Optimal forecast reconciliation with time series selection

Xiaoqian Wang*
and
Rob J Hyndman
and
Shanika L Wickramasuriya
Department of Econometrics & Business Statistics, Monash University

December 11, 2023

Abstract

Forecast reconciliation ensures forecasts of time series in a hierarchy adhere to aggregation constraints, enabling aligned decision making. While forecast reconciliation can improve overall accuracy in hierarchical or grouped structures, the most substantial improvements occur in series with initially poor-performing base forecasts, and some series may still experience a deterioration in reconciled forecasts. However, in practice, some series in a structure often have poor base forecasts due to model misspecification or low forecastability. To address this, we propose two categories of forecast reconciliation methods that incorporate time series selection based on out-of-sample and in-sample information, respectively. Our methods keep "poor" base forecasts of some series unused in forming reconciled forecasts, preventing their negative impact on reconciliation. This process adjusts weights allocated to the remaining series accordingly when generating bottom-level reconciled forecasts. Additionally, our methods mitigate disparities arising from using different estimates of the base forecast error covariance matrix, thus alleviating the challenge of estimator selection. We evaluate the proposed methods through two simulation studies and empirical applications using Australian labour force data and domestic tourism data, showing improved accuracy compared to alternative methods, especially for higher aggregation levels, longer forecast horizons, and model misspecification.

Keywords: Coherent, Hierarchical time series, Grouped time series, Linear forecast reconciliation, Optimization problem

*Corresponding author.

1 Introduction

Forecast reconciliation is a post-processing method that ensures forecasts of multivariate time series adhere to known linear constraints (Hyndman et al. 2011, Wickramasuriya et al. 2019). For example, the sum of regional unemployment forecasts should be equal to the national unemployment forecast.

Hyndman et al. (2011) introduced optimal forecast reconciliation, whereby "base" forecasts of all series are generated independently, and then adjusted to satisfy the constraints, leading to a set of coherent reconciled forecasts. Subsequent research has extended and developed the idea in the context of cross-sectional data (Hyndman et al. 2016, Wickramasuriya et al. 2019, Panagiotelis et al. 2021), temporal data (Athanasopoulos et al. 2017), and cross-temporal data (Di Fonzo & Girolimetto 2023). Athanasopoulos et al. (2024) provided a comprehensive introduction to the forecast reconciliation literature.

Reconciliation is known to improve overall forecast accuracy in collections of time series with aggregation constraints. On average, when the base forecasts are unbiased, the mean squared reconciled forecast error from the minimum trace reconciliation method (Wickramasuriya et al. 2019) is lower than that from the base forecasts (Wickramasuriya 2021). Most of the improvements attributed to reconciliation are observed in series with initially poor-performing base forecasts (Athanasopoulos et al. 2017). In practice, it is not uncommon for some series to have poor base forecasts due to challenges such as model misspecification or low signal-to-noise ratio (SNR). In such cases, it may be advantageous to exclude the worst base forecasts when performing reconciliation. This is the motivation for our proposed methods.

First, we propose forecast reconciliation methods that incorporate time series selection based on out-of-sample information, assuming unbiased base forecasts. We formulate this as an optimization problem, using diverse penalty functions to control the number of nonzero column entries in the weighting matrix for linear forecast reconciliation. We show that the number of selected time series is at least equal to the number of series at the bottom level, and we can reconstruct the entire structure by aggregating/disaggregating the selected series. Second, we relax the unbiasedness assumption and introduce an additional reconciliation method with selection, utilizing in-sample observations and their fitted values. This enables us to use the

in-sample reconciliation performance for selection purposes. In this case, it may happen that fewer than the number of series at the bottom level are used for reconciliation. Through simulation experiments and two empirical applications, we demonstrate that our proposed methods guarantee coherent forecasts that outperform or match their respective benchmark methods. The improvements are particularly pronounced when focusing on higher aggregation levels, longer forecast horizons, and cases of model misspecification. A remarkable feature of the proposed methods is their ability to diminish disparities arising from using different estimates of the base forecast error covariance matrix, thereby mitigating challenges associated with estimator selection, which is a prominent concern in the field of forecast reconciliation research.

The remainder of the paper is structured as follows. Section 2 presents the notations and a review of linear forecast reconciliation methods. Section 3 introduces our proposed methods to achieve time series selection in reconciliation, and provides some theoretical insights. Section 4 and Section 5 show the results from simulations and two real-world datasets, respectively, followed by concluding remarks in Section 6. The R code for reproducing the results is available at https://github.com/xqnwang/hfs.

2 Preliminaries

2.1 Notation

A *hierarchical time series* is an n-dimensional multivariate time series that adheres to known linear constraints. Let $\boldsymbol{y}_t \in \mathbb{R}^n$ be a vector comprising observations from all time series in the hierarchy at time t, and $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$ be a vector comprising observations of only the most disaggregated ("bottom-level") time series at time t. The full hierarchy at time t can be written as

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

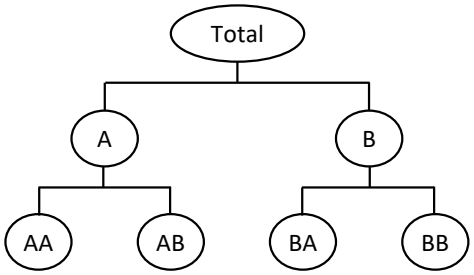

Figure 1: *An example of a two-level hierarchical time series.*

for $t = 1, 2, \ldots, T$, where T is the length of the time series, and $\boldsymbol{S}$ is an $n \times n_b$ *summing matrix* that defines the aggregation constraints. We can write the summing matrix as $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_{n_b} \end{bmatrix}$, where $\boldsymbol{A}$ is an $n_a \times n_b$ *aggregation matrix* with $n = n_a + n_b$, and $\boldsymbol{I}_{n_b}$ is an n_b-dimensional identity matrix.

For example, Figure 1 shows a simple hierarchy with $n = 7$, $n_b = 4$, $n_a = 3$, $\boldsymbol{y}_t = [y_{\text{Total},t}, y_{\text{A},t}, y_{\text{B},t}, y_{\text{AA},t}, y_{\text{AB},t}, y_{\text{BA},t}, y_{\text{BB},t}]'$, $\boldsymbol{b}_t = [y_{\text{AA},t}, y_{\text{AB},t}, y_{\text{BA},t}, y_{\text{BB},t}]'$, and

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \multicolumn{4}{c}{\boldsymbol{I}_4} \end{bmatrix}.$$

The notation is general enough to include aggregation constraints that are non-hierarchical. Please refer to Hyndman & Athanasopoulos (2021) for further details.

2.2 Linear forecast reconciliation

Let $\hat{\boldsymbol{y}}_{T+h|T} \in \mathbb{R}^n$ be a vector of h-step-ahead *base forecasts* for all time series in the structure, given observations up to and including time T, and stacked in the same order as $\boldsymbol{y}_t$. We can use any method to generate these forecasts, but in general they will not be coherent (i.e., they won't satisfy the constraints). Let $\tilde{\boldsymbol{y}}_{T+h|T} \in \mathbb{R}^n$ denote a vector of h-step-ahead *reconciled forecasts* given by

$$\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{y}}_{T+h|T}, \tag{1}$$

where $\boldsymbol{G}_h$ is an $n_b \times n$ weighting matrix and $\boldsymbol{S}$ is an $n \times n_b$ summing matrix (Wickramasuriya et al. 2019).

Minimum trace reconciliation

Let $\hat{\boldsymbol{e}}_{t+h|t} = \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}$ denote the h-step-ahead in-sample *base forecast errors*, and $\tilde{\boldsymbol{e}}_{t+h|t} = \boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t}$ denote the h-step-ahead *reconciled forecast errors*, for $t = 1, 2, \ldots, T-h$. Wickramasuriya et al. (2019) formulated the reconciliation problem as minimizing the trace (MinT) of the h-step-ahead covariance matrix of the reconciled forecast errors, $\text{Var}(\tilde{\boldsymbol{e}}_{t+h|t})$. Assuming the base forecasts are unbiased, and that we want the reconciled forecasts to be unbiased, the unique solution to the minimization problem is

$$\boldsymbol{G}_h = \left(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}, \tag{2}$$

where $\boldsymbol{W}_h$ is the positive definite covariance matrix of the h-step-ahead base forecast errors.

The trace minimization problem can be reformulated as a least squares problem with linear constraints:

$$\min_{\tilde{\boldsymbol{y}}_{T+h|T}} \quad \frac{1}{2}(\hat{\boldsymbol{y}}_{T+h|T} - \tilde{\boldsymbol{y}}_{T+h|T})'\boldsymbol{W}_h^{-1}(\hat{\boldsymbol{y}}_{T+h|T} - \tilde{\boldsymbol{y}}_{T+h|T}) \qquad \text{s.t.} \quad \tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{T+h|T}, \tag{3}$$

where $\tilde{\boldsymbol{b}}_{T+h|T} \in \mathbb{R}^{n_b}$ comprises the h-step-ahead bottom-level reconciled forecasts, made at time T. The intuition behind MinT reconciliation is that **the larger the estimated variance of the base forecast errors, the larger the range of adjustments permitted for forecast reconciliation**.

It is challenging to estimate $\boldsymbol{W}_h$, especially for $h > 1$. It is common to assume $\boldsymbol{W}_h = k_h\boldsymbol{W}_1, \forall h$, where $k_h > 0$; then the MinT solution of $\boldsymbol{G}$ does not change with the forecast horizon, h. Hence, we will drop the subscript h for ease of exposition. Table 1 lists the most popularly used candidate estimators for $\boldsymbol{W}_h$.

Relaxation of the unbiasedness assumptions

Both Hyndman et al. (2011) and Wickramasuriya et al. (2019) imposed unbiasedness conditions on the base forecasts and the reconciled forecasts. Ben Taieb & Koo (2019) proposed a reconciliation method relaxing the assumption of unbiasedness. Specifically, by expanding the training window incrementally, one observation at a time, they formulated the reconciliation problem as a regularized empirical risk

Table 1: *Forecast reconciliation methods for which different estimators of $\boldsymbol{W}$ are used.*

Reconciliation method	$\boldsymbol{W}_h \propto$
OLS (Hyndman et al. 2011)	$\boldsymbol{I}$
WLSs (Athanasopoulos et al. 2017)	$\text{diag}(\boldsymbol{S}\mathbf{1})$
WLSv (Hyndman et al. 2016)	$\text{Diag}(\hat{\boldsymbol{W}}_1)$
MinT (Wickramasuriya et al. 2019)	$\hat{\boldsymbol{W}}_1$
MinTs (Wickramasuriya et al. 2019)	$\lambda\,\text{Diag}(\hat{\boldsymbol{W}}_1)+(1-\lambda)\hat{\boldsymbol{W}}_1$

Note: $\mathbf{1}$ is a vector of 1s of size n_b, $\text{diag}(\cdot)$ constructs a diagonal matrix using a given vector, $\hat{\boldsymbol{W}}_1$ denotes the unbiased covariance estimator based on the in-sample one-step-ahead base forecast errors (i.e., residuals), and $\text{Diag}(\cdot)$ forms a diagonal matrix using the diagonal elements of the input matrix.

minimization (RERM) problem given by

$$\min_{\boldsymbol{G}_h} \frac{1}{(T-T_1-h+1)n} \left\|\boldsymbol{Y}_h^* - \hat{\boldsymbol{Y}}_h^* \boldsymbol{G}_h' \boldsymbol{S}'\right\|_F^2 + \lambda \|\,\text{vec}(\boldsymbol{G}_h)\|_1,$$

where T_1 denotes the minimum number of observations used for model training, $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_1$ is the L_1 norm, $\text{vec}(\cdot)$ denotes the vectorization of a matrix (stacking the columns of the matrix), $\boldsymbol{Y}_h^* = [\boldsymbol{y}_{T_1+h},\dots,\boldsymbol{y}_T]'$, $\hat{\boldsymbol{Y}}_h^* = \left[\hat{\boldsymbol{y}}_{T_1+h|T_1},\dots,\hat{\boldsymbol{y}}_{T|T-h}\right]'$, and $\lambda \geq 0$ is a regularization parameter.

