from the hierarchical structure and data to adjust the initially produced base forecasts, so that the resulting *reconciled forecasts* are *coherent* (i.e., respecting the aggregation constraints). It was first introduced by Hyndman et al. (2011), and later developed by Erven & Cugliari (2015), Hyndman et al. (2016), Ben Taieb & Koo (2019), Wickramasuriya et al. (2019), Wickramasuriya et al. (2020), and others that focus point forecast reconciliation. Recognising the important of probabilistic forecasting, probabilistic forecast reconciliation was studied by Shang & Hyndman (2017), Jeon et al. (2019), and Ben Taieb et al. (2021), and later formally defined by Panagiotelis et al. (2023). Athanasopoulos et al. (2024) provide a comprehensive review of the literature on forecast reconciliation.

Among the modern methods, the Min Trace (MinT) approach developed by Wickramasuriya et al. (2019) is is widely used due to its strong theoretical properties for minimising total reconciled forecast error variance, computational efficiency, and robust empirical performance. MinT is later extended to probabilistic reconciliation by Wickramasuriya (2024), showing

that it also minimises the negative log score of the reconciled distribution under Gaussian assumptions. Wickramasuriya et al. (2019) also argued that modelling spatial autocorrelations directly from the start would be challenging as in this case of a large collection of time series. Post-processing reconciliation has the advantage to implicitly model this spatial autocorrelation structure, especially true for MinT. MinT is implemented in popular R and Python software ecosystems (Nixtla, 2025; O'Hara-Wild et al., 2024).

A central difficulty for MinT is estimating the covariance matrix of base-forecast errors, particularly beyond one-step forecast horizons. This is a high-dimensional estimation problem in which the number of series often exceeds the time dimension. A common practice, following Wickramasuriya et al. (2019), is to estimate the 1-step-ahead covariance from the residuals, using linear shrinkage toward a diagonal target (Schäfer & Strimmer, 2005). Then, proportionally scale this estimate to approximate the multi-step-ahead covariance matrix. While convenient and guaranteed to produce a positive-definite estimate, this practice has three important shortcomings. First, the shrinkage is uniform across off-diagonals, applying a single penalty that may over-shrink genuine dependence and under-shrink noise. Second, many hierarchical data sets exhibit strong latent low-rank structures that can be explicitly exploited. Third, the proportionality relationship between the h-step-ahead and 1-step-ahead forecast error covariance matrices might not hold when error dynamics change with horizon.

These limitations matter because the theoretical advantages of MinT depend on the quality of the covariance estimate available in finite samples. Despite its central role, there has been limited work on tailored covariance estimation for reconciliation. An exception is the recent double-shrinkage proposal in Carrara et al. (2025), which introduces an additional target designed to encode conditional dependence suggested by the hierarchy. This line of work still remains at an early stage and does not fully address the aforementioned limitations.

Meanwhile, there has been substantial progress in high-dimensional covariance estimation that can be leveraged for MinT. Building on shrinkage, Huang & Fryzlewicz (2019) proposed NOVELIST, which shrinks toward a soft-thresholded target rather than a diagonal one, improving flexibility to sparse targets. Extending beyond linear shrinkage, Ledoit & Wolf (2012) (and further developed by Ledoit & Wolf (2020)) introduced nonlinear shrinkage that replaces sample eigenvalues with data-driven nonlinear transformation of themselves. In a complementary direction, Fan et al. (2013) introduce factor-based estimators that explicitly preserve latent low-rank structure while thresholding the idiosyncratic component. Related thresholding approaches, including hard thresholding (Bickel & Levina, 2008), generalised thresholding (Rothman et al., 2009), and adaptive thresholding (Cai & Liu, 2011), have also been developed. This growing toolkit provides multiple pathways to maximise the performance of MinT.

This paper focuses on covariance estimation for MinT. We examine the shortcomings of the standard shrinkage plus scaling practice and introduce alternative estimators that address these issues individually and in combination. Specifically, we study NOVELIST as a more adaptive target-based shrinkage; principal-component-adjusted (PC-adjusted) estimators that exploit latent factor structures; and horizon-specific estimators that relax proportional scaling, including direct estimation from multi-step residuals. We evaluate MinT under these alternatives for both point and probabilistic reconciliation. In a large, complex real-world hierarchy, our findings reveal three main insights. First, NOVELIST improves probabilistic performance relative to linear shrinkage, including higher empirical coverage. Second, PC-adjusted estimators consistently outperform other estimators when common components are strong, for both point and probabilistic metrics. Third, proportional scaling of the one-step covariance remains a competitive baseline, indicating that the assumption may not be universally invalid.

The remainder of the paper is organised as follows. Section 2 sets out the framework for hierarchical time series, forecast reconciliation, and MinT, and motivates the need for improved covariance estimation. Section 3 presents the covariance estimators considered, outlining their strengths and limitations in the reconciliation context. Section 4 defines the evaluation metrics for point and probabilistic reconciliation. Section 5 details the simulation design and examines the performance of NOVELIST compared to shrinkage. Section 6 dives into a real-world application with all proposed methods and highlights behaviours not observed in the simplified simulations.

2 Theoretical Framework

2.1 Hierarchical Time Series

Hierarchical time series are multivariate time series $\boldsymbol{y}_t \in \mathbb{R}^n$ organised in a structure where the series adheres to some constraints. Figure 1 illustrates a simple two-level hierarchical structure with one top-level series $y_{Tot,t}$, disaggregating down to two level-1 series $(y_{A,t}, y_{B,t})'$, and to four bottom-level series $(y_{A1,t}, y_{A2,t}, y_{B1,t}, y_{B2,t})'$. Here, the *aggregation constraints* imply that $y_{Tot,t} = y_{A,t} + y_{B,t}$, $y_{A,t} = y_{A1,t} + y_{A2,t}$, and $y_{B,t} = y_{B1,t} + y_{B2,t}$.

The bottom-level (or most disaggregated) series are denoted as $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$. Thus, the full vector of all series in the hierarchy can be represented as:

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

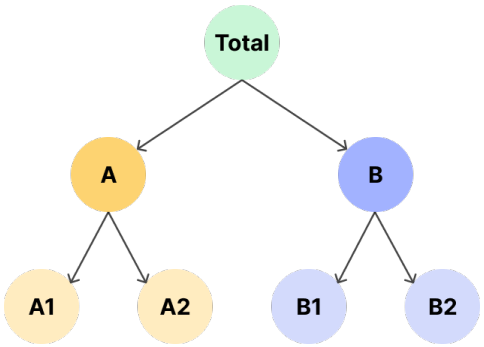

Figure 1: A 2-level hierarchical tree structure

where $\boldsymbol{S} \in \mathbb{R}^{n \times n_b}$ is a summing matrix that aggregates the bottom-level to all-level series. The summing matrix $\boldsymbol{S}$ for the tree structure in Figure 1 is:

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & & \boldsymbol{I_4} & \end{bmatrix}.$$

The matrix $\boldsymbol{S}$ encodes the aggregation constraints implied by the structure. Hence, the columns of $\boldsymbol{S}$ span a linear subspace. Any observation $\boldsymbol{y}_t$ that lies inside this subspace is called *coherent*, while those outside are *incoherent*. We refer to the subspace spanned by $\boldsymbol{S}$ as the *coherent subspace* $\mathfrak{s} \in \mathbb{R}^{n_b}$.

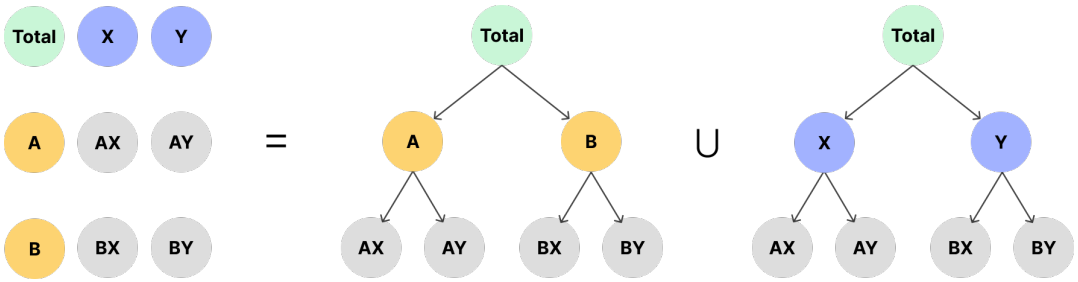

Figure 2: A 2-level grouped structure, which can be considered as the union of two hierarchical trees with common top and bottom level series

This representation extends beyond hierarchical (nested) structures. When attributes of interest are crossed, such as the company sales at any aggregation level (company-wise, city-wise, or outlet-wise) is also considered by kinds of products, the structure is described as a *grouped structure*. In grouped systems, as illustrated in Figure 2, aggregation and disaggregation paths

are not unique, but the linear constraints can still be written compactly through a summing matrix $\boldsymbol{S}$. For simplicity, we refer to both structures as hierarchical structure and distinguish between them when needed.

In practice, when we produce forecasts for each individual series, referred to as *base forecasts* $\hat{\boldsymbol{y}}_{t+h|t}$, they generally violate the aggregation constraints, and thus are incoherent. Coherency can be restored by linearly projecting the base forecasts onto the coherent subspace $\mathfrak{s}$ using a projection matrix $\boldsymbol{P}$: $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{P}\hat{\boldsymbol{y}}_{t+h|t}$, where $\tilde{\boldsymbol{y}}_{t+h|t}$ are the *reconciled forecasts*.

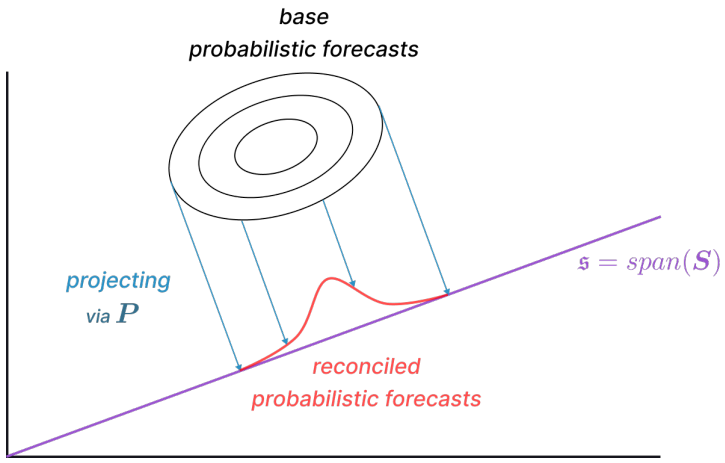

Figure 3: Geometry of probabilistic forecast reconciliation. The base forecast distribution is projected orthogonally onto the coherent subspace (purple line), resulting in the reconciled forecast distribution (red). The projection is defined by the projection matrix $\boldsymbol{P}$. Note that this figure is schematic since most applications are high-dimensional.

Many existing reconciliation methods including the OLS (Hyndman et al., 2011), WLS (Hyndman et al., 2016), and MinT (Wickramasuriya et al., 2019) express the projection matrix as $\boldsymbol{P} = \boldsymbol{SG}$, for a suitable $n_b \times n$ mapping matrix $\boldsymbol{G}$. The idea is to map the base forecasts of all levels $\hat{\boldsymbol{y}}_{t+h|t}$ down into the bottom level, which is then aggregated to the higher levels by $\boldsymbol{S}$. Since the projection matrix $\boldsymbol{P}$ is idempotent, $\boldsymbol{G}$ must satisfy the condition $\boldsymbol{SGS} = \boldsymbol{S}$. Within this class, a broad family of mapping matrix is given by: $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{M}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{M}^{-1}$, for some positive definite matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ (Gamakumara, 2020).