When $\lambda = 0$, the problem reduces to an empirical risk minimization (ERM) problem without regularization. Assuming that the series in the structure are jointly weakly stationary and $\hat{\boldsymbol{Y}}_h^{*\prime}\hat{\boldsymbol{Y}}_h^*$ is invertible, it has a closed-form solution given by

$$\hat{\boldsymbol{G}}_h = \boldsymbol{B}_h^{*\prime}\hat{\boldsymbol{Y}}_h^* \left(\hat{\boldsymbol{Y}}_h^{*\prime}\hat{\boldsymbol{Y}}_h^*\right)^{-1},$$

where $\boldsymbol{B}_h^* = [\boldsymbol{b}_{T_1+h},\dots,\boldsymbol{b}_T]'$. If $\hat{\boldsymbol{Y}}_h^{*\prime}\hat{\boldsymbol{Y}}_h^*$ is not invertible, Ben Taieb & Koo (2019) suggested using a generalized inverse. When $\lambda > 0$, imposing the L_1 penalty on $\boldsymbol{G}_h$ will introduce sparsity and reduce estimation variance, albeit at the cost of introducing some bias.

Wickramasuriya (2021) proposed an empirical MinT (**EMinT**) solution without the unbiasedness constraint by minimizing the trace of the covariance matrix of the reconciled forecast errors, $\text{Var}(\tilde{\boldsymbol{e}}_{T+h|T})$. Assuming the series are jointly weakly stationary, she derived the solution

$$\hat{\boldsymbol{G}}_h = \boldsymbol{B}_h'\hat{\boldsymbol{Y}}_h \left(\hat{\boldsymbol{Y}}_h'\hat{\boldsymbol{Y}}_h\right)^{-1},$$

where $\boldsymbol{B}_h = [\boldsymbol{b}_h, \ldots, \boldsymbol{b}_T]'$, and $\hat{\boldsymbol{Y}}_h = \left[\hat{\boldsymbol{y}}_{h|0}, \ldots, \hat{\boldsymbol{y}}_{T|T-h}\right]'$.

The difference between EMinT and ERM lies in the data sources, as EMinT uses in-sample observations and base forecasts, while ERM relies on observations and base forecasts from a holdout validation set. Both ERM and EMinT consider an estimate of $\boldsymbol{G}$ that changes over the forecast horizon, which is why we keep the subscript h here.

A challenge in forecast reconciliation arises when some base forecasts perform poorly, as the role of the weighting matrix $\boldsymbol{G}$ is to assimilate *all* base forecasts and map them into bottom-level disaggregated forecasts, which are subsequently summed by $\boldsymbol{S}$. While the RERM method proposed by Ben Taieb & Koo (2019) introduces sparsity by shrinking some elements of $\boldsymbol{G}$ towards zero, it remains incapable of mitigating the adverse impact of underperforming base forecasts. Moreover, the method is time-consuming because it uses expanding windows to recursively generate out-of-sample base forecasts.

We therefore propose two new forecast reconciliation methods involving time series selection: constrained out-of-sample (under the unbiasedness assumption) and unconstrained in-sample (without the unbiasedness assumption). These methods aim to address the negative effect of some poor base forecasts on the overall performance of the reconciled forecasts. Additionally, through the incorporation of regularization in the objective function, our method improves reconciliation outcomes produced with a "poor" choice of $\boldsymbol{W}$.

3 Forecast reconciliation with time series selection

In this section, we introduce our methods for forecast reconciliation while automatically avoiding the use of poor base forecasts. Section 3.1 introduces constrained reconciliation methods in a formulation of the problem based on out-of-sample base forecasts, while Section 3.2 presents an unconstrained reconciliation method, where we formulate the problem based on in-sample observations and base forecasts.

3.1 Series selection with unbiasedness constraint

As $\boldsymbol{S}$ is fixed and $\hat{\boldsymbol{y}}_{T+h|T}$ is given, $\boldsymbol{G}_h$ determines the linear reconciliation performance, as shown in Equation 1. We drop the subscript h here as we assume $\boldsymbol{W}$ and $\boldsymbol{G}$ do not vary with the forecast horizon. A natural way to remove forecasts of some series is by controlling the number of nonzero column entries in $\boldsymbol{G}$. This leads to a generalization of the MinT optimization problem with an additional penalty term:

$$\min_{\boldsymbol{G}} \quad \frac{1}{2}(\hat{\boldsymbol{y}} - \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}})' \boldsymbol{W}^{-1} (\hat{\boldsymbol{y}} - \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}) + \lambda \mathfrak{g}(\boldsymbol{G}) \qquad \text{s.t.} \quad \boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}, \tag{4}$$

where $\hat{\boldsymbol{y}} := \hat{\boldsymbol{y}}_{T+1|T}$, $\mathfrak{g}(\cdot)$ penalizes the columns of $\boldsymbol{G}$ towards zero, and λ is a penalty parameter. This can be considered *a grouped variable selection problem*, with each group corresponding to a column of $\boldsymbol{G}$. The constraint, $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$, ensures that the reconciled forecasts are unbiased if the base forecasts are unbiased (Wickramasuriya et al. 2019). When $\lambda = 0$, the problem reduces to the MinT optimization problem in Equation 3 with a closed-form solution given by Equation 2.

Proposition 1. *Under the assumption of unbiasedness, the count of nonzero column entries of* $\boldsymbol{G}$ (*i.e., the number of time series selected for reconciliation*), *derived through solving Equation 4, is at least equal to the number of time series at the bottom level. In addition, we can restore the full hierarchical structure by aggregating/disaggregating the selected time series.*

Proof. Under unbiasedness, $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$, and

$$\min\left(\text{rank}(\boldsymbol{G}), \text{rank}(\boldsymbol{S})\right) \geq \text{rank}(\boldsymbol{I}_{n_b}) = n_b,$$

which indicates that the count of nonzero column entries of $\boldsymbol{G}$ is at least equal to n_b.

Let $\boldsymbol{X}_{\cdot\mathbb{S}} \in \mathbb{R}^{r\times|\mathbb{S}|}$ denote the submatrix of the $r \times c$ matrix $\boldsymbol{X}$ with column indices forming a set $\mathbb{S}$ (and when $\mathbb{S} = \{j\}$, we simply use $\boldsymbol{X}_{\cdot j}$), where $|\mathbb{S}|$ denotes the size of the set $\mathbb{S}$. Similarly, let $\boldsymbol{X}_{\mathbb{S}\cdot} \in \mathbb{R}^{|\mathbb{S}|\times c}$ denote the submatrix of $\boldsymbol{X}$ whose rows are indexed by a set $\mathbb{S}$ (and when $\mathbb{S} = \{i\}$, we simply use $\boldsymbol{X}_{i\cdot}$). Assuming that the set $\mathbb{S}$ consists of the indices of nonzero columns in the solution to Equation 4, $\hat{\boldsymbol{G}}$, the

following equations hold:

$$\boldsymbol{G}\boldsymbol{S} = \hat{\boldsymbol{G}}_{\cdot\mathbb{S}}\boldsymbol{S}_{\mathbb{S}\cdot} = \boldsymbol{I}, \qquad \text{and} \qquad \min\left(\text{rank}(\hat{\boldsymbol{G}}_{\cdot\mathbb{S}}), \text{rank}(\boldsymbol{S}_{\mathbb{S}\cdot})\right) \geq \text{rank}(\boldsymbol{I}_{n_b}) = n_b.$$

Additionally, we have $\text{rank}(\boldsymbol{S}_{\mathbb{S}\cdot}) \leq n_b$ as $\boldsymbol{S}$ has n_b columns. Therefore, we can conclude that $\text{rank}(\boldsymbol{S}_{\mathbb{S}\cdot}) = n_b$. Moreover, we have

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t = \boldsymbol{S}\boldsymbol{G}\boldsymbol{S}\boldsymbol{b}_t = \boldsymbol{S}\hat{\boldsymbol{G}}_{\cdot\mathbb{S}}\boldsymbol{S}_{\mathbb{S}\cdot}\boldsymbol{b}_t = \boldsymbol{S}\hat{\boldsymbol{G}}_{\cdot\mathbb{S}}(\boldsymbol{y}_t)_{\mathbb{S}},$$

which implies that the hierarchical structure can be fully restored by aggregating/disaggregating the selected time series denoted by $(\boldsymbol{y}_t)_{\mathbb{S}}$.

For example, consider the simple hierarchy shown in Figure 1, it is not possible for our constrained reconciliation methods with selection to simultaneously zero out columns of $\boldsymbol{G}$ associated with series AA and AB. However, it is possible to zero out columns related to series AA and BA simultaneously.

Proposition 2. *The optimization problem in Equation 4 can be reformulated as a least squares problem with regularization and linear equality constraint as follows:*

$$\begin{aligned} \min_{\text{vec}(\boldsymbol{G})} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right)\text{vec}(\boldsymbol{G})\right)' \boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right)\text{vec}(\boldsymbol{G})\right) + \lambda\, \mathfrak{g}\left(\text{vec}(\boldsymbol{G})\right) \\ \text{s.t.} \quad & \left(\boldsymbol{S}' \otimes \boldsymbol{I}_{n_b}\right)\text{vec}(\boldsymbol{G}) = \text{vec}(\boldsymbol{I}_{n_b}), \end{aligned} \tag{5}$$

which is characterized as a high-dimensional problem in which the number of features, denoted as $p = n_b \times n$, is much larger than the number of observations, n.

Proof. We have

$$\begin{aligned} &\text{vec}\left(\hat{\boldsymbol{y}}\right) = \hat{\boldsymbol{y}}, \\ &\text{vec}\left(\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}\right) = \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right)\text{vec}(\boldsymbol{G}), \\ &\text{vec}\left(\boldsymbol{G}\boldsymbol{S}\right) = \text{vec}\left(\boldsymbol{I}_{n_b}\boldsymbol{G}\boldsymbol{S}\right) = \left(\boldsymbol{S}' \otimes \boldsymbol{I}_{n_b}\right)\text{vec}(\boldsymbol{G}). \end{aligned}$$

Substituting these into Equation 4, the previous problem now takes the form of a regression problem with an additional regularization term and an equality constraint on the coefficients, as shown in Equation 5.

Next, we present three classes of regularizations that allow forecast reconciliation with series selection, resulting in three optimization problems: (i) group best-subset selection with ridge regularization, (ii) an intuitive method with L_0 regularization, and (iii) a group lasso method.

Group best-subset selection with ridge regularization

In a high-dimensional context with $p \gg n$, it is common to assume that the true regression coefficient (i.e., $\text{vec}(\boldsymbol{G})$ in our problem) is sparse. We apply a combination of L_0 and L_2 regularization to control the nonzero column entries in $\boldsymbol{G}$:

$$\begin{aligned} \min_{\text{vec}(\boldsymbol{G})} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right)\text{vec}(\boldsymbol{G})\right)' \boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right)\text{vec}(\boldsymbol{G})\right) + \lambda_0 \sum_{j=1}^{n} 1\left(\boldsymbol{G}_{\cdot j} \neq \boldsymbol{0}\right) + \lambda_2 \left\|\text{vec}\left(\boldsymbol{G}\right)\right\|_2^2 \\ \text{s.t.} \quad & \left(\boldsymbol{S}' \otimes \boldsymbol{I}_{n_b}\right)\text{vec}(\boldsymbol{G}) = \text{vec}(\boldsymbol{I}_{n_b}), \end{aligned} \tag{6}$$

where $1(\cdot)$ is the indicator function, $\lambda_0 \geq 0$ controls the number of nonzero columns of $\boldsymbol{G}$, $\lambda_2 \geq 0$ controls the strength of the ridge regularization, and $\|\cdot\|_2$ is the L_2 norm. In a hierarchical or grouped time series context, $\text{vec}(\boldsymbol{G})$ has an inherent non-overlapping grouping structure, wherein each group corresponds to a single column of $\boldsymbol{G}$, each of size n_b. Therefore, we call this reconciliation method *group best-subset selection with ridge regularization*. In the results that follow, we label the **Subset** method differently based on various estimators for $\boldsymbol{W}$, referring to them as **OLS-subset**, **WLSs-subset**, **WLSv-subset**, **MinT-subset**, and **MinTs-subset**, respectively.

The inclusion of the ridge term in Equation 6 is motivated by earlier work on best-subset selection (e.g., Hazimeh & Mazumder 2020, Mazumder et al. 2022), which suggests that additional ridge regularization can mitigate the poor predictive performance of best-subset selection method in low SNR regimes.

We present a Big-M based mixed integer programming (MIP) formulation for the problem in Equation 6:

$$\min_{\text{vec}(\boldsymbol{G}),\boldsymbol{z},\check{\boldsymbol{e}},\boldsymbol{g}^+} \frac{1}{2}\check{\boldsymbol{e}}'\boldsymbol{W}^{-1}\check{\boldsymbol{e}} + \lambda_0 \sum_{j=1}^{n} z_j + \lambda_2 \boldsymbol{g}^{+\prime}\boldsymbol{g}^+ \tag{7}$$

$$
\begin{aligned}
\text{s.t.} \quad & \left(\boldsymbol{S}' \otimes \boldsymbol{I}_{n_b}\right) \text{vec}(\boldsymbol{G}) = \text{vec}\left(\boldsymbol{I}_{n_b}\right) \\
& \hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right) \text{vec}(\boldsymbol{G}) = \check{\boldsymbol{e}} \\
& \sum_{i=1}^{n_b} g^{+}_{i+(j-1)n_b} \leqslant \mathscr{M} z_j, \quad j \in [n] \\
& \boldsymbol{g}^{+} \geqslant \text{vec}(\boldsymbol{G}) \\
& \boldsymbol{g}^{+} \geqslant -\text{vec}(\boldsymbol{G}) \\
& z_j \in \{0,1\}, \quad j \in [n],
\end{aligned}
$$

where $\mathscr{M}$ is a Big-M parameter (specified a-priori) that is sufficiently large that the optimal solution to Equation 7, $\boldsymbol{g}^{+*}$, satisfies $\max_{j\in[n]} \sum_{i=1}^{n_b} g^{+}_{i+(j-1)n_b} \leqslant \mathscr{M}$. The binary variable $z_j = 0$ implies that $\boldsymbol{G}_{\cdot j} = \boldsymbol{0}$, and $z_j = 1$ implies that $\sum_{i=1}^{n_b} g^{+}_{i+(j-1)n_b} \leqslant \mathscr{M}$. Such Big-M formulations are commonly used in MIP problems to model relations between discrete and continuous variables, and have been recently explored in regression with L_0 regularization (Bertsimas et al. 2016). The problem is a mixed integer quadratic program (MIQP) that can be solved using commercial MIP solvers, e.g., Gurobi and CPLEX.

Parameter tuning. To avoid computationally expensive cross-validation, we tune the parameters to minimize the sum of squared reconciled forecast errors on the truncated training set, comprising only the $\max\{h,s\}$ observations closest to the forecast origin, where s is the seasonal period for seasonal data and $s = T$ for non-seasonal data. Let $\lambda_0^1 = \frac{1}{2}\left(\hat{\boldsymbol{y}} - \tilde{\boldsymbol{y}}^{\text{bench}}\right)' \boldsymbol{W}^{-1} \left(\hat{\boldsymbol{y}} - \tilde{\boldsymbol{y}}^{\text{bench}}\right)$, which captures the scale of the first term in the objective function, where $\tilde{\boldsymbol{y}}^{\text{bench}}$ is a vector of reconciled forecasts obtained using Equation 2 with the same estimator of $\boldsymbol{W}$, and define $\lambda_0^k = 0.0001\lambda_0^1$. For the parameter λ_0, we consider a grid of $k+1$ values, $\{\lambda_0^1, \ldots, \lambda_0^k, 0\}$, where $\lambda_0^j = \lambda_0^1 \left(\lambda_0^k/\lambda_0^1\right)^{(j-1)/(k-1)}$ for $j \in [k]$. So $\lambda_0^1, \ldots, \lambda_0^k$ is a sequence decreasing on the log scale. We use a grid of six values for the parameter λ_2, $\{0, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$. Therefore, we tune over a two-dimensional grid of $(k+1)\times 6$ values to find the optimal combination of λ_0 and λ_2.

Computation details. The MIQP problem in Equation 7 is NP-hard and computationally intensive. Bertsimas et al. (2016) showed that commercial MIP solvers are capable of tackling problem instances for

p up to a thousand. To address larger instances, there has been impressive work on developing MIP-based approaches for solving L_0-regularized regression problem; e.g., Bertsimas et al. (2016), Hazimeh & Mazumder (2020), and Hazimeh et al. (2022). However, it is challenging to extend these approaches to accommodate additional constraints within the optimization problem. Despite the potential sluggishness of handling large instances with commercial MIP solvers, in our experiments, we use Gurobi to solve our problem in Equation 7 by configuring parameters such as MIPGap = 0.001 and TimeLimit = 600 seconds for cases with $p > 1000$. This enables us to terminate the solver before reaching the global optimum and return a suboptimal solution instead. This strategy is motivated by our need to consider numerous parameter candidates, and the final solution will be validated against the training set, which prevents the utilization of a poor estimate of $\boldsymbol{G}$.

Intuitive method with L_0 regularization

Instead of estimating the entire matrix $\boldsymbol{G}$ as above, we leverage the MinT solution in Equation 2 to streamline the optimization problem under consideration. Specifically, we define $\bar{\boldsymbol{S}} = \boldsymbol{A}\boldsymbol{S}$, where $\boldsymbol{A} = \text{diag}(\boldsymbol{z})$ is an $n \times n$ diagonal matrix, and $\boldsymbol{z}$ is an n-dimensional vector with elements either equal to 0 or 1. Taking the MinT solution in Equation 2, we have $\bar{\boldsymbol{G}} = (\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}$. Given fixed $\boldsymbol{S}$ and estimation of $\boldsymbol{W}$, $\bar{\boldsymbol{G}}$ is entirely determined by $\boldsymbol{A}$. By this way, when the jth diagonal element of $\boldsymbol{A}$ equals zero, the jth column of $\bar{\boldsymbol{G}}$ becomes entirely composed of zeros. Therefore, the optimization problem can be reduced to an integer quadratic programming (IQP) problem in which all of the variables are restricted to being integers:

$$\begin{aligned} \min_{\boldsymbol{A}} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}} - \boldsymbol{S}\bar{\boldsymbol{G}}\hat{\boldsymbol{y}}\right)' \boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}} - \boldsymbol{S}\bar{\boldsymbol{G}}\hat{\boldsymbol{y}}\right) + \lambda_0 \sum_{j=1}^{n} \boldsymbol{A}_{jj} \\ \text{s.t.} \quad & \bar{\boldsymbol{G}} = (\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1} \qquad \text{and} \qquad \bar{\boldsymbol{G}}\boldsymbol{S} = \boldsymbol{I}, \end{aligned}$$

where $\lambda_0 \geq 0$ controls the number of nonzero diagonal elements in $\boldsymbol{A}$, consequently affecting the number of nonzero columns (i.e., selected time series) in $\boldsymbol{G}$. We call this reconciliation method the *intuitive*

method with L_0 regularization. In the results that follow, we label the **Intuitive** method differently based on various estimators for $\boldsymbol{W}$, referring to them as **OLS-intuitive**, **WLSs-intuitive**, **WLSv-intuitive**, **MinT-intuitive**, and **MinTs-intuitive**, respectively.

We should note that implementing grouped variable selection with this optimization problem can be challenging because it imposes restrictions $\bar{\boldsymbol{G}}$ to ensure it adheres rigorously to the analytical solution of MinT while making the selection. Therefore, the resulting solution tends to be dense and may not have zero columns.

To ensure the invertibility of $\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{S}$, and make the problem compatible with Gurobi, we reformulate the problem as

$$\begin{aligned}
\min_{\boldsymbol{A},\bar{\boldsymbol{G}},\boldsymbol{C},\check{\boldsymbol{e}},\boldsymbol{z}} \quad & \frac{1}{2}\check{\boldsymbol{e}}'\boldsymbol{W}^{-1}\check{\boldsymbol{e}} + \lambda_0 \sum_{j=1}^{n} z_j \\
\text{s.t.} \quad & \bar{\boldsymbol{G}}\boldsymbol{S} = \boldsymbol{I} \\
& \hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right) \operatorname{vec}(\bar{\boldsymbol{G}}) = \check{\boldsymbol{e}} \\
& \bar{\boldsymbol{G}}\boldsymbol{A}\boldsymbol{S} = \boldsymbol{I} \\
& \bar{\boldsymbol{G}} = \boldsymbol{C}\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1} \\
& z_j \in \{0,1\}, \quad j \in [n].
\end{aligned} \tag{8}$$

Parameter tuning. Similar to the setup in the group best-subset selection, we select the tuning parameter, λ_0, by minimizing the sum of squared reconciled forecast errors on a truncated training set, comprising only the $\max\{h,s\}$ observations that occurred prior to the forecast origin. Let $\lambda_0^1 = \frac{1}{2}\left(\hat{\boldsymbol{y}} - \tilde{\boldsymbol{y}}^{\text{bench}}\right)' \boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}} - \tilde{\boldsymbol{y}}^{\text{bench}}\right)$, and $\lambda_0^k = 0.0001\lambda_0^1$, the collection of candidate values for λ_0 we consider is $\{\lambda_0^1, \ldots, \lambda_0^k, 0\}$, where $\lambda_0^j = \lambda_0^1\left(\lambda_0^k/\lambda_0^1\right)^{(j-1)/(k-1)}$ for $j \in [k]$.

Computation details. Following a setup akin to that in the group best-subset selection, we employ Gurobi to solve Equation 8 by configuring parameters such as MIPGap = 0.001 and TimeLimit = 600 seconds for problems with $p > 1000$.

Group lasso method

Lasso is another popular method for the selection and estimation of parameters in the context of linear regression. Yuan & Lin (2006) introduced the group lasso method that can be used when there is a grouped structure among the variables. Here, we consider *a group lasso problem under the unbiasedness assumption* given by

$$\begin{aligned} \min_{\boldsymbol{G}} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right)\mathrm{vec}(\boldsymbol{G})\right)' \boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right)\mathrm{vec}(\boldsymbol{G})\right) + \lambda \sum_{j=1}^{n} w_j \left\|\boldsymbol{G}_{\cdot j}\right\|_2 \\ \text{s.t.} \quad & \left(\boldsymbol{S}' \otimes \boldsymbol{I}_{n_b}\right)\mathrm{vec}(\boldsymbol{G}) = \mathrm{vec}\left(\boldsymbol{I}_{n_b}\right), \end{aligned} \tag{9}$$

where $\lambda \geq 0$ is a tuning parameter, $w_j \neq 0$ is the penalty weight assigned in $\boldsymbol{G}_{\cdot j}$ to make the model more flexible, and the second term in the objective is the penalty function that is intermediate between the L_1-penalty that is used in the lasso and the L_2-penalty that is used in ridge regression. In the results that follow, we label the **Lasso** method based on various estimators for $\boldsymbol{W}$, referring to them as **OLS-lasso**, **WLSs-lasso**, **WLSv-lasso**, **MinT-lasso**, and **MinTs-lasso**, respectively.