When setting $\boldsymbol{M} = \boldsymbol{I}_n$, the identity matrix, this reduces to the OLS reconciliation, which corresponds to an orthogonal projection onto the coherent subspace (similar to Figure 3). A schematic illustration of this projection is depicted in Figure 3. Reconciliation takes the (possibly elliptical) base forecast distribution and projects it onto the coherent subspace, resulting in the reconciled forecast distribution. The choice of $\boldsymbol{M}$ determines the direction of the projection, which can yield oblique projection to deliver better-performing forecasts.

2.2 The Minimum Trace (MinT) Reconciliation

Wickramasuriya et al. (2019) showed that by setting $\boldsymbol{M} = \boldsymbol{W}_h = \mathbb{E}(\hat{\boldsymbol{e}}_{t+h|t}\ \hat{\boldsymbol{e}}'_{t+h|t})$, the covariance matrix of the h-step-ahead base forecast errors $\hat{\boldsymbol{e}}_{t+h|t} = \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}$, we essentially minimise the total variance of the reconciled forecast errors across all series. Equivalently, MinT is the unique linear-unbiased reconciler that minimises the trace of $\text{Var}[y_{t+h} - \tilde{y}_{t+h|t}] = \boldsymbol{S}\boldsymbol{G}_h\boldsymbol{W}_h\boldsymbol{G}'_h\boldsymbol{S}'$. This method is thus called Minimum Trace (MinT) reconciliation. The matrix $\boldsymbol{G}_h$ is thus given by:

$$\boldsymbol{G}_h = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1},$$

provided that $\boldsymbol{W}_h$ is positive definite. Intuitively, $\boldsymbol{W}_h^{-1}$ down-weights the base forecasts of series with high uncertainty, and up-weights those with low uncertainty, when mapping to the bottom level.

Although originally developed for point reconciliation, Wickramasuriya (2024) showed that MinT extends naturally to the probabilistic setting when base predictive distributions are Gaussian. If h-step-ahead base forecast distribution be $\hat{\boldsymbol{y}}_{t+h|t} \sim \mathcal{N}(\hat{\boldsymbol{y}}_{t+h|t}, \boldsymbol{W}_h)$, then reconciliation gives $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{y}}_{t+h|t} \sim \mathcal{N}(\tilde{\boldsymbol{y}}_{t+h|t}, \boldsymbol{S}\boldsymbol{G}_h\boldsymbol{W}_h\boldsymbol{G}'_h\boldsymbol{S}')$. Within the class of linear-unbiased reconcilers, MinT also minimises the negative log score (equivalently, the Gaussian log predictive score) of the reconciled distribution.

In this paper we adopt this Gaussian framework to compare covariance estimators for MinT in both point and probabilistic reconciliation. It is also worth to mention that the methods can be extended to non-Gaussian settings by bootstrapping (Gamakumara, 2020; Panagiotelis et al., 2023), which is beyond the scope of this paper.

2.3 Shrinkage Estimator for MinT

The performance of MinT hinges on a reliable, positive-definite estimate of $\boldsymbol{W}_h$, which comes in both the mapping matrix $\boldsymbol{G}_h$ and the reconciled forecast variance $\boldsymbol{S}\boldsymbol{G}_h\boldsymbol{W}_h\boldsymbol{G}'_h\boldsymbol{S}'$.

However, the covariance matrix $\boldsymbol{W}_h$ is often not available in closed-form, and is challenging to estimate in high-dimensional setting where the number of series n is larger than the time dimension T. To tackle this issue, the original paper Wickramasuriya et al. (2019) assumed a proportionality relationship between $\hat{\boldsymbol{W}}_h^g = k_h g(\hat{\boldsymbol{W}}_1)$, where $\hat{\boldsymbol{W}}_1$ is the covariance matrix of the in-sample 1-step-ahead base residuals (to approximate $\boldsymbol{W}_1$) and $k_h > 0$ is a scaling constant

(which will be algebraically cancelled out in point-forecast reconciliation). The function $g(.)$ is a covariance estimator that produces a positive-definite matrix, the main focus of this paper.

The recommended choice for $g(.)$ in the original work is the shrinkage estimator with diagonal target from Schäfer & Strimmer (2005):

$$\hat{\boldsymbol{W}}_1^S = \lambda_S \operatorname{diag}(\hat{\boldsymbol{W}}_1) + (1 - \lambda_S)\hat{\boldsymbol{W}}_1 \,,$$

where $\operatorname{diag}(\hat{\boldsymbol{W}}_1)$ comprises only the diagonal elements of $\hat{\boldsymbol{W}}_1$. We refer to any $\lambda_S \in [0, 1]$ as the shrinkage intensity of the shrinkage estimator. This approach shrinks the covariance matrix $\hat{\boldsymbol{W}}_1$ towards its diagonal matrix, meaning the off-diagonal elements are shrunk towards zero while the diagonal ones remain unchanged.

Schäfer & Strimmer (2005) also provided an closed-form estimate of the optimal shrinkage intensity parameter λ_S:

$$\hat{\lambda}_S = \frac{\sum_{i \neq j} \widehat{Var}(\hat{r}_{ij})}{\sum_{i \neq j} \hat{r}_{ij}^2} \,,$$

where $\hat{r}_{ij}$ is the i, j-th element of $\hat{\boldsymbol{R}}_1$, the 1-step-ahead sample correlation matrix (obtained from $\hat{\boldsymbol{W}}_1$). The optimal estimate is obtained by minimising $MSE(\hat{\boldsymbol{W}}_1) = Bias(\hat{\boldsymbol{W}}_1)^2 + Var(\hat{\boldsymbol{W}}_1)$. More specifically, we trade a small bias for a substantial variance reduction, which is especially valuable in high dimension.

Despite its simplicity and guaranteed positive-definiteness, MinT coupled with diagonal-target shrinkage presents three important limitations.

Problem 1: Uniform shrinkage

Linear shrinkage shrinks all off-diagonal elements towards zeros with equal weights λ_S. The resulting penalty is global and non-adaptive: strong, genuinely systematic correlations are shrunk at the same rate as weak, noisy ones. In hierarchical and grouped systems, this can be problematic. Aggregation naturally induces stronger dependence among related nodes (for example, a region and a neighbouring region, or a region and its parent state), while many unrelated pairs exhibit near-zero correlations. A uniform penalty may over-shrink informative co-movements and under-shrink idiosyncratic noise, reducing reconciliation efficiency.

Problem 2: Latent factors

Many real-world hierarchical data sets exhibit a prominent low-rank (factor) structure.

In Australian domestic tourism, for example, national and state-level movements often load on a small number of common components, with residual dependence concentrated in local idiosyncrasies. A scree plot of the one-step residual covariance typically shows a marked elbow after a handful of principal components, indicating strong latent structure. Diagonal-target shrinkage is factor-unaware: it neither preserves the low-rank common subspace nor differentially shrinks the idiosyncratic remainder, and so can be inefficient when common factors dominate.

Taking an example of the Australian domestic overnight trips data set (*Tourism Research Australia*, 2024), where the national trips are disaggregated into states and territories, and further into regions. The tourism activities might be driven by a few common factors, such as economic conditions, fuel prices, or major events, which might be left in the residuals after fitting the base models. From Figure 4, a scree plot of the one-step residual covariance $\hat{\boldsymbol{W}}_1$ shows a marked elbow after a handful of principal components (largest eigenvalues), indicating strong latent structure.

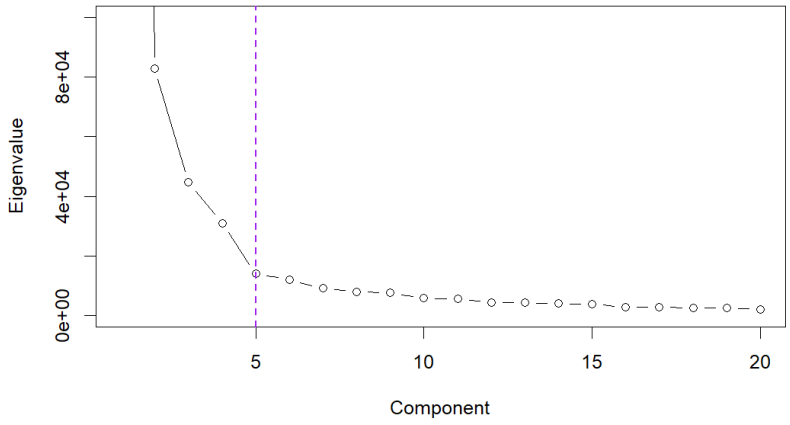

Figure 4: Twenty largest eigenvalues of one-step-ahead in-sample base forecast error covariance, Australian domestic overnight trips. The point of inflection occurs around the fifth largest eigenvalue.

Problem 3: Proportional scaling for h-step-ahead

The proportionality relationship $\hat{\boldsymbol{W}}_h^g = k_h g(\hat{\boldsymbol{W}}_1)$ enforces horizon-invariant cross-series dependence of the forecast errors up to a scalar factor. In many applications, the shape of the error covariance might change with horizon due to evolving error dynamics or horizon-specific interactions among series. Proportional scaling may therefore misrepresent multi-step dependence.

Additionally, while the scalar factor cancels in point reconciliation, it directly controls dispersion in probabilistic reconciliation and can lead to miscalibrated predictive distributions. However, we will not explore this issue in this paper, and leave it for future research.

In the next sections, we introduce alternative covariance estimators that address these limitations individually and in combination.

3 Covariance Estimation Approaches

3.1 NOVELIST Estimator

NOVELIST (NOVEL Integration of the Sample and Thresholded Covariance) estimator, proposed by Huang & Fryzlewicz (2019), introduces a parameter to control the sparsity level in the target toward which the sample matrix is shrunk. By soft-thresholding small correlations before shrinkage, NOVELIST preserves strong signals while attenuating weak, noisy ones, addressing the uniform, non-adaptive nature of diagonal-target shrinkage.

The construction proceeds in two steps. First, apply elementwise soft-thresholding to the sample correlation matrix; second, shrink the sample correlation toward this thresholded target. Working in correlation space avoids rescaling issues and keeps diagonal entries at one. The method introduces an extra parameter, the threshold δ, which is used to control the amount of soft-thresholding. The NOVELIST estimator for covariance matrix is given by:

$$\hat{\boldsymbol{W}}_1^N = \lambda_\delta \hat{\boldsymbol{W}}_{1,\delta} + (1 - \lambda_\delta)\hat{\boldsymbol{W}}_1, \tag{1}$$

where $\hat{\boldsymbol{W}}_{1,\delta}$ is the thresholded version of $\hat{\boldsymbol{W}}_1$. By convenient setting as discussed above, we rewrite it in terms of correlation:

$$\hat{\boldsymbol{R}}_1^N = \lambda_\delta \hat{\boldsymbol{R}}_{1,\delta} + (1 - \lambda_\delta)\hat{\boldsymbol{R}}_1, \tag{2}$$

where, $\hat{\boldsymbol{R}}_{1,\delta}$ is the thresholded correlation matrix, where each element is regularised by:

$$\hat{r}^{\delta}_{1,ij} = \text{sign}(\hat{r}_{1,ij}) \max(|\hat{r}_{1,ij}| - \delta,\ 0), \tag{3}$$

where $\delta \in [0, 1]$ is the threshold parameter. For a given threshold δ, Huang & Fryzlewicz (2019) derived an analytical expression for the optimal shrinkage intensity parameter $\lambda(\delta)$, following similar logic to Schäfer & Strimmer (2005). It can be computed as:

$$\hat{\lambda}(\delta) = \frac{\sum_{i \neq j} \widehat{Var}(\hat{r}_{1,ij})\ \mathbf{1}(|\hat{r}_{1,ij}| \leq \delta)}{\sum_{i \neq j} (\hat{r}_{1,ij} - \hat{r}^{\delta}_{1,ij})^2}, \tag{4}$$

where $\mathbf{1}(.)$ is the indicator function.

On the other hand, the optimal threshold $\hat{\delta}$ does not have a closed-form solution, and is typically obtained by rolling-window cross-validation procedure. The idea is to find the threshold δ^*, with the corresponding λ^* and $\hat{\boldsymbol{R}}^{N}_{1}(\delta^*, \lambda^*)$, that minimises the average out-of-sample 1-step-ahead reconciled forecast mean squared error over all windows. The formal algorithm is given in Section 3.1.1.

Note that when $\delta \in \left[\max_{i \neq j} |\hat{r}_{1,ij}|,\ 1\right]$, the NOVELIST estimator collapses to the shrinkage estimator, and when $\delta = 0$, it becomes the sample covariance matrix. An additional concern is that the estimator does not guarantee to be positive definite, but we can use Higham (2002) algorithm to compute the nearest positive definite matrix if needed.

3.1.1 NOVELIST cross-validation algorithm

It is only required to fit the base models once on the whole training data $\{\boldsymbol{y}_t\}_{t=1}^{T}$, and obtain the in-sample fitted values $\{\hat{\boldsymbol{y}}_t\}_{t=1}^{T}$.

Algorithm 1 Cross-validation procedure

Input: Observations and fitted values $\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t \in \mathbb{R}^n$ for $t = 1, \ldots, T$, set of threshold candidates Δ, window size v.
$\hat{\boldsymbol{e}}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ for $t = 1, \ldots, T$
for $i = v : T - 1$ **do**
    $j = i - v + 1$
    $\hat{\boldsymbol{W}}_j = \frac{1}{v} \sum_{t=j}^{i} \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}_t'$
    $\hat{\boldsymbol{D}}_j = \operatorname{diag}(\hat{\boldsymbol{W}}_j)$
    $\hat{\boldsymbol{R}}_j = \hat{\boldsymbol{D}}_j^{-1/2} \hat{\boldsymbol{W}}_j \hat{\boldsymbol{D}}_j^{-1/2}$
    for $\delta \in \Delta$ **do**
        Compute thresholded correlation $\hat{\boldsymbol{R}}_{j,\delta}$ using Equation 5
        Compute $\hat{\lambda}_{j,\delta}$ using Equation 6
        Compute $\hat{\boldsymbol{R}}_{j,\delta}^{N}$ using Equation 4
        $\hat{\boldsymbol{W}}_{j,\delta}^{N} = \hat{\boldsymbol{D}}_j^{1/2} \hat{\boldsymbol{R}}_{j,\delta}^{N} \hat{\boldsymbol{D}}_j^{1/2}$
        $\boldsymbol{G} = (\boldsymbol{S}' \hat{\boldsymbol{W}}_{j,\delta}^{N^{-1}} \boldsymbol{S})^{-1} \boldsymbol{S}' \hat{\boldsymbol{W}}_{j,\delta}^{N^{-1}}$
        Reconciled forecasts $\tilde{\boldsymbol{y}}_{i+1|\delta} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{i+1}$
        $\tilde{\boldsymbol{e}}_{i+1|\delta} = \boldsymbol{y}_{i+1} - \tilde{\boldsymbol{y}}_{i+1|\delta}$
    end for
end for
$\mathrm{MSE}_\delta = \frac{1}{T-v} \sum_{i=v}^{T-1} (\tilde{\boldsymbol{e}}_{i+1|\delta})^2$ for each $\delta \in \Delta$
$\hat{\delta}^* = \arg\min_{\delta \in \Delta} \mathrm{MSE}_\delta$
Compute $\hat{\lambda}^*$ on all training data using $\hat{\delta}^*$
Compute $\hat{\boldsymbol{R}}_1^*$ using $\hat{\delta}^*$ and $\hat{\lambda}^*$ on all training data, using Equation 3
Output: Estimate of optimal $\hat{\delta}^*$

Remark. Minimising a multi-step-ahead forecast error metric in the cross-validation procedure oftern yields a $\hat{\delta}^*$ that is close to the 1-step-ahead case.

3.2 PC-adjusted Estimator

To exploit latent common components explicitly, the PC-adjusted method takes the latent factors directly into its construction. It starts by decomposing the covariance matrix $\hat{\boldsymbol{W}}_1$ into a prominent principle components part (low-rank) and a orthogonal complement part $\hat{\boldsymbol{W}}_1^K$ (the correlation matrix after removing the first K principal components). Then we can apply either shrinkage or NOVELIST estimator to $\hat{\boldsymbol{W}}_1^K$:

$$\hat{\boldsymbol{W}}_1^{g,K} = \sum_{k=1}^{K} \hat{\gamma}_k \hat{\boldsymbol{\xi}}_k \hat{\boldsymbol{\xi}}_k' + g(\hat{\boldsymbol{W}}_1^K)$$

where $g(.)$ is either the shrinkage or NOVELIST estimator, $\hat{\gamma}_k$ and $\hat{\boldsymbol{\xi}}_k$ are the k-th largest eigenvalue and the corresponding eigenvector of the sample covariance matrix, respectively. The number of principal components K can be selected using information criteria.

Similar to the NOVELIST estimator, its PC-adjusted variant $\hat{\boldsymbol{W}}_1^{N,K}$ requires a cross-validation procedure to select the threshold paramter and adjustment to obtain positive definiteness.

3.3 Scaled Variance

To address the potential issue of the proportionality relationship $\hat{\boldsymbol{W}}_h^g = k_h g(\hat{\boldsymbol{W}}_1)$ not holding in practice. A simple relaxation retains the 1-step-ahead correlation shape but allows the variances to scale differently with horizon. The scaled variance estimator is given by:

$$\hat{\boldsymbol{W}}_h^{g,sv} = \boldsymbol{D}_h^{1/2} g(\hat{\boldsymbol{R}}_1) \boldsymbol{D}_h^{1/2},$$

where $\boldsymbol{D}_h = \text{diag}(\hat{\sigma}_{1,h}^2, \dots, \hat{\sigma}_{n,h}^2)$, and $\hat{\sigma}_{i,h}^2$ is the variance of the i-th series' h-step-ahead base forecast errors. Similarly, $g(.)$ is either the shrinkage or NOVELIST estimator.

When NOVELIST is used to produce h-step-ahead reconciled forecasts, the cross-validation procedure is slightly modified to evaluate the out-of-sample reconciled forecast MSE using $\hat{\boldsymbol{W}}_h^{N,sv}$ instead of $\hat{\boldsymbol{W}}_1^N$.

3.4 Constructing from h-step-ahead Residuals

Another alternative is to directly estimate the covariance matrix from the h-step-ahead base forecast errors, without assuming any proportionality relationship:

$$\hat{\boldsymbol{W}}_h^g = g(\hat{\boldsymbol{W}}_h)$$

where $\hat{\boldsymbol{W}}_h$ is the covariance matrix of in-sample h-step-ahead base forecast residuals, and $g(.)$ is either the shrinkage or NOVELIST estimator.

Similar to the scaled variance approach, if NOVELIST is used, the cross-validation procedure is modified to compute $\hat{\boldsymbol{W}}_h^N$ and evaluate the out-of-sample reconciled forecast MSE using it.

Summary of MinT with Covariance Estimators

The covariance estimators explored in this paper are summarised in the table below. The abbreviations will be used in the following sections.

Table 1: Summary of covariance estimators for MinT reconciliation

Covariance estimators used	Abbreviation
Shrinkage	MinT-S
NOVELIST	MinT-N
PC-adjusted Shrinkage with K PCs	MinT-S(PCK)
PC-adjusted NOVELIST with K PCs	MinT-N(PCK)
Scaled Variance Shrinkage	MinT-S(sv)
Scaled Variance NOVELIST	MinT-N(sv)
Constructed from h-step-ahead Shrinkage	MinT-S(hcov)
Constructed from h-step-ahead NOVELIST	MinT-N(hcov)

4 Evaluation of Point and Probabilistic Forecasts

This section briefly introduces the scoring rules used to evaluate the point and probabilistic forecast accuracy in Section 5 and Section 6.

For point forecasts, we use the mean squared error (MSE) to evaluate the accuracy of different reconciliation methods: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{i,t+h} - \tilde{y}_{i,t+h|t})^2$, where $y_{i,t+h}$ is the realised value of series i at time $t+h$, and $\tilde{y}_{i,t+h|t}$ is the reconciled point forecast.

Meanwhile, to assess the quality of the probabilistic forecasts, it is common to use proper scoring rules. A scoring rule is a function $S(.,.)$ taking a predictive distribution as its first argument and a realisation as its second argument, then returns a numerical score. We follow a convention that lower scores are better. A scoring rule is said to be *proper* if $\mathbb{E}_Q[S(Q,y)] \leq \mathbb{E}_Q[S(F,y)]$ for all F, where F is a predictive distribution produced by forecasting model, Q is the true distribution of the realisation y, and $\mathbb{E}_Q$ is the expectation with respect to Q. Hence, the expected score is minimised when the forecast distribution matches the true distribution.

We employ Winkler score and continuous ranked probability score as our univariate scoring rules, and energy score as the multivariate scoring rule. All three are proper scoring rules. In this paper, we only evaluate 1-step-ahead probabilistic forecasts, and thus we drop the subscript t and h for simplicity.

Winkler score (WS). If the $100(1-\alpha)\%$ prediction interval of i-th series is $[l_i, u_i]$ (the $\alpha/2$ and $1-\alpha/2$ quantiles), then the Winkler score is defined as:

$$WS_\alpha(l_i, u_i; y_i) = (u_i - l_i) + \frac{2}{\alpha}(l_i - y_i)\mathbf{1}(y_i < l_i) + \frac{2}{\alpha}(y_i - u_i)\mathbf{1}(y_i > u_i),$$

where y_i is the observed value of i-th series, and $\mathbf{1}(.)$ is the indicator function. The Winkler score rewards narrow intervals that contain the observation, and penalises intervals that do not contain the observation.

Continuous ranked probability score (CRPS). The CRPS is defined as the squared difference between the predictive cumulative distribution function (CDF) F_i and the empirical CDF of the observation y_i of series i:

$$CRPS(F_i, y_i) = \int_{-\infty}^{\infty} (F_i(x) - \mathbf{1}(x \geq y_i))^2 dx.$$

When the predictive distribution is Gaussian with mean μ_i and standard deviation σ_i, the CRPS has a closed-form expression:

$$CRPS(F_i, y_i) = \sigma_i \left[z_i \left(2\Phi(z_i) - 1\right) + 2\phi(z_i) - \frac{1}{\sqrt{\pi}} \right],$$

where $z_i = \frac{y_i - \mu_i}{\sigma_i}$, and $\Phi(.)$ and $\phi(.)$ are the CDF and probability density function (PDF) of a standard normal distribution, respectively.

Energy score (ES). The energy score is a multivariate generalisation of the CRPS. It is defined as:

$$ES(F, \boldsymbol{y}) = \mathbb{E}_F ||\boldsymbol{X} - \boldsymbol{y}||^\beta - \frac{1}{2}\mathbb{E}_F ||\boldsymbol{X} - \boldsymbol{X}'||^\beta,$$

where $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random vectors with multivariate distribution F, $\boldsymbol{y}$ is the observed vector, $||.||$ is the Euclidean norm, and $\beta \in (0, 2]$. We set $\beta = 1$ following common convention.

Since the closed-form expression of the ES may not available, we approximate it using Monte Carlo samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ drawn from P:

$$\widehat{ES}(F, \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} ||\boldsymbol{x}_m - \boldsymbol{y}|| - \frac{1}{2M(M-1)} \sum_{m=1}^{M} ||\boldsymbol{x}_m - \boldsymbol{x}_m^*||,$$

where $\boldsymbol{x}_m^*$ is a randomly selected sample from $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \setminus \{\boldsymbol{x}_m\}$. In our experiments, we use $M = 10000$ samples to approximate the ES.