Next, we present the second order cone programming (SOCP) formulation for the group lasso based estimators given by

$$\begin{aligned} \min_{\mathrm{vec}(\boldsymbol{G}),\check{\boldsymbol{e}},\boldsymbol{g}^{+}} \quad & \frac{1}{2}\check{\boldsymbol{e}}'\boldsymbol{W}_h^{-1}\check{\boldsymbol{e}} + \lambda \sum_{j=1}^{n} w_j c_j \\ \text{s.t.} \quad & \left(\boldsymbol{S}' \otimes \boldsymbol{I}_{n_b}\right)\mathrm{vec}(\boldsymbol{G}) = \mathrm{vec}\left(\boldsymbol{I}_{n_b}\right) \\ & \hat{\boldsymbol{y}} - \left(\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\right)\mathrm{vec}(\boldsymbol{G}) = \check{\boldsymbol{e}} \\ & c_j = \sqrt{\sum_{i=1}^{n_b} g_{i+(j-1)n_b}^{+2}}, \quad j \in [n]. \end{aligned} \tag{10}$$

Equation 10 includes additional auxiliary variables $c_j \in \mathbb{R}_{\geq 0}$, $j \in [n]$, and second order cone constraints, $c_j = \sqrt{\sum_{i=1}^{n_b} g_{i+(j-1)n_b}^{+2}}$ for $j \in [n]$.

Compared to the previous two methods above, the group lasso method is computationally friendlier.

Nonetheless, Hazimeh et al. (2023) demonstrated, both empirically and theoretically, that the group L_0-regularized method exhibits advantages over its group lasso counterpart across a range of regimes. Group lasso can either be highly dense or possess non-zero coefficients that are overly shrunk. This issue becomes more pronounced when the groups are correlated with each other, as group lasso tends to retain all correlated groups instead of seeking a more concise model.

Penalty weights and parameter tuning. In the context of group lasso, the default choice for the penalty weight, w_j, is $\sqrt{p_j}$, where p_j is the size of each group (in our case, $p_j = n_b$). In our experiments, we allocate different penalty weights to each group using $w_j = 1/\|\boldsymbol{G}_{\cdot j}^{\text{bench}}\|_2$, which allows us to account for variations in scale across different time series in the structure.

We compute the group lasso over $k+1$ values of the tuning parameter λ, and select the tuning parameter by optimizing the sum of squared reconciled forecast errors on a truncated training set, consisting only of $\max\{h,s\}$ observations occurred prior to the forecast origin. The collection of candidate values for λ under consideration is $\{\lambda^1,\ldots,\lambda^k,0\}$, where $\lambda^1 = \max_{j=1,\ldots,n} \big\| -\big((\hat{\boldsymbol{y}}' \otimes \boldsymbol{S})_{\cdot j^*}\big)' \boldsymbol{W}^{-1}\hat{\boldsymbol{y}}\big\|_2 / w_j$, $\lambda^k = 0.0001\lambda^1$, and $\lambda^j = \lambda^1(\lambda^k/\lambda^1)^{(j-1)/(k-1)}$ for $j \in [k]$.

Proposition 3. *Ignoring the unbiasedness constraint, we define λ^1 as the smallest λ value such that all predictors in the group lasso problem have zero coefficients. Then we have*

$$\lambda^1 = \max_{j=1,\ldots,n} \big\| -\big((\hat{\boldsymbol{y}}' \otimes \boldsymbol{S})_{\cdot j^*}\big)' \boldsymbol{W}^{-1}\hat{\boldsymbol{y}}\big\|_2 / w_j,$$

where j^* denotes the column index of $\hat{\boldsymbol{y}}' \otimes \boldsymbol{S}$ that corresponds to the jth column of $\boldsymbol{G}$.

Proof. Denote $\beta = \text{vec}(\boldsymbol{G})$, and the first term in the objective of Equation 9 as $L(\beta \mid \boldsymbol{D})$, where $\boldsymbol{D}$ is the working data $\{\hat{\boldsymbol{y}}, \hat{\boldsymbol{y}}' \otimes \boldsymbol{S}\}$. Ignoring the unbiasedness constraint, we define λ^1 as the smallest λ value such that all predictors in the group lasso problem have zero coefficients, i.e., the solution at λ^1 is $\hat{\beta}^1 = \boldsymbol{0}$. (Note that there is no intercept in our problem.) Under the Karush-Kuhn-Tucker conditions, we have

$$\lambda^1 = \max_{j=1,\ldots,n} \big\| \big[\nabla L(\hat{\beta}^1 \mid \boldsymbol{D})\big]^{(j)} \big\|_2 / w_j = \max_{j=1,\ldots,n} \big\| -\big((\hat{\boldsymbol{y}}' \otimes \boldsymbol{S})_{\cdot j^*}\big)' \boldsymbol{W}^{-1}\hat{\boldsymbol{y}}\big\|_2 / w_j.$$

Computation details. Due to the incorporation of the unbiasedness constraint, we can not directly use some open-source packages designed for group lasso. Consequently, we employ Gurobi to solve the SOCP problem, configuring it by setting OptimalityTol = 0.0001.

3.2 Series selection method without unbiasedness constraint

In this section, we relax the unbiasedness constraint, $\boldsymbol{GS} = \boldsymbol{I}$, and introduce a reconciliation method with selection that relies on in-sample observations and fitted values. Let $\boldsymbol{Y} \in \mathbb{R}^{T\times n}$ denote a matrix comprising observations from all time series on the training set in the structure, and $\hat{\boldsymbol{Y}} \in \mathbb{R}^{T\times n}$ denote a matrix of in-sample one-step-ahead forecasts (i.e., fitted values) for all time series. The proposed *empirical group lasso* method considers the optimization problem

$$\min_{\boldsymbol{G}} \quad \frac{1}{2T}\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\boldsymbol{G}'\boldsymbol{S}'\right\|_F^2 + \lambda \sum_{j=1}^{n} w_j \left\|\boldsymbol{G}_{\cdot j}\right\|_2,$$

where $\lambda \geq 0$ is a tuning parameter, $w_j \neq 0$ is the penalty weight assigned in $\boldsymbol{G}_{\cdot j}$ to make a more flexible model. We rewrite the problem as

$$\min_{\text{vec}(\boldsymbol{G})} \quad \frac{1}{2T}\left\|\text{vec}(\boldsymbol{Y}) - (\boldsymbol{S}\otimes\hat{\boldsymbol{Y}})\,\text{vec}\left(\boldsymbol{G}'\right)\right\|_2^2 + \lambda \sum_{j=1}^{n} w_j \left\|\boldsymbol{G}_{\cdot j}\right\|_2,$$

which becomes a standard group lasso problem, with $\text{vec}(\boldsymbol{Y})$ serving as the dependent variable and $\boldsymbol{S}\otimes\hat{\boldsymbol{Y}}$ as the covariate matrix. We denote this as **Elasso** in the results that follow.

Upon relaxing the unbiasedness constraint, the number of non-zero column entries in the solution for $\boldsymbol{G}$ may be less than the number of time series at the bottom level. This differs from the series selection methods with an unbiasedness constraint that we introduced in Section 3.1. In an extreme scenario, it can happen that the solution takes the form of a top-down $\boldsymbol{G}_{TD} = [\boldsymbol{p} \mid \boldsymbol{O}_{n_b\times(n-1)}]$, where only the column corresponding to the top level (most aggregated level) retains non-zero values, and $\boldsymbol{p} = (p_1, p_2, \ldots, p_{n_b})$ is a proportionality vector obtained based on in-sample reconciled forecast errors.

We also explored the empirical version of group best-subset selection with ridge regularization and the intuitive method with L_0 regularization in which we do not impose the unbiasedness constraint. It is worth mentioning that Hazimeh et al. (2023) presented a new algorithmic framework for formulating the group L_0 problem with ridge regularization and provided the **L0Group** Python package for implementation. However, our experiments showed that this algorithm can not terminate within five hours for typical instances with $p \sim 10^4$. Therefore, in this paper, we only present the empirical group lasso method for series selection without the unbiasedness constraint.

Penalty weights and parameter tuning. Similarly to the setup in the group lasso method, we assign different penalty weights to each group by setting $w_j = 1/\|\boldsymbol{G}_{\cdot j}^{\text{OLS}}\|_2$, where $\boldsymbol{G}^{\text{OLS}}$ is the solution obtained by the OLS estimator of $\boldsymbol{W}$. Given a fixed tuning parameter value, we solve the target optimization problem by considering the initial $T - T_v$ observations, where $T_v = \max\{h, s\}$ for seasonal time series and $T_v = \lfloor \frac{1}{10} T \rfloor$ for non-seasonal time series. Then we select the tuning parameter, λ, by minimizing the sum of squared reconciled forecast errors on a truncated training set, comprising only the T_v observations closest to the forecast origin. Specifically, we form the set of candidate values for λ as $\{\lambda^1, \ldots, \lambda^k, 0\}$, where $\lambda^1 = \max_{j=1,\ldots,n} \left\| -\frac{1}{N} \left(\left(\boldsymbol{S} \otimes \hat{\boldsymbol{Y}} \right)_{\cdot j*} \right)' \text{vec}(\boldsymbol{Y}) \right\|_2 / w_j$, $\lambda^k = 0.0001\lambda^1$, and $\lambda^j = \lambda^1 \left(\lambda^k / \lambda^1 \right)^{(j-1)/(k-1)}$ for $j \in [k]$. Following the same derivation as in the proof of **Proposition 3**, λ^1 is the smallest λ value such that all predictors in the empirical group lasso problem have zero coefficients, i.e., $\boldsymbol{G} = \boldsymbol{O}$. Note that we need to resolve the optimization problem based on the whole training set by using the optimal tuning parameter to obtain the final solution.

Computation details. While there are open-source packages available for solving group lasso problems, they are still relatively slow when handling large instances. For example, given a specific value for the parameter, λ, our experiments observed that, using the **gglasso** R package, we can not obtain a solution within five hours for typical instances with $p \sim 10^4$. Instead, we use Gurobi to solve the problem using the SOCP formulation for the empirical group lasso which aligns with Equation 10 but omits the unbiasedness constraint.

4 Monte Carlo simulations

To evaluate the performance of the proposed reconciliation methods with time series selection outlined in Section 3, we carry out two simulations with different designs. In both simulations, we consider a hierarchy comprising two levels of aggregation, as shown in Figure 1. Specifically, the structure has four series at the bottom level, and seven series in total; i.e., $n_b = 4$, and $n = 7$. The bottom-level series are first generated and then summed to obtain aggregated series at higher levels.

Section 4.1 considers a setup where the bottom-level series are generated using a structural time series model, but model misspecification exists for some series within the structure. Section 4.2 explores the impact of the correlation between series on the performance of reconciled forecasts.

4.1 Setup 1: Exploring the effect of model misspecification

We follow a simulation setup similar to Wickramasuriya et al. (2019), assuming that the bottom-level time series are generated using the basic structural time series model

$$\boldsymbol{b}_t = \boldsymbol{\mu}_t + \boldsymbol{\gamma}_t + \boldsymbol{\eta}_t,$$

where $\boldsymbol{\mu}_t$ and $\boldsymbol{\gamma}_t$ are trend and seasonal components defined by

$$\begin{aligned}
\boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t-1} + \boldsymbol{v}_t + \boldsymbol{\varrho}_t, & \boldsymbol{\varrho}_t &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\rho^2 \boldsymbol{I}_4\right), \\
\boldsymbol{v}_t &= \boldsymbol{v}_{t-1} + \boldsymbol{\zeta}_t, & \boldsymbol{\zeta}_t &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\zeta^2 \boldsymbol{I}_4\right), \\
\boldsymbol{\gamma}_t &= -\sum_{i=1}^{s-1} \boldsymbol{\gamma}_{t-i} + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\omega^2 \boldsymbol{I}_4\right),
\end{aligned}$$

$\boldsymbol{\varrho}_t$, $\boldsymbol{\zeta}_t$, and $\boldsymbol{\omega}_t$ are error terms independent of each other and over time, and the error term $\boldsymbol{\eta}_t$ is generated independently from an ARIMA$(p,0,q)$ process, where p and q take values of 0 or 1 with equal probability. The coefficients in the ARIMA process are sampled randomly from a uniform distribution within the range

$[0.5, 0.7]$, and the contemporaneous error covariance matrix is given by

$$\begin{bmatrix} 5 & 3 & 2 & 1 \\ 3 & 4 & 2 & 1 \\ 2 & 2 & 5 & 3 \\ 1 & 1 & 3 & 4 \end{bmatrix},$$

which enables correlations among time series in a hierarchical structure.