5 Simulation

5.1 General Design

The general design of data generating process for bottom-level series is a stationary VAR(1) process, with the following structure:

$$\boldsymbol{b}_t = \boldsymbol{A}\boldsymbol{b}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{A}$ is a $n_b \times n_b$ block diagonal matrix of autoregressive coefficients $\boldsymbol{A} = diag(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_M)$, with each $\boldsymbol{A}_m$ being a $n_{b,m} \times n_{b,m}$ matrix. The block diagonal structure ensures that the time series are grouped into M groups (blocks), with each group having its own autoregressive coefficients. This aims to simulate the interdependencies between the time series within each group, where reconciliation will be expected to better capture these compared to independent base forecasts.

$\boldsymbol{\epsilon}_t$ is a Gaussian innovation process, with covariance matrix $\boldsymbol{\Sigma}$. The covariance matrix $\boldsymbol{\Sigma}$ is generated specifically using the Algorithm 1 in Hardin et al. (2013):

1. A compound symmetric correlation matrix is used for each block of size $n_{b,m}$ in $\boldsymbol{A}_m$, where the correlation entries ρ_j for each block m are sampled from a uniform distribution between 0 and 1. They are within group baseline correlations.
2. A constant correlation, which is smaller than $\min\{\rho_1, \rho_2, \ldots, \rho_M\}$, is imposed on the entries between different blocks. It serves as between group baseline correlations.

3. We select a constant ε such that $0 \leq \varepsilon < 1 - \max\{\rho_1, \rho_2, \ldots, \rho_M\}$ to control the noise intensity. We then generate n_b unit vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_b}$ and obtain a entry-wise noise element as $\varepsilon \boldsymbol{u}_i' \boldsymbol{u}_j$ for the i, j-th entry. These entry-wise noise elements are added on top of the baseline correlation matrix (except the diagonal entries), creating a noisy positive correlation matrix $\boldsymbol{R}^+$.

4. The covariance matrix $\boldsymbol{\Sigma}^+$ is then constructed by $\boldsymbol{\Sigma}^+ = \boldsymbol{D}\boldsymbol{R}^+\boldsymbol{D}$, where $\boldsymbol{D} = diag(\sigma_1, \ldots, \sigma_{n_b})$ is a diagonal matrix of standard deviations. Each σ_i is uniformly sampled from the range of $[\sqrt{2}, \sqrt{6}]$, for all n_b series.

These steps result in a positive covariance matrix $\boldsymbol{\Sigma}^+$ where all the elements are positive. To allow a mixture of positive and negative covariance, we randomly flip the signs of them but need to maintain the positive definiteness. This can be done by pre- and post-multiplying $\boldsymbol{\Sigma}$ by a random diagonal matrix $\boldsymbol{V}$ with diagonal entries sampled from $\{-1, 1\}$ with equal probability. The final covariance matrix is given by $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{V}$.

For all hierarchies in our experiments, we simulate two panel lengths, $T = 54$ ("short") and $T = 304$ ("long"), reserving the final four observations as an out-of-sample test set. We perform 500 Monte Carlo replications for each configuration. In each replication, we fit univariate ARIMA models to the training observations using the automatic AICc minimization algorithm of Hyndman & Khandakar (2008), implemented in the *fabletools* package (O'Hara-Wild et al., 2024), generating 1–4-step base forecasts. We then reconcile the base forecasts using different reconciliation methods, and evaluate their point and probabilistic forecast accuracy on the test set.

5.2 Exploring Effects of Hierarchy's Size

In our main experiments, we examine how MinT combined with the different estimators perform as the hierarchy expands. We generate synthetic data from the VAR(1) framework described earlier, varying the number of bottom-level series, n_b, across two structures: a "small" structure with six groups of six bottom-level series ($n_b = 6\text{x}6 = 36$), and a "large" configuration with two groups of fifty bottom-level series ($n_b = 2\text{x}50 = 100$).

In the 36-series case, each block of six forms a level-1 aggregate, and those six aggregates form the total. The 100-series design employs a deliberately intricate aggregation path to stress-test reconciliation methods. We first sum the one hundred bottom series into ten intermediate series by grouping them in contiguous blocks of ten. These ten series are then organised into three level-2 aggregates—four, three, and four series, respectively—before finally summing to a single top node. This asymmetric hierarchy creates overlapping correlation patterns: some

level-2 series share bottom-level groups, while others draw from both, emulating practical scenarios such as regional sales aggregations that span multiple product categories or overlapping territories. The aggregation paths for both structures are illustrated in Figure 5.

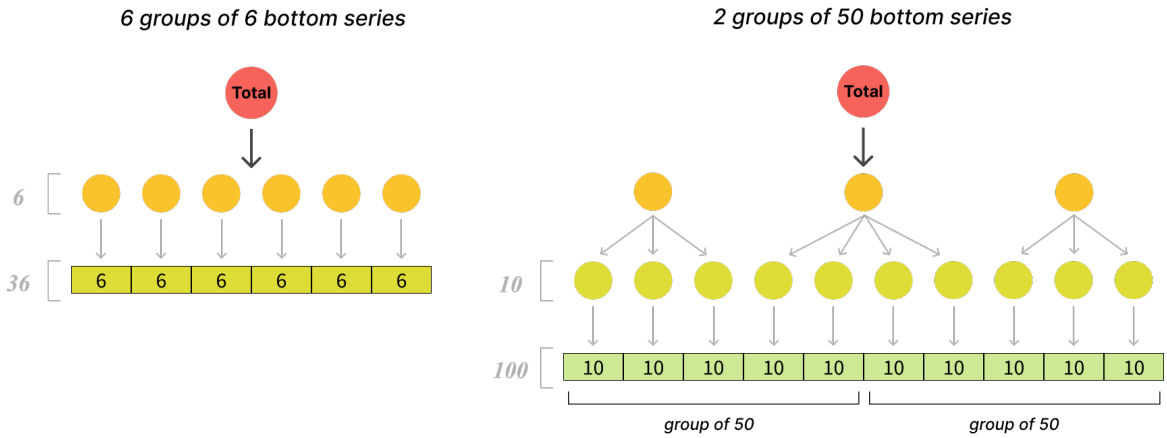

Figure 5: Aggregation structures used in the simulation experiments: 6 groups of 6 (left) and 2 groups of 50 (right)

The VAR(1) and correlation configurations for the 6 by 6 case and 2 by 50 case are illustrated in Figure 6 and Figure 7, respectively. The block diagonal structure of the VAR(1) coefficient matrices $\boldsymbol{A}$ reflects the grouping of series, and the correlation matrices- show higher correlations among series within the same group.

Figure Figure 8 illustrates the relative improvements in mean squared error (MSE) of reconciled forecasts over the incoherent base forecasts, across the two structures (small and large) and time series lengths (short and long). We evaluate ten different MinT variants based on the covariance estimators discussed in Table 1, with visual distinctions made through colors, line types, and point shapes. MinT with shrinkage (*MinT-S*) and its variants are colored in mint green, while MinT with NOVELIST (*MinT-N*) and its variants are in purple. The solid lines represent vanilla *MinT-S* and *MinT-N*; the dashed lines with dot points denote the PC-adjusted variants (e.g. *MinT-S(PC1)*, *MinT-N(PC2)*); and the dotted or dashed-dotted lines indicate the scaled variance and h-step-ahead residual versions (e.g. *MinT-S(SV)*, *MinT-N(hcov)*).

The first key observation is that methods with shrinkage slightly outperform those with NOVELIST across all scenarios, although the differences are small. Second, the PC-adjusted variants (using one and two principal components) do not yield improvements over the vanilla versions. This is expected since the synthetic data generating process does not simulate from strong latent factors. Lastly, the scaled variance and h-step-ahead residual approaches do

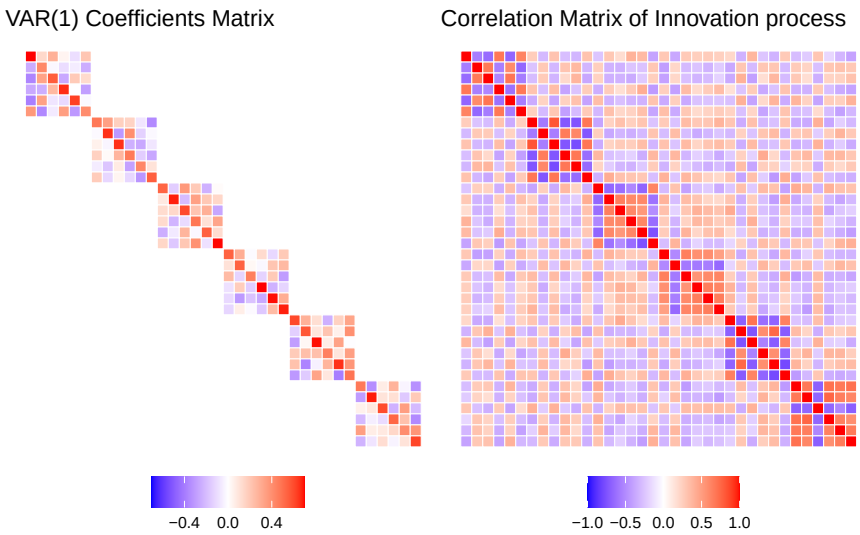

Figure 6: The VAR(1) coefficients matrix (left) and correlation matrix of the innovation process (right) for the 6 groups of 6 structure.

not enhance performance as the forecast horizon increases, suggesting that the proportionality assumption may not be severely violated in this VAR(1) setup.

When looking into each Monte Carlo replication, we find that there are instances where the NOVELIST estimator collapses to the shrinkage estimator due to a large optimal threshold $\hat{\delta}$ being selected in the cross-validation step, resulting in a diagonal shrinkage target.

Moving on to probabilistic forecasts, we evaluate the performance of reconciliation methods using the energy score, a proper scoring rule for multivariate predictive distributions. Figure 9 presents the percentage relative improvement in energy score for 1-step-ahead forecasts. Across both hierarchies and time dimensions, the MinT methods consistently outperform the base forecasts. Meanwhile, the differences among *MinT-S* and *MinT-N* variants are small, except for the 6 by 6 case with 50 observations, where shrinkage has a pronounced edge. The PC-adjusted variants again degrade performance as we add more principal components. The scaled variance and h-step-ahead residuals approaches are not available since we only evaluate 1-step-ahead covariance estimates.

Other scoring rules, such as the Winkler score and CRPS, yield similar conclusions.

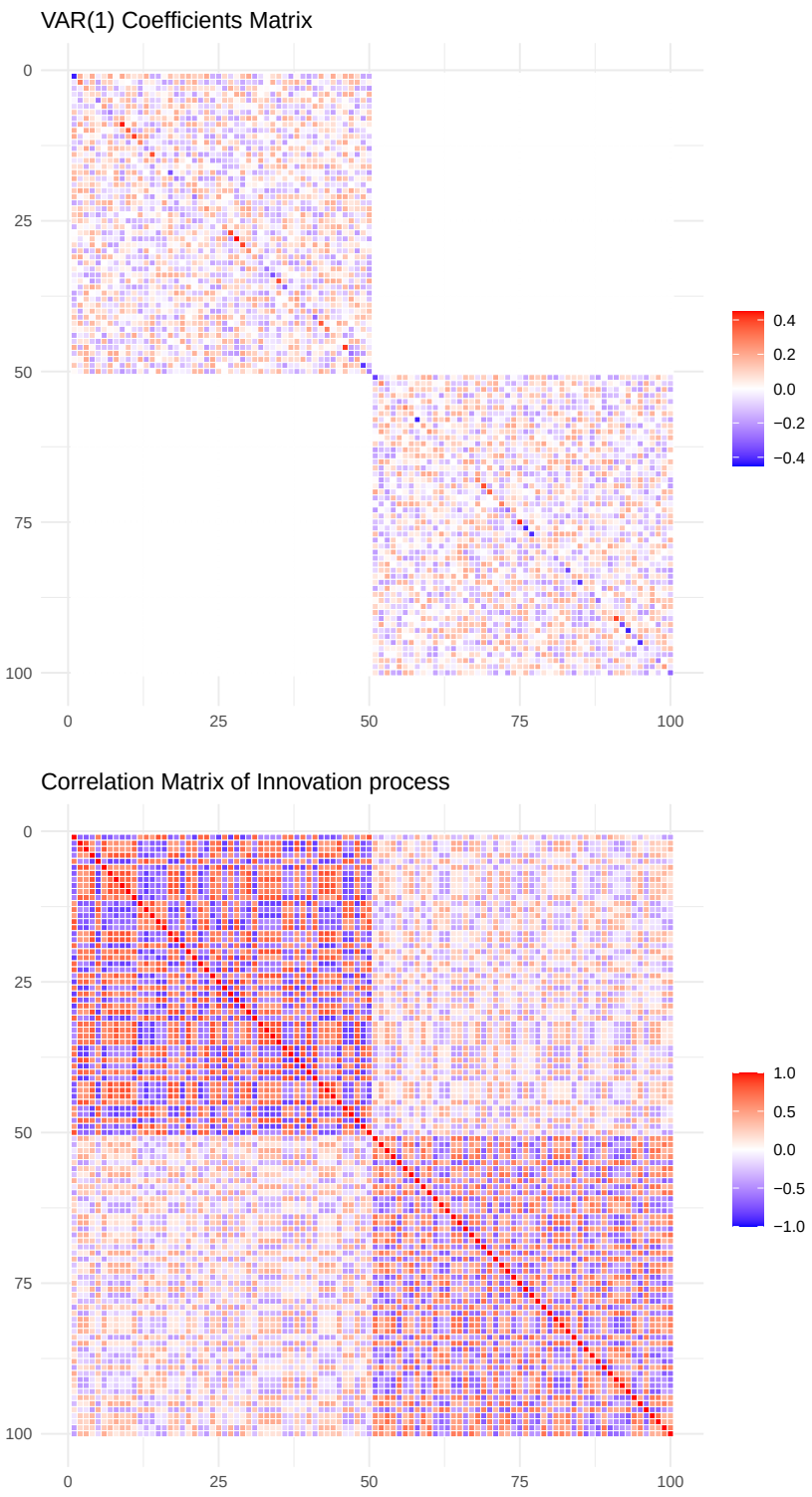

Figure 7: The VAR(1) coefficients matrix (top) and correlation matrix of the innovation process (bottom) for the 2 groups of 50 structure.

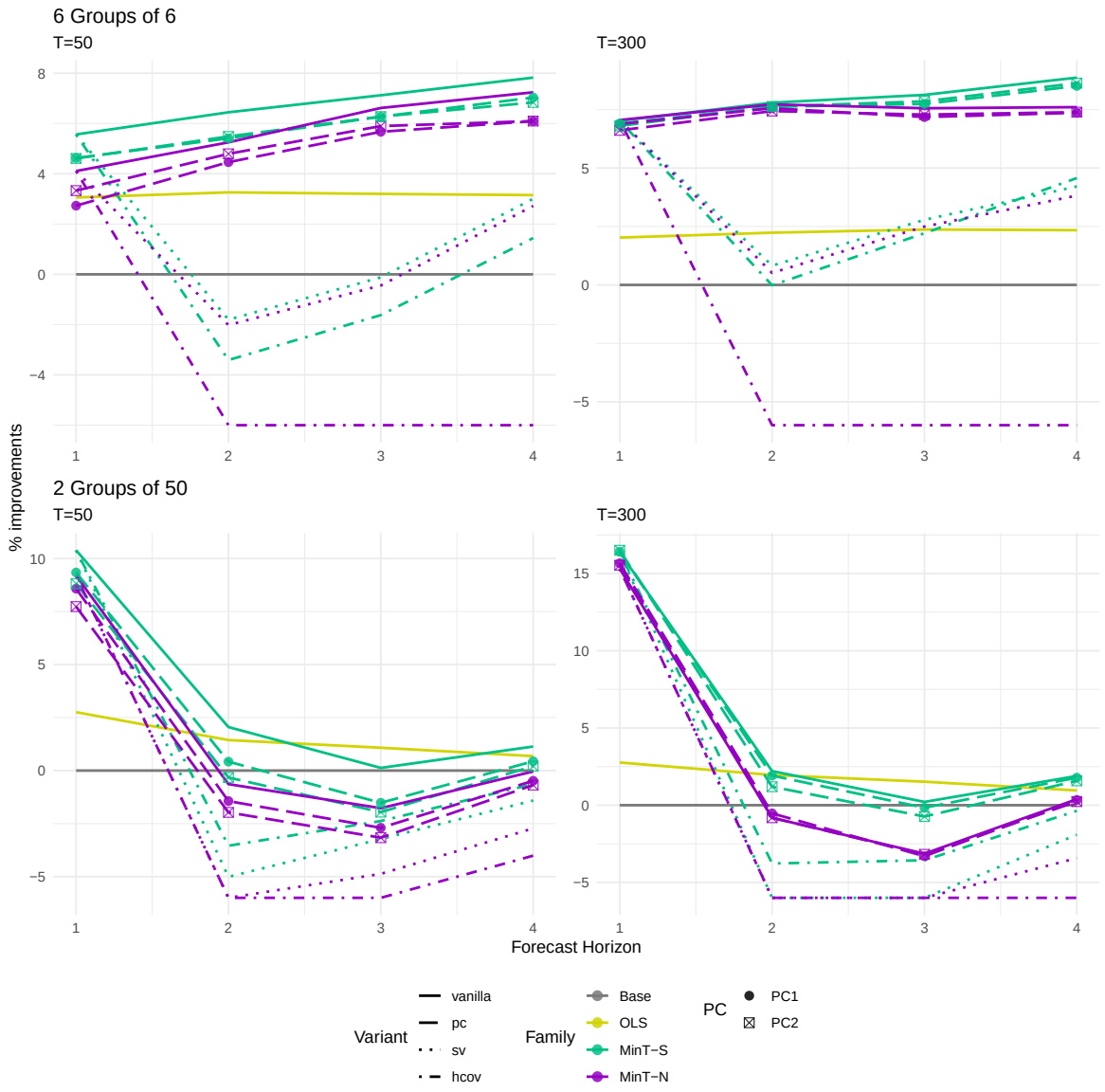

Figure 8: Percentage relative improvement in MSE of reconciled forecasts over the base forecasts in the 6 by 6 case (top row) and the 2 by 50 case (bottom row), T=50 (left column) and T=300 (right column), for 1- to 4-step-ahead forecasts. The positive (negative) entries indicate a decrease (increase) in MSE relative to base. The negative improvements beyond 6% are capped.

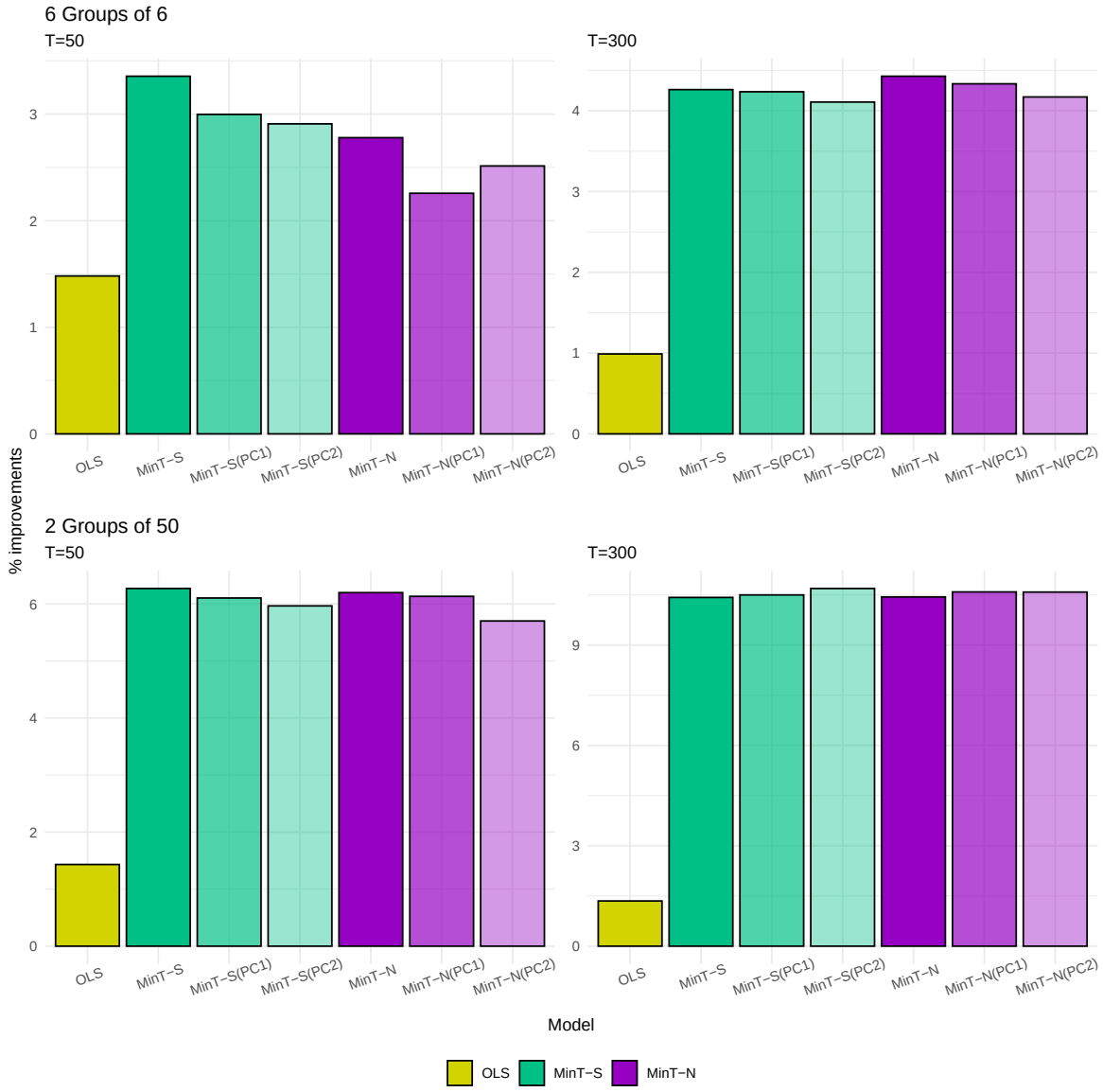

Figure 9: Percentage relative improvement in Energy score in both hierarchies and time dimensions, for 1-step-ahead forecasts. The positive entries indicate a decrease in Energy score relative to base.

5.3 Other Data Generating Processes

In attempts to differentiate the performance of NOVELIST from the shrinkage estimator, we also simulate from a sparse covariance matrix for the previous section's 2 groups of 50 bottom series setting, as illustrated in Figure 10. The sparse covariance matrix is obtained by randomly choosing 40% of the bottom series and setting their correlations with all other series to zero, resulting in a grid-like sparse structure. The VAR(1) coefficient matrix remains the same as in the dense case. The idea is to allow NOVELIST to exploit the sparsity in the covariance structure, since it can control the sparsity of the shrinkage target via the thresholding parameter δ. However, no profound insights can be drawn from the results.