We set $s = 4$ for quarterly data, with error variances $\sigma_\rho^2 = 2$, $\sigma_\zeta^2 = 0.007$, and $\sigma_\omega^2 = 7$. The initial values for $\boldsymbol{\mu}_0$, $\boldsymbol{v}_0$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$, and $\boldsymbol{\gamma}_2$ are generated independently from a multivariate normal distribution with zero mean and identity covariance matrix. For each series at the bottom level, we generate a total of $T + h = 180$ observations, with the last $h = 16$ observations serving as the test set. Recall that the bottom-level series are aggregated to obtain the data for the aggregated levels. This process is repeated 500 times.

We use ETS models to generate base forecasts for all time series in the hierarchy, using the default settings as implemented in the **forecast** R package (Hyndman et al. 2023). To introduce model misspecification into our experiment, we deliberately undermine the quality of in-sample and out-of-sample forecasts (i.e., fitted values and base forecasts) for some specific time series. Specifically, we investigate three scenarios characterized by artificial model misspecifications, where a 1.5 multiplier is applied to in-sample and out-of-sample forecasts for a single series in each scenario; i.e., series AA at the bottom level, series A at the middle level, and series Total at the top level, resulting in Scenarios A–C, respectively.

The results are summarized in Table 2, Table A1, and Table A2, with each table reporting the average root mean squared error (RMSE) for each level as well as the whole structure (denoted as *Average*). The *Base* row shows the average RMSE of base forecasts, while entries below this row report the percentage decrease (negative) or increase (positive) in the average RMSE of reconciled forecasts compared to base forecasts. The *BU* row uses a "bottom-up" approach, aggregating bottom-level base forecasts to form forecasts for the aggregated series. Notably, in each scenario, the largest improvements occur at the respective hierarchical level where model misspecification is introduced, while slightly deteriorating the performance at other levels.

Table 2: *Out-of-sample forecast results for the simulated data in Scenario A, Setup 1.*

	Top				Middle				Bottom				Average			
Method	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16
Base	9.6	10.7	12.6	15.6	6.3	7.3	8.6	10.8	6.4	7.5	8.3	9.8	6.8	7.9	9.0	10.9
BU	57.8	68.5	53.7	38.9	58.2	61.8	48.1	34.4	0.0	0.0	0.0	0.0	27.0	29.6	23.8	17.7
OLS	0.6	2.2	1.8	1.4	7.1	6.4	4.6	3.1	–7.6	–8.6	–8.2	–7.3	–2.1	–2.5	–2.7	–2.6
OLS-subset	0.6	**1.8**	**1.5**	**1.3**	7.2	**5.2**	**3.8**	**2.6**	**–8.3**	**–12.9**	**–11.6**	**–9.9**	**–2.4**	**–5.2**	**–4.8**	**–4.1**
OLS-intuitive	0.8	2.6	2.1	1.8	7.5	**6.1**	**4.4**	**3.0**	**–9.0**	**–12.8**	**–11.6**	**–9.9**	**–2.7**	**–4.8**	**–4.5**	**–3.8**
OLS-lasso	0.6	2.2	1.8	1.6	7.4	6.7	4.8	3.2	–7.6	–8.5	–8.1	–7.2	–2.0	–2.4	–2.6	–2.5
WLSs	7.3	10.6	8.1	5.9	15.6	16.0	11.8	8.0	–6.9	–7.8	–7.4	–6.4	1.9	2.0	1.0	0.2
WLSs-subset	**5.0**	**5.7**	**4.6**	**3.6**	**12.3**	**10.0**	**7.5**	**5.2**	**–7.6**	**–10.5**	**–9.6**	**–8.2**	**0.2**	**–2.0**	**–2.1**	**–2.0**
WLSs-intuitive	**7.1**	**9.2**	**7.1**	**5.2**	16.5	**15.5**	**11.5**	**7.9**	–6.8	**–9.2**	**–8.4**	**–7.3**	2.1	**0.9**	**0.1**	**–0.4**
WLSs-lasso	7.3	**10.3**	**8.0**	5.9	15.7	16.1	11.8	8.1	**–7.0**	–7.8	–7.3	–6.4	1.9	2.0	1.0	0.2
WLSv	1.0	2.9	2.3	1.9	4.5	4.3	3.2	2.1	–25.8	–26.4	–22.7	–18.3	–12.4	–12.6	–10.7	–8.4
WLSv-subset	**–1.0**	**0.3**	**0.4**	**0.5**	**0.6**	**0.6**	**0.5**	**0.3**	**–32.3**	**–32.2**	**–27.3**	**–21.7**	**–17.3**	**–17.3**	**–14.2**	**–10.9**
WLSv-intuitive	**–0.5**	**0.2**	**0.3**	**0.5**	**0.9**	**0.7**	**0.5**	**0.3**	**–32.3**	**–32.3**	**–27.4**	**–21.7**	**–17.1**	**–17.3**	**–14.2**	**–10.9**
WLSv-lasso	**0.4**	**1.5**	**1.5**	**1.4**	**3.0**	**2.5**	**2.0**	**1.3**	**–28.5**	**–29.2**	**–24.9**	**–19.9**	**–14.4**	**–14.9**	**–12.3**	**–9.5**
MinT	–0.4	0.7	0.9	0.6	0.7	0.7	0.6	0.3	–32.9	–33.4	–28.3	–22.5	–17.5	–17.8	–14.6	–11.3
MinT-subset	**–0.6**	0.7	**0.8**	0.7	**0.6**	0.8	0.6	0.3	**–33.0**	–33.1	–28.0	–22.3	**–17.6**	–17.6	–14.5	–11.2
MinT-intuitive	–0.4	0.7	0.9	0.6	0.7	0.7	0.6	0.3	–32.9	–33.4	–28.3	–22.5	–17.5	–17.8	–14.6	–11.3
MinT-lasso	**–0.7**	**0.3**	**0.6**	**0.4**	**0.3**	**0.4**	**0.4**	**0.1**	**–33.2**	**–33.7**	**–28.5**	**–22.6**	**–17.8**	**–18.1**	**–14.8**	**–11.4**
MinTs	–0.9	0.6	0.7	0.5	0.6	0.6	0.5	0.2	–32.9	–33.5	–28.3	–22.5	–17.6	–17.9	–14.6	–11.3
MinTs-subset	–0.7	0.9	1.1	1.0	0.7	0.8	0.7	0.4	**–33.0**	–33.1	–27.9	–22.2	–17.6	–17.5	–14.3	–11.0
MinTs-intuitive	–0.9	0.6	0.7	0.5	0.6	0.6	0.5	0.2	–32.9	–33.5	–28.3	–22.5	–17.6	–17.9	–14.6	–11.3
MinTs-lasso	–0.9	**0.4**	**0.6**	0.5	0.6	**0.4**	**0.4**	**0.1**	**–33.2**	**–33.6**	**–28.4**	**–22.6**	**–17.7**	**–18.0**	**–14.8**	**–11.4**
EMinT	2.2	2.9	2.5	1.7	2.5	2.9	2.3	1.3	–31.9	–32.3	–27.5	–22.0	–15.9	–16.2	–13.4	–10.5
Elasso	**1.5**	**2.8**	**2.4**	1.7	**2.1**	**2.8**	2.3	1.3	**–32.1**	–32.2	–27.4	–21.9	**–16.3**	–16.2	–13.3	–10.5

Note: The Base row shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, the proposed methods are indicated with a gray background, and methods that outperform the benchmark method are marked in bold.

Focusing on the results of benchmark reconciliation methods, we find that the BU approach performs the best in Scenarios B and C but ranks as the worst overall in Scenario A. This is not surprising, as bottom-level base forecasts are deteriorated in Scenario A, while higher-level base forecasts are deteriorated in Scenarios B and C. Moreover, WLSv, MinT, and MinTs perform especially well in Setup 1, benefiting from their ability to consider in-sample covariance of base forecast errors, allowing for a larger range of adjustments in reconciliation for base forecasts with higher estimated error variance. EMinT also provides accurate reconciled forecasts in our setup, where the in-sample forecasts for specific series are intentionally undermined, a situation that can be detected by the in-sample information based EMinT method. However, OLS and WLSs significantly underperform other benchmark methods in this simulation design.

In all three scenarios, our proposed methods consistently produce either improved or comparable reconciled forecasts compared to their respective benchmarks. The improvements are particularly pronounced when using OLS and WLSs estimators of $\boldsymbol{W}$ in the benchmark methods, which do not take into account the

in-sample covariance of base forecast errors. One advantage of using our proposed forecast reconciliation methods with selection is their ability to reduce the difference introduced by using different estimates of $\boldsymbol{W}$, thereby mitigating the risk of estimator selection. In some cases, such as Scenarios B and C, we can align the forecast accuracy achieved using different estimators, and make them approach the best results we can obtain. Dropping the unbiasedness assumption, Elasso performs similarly to EMinT overall while achieving improvements at the top level, which is typically the aspect of greatest concern to practitioners.

Table 3: *Proportion of time series being selected after using the proposed reconciliation methods with selection in Scenario A, Setup 1.*

	Top	A	B	AA	AB	BA	BB	Summary
OLS-subset	0.52	0.79	0.57	0.79	1	0.91	0.85	
OLS-intuitive	0.80	0.90	0.81	0.80	1	0.85	0.86	
OLS-lasso	0.90	1.00	0.68	1.00	1	1.00	1.00	
WLSs-subset	0.85	0.91	0.86	0.90	1	0.97	0.97	
WLSs-intuitive	0.92	0.95	0.67	0.92	1	0.92	0.95	
WLSs-lasso	0.72	1.00	0.72	1.00	1	1.00	1.00	
WLSv-subset	0.50	0.62	0.42	0.19	1	0.81	0.87	
WLSv-intuitive	0.59	0.55	0.49	0.17	1	0.76	0.86	
WLSv-lasso	0.40	1.00	0.41	0.77	1	1.00	1.00	
MinT-subset	0.66	0.90	0.61	0.72	1	0.91	0.93	
MinT-intuitive	1.00	1.00	1.00	1.00	1	1.00	1.00	
MinT-lasso	0.80	0.96	0.84	0.72	1	0.98	0.97	
MinTs-subset	0.57	0.88	0.52	0.67	1	0.89	0.92	
MinTs-intuitive	1.00	1.00	1.00	1.00	1	1.00	1.00	
MinTs-lasso	0.68	1.00	0.66	0.74	1	1.00	1.00	
Elasso	0.82	0.63	0.69	1.00	1	1.00	1.00	

Note: the last column displays a stacked barplot for each method, based on the total number of selected series data from 500 simulation instances, with a darker sub-bar indicating a larger number.

In addition, we report the proportion of time series being selected from the implementation of our proposed methods in 500 simulation instances, as shown in Table 3, Table A3, and Table A4. Clearly, our proposed methods select fewer time series, and generally improve forecast accuracy, compared to the benchmark methods. Furthermore, we observe that the Subset methods tend to return fewer time series compared to the Intuitive and Lasso methods, which aligns with our expectations that the Intuitive and Lasso methods tend to produce dense estimates. Most importantly, depending on the scenario considered, the time series with model misspecification has been selected less often than others. For example, considering Scenario A, series AA is expected to be removed, while AB is expected to be retained. This allows us to obtain

series AA via operations such as A−AB, Total−B−AB, or Total−AB−BA−BB. The results in Table 3 align with our expectations, and show that series AA is dropped often, whereas AB is selected all the time.