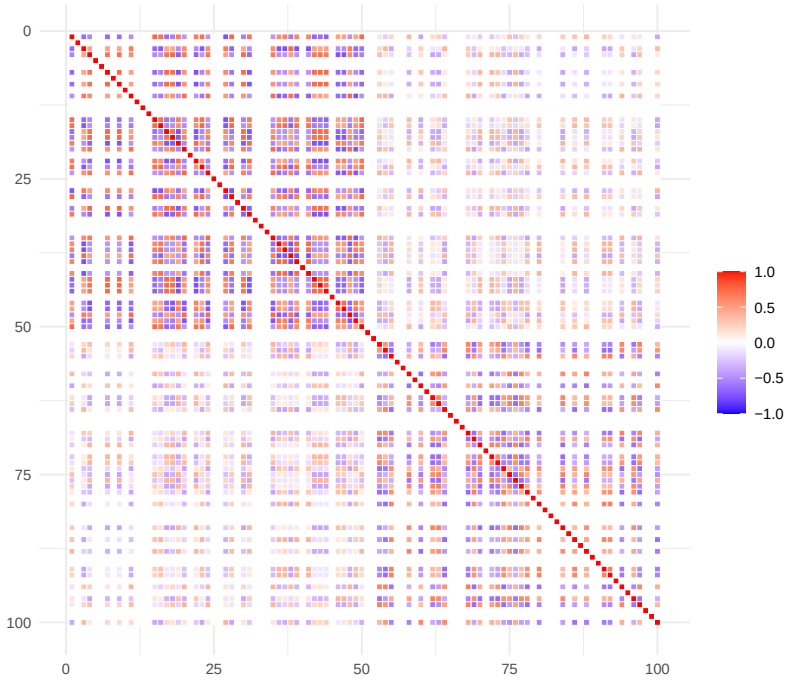

Figure 10: Sparse correlation matrix of the innovation process for 2 groups of 50 structure.

Additional designs (varying block sizes, grouped structure, aggregation paths, correlation configurations) also failed to separate NOVELIST from Shrinkage. Their nearly identical performance under these synthetic scenarios suggests that our current simulation may not unveil the full advantages of the thresholding estimators. Nevertheless, we have not explored settings where PC variants or using h-step-ahead residuals approaches would have an edge.

These findings motivates our turn to empirical data in the next section, where latent structural features, regime shifts, and noisy, intermittent series may possibly reveal performance

differences.

6 Forecasting Australian Domestics Tourism

Forecasting domestic tourism flows is of great economic and policy significance, as policymakers and stakeholders rely on accurate forecasts to make informed decisions regarding resource allocation, infrastructure development, and marketing strategies. The domestic tourism flows in Australia exhibit a natural hierarchical and grouped structure, driven both by geography and by purpose of travel. At the top of this hierarchy lies the national total, which splits into the seven states and territories. Each state is further sub-divided into tourism zones, which in turn break down into 77 regions. A complete illustration of this geographic hierarchy appears in Appendix Section 8.2. Intersecting this geographic hierarchy is a second dimension: policymakers are also interested in travel motives. This partitions tourism flows into four categories: holiday, business, visiting friends and relatives, and other. Altogether, this yields a grouped structure of 560 series, from the most disaggregated regional-purpose cells up to the full national aggregate. Table 2 depicts this structure.

Table 2: Hierarchical and grouped structure of Australian domestic tourism flows

Geographical division	Number of series per geographical division	Number of series per purpose	Total number of series
Australia	1	4	5
States	7	28	35
Zones	27	108	135
Regions	77	308	385
Total	112	448	560

We quantify tourism demand via "visitor nights", the total number of nights spent by Australians away from home. The data is collected via the National Visitor Survey, managed by Tourism Research Australia, using computer assisted telephone interviews from nearly 120,000 Australian residents aged 15 years and over (*Tourism Research Australia*, 2024).

The data are monthly time series spanning from January 1998 to December 2016, resulting in 228 observations per series. This gives a challenging high-dimensional forecasting problem with number of series (560) doubled the number of observations, which is ideal for evaluating reconciliation approaches that rely on high-dimensional covariance estimation. The extreme dimensionality over sample size mirrors many contemporary business problems, for instance,

Starbucks Corporation sales. Tourism demand is also economically vital yet highly volatile, with geographical and purpose-specific patterns create a realistic stress-test for reconciliation algorithms.

Figure 11: Rolling-window cross-validation scheme for evaluating forecasting performance in Australia tourism data

To assess forecasting performance between models, we adopt a rolling-window cross-validation scheme. Beginning with the first 120 monthly observations (January 1998-December 2005) as the initial training set, we obtain the best-fitted ARIMA model for each of the 560 series via the automatic algorithm by minimising AICc from Hyndman & Khandakar (2008), implemented in the *fabletools* package (O'Hara-Wild et al., 2024). The 1- to 12-step-ahead base forecasts are then generated by these ARIMA models, and then reconciled using multiple approaches. To estimate the NOVELIST and its variants, we implement an extra cross-validation prodecure within this training window, as described in Section 3.1.1. We then roll the training window forward by one month and refit all models, re-estimate reconciliation variants, and produce another batch of 1- to 12-step-ahead forecasts, repeating until the training set reaches December 2015. In total, this results in 97 out-of-sample windows. The entire procedure is illustrated in Figure 11.

Figure 12 reports percentage relative improvements in MSE of reconciled forecasts over the incoherent base ARIMA forecasts, across horizons from 1 to 12 months. We evaluate ten MinT variants (as defined in Table 1), similar to the simulation's Figure 8. The MinT with shrinkage (*MinT-S*) and its variants are colored in mint green, while MinT with NOVELIST (*MinT-N*) and its variants are in purple. Solid lines represent "vanilla" *MinT-S* and *MinT-N*; dashed lines with dot points denote the PC-adjusted variants (e.g. *MinT-S(PC1)*, *MinT-N(PC2)*); and

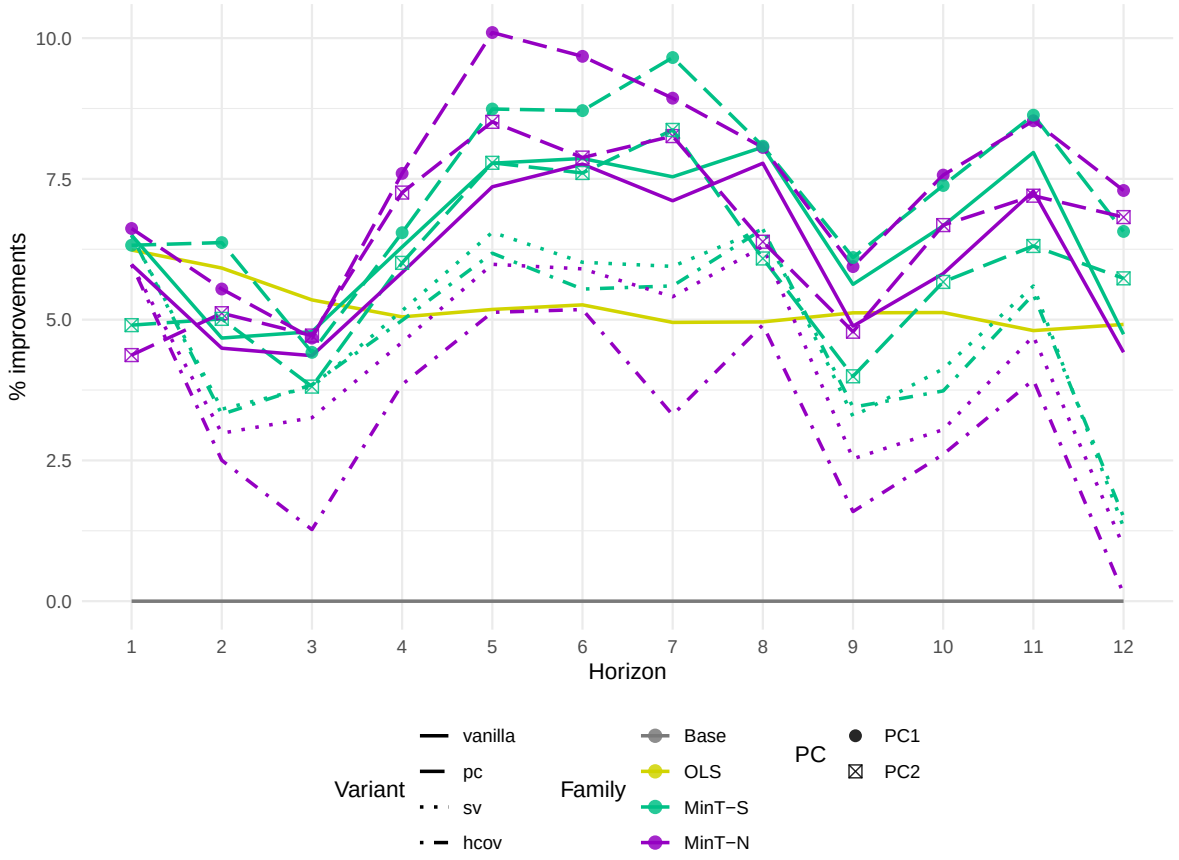

Figure 12: Percentage relative improvement in the mean squared error (MSE) of different reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1 to 12 steps ahead forecasts. The positive entries indicate an decrease in MSE.

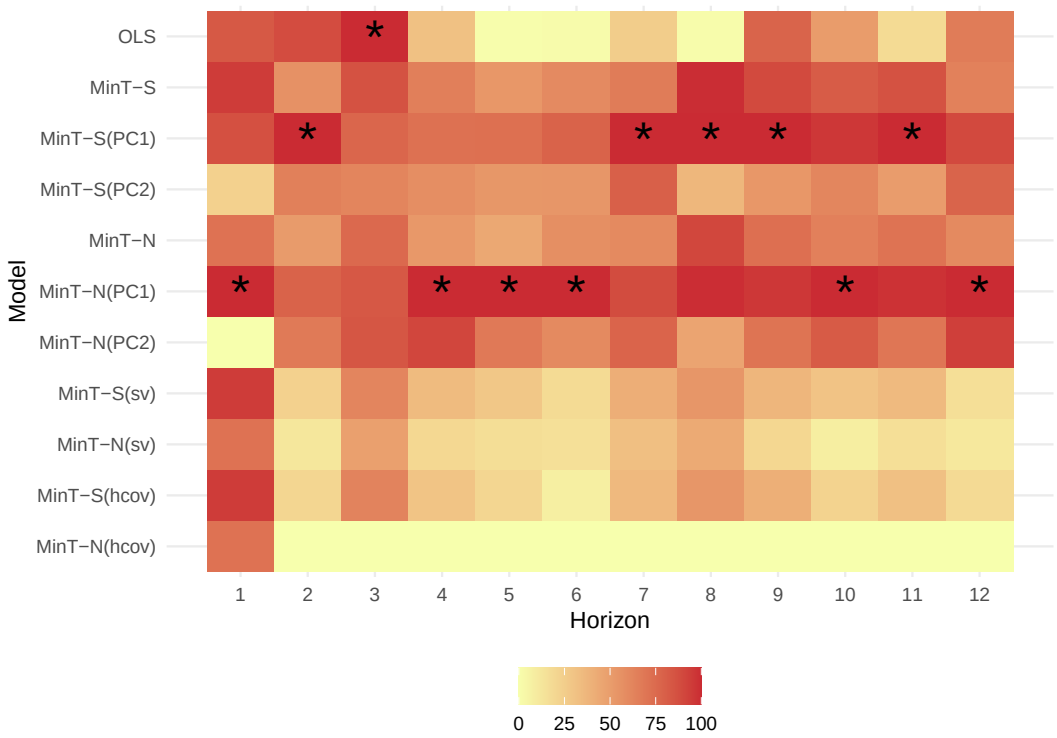

Figure 13: Heatmap of relative improvement in the mean squared error (MSE) of different reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1 to 12 steps ahead forecasts. The values are scaled to the range of 0 to 100 for better visualisation, with darker colors indicating greater improvement and best performance is noted by a star. This re-scaling does not affect relative rankings among methods.

dotted or dashed-dotted lines indicate the scaled variance and h-step-ahead residual versions (e.g. *MinT-S(SV)*, *MinT-N(hcov)*).