4.2 Setup 2: Exploring the effect of correlation

We now simulate a hierarchical structure with correlated series, using a similar simulation to Wickramasuriya (2021), and the same hierarchical structure as shown in Figure 1. We use a stationary VAR(1) data generating process for the time series at the bottom level:

$$\boldsymbol{b}_t = \boldsymbol{c} + \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{bmatrix} \boldsymbol{b}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{c}$ is a constant vector with all entries set to 1, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are 2×2 matrices with eigenvalues $z_{1,2} = 0.6[\cos(\pi/3) \pm i\sin(\pi/3)]$ and $z_{3,4} = 0.9[\cos(\pi/6) \pm i\sin(\pi/6)]$, respectively, $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 \\ 0 & \boldsymbol{\Sigma}_2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & \sqrt{6}\rho \\ \sqrt{6}\rho & 3 \end{bmatrix},$$

and $\rho \in \{0, \pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8\}$ controls the error correlation in the simulated hierarchy.

For each time series at the bottom level, we generate a total of 101 observations, with the last observation serving as the test set, i.e., $T = 100$ and $h = 1$. Once again, the data at the higher levels are obtained by aggregating the bottom-level series. The process is repeated 500 times for each candidate correlation, ρ.

For each series, base forecasts are generated from ARMA models. We identify the best ARMA model using the automated algorithm implemented in the **forecast** R package (Hyndman & Khandakar 2008). Additionally, when fitting ARMA models for time series Total, A, and BA, we introduce a slight bias by omitting the constant term. Figure 2 presents an illustrative example of a simulated hierarchical time series. The left panels depict time plots for each series at different levels of the structure, while the right panels show the residuals obtained from forecasting each series using the fitted ARMA model. Notably, despite our omission of the constant term when fitting ARMA models to series Total, A, and BA, the

residuals derived from the identified optimal models still exhibit fluctuations around zero and do not display significant deviations in comparison to the residuals from other series. This is because the influence of the constant term is minimal, i.e., it is much smaller compared to the data variability. Thus, it may be challenging to identify the "poor" base forecasts and exclude them from reconciliation in this setup.

Table 4 summarizes the average RMSE of the base forecasts across various error correlations and the percentage relative improvements in RMSE achieved by reconciliation methods relative to the base forecasts. The results show that, for OLS, WLSs, and WLSv estimators, our proposed methods consistently dominate or are equivalent to their respective benchmark methods at all levels. We should highlight the challenge of identifying the "poor" base forecasts in this simulation design, given that the omission of

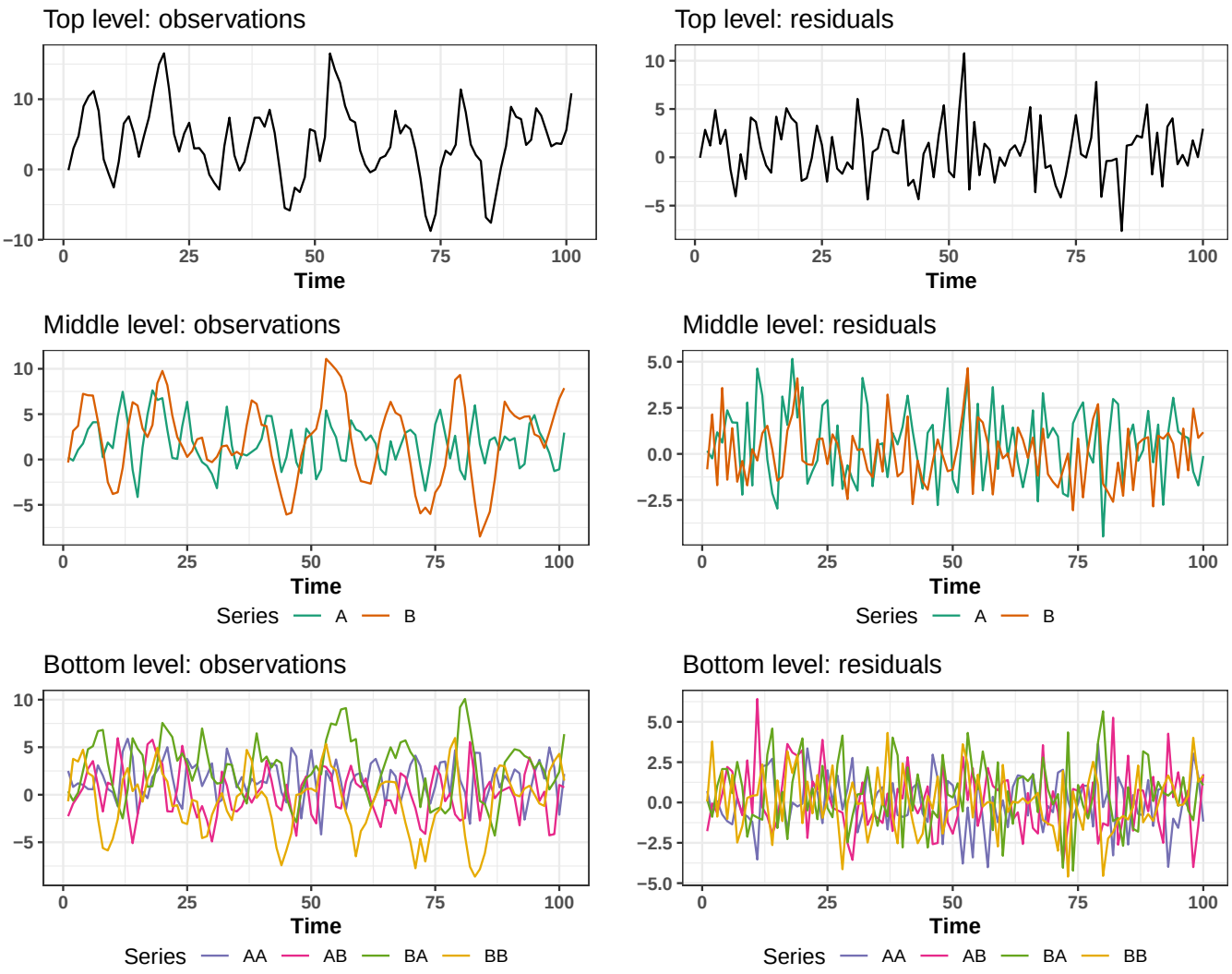

Figure 2: *An example hierarchical time series and its in-sample residuals in Setup 2.*

Table 4: *Out-of-sample forecast results across various error correlations for simulation in Setup 2.*

	Top					Middle					Bottom					Average				
Method	ρ=–0.8	–0.4	0	0.4	0.8	ρ=–0.8	–0.4	0	0.4	0.8	ρ=–0.8	–0.4	0	0.4	0.8	ρ=–0.8	–0.4	0	0.4	0.8
Base	2.4	2.9	3.4	4.1	4.0	1.5	1.8	2.1	2.4	2.5	1.5	1.5	1.5	1.5	1.4	1.6	1.8	2.0	2.1	2.1
BU	–17.0	–9.0	–6.7	–7.0	–7.4	–6.8	0.4	4.8	5.7	2.8	0.0	0.0	0.0	0.0	0.0	–5.3	–1.9	–0.2	–0.1	–1.0
OLS	–11.0	–8.2	–7.7	–8.2	–8.0	–3.5	–0.7	3.1	2.5	0.8	0.7	–0.6	–2.0	–2.3	–2.1	–2.8	–2.4	–1.8	–2.4	–2.7
OLS-subset	**–11.4**	**–8.4**	**–8.1**	**–8.4**	**–8.8**	**–3.7**	–0.7	3.2	2.5	**0.4**	**0.3**	**–0.8**	–2.0	–1.7	**–2.6**	**–3.2**	**–2.5**	**–1.9**	–2.2	**–3.2**
OLS-intuitive	**–11.6**	–8.0	**–7.8**	–8.0	**–8.4**	**–3.6**	–0.4	3.7	2.5	**0.3**	**0.6**	–0.2	–1.3	–0.4	–1.5	**–3.0**	–2.0	–1.3	–1.6	**–2.8**
OLS-lasso	**–19.2**	**–9.8**	–7.2	**–8.7**	**–8.2**	**–10.5**	**–1.7**	**2.9**	**2.4**	0.8	**–0.8**	**–0.8**	–1.6	–2.3	–2.1	**–7.1**	**–3.1**	–1.6	**–2.5**	**–2.8**
WLSs	–16.8	–11.1	–9.6	–10.4	–10.2	–8.1	–2.8	1.5	1.2	–0.4	–0.3	–1.1	–2.4	–2.9	–2.9	–5.7	–3.9	–3.0	–3.6	–4.0
WLSs-subset	**–17.3**	**–11.4**	**–9.9**	**–11.1**	**–10.8**	**–8.3**	–2.8	**1.4**	**0.7**	**–0.9**	**–0.7**	**–1.3**	–2.4	**–3.2**	**–3.3**	**–6.1**	**–4.0**	**–3.1**	**–4.1**	**–4.5**
WLSs-intuitive	**–16.9**	**–11.5**	**–9.8**	–10.0	**–10.6**	**–8.5**	–2.8	**1.4**	1.5	**–0.7**	**–0.7**	**–1.2**	–2.3	–2.7	**–3.0**	**–6.1**	**–4.0**	–3.0	–3.3	**–4.3**
WLSs-lasso	**–18.3**	–11.1	–9.2	**–10.5**	–9.8	**–9.3**	–2.4	**1.4**	1.2	–0.1	**–0.8**	–1.0	–2.4	–2.9	–2.8	**–6.6**	–3.7	–2.9	**–3.7**	–3.7
WLSv	–16.5	–11.9	–10.0	–10.6	–10.6	–7.6	–3.4	0.9	1.1	–0.5	–0.5	–1.2	–2.3	–2.9	–3.0	–5.7	–4.3	–3.2	–3.7	–4.2
WLSv-subset	**–16.8**	**–12.1**	–9.8	**–10.8**	**–10.7**	**–7.8**	**–3.5**	1.1	1.2	**–1.0**	**–1.1**	**–1.3**	–2.2	–2.9	**–3.2**	**–6.1**	**–4.4**	–3.0	–3.7	**–4.4**
WLSv-intuitive	**–17.6**	**–12.6**	**–10.1**	–10.5	–10.6	**–8.7**	**–3.8**	**0.7**	1.1	**–0.8**	**–1.9**	**–1.5**	–2.3	**–3.0**	–3.0	**–7.0**	**–4.7**	**–3.3**	–3.7	**–4.3**
WLSv-lasso	**–19.8**	–11.6	–9.7	–10.5	–10.6	**–10.5**	–3.0	1.2	1.2	–0.5	**–1.2**	–1.1	–2.2	–2.9	–3.0	**–7.5**	–4.1	–3.0	–3.7	–4.2
MinT	–25.4	–18.8	–12.4	**–15.3**	**–12.6**	–15.5	–7.0	0.0	–2.0	–2.0	–4.0	–4.6	–4.3	–5.8	–5.1	–11.4	–8.5	–5.0	–7.2	–6.0
MinT-subset	–25.4	–18.8	–12.4	**–15.3**	**–12.6**	–15.5	–7.0	0.0	–2.0	–2.0	–4.0	–4.6	–4.3	–5.8	–5.1	–11.4	–8.5	–5.0	–7.2	–6.0
MinT-intuitive	–25.4	–18.8	–12.4	**–15.3**	**–12.6**	–15.5	–7.0	0.0	–2.0	–2.0	–4.0	–4.6	–4.3	–5.8	–5.1	–11.4	–8.5	–5.0	–7.2	–6.0
MinT-lasso	–25.4	–18.8	–12.4	**–15.3**	**–12.6**	–15.5	–7.0	0.0	–2.0	–2.0	–4.0	–4.6	–4.3	–5.8	–5.1	–11.4	–8.5	–5.0	–7.2	–6.0
MinTs	–25.4	–17.7	–12.1	–14.2	–12.5	–16.1	–6.8	–0.8	–1.6	**–2.4**	–4.0	–4.6	–4.9	–5.9	**–5.2**	–11.6	–8.2	–5.4	–6.8	**–6.2**
MinTs-subset	–25.2	–17.6	–12.1	–14.2	–12.5	–16.1	–6.8	–0.8	–1.6	**–2.4**	–3.9	–4.6	–4.9	–5.9	**–5.2**	–11.5	–8.2	–5.4	–6.8	**–6.2**
MinTs-intuitive	–25.4	–17.7	–12.1	–14.2	–12.5	–16.1	–6.8	–0.8	–1.6	**–2.4**	–4.0	–4.6	–4.9	–5.9	**–5.2**	–11.6	–8.2	–5.4	–6.8	**–6.2**
MinTs-lasso	–25.4	–17.6	–12.1	–14.2	–12.5	–16.1	–6.7	–0.8	–1.6	**–2.4**	–4.0	–4.6	–4.9	–5.9	**–5.2**	–11.6	–8.2	–5.4	–6.8	**–6.2**
EMinT	**–31.2**	**–19.8**	**–12.5**	–14.1	–11.1	**–22.9**	**–10.9**	**–2.4**	**–3.2**	–1.0	**–7.4**	**–7.3**	**–6.9**	**–7.5**	–5.1	**–16.4**	**–11.2**	**–6.9**	**–7.9**	–5.3
Elasso	–31.0	–19.1	–11.1	–13.6	**–11.2**	–22.7	–9.7	–1.8	–2.4	**–1.7**	**–7.4**	–7.2	–6.1	–5.7	–3.5	–16.3	–10.6	–6.0	–6.8	–4.9