Three patterns stand out. First, the vanilla *MinT-S* (solid green) slightly outperform *MinT-N* (solid purple) at most horizons, although the gap is small. Second, variants that alter the multi-step covariance, either via scaled variance or direct h-step residual covariances, underperform standard MinT, suggesting that extra estimation at horizon $h > 1$ is not rewarded in this empirical analysis. Most importantly, incorporating a single dominant factor via PC-adjustment (dashed line with round points) delivers best performance among all methods. Both *MinT-S(PC1)* and *MinT-N(PC1)* consistently beat their unadjusted counterparts across horizons. However, adding more than one PC brings no additional benefit and can erode performance, likely due to injecting idiosyncratic noise from weaker components.

Figure 13 complements the line plot with a heatmap that standardises each method's MSE improvement to a 0–100 scale to enhance visual discrimination across horizons. Darker shades indicate larger improvements. The heatmap highlights the pattern from the line plot: *MinT-N(PC1)* and *MinT-S(PC1)* exhibit the highest concentration of dark cells and are repeatedly marked as the best method across horizons (star markers). Overall, the results underscores the consistent gains from PC adjustment and highlights the diminishing returns from more complex covariance treatments (lighter cells for scaled variance and h-step residuals).

Turning to probabilistic forecasts (1-step-ahead forecasts), Figure 14 shows that *MinT-N* variants (purple bars) consistently outperform *MinT-S* variants (green bars) across univariate and multivariate scores. The PC1-adjusted variants again yield improvements over their vanilla counterparts, with *MinT-N(PC1)* leading overall. In the multivariate evaluation, the Energy score places *OLS* close to the PC1-adjusted MinT methods. One surprising finding is that all MinT variants produce inferior forecasts compared to the base forecasts when considering 95% Winkler score, while *OLS* performs best.

To dissect the 95% Winkler score results, Figure 15 breaks down performance by hierarchical level, and examines empirical coverage of the 95% prediction intervals. The left panel shows that *MinT-S* produces inferior forecasts compared to the base forecasts at all levels, especially in higher aggregated levels. From our inspection, this is due to the overly shrunk variances from the shrinkage estimator, leading to narrow prediction intervals. The right panel confirms this, as *MinT-S* has the lowest empirical coverage across all levels, indicating that its prediction intervals fail to capture the true observations adequately and resulting in poor Winkler scores. *MinT-N* has relatively good coverage and improved Winkler score at bottom levels, but still is inferior at higher levels. The PC-adjusted variants seem to strike a better balance, improving overall coverage and relative Winkler scores over the base.

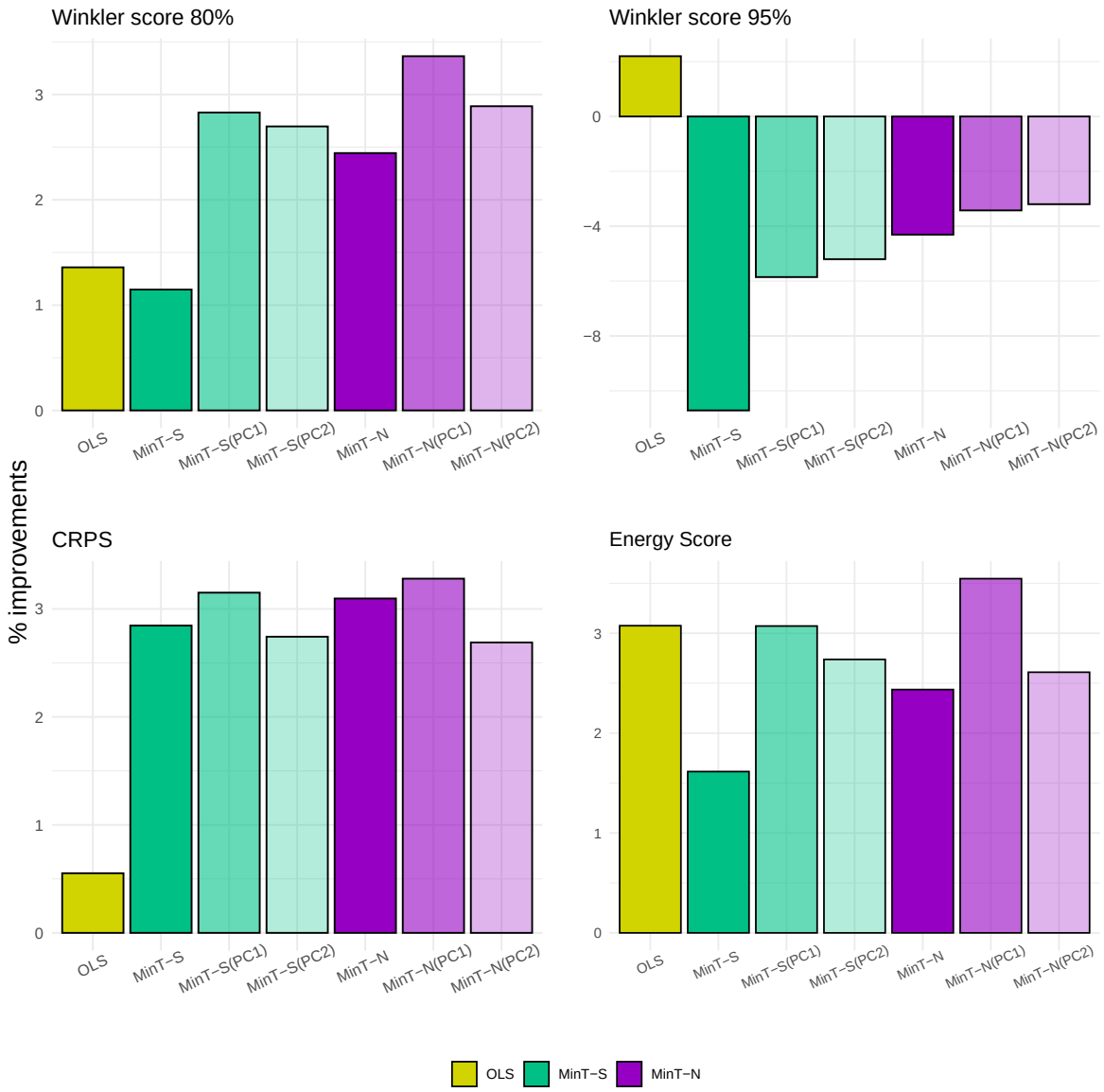

Figure 14: Percentage relative improvement in the Winkler score at 80% and 95% nominal coverage, CRPS, and Energy score of multiple reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1-step-ahead forecasts. The positive (negative) entries indicate a decrease (increase) in the probabilistic scores relative to base.

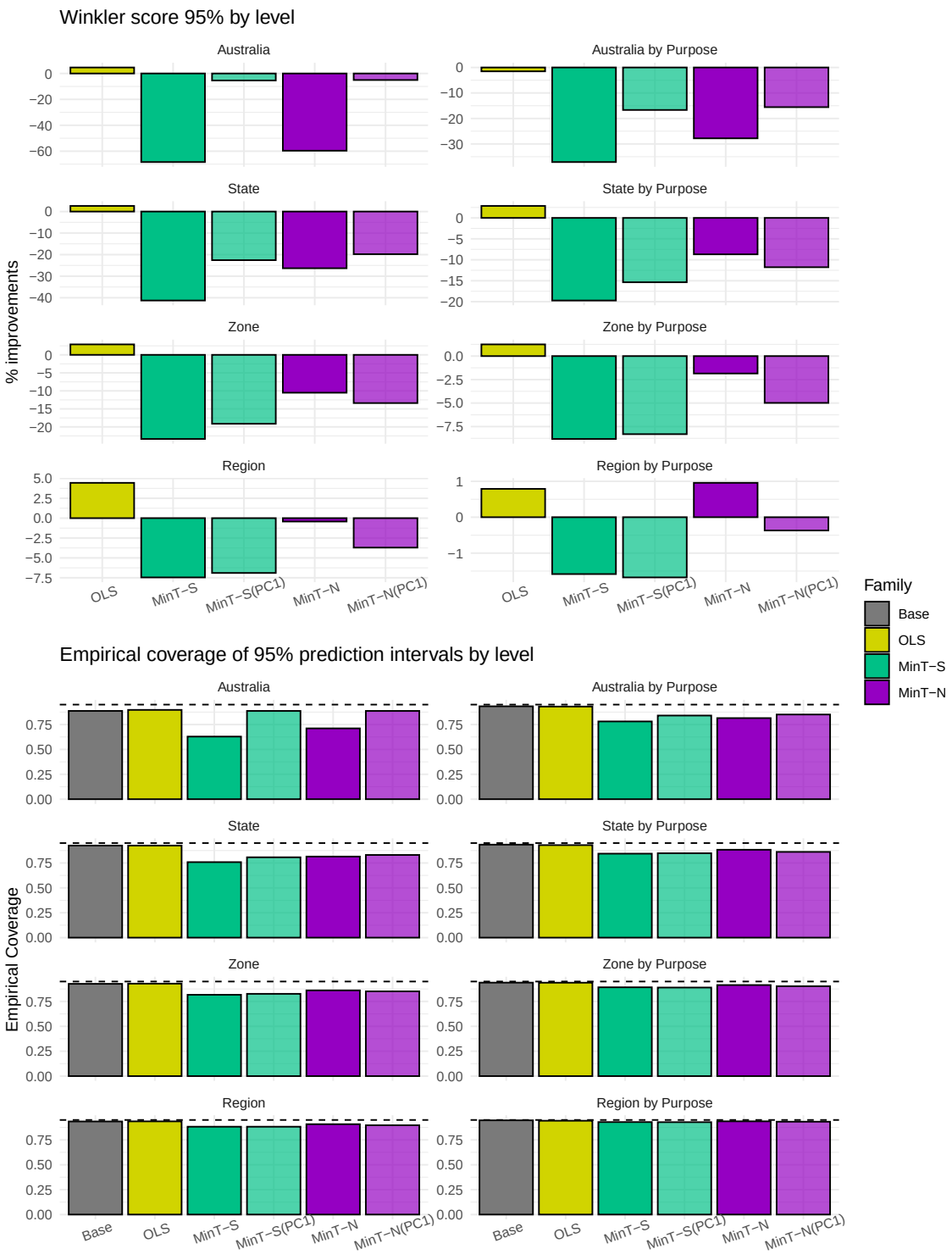

Figure 15: Percentage relative improvement in Winkler score at 95% nominal coverage by aggregation level (left), and empirical coverage of 95% prediction intervals by aggregation level (right).

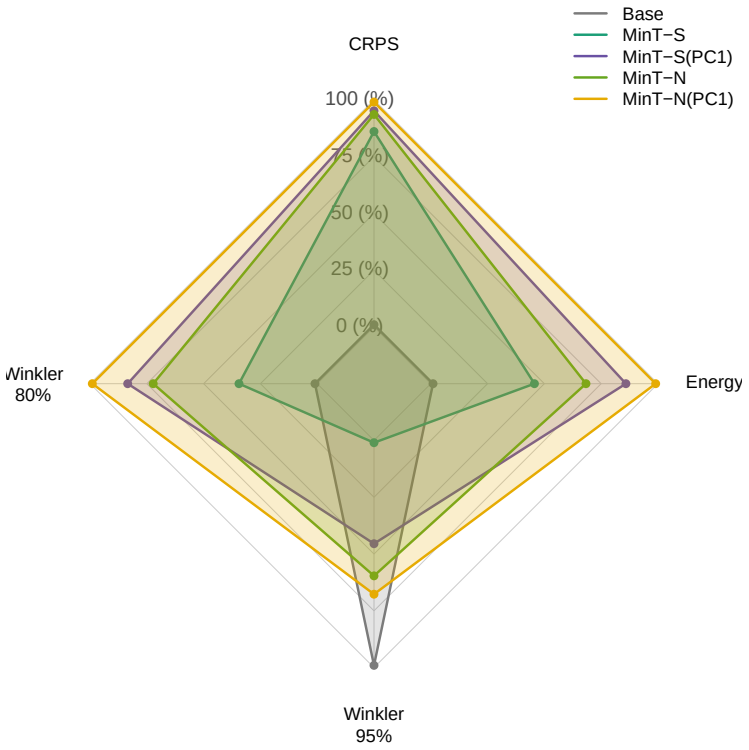

Figure 16: Radar plot of relative improvements in probabilistic scores (Winkler score at 80% and 95% intervals, CRPS, and Energy) over the base forecasts. The scores are scaled to a range of 0 to 100, with larger values indicating better performance. The outermost polygon represents the best possible score (100) and the innermost polygon represents the worst possible score (0). Only the top 4 MinT approaches are shown, together with the base forecasts.