Note: The Base row shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, the proposed methods are indicated with a gray background, and methods that outperform the benchmark method are marked in bold.

the constant term has minimal impact relative to the data variability. In addition, we observe that the MinT and MinTs methods perform especially well and our methods provide results similar to these benchmark methods. This is attributed to the use of in-sample covariance by MinT and MinTs, which allows for large adjustments in reconciliation for base forecasts with high estimated error variance. Elasso forecasts are slightly worse than EMinT, possibly due to the difficulty of identifying underperforming base forecasts in this simulation setup. We have also considered alternative error correlation values, $\rho = -0.6, -0.2, 0.2, 0.4$, for this simulation setting, but to save space, we do not present all results. The omitted results follow a similar pattern and are available upon request.

In Table 5 and Table A5, we present the proportion of time series being selected by applying our proposed methods. As observed in Table 5, for OLS, WLSs, and WLSv estimators, the Subset and Intuitive methods are still able to exclude the series Total, A, and BA in some instances, in which small biases are introduced in model fitting, while essentially retaining the remaining series in the hierarchy. The Subset methods perform better than the Intuitive method in selection. The Lasso methods typically select all bottom-level series since they tend to yield dense estimates as discussed in Section 3.1. Elasso also selects

all bottom-level series. When dealing with a high positive error correlation, Table A5 shows that our methods still have the potential to do some selection but it becomes somewhat challenging to identify and exclude the series that should be omitted in reconciliation. Hence, our methods are preferred, particularly when the error correlation within the hierarchical structure is negative.

Table 5: *Proportion of time series being selected after using the proposed reconciliation methods with selection in Setup 2, with the error correlation being -0.8.*

	Top	A	B	AA	AB	BA	BB	Summary
OLS-subset	0.32	0.34	0.95	0.98	1	0.74	1.00	
OLS-intuitive	0.58	0.52	0.93	0.97	1	0.61	0.97	
OLS-lasso	0.61	0.34	0.38	1.00	1	1.00	1.00	
WLSs-subset	0.27	0.40	0.98	1.00	1	0.73	1.00	
WLSs-intuitive	0.49	0.57	0.96	1.00	1	0.74	0.99	
WLSs-lasso	0.48	0.62	0.72	1.00	1	1.00	1.00	
WLSv-subset	0.30	0.42	1.00	1.00	1	0.68	1.00	
WLSv-intuitive	0.49	0.53	0.99	1.00	1	0.47	1.00	
WLSv-lasso	0.35	0.70	0.85	1.00	1	1.00	1.00	
MinT-subset	1.00	1.00	1.00	1.00	1	1.00	1.00	
MinT-intuitive	1.00	1.00	1.00	1.00	1	1.00	1.00	
MinT-lasso	1.00	1.00	1.00	1.00	1	1.00	1.00	
MinTs-subset	0.87	0.85	1.00	1.00	1	0.85	1.00	
MinTs-intuitive	1.00	1.00	1.00	1.00	1	1.00	1.00	
MinTs-lasso	0.86	0.84	1.00	1.00	1	0.85	1.00	
Elasso	0.94	0.79	0.93	1.00	1	1.00	1.00	

Note: the last column displays a stacked barplot for each method, based on the total number of selected series data from 500 simulation instances, with a darker sub-bar indicating a larger number.

5 Applications

In this section we describe two empirical applications: Section 5.1 focuses on a grouped hierarchy built using the Australian labour force survey data released by the Australian Bureau of Statistics, while Section 5.2 considers Australian domestic tourism flows with a natural geographic hierarchy.

5.1 Forecasting Australian labour force

The dataset from the Labour Force Survey was released by the Australian Bureau of Statistics, and comprises monthly data on the number of unemployed persons in Australia for the period from January

2010 to July 2023[1]. To deal with the few missing observations, we use linear interpolation between observations. Analysis of unemployment data by labour market region and duration of job search can provide valuable insights into regional disparities, and the structural nuances underlying unemployment. Forecast reconciliation is crucial in such a case to ensure aligned decision making.

We construct a grouped hierarchy by disaggregating the number of unemployed persons over two independent attributes, duration of job search (referred to as *Duration*), and State and Territory (referred to as *STT*). At the bottom level, the data are disaggregated by both attributes. We refer to the bottom level as the *Duration* $\times$ *STT* level. Specifically, there are six different groups of job search duration, under 1 month, 1–3 months, 3–6 months, 6–12 months, 1–2 years, and 2 years and over. Additionally, the number of unemployed persons in Australia can be disaggregated by eight states and territories, i.e., NSW (New South Wales), VIC (Victoria), QLD (Queensland), SA (South Australia), WA (Western Australia), TAS (Tasmania), NT (Northern Territory), and ACT (Australian Capital Territory). So the final grouped hierarchy consists of the top series, six series at the Duration level, eight series at the STT level, and 48 series at the Duration $\times$ STT level, giving 63 time series in total, each of length 163 observations.

The top panel in Figure 3 shows the total number of unemployed persons in Australia from January 2010 to July 2023, representing the top-level series in the hierarchical structure. The monthly series shows strong seasonality within each year, marked by prominent peaks occurring every January, attributable to school-leavers. Lower peaks occur in July, perhaps impacted by the start of the financial year. Amidst the backdrop of COVID-19's non-essential service shutdowns and trading restrictions, March and April 2020 saw a notable surge in unemployment. However, as coronavirus cases dwindled significantly and restrictions eased in the aftermath, employment made a remarkable recovery, leading to a subsequent decline in unemployment. The bottom-left panel displays the breakdown of unemployed individuals by state and territory, while the bottom-right panel presents the breakdown by the duration of job search. The plots display diverse and rich dynamics both within and between different levels of the hierarchy. For

[1]The Labour Force Survey data is publicly available at https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia-detailed/aug-2023.

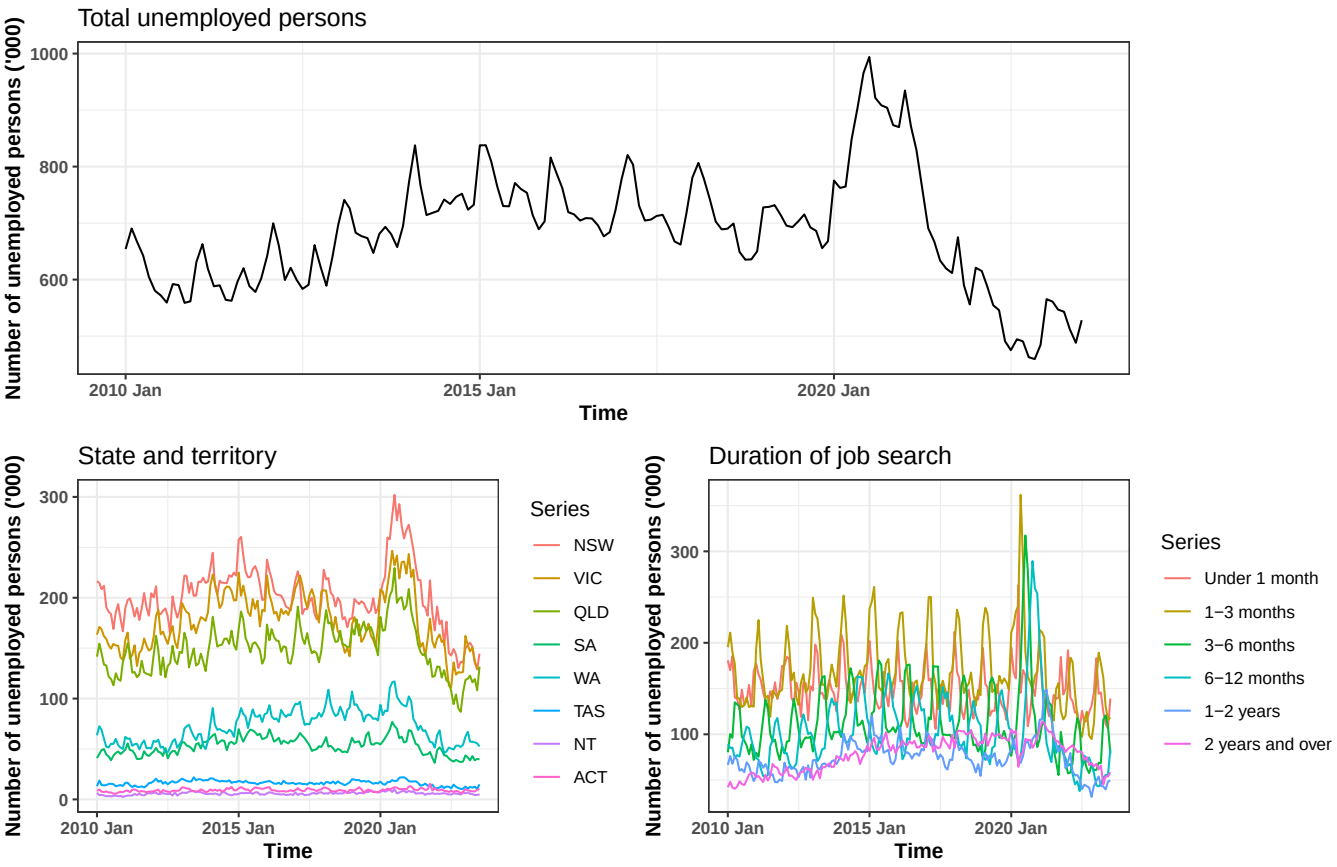

Figure 3: *Australia unemployed persons, disaggregated by state and territory, and by duration of job search.*

example, there was noticeable growth observed during 2020 for some states such as NSW, VIC, and QLD, whereas other states did not experience such significant growth. Additionally, there is a resemblance in the seasonal patterns between NSW and QLD, while the seasonal pattern in VIC differs. When comparing the series at the STT level and Duration level, the seasonal patterns in the Duration-level series are more consistent and potentially easier to forecast.