The summary radar graph in Figure 16 consolidates these findings. Among the selected MinT variants considering probabilistic criteria, *MinT-N(PC1)* (yellow polygon) clearly leads across CRPS, Winkler score at 80% and 95% intervals, and Energy score, extending the improvements seen in the single-metric panels. Except for the Winkler score at 95% interval, all MinT variants outperform the base forecasts.

7 Conclusions and Future Work

This paper tackles a central bottleneck in Minimum Trace (MinT) reconciliation: obtaining reliable, positive-definite estimates of the base-forecast error covariance. The original practice, following Wickramasuriya et al. (2019), estimates the one-step covariance via diagonal-target shrinkage and obtains multi-step horizons by proportional scaling. We identified three limitations of this approach–uniform shrinkage, lack of factor awareness, and horizon-invariant dependence–and proposed alternatives that address them individually and in combination. Specifically, we adopt NOVELIST to introduce a more flexible shrinkage target; develop PC-adjusted variants of shrinkage and NOVELIST to preserve dominant latent components in the data; and consider multi-step covariance constructions that relax proportional scaling assumption.

Empirical evidence from Australian domestic tourism yields several insights. For point reconciliation, shrinkage (and its PC-adjusted version) remains a robust default choice, while NOVELIST demonstrates improved performance for probabilistic forecasts. Across both settings, adjusting for a single dominant principal component consistently enhances performance, indicating its utility when such latent structures are present. More complex adjustments, such as incorporating multiple principal components or horizon-specific covariances, tend to introduce additional idiosyncratic noise without notable benefits.

Several limitations and directions for future work remain. First, more intricate simulation designs are needed to further highlight the differences between shrinkage and NOVELIST, particularly under explicit factor models and varying sparsity/noise levels. Second, our current PC-adjusted implementations do not standardise series prior to factor extraction, which may lead to higher-level series largely influencing the principal components. Third, we can explore alternative methods that leverage cross-series information or eigenvalue/eigenvector structures. Fourth, the observed inferior performance of MinT with shrinkage under the Winkler 95% score at higher aggregation levels needs detailed investigation. Finally, evaluating genuinely h-step probabilistic reconciliation remains an open avenue, including how to best regularise multi-step covariances given limited effective samples.

All data generation, covariance estimation, and reconciliation routines were implemented in the ReconCov R package and is available under an open-source license on GitHub (Su, 2025).

8 Appendix

8.1 Appendix: Simulation Supplementary

8.2 Appendix: Australian Domestic Tourism Geographical Hierarchy

Table 3: Geographical divisions of Australia.

Series	Name	Label	Series	Name	Label
1	Australia	Total	57	Bundaberg	CAA
2	NSW	A	58	Capricorn	CAB
3	NT	B	59	Fraser Coast	CAC
4	QLD	C	60	Gladstone	CAD
5	SA	D	61	Mackay	CAE
6	TAS	E	62	Southern Queensland Country	CAF
7	VIC	F	63	Outback Queensland	CBA
8	WA	G	64	Brisbane	CCA
9	ACT	AA	65	Gold Coast	CCB
10	Metro NSW	AB	66	Sunshine Coast	CCC
11	Nth Coast NSW	AC	67	Townsville	CDA
12	Nth NSW	AD	68	Tropical North Queensland	CDB
13	Sth Coast NSW	AE	69	Whitsundays	CDC
14	Sth NSW	AF	70	Clare Valley	DAA
15	Central NT	BA	71	Flinders Ranges and Outback	DAB
16	Nth Coast NT	BB	72	Murray River, Lakes and Coorong	DAC
17	Central Coast QLD	CA	73	Riverland	DAD
18	Inland QLD	CB	74	Adelaide	DBA
19	Metro QLD	CC	75	Adelaide Hills	DBB
20	Nth Coast QLD	CD	76	Barossa	DBC
21	Inland SA	DA	77	Fleurieu Peninsula	DCA
22	Metro SA	DB	78	Kangaroo Island	DCB
23	Sth Coast SA	DC	79	Limestone Coast	DCC
24	West Coast SA	DD	80	Eyre Peninsula	DDA
25	Nth East TAS	EA	81	Yorke Peninsula	DDB
26	Nth West TAS	EB	82	East Coast	EAA
27	Sth TAS	EC	83	Launceston and the North	EAB
28	East Coast VIC	FA	84	North West	EBA
29	Metro VIC	FB	85	West Coast	EBB
30	Nth East VIC	FC	86	Hobart and the South	ECA
31	Nth West VIC	FD	87	Gippsland	FAA
32	West Coast VIC	FE	88	Lakes	FAB
33	Nth WA	GA	89	Phillip Island	FAC
34	Sth WA	GB	90	Geelong and the Bellarine	FBA
35	West Coast WA	GC	91	Melbourne	FBB

36	Canberra	AAA	92	Peninsula	FBC
37	Central Coast	ABA	93	Central Murray	FCA
38	Sydney	ABB	94	Goulburn	FCB
39	Hunter	ACA	95	High Country	FCC
40	North Coast NSW	ACB	96	Melbourne East	FCD
41	Blue Mountains	ADA	97	Murray East	FCE
42	Central NSW	ADB	98	Upper Yarra	FCF
43	New England North West	ADC	99	Ballarat	FDA
44	Outback NSW	ADD	100	Bendigo Loddon	FDB
45	South Coast	AEA	101	Central Highlands	FDC
46	Capital Country	AFA	102	Macedon	FDD
47	Riverina	AFB	103	Mallee	FDE
48	Snowy Mountains	AFC	104	Spa Country	FDF
49	The Murray	AFD	105	Western Grampians	FDG
50	Alice Springs	BAA	106	Wimmera	FDH
51	Barkly	BAB	107	Great Ocean Road	FEA
52	Lasseter	BAC	108	Australia's North West	GAA
53	MacDonnell	BAD	109	Australia's Golden Outback	GBA
54	Darwin	BBA	110	Australia's Coral Coast	GCA
55	Katherine Daly	BBB	111	Australia's South West	GCB
56	Litchfield Kakadu Arnhem	BBC	112	Destination Perth	GCC

References

Angam, B., Beretta, A., De Poorter, E., Duvinage, M., & Peralta, D. (2025). Forecast reconciliation for vaccine supply chain optimization. In *Communications in computer and information science* (pp. 101–118). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-74650-5/_6

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, *25*(1), 146–166. https://doi.org/10.1016/j.ijforecast.2008.07.004

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Panagiotelis, A. (2024). *Forecast reconciliation: A review*. *40*(2), 430–456. https://www.sciencedirect.com/science/article/pii/S0169207023001097

Ben Taieb, S., & Koo, B. (2019). Regularized regression for hierarchical forecasting without unbiasedness conditions. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3292500.3330976

Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2021). Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*, *116*(533), 27–43. https://doi.org/10.1080/01621459.2020.1736081

Bickel, P. J., & Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, *36*(6), 2577–2604. https://doi.org/10.1214/08-aos600

Cai, T., & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, *106*(494), 672–684. https://doi.org/10.1198/jasa.2011.tm10560

Carrara, C., Zambon, L., Azzimonti, D., & Corani, G. (2025). A novel shrinkage estimator of the covariance matrix for hierarchical time series. In *Italian statistical society series on advances in statistics* (pp. 140–145). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-96736-8/_24

Di Modica, C., Pinson, P., & Ben Taieb, S. (2021). Online forecast reconciliation in wind power prediction. *Electric Power Systems Research*, *190*(106637), 106637. https://doi.org/10.1016/j.epsr.2020.106637

El Gemayel, J., Lafarguette, R., Itd, K. M., et al. (2022). *United arab emirates: Technical assistance reportliquidity management and forecasting*.

Erven, T. van, & Cugliari, J. (2015). Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. In *Modeling and stochastic learning for forecasting in high dimensions* (pp. 297–317). Springer International Publishing. https://doi.org/10.1007/978-3-319-18732-7/_15

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding princi-

pal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *75*(4), 603–680. https://doi.org/10.1111/rssb.12016

Gamakumara, P. (2020). *Probabilistic forecast reconciliation: Theory and applications* [PhD thesis, Monash University]. https://doi.org/10.26180/5e4ca9d0c4b9d

Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, *7*(3), 1733–1762. https://www.jstor.org/stable/23566492

Higham, N. (2002). Computing the nearest correlation matrix—a problem from finance. *Ima Journal of Numerical Analysis*, *22*, 329–343. https://doi.org/10.1093/IMANUM/22.3.329

Huang, N., & Fryzlewicz, P. (2019). NOVELIST estimator of large correlation and covariance matrices and their inverses. *Test (Madrid, Spain)*, *28*(3), 694–727. https://doi.org/10.1007/s11749-018-0592-4

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589. https://doi.org/10.1016/j.csda.2011.03.006

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*, 1–22. https://doi.org/10.18637/JSS.V027.I03

Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, *97*, 16–32. https://doi.org/10.1016/j.csda.2015.11.007

Jeon, J., Panagiotelis, A., & Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research*, *279*(2), 364–379. https://doi.org/10.1016/j.ejor.2019.05.020

Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*. https://doi.org/10.1214/12-AOS989

Ledoit, O., & Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*, *48*(5), 3043–3065. https://doi.org/10.1214/19-AOS1921

Li, H., Li, H., Lu, Y., & Panagiotelis, A. (2019). A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance, Mathematics & Economics*, *86*, 122–133. https://doi.org/10.1016/j.insmatheco.2019.02.011

Nixtla. (2025). *Time series forecasting software*. https://www.nixtla.io/

O'Hara-Wild, M., Hyndman, R. J., & Wang, E. (2024). *Fabletools R package* (Version v0.5.0). https://fabletools.tidyverts.org/

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., & Hyndman, R. J. (2023). Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research*, *306*(2), 693–706. https://doi.org/10.1016/j.ejor.2022.07.040

Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, *104*(485), 177–186. https://doi.org/10.1198/jasa.2009.0101

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), Article32. https://doi.org/10.2202/1544-6115.1175

Seaman, B., & Bowman, J. (2022). Applicability of the M5 to forecasting at walmart. *International Journal of Forecasting*, *38*(4), 1468–1472. https://doi.org/10.1016/j.ijforecast.2021.06.002

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: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *26*(2), 330–343. https://doi.org/10.1080/10618600.2016.1237877

Su, V. (2025). *ReconCov R package* (Version beta). https://github.com/lordtahdus/ReconCov

Tourism research australia. (2024). https://www.tra.gov.au/.

Wickramasuriya, S. L. (2024). Probabilistic forecast reconciliation under the gaussian framework. *Journal of Business & Economic Statistics: A Publication of the American Statistical Association*, *42*(1), 272–285. https://doi.org/10.1080/07350015.2023.2181176

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. https://doi.org/10.1080/01621459.2018.1448825

Wickramasuriya, S. L., Turlach, B. A., & Hyndman, R. J. (2020). Optimal non-negative forecast reconciliation. *Statistics and Computing*, *30*(5), 1167–1182. https://doi.org/10.1007/s11222-020-09930-0