We assess the forecast accuracy of base forecasts and various reconciliation methods through a rolling forecast origin approach. Our aim is to generate 1- to 12-step-ahead forecasts for each of the 63 series while ensuring coherence. Given the limited data compared to the forecast horizon, we initiate the process with a training set of 139 observations for each series. The training set is used to select the optimal ETS model with the automatic algorithm implemented in the **forecast** package for R (Hyndman & Khandakar 2008). Using these fitted ETS models, we generate base forecasts, and perform diverse

forecast reconciliation methods. Then we roll the forecast origin forward by one month and repeat the process, until July 2022. We note that it may be challenging to identify the series with "poor" forecasts due to structural changes in the data caused by the COVID-19 pandemic, which affect the accuracy of forecasts across all time series.

Table 6: *Average out-of-sample forecast results for Australian labour force data.*

	Top				Duration				STT				Duration x STT				Average			
Method	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12
Base	29.4	44.9	58.6	67.6	10.1	14.2	16.3	18.1	6.6	8.4	9.9	10.7	2.3	2.9	3.1	3.3	4.0	5.3	6.1	6.6
BU	46.7	34.1	29.3	24.2	7.4	2.3	0.8	0.8	5.1	9.8	10.5	10.4	0.0	0.0	0.0	0.0	8.4	7.1	6.8	6.3
OLS	2.0	1.7	1.5	1.0	0.6	–4.2	–4.3	–3.3	–0.7	0.4	0.0	–0.1	1.9	0.7	0.8	0.7	1.0	–0.5	–0.6	–0.5
OLS-subset	2.1	**1.0**	**–1.2**	**–2.0**	0.6	–4.1	**–5.2**	**–4.4**	**–1.0**	0.6	**–0.6**	**–1.2**	1.9	0.8	**0.3**	**0.2**	1.0	–0.5	**–1.5**	**–1.6**
OLS-intuitive	**–1.3**	**1.5**	**1.0**	**0.2**	**–0.9**	–3.9	–4.2	–3.2	**–1.3**	0.8	0.5	0.6	**1.8**	1.2	1.3	1.2	**0.1**	–0.1	–0.3	–0.2
OLS-lasso	2.0	1.7	1.5	1.0	0.6	–4.2	–4.3	–3.3	–0.7	0.4	0.0	–0.1	1.9	0.7	0.8	0.7	1.0	–0.5	–0.6	–0.5
WLSs	16.9	10.8	9.0	7.0	–0.7	–4.8	–5.0	–4.3	–3.0	0.6	1.4	1.6	–1.3	**–1.7**	**–1.5**	**–1.6**	0.6	–0.3	–0.2	–0.3
WLSs-subset	**14.9**	**9.7**	**6.4**	**4.2**	**–1.6**	–3.7	–4.1	–3.4	–2.3	1.5	1.6	**1.0**	–0.6	–0.2	0.5	0.3	0.6	0.6	0.4	0.1
WLSs-intuitive	16.9	10.8	9.0	7.0	–0.7	–4.8	–5.0	–4.3	–3.0	0.6	1.4	1.6	–1.3	**–1.7**	**–1.5**	**–1.6**	0.6	–0.3	–0.2	–0.3
WLSs-lasso	16.9	10.8	9.0	7.0	–0.7	–4.8	–5.0	–4.3	–3.0	0.6	1.4	1.6	–1.3	**–1.7**	**–1.5**	**–1.6**	0.6	–0.3	–0.2	–0.3
WLSv	15.6	9.8	8.4	6.6	1.1	–5.0	**–5.6**	–4.5	–2.6	–0.5	–0.1	0.2	**–1.5**	–1.5	**–1.5**	–1.3	0.9	–0.7	–0.8	–0.6
WLSv-subset	**10.2**	**5.1**	**2.9**	**1.8**	**–0.6**	**–5.4**	–5.5	**–4.6**	–1.8	**–0.8**	**–1.1**	**–1.0**	–1.1	–1.1	–0.7	–0.5	**0.2**	**–1.3**	**–1.5**	**–1.3**
WLSv-intuitive	15.6	9.8	8.4	6.6	1.1	–5.0	**–5.6**	–4.5	–2.6	–0.5	–0.1	0.2	**–1.5**	–1.5	**–1.5**	–1.3	0.9	–0.7	–0.8	–0.6
WLSv-lasso	15.6	9.8	8.4	6.6	1.1	–5.0	**–5.6**	–4.5	–2.6	–0.5	–0.1	0.2	**–1.5**	–1.5	**–1.5**	–1.3	0.9	–0.7	–0.8	–0.6
MinTs	9.9	5.8	6.4	5.3	0.6	–5.0	–5.0	–3.7	–3.4	–2.4	–1.4	–0.8	–0.4	–0.9	–0.8	–0.7	0.4	**–1.3**	–0.9	–0.5
MinTs-subset	**9.1**	6.1	6.5	**4.5**	**0.3**	–4.8	**–5.1**	**–3.9**	**–3.6**	–2.4	–1.3	**–0.9**	**–0.6**	–0.8	–0.7	–0.7	**0.1**	–1.2	–0.9	**–0.7**
MinTs-intuitive	9.9	5.8	6.4	5.3	0.6	–5.0	–5.0	–3.7	–3.4	–2.4	–1.4	–0.8	–0.4	–0.9	–0.8	–0.7	0.4	**–1.3**	–0.9	–0.5
MinTs-lasso	9.9	5.8	6.4	5.3	0.6	–5.0	–5.0	–3.7	–3.4	–2.4	–1.4	–0.8	–0.4	–0.9	–0.8	–0.7	0.4	**–1.3**	–0.9	–0.5
EMinT	43.1	17.6	9.9	10.2	36.8	25.2	27.9	24.1	16.8	15.1	6.2	6.2	32.3	27.9	29.8	27.9	31.4	23.3	21.4	19.6
Elasso	**–5.8**	**–2.0**	**–2.5**	**–2.3**	**33.5**	**11.1**	**1.1**	**–3.6**	**–17.4**	**–8.9**	**–10.0**	**–8.8**	**20.6**	**6.4**	**2.6**	**0.5**	**12.7**	**3.4**	**–1.1**	**–2.9**

Note: The Base row shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, the proposed methods are indicated with a gray background, and methods that outperform the benchmark method are marked in bold.

The average results are presented in Table 6. The poor performance of the MinT method and associated methods can be attributed to the poor sample covariance estimator when the sample size is only slightly larger than the number of series in the structure. The Subset methods using different estimators of $\boldsymbol{G}$ generally improve forecast accuracy over their benchmark methods, particularly when focusing on aggregation levels, which are typically of paramount concern to practitioners. The only exception is the WLSs-subset method, which returns reduced accuracy for longer horizons. However, it still demonstrates improvements in top-level forecasts. Moreover, the Intuitive and Lasso methods almost always yield results identical to the corresponding benchmark methods, because they tend to provide dense estimates, and ETS models typically do not result in extremely poor forecasts. The only exception is OLS-intuitive, which shows improved forecast accuracy at the top level but deterioration at other levels. When we drop the unbiasedness assumption, EMinT is the worst-performing method across all levels because it relies on

the assumption that the series in the hierarchy are jointly weakly stationary, which is evidently not the case in the application. Elasso significantly improves the quality of forecasts over EMinT, with the most accurate coherent forecasts observed at the top level and STT level. Overall, Elasso performs well for longer forecast horizons, but it is less effective for one-step-ahead forecasts.

Table 7: *Out-of-sample forecast results on a single test set (from August 2022 to July 2023) for Australian labour force data.*

	Top				Duration				STT				Duration x STT				Average			
Method	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12
Base	18.5	13.6	18.3	28.3	11.8	12.7	13.9	16.9	6.7	6.0	6.0	6.3	2.3	2.6	2.7	2.9	4.1	4.1	4.4	5.1
BU	**–81.5**	33.4	–19.9	–45.0	**–30.7**	–9.2	–7.9	–10.1	–12.9	–10.4	–13.4	–13.5	0.0	0.0	0.0	0.0	**–17.1**	–2.8	–5.9	–9.3
OLS	–16.2	–14.2	–13.4	–10.4	2.5	–2.6	–2.7	–0.6	–1.8	–0.9	–1.9	0.3	6.7	5.1	5.1	4.9	2.1	0.7	0.4	1.1
OLS-subset	**–17.0**	–2.1	**–31.2**	**–38.4**	**2.0**	–1.7	**–5.0**	**–2.7**	**–2.7**	**–4.2**	**–8.6**	**–7.3**	6.7	5.2	**3.3**	**3.7**	**1.7**	1.1	**–3.5**	**–3.8**
OLS-intuitive	**–79.6**	**–23.9**	**–31.9**	**–32.1**	**–13.0**	0.4	–0.8	0.3	**–8.9**	4.9	7.0	13.2	**6.3**	12.3	12.4	11.6	**–8.5**	5.6	4.7	4.4
OLS-lasso	–16.2	–14.2	–13.4	–10.4	2.5	–2.6	–2.7	–0.6	–1.8	–0.9	–1.9	0.3	6.7	5.1	5.1	4.9	2.1	0.7	0.4	1.1
WLSs	–60.6	–29.4	–44.0	–38.6	–12.0	–7.6	–6.9	–5.9	–6.5	–8.1	–9.4	–8.0	3.2	1.7	1.7	1.6	–7.7	–4.4	–5.8	–5.9
WLSs-subset	**–61.6**	–22.4	**–47.3**	**–50.4**	–12.0	**–8.0**	**–10.5**	**–7.8**	**–6.6**	**–10.7**	**–14.3**	**–12.8**	3.2	5.6	4.1	5.9	**–7.8**	–2.8	**–6.7**	**–6.4**
WLSs-intuitive	–60.6	–29.4	–44.0	–38.6	–12.0	–7.6	–6.9	–5.9	–6.5	–8.1	–9.4	–8.0	3.2	1.7	1.7	1.6	–7.7	–4.4	–5.8	–5.9
WLSs-lasso	–60.6	–29.4	–44.0	–38.6	–12.0	–7.6	–6.9	–5.9	–6.5	–8.1	–9.4	–8.0	3.2	1.7	1.7	1.6	–7.7	–4.4	–5.8	–5.9
WLSv	–60.6	–29.1	–41.4	–36.6	–14.5	–8.7	–5.6	–4.8	–3.3	–6.8	–8.0	–7.0	5.5	2.6	2.6	3.1	–6.7	–4.0	–4.5	–4.6
WLSv-subset	–51.6	**–32.7**	–36.6	–29.6	**–18.3**	**–9.8**	**–10.5**	**–10.9**	–1.1	–4.3	**–8.1**	**–7.3**	**2.5**	**2.1**	**2.3**	**1.8**	**–7.9**	**–4.3**	**–5.8**	**–6.5**
WLSv-intuitive	–60.6	–29.1	–41.4	–36.6	–14.5	–8.7	–5.6	–4.8	–3.3	–6.8	–8.0	–7.0	5.5	2.6	2.6	3.1	–6.7	–4.0	–4.5	–4.6
WLSv-lasso	–60.6	–29.1	–41.4	–36.6	–14.5	–8.7	–5.6	–4.8	–3.3	–6.8	–8.0	–7.0	5.5	2.6	2.6	3.1	–6.7	–4.0	–4.5	–4.6
MinTs	–27.0	–21.1	–22.9	–21.8	–9.1	–7.9	–6.7	–4.6	–3.6	–9.0	–10.5	–7.6	7.7	3.7	3.0	3.4	–1.9	–3.3	–3.9	–3.1
MinTs-subset	**–41.4**	–9.3	–17.8	**–45.1**	**–12.2**	–5.0	**–8.2**	**–6.8**	**–6.1**	**–9.8**	–7.1	**–9.3**	**5.3**	4.7	3.8	3.6	**–5.3**	–1.5	–3.0	**–6.1**
MinTs-intuitive	–27.0	–21.1	–22.9	–21.8	–9.1	–7.9	–6.7	–4.6	–3.6	–9.0	–10.5	–7.6	7.7	3.7	3.0	3.4	–1.9	–3.3	–3.9	–3.1
MinTs-lasso	–27.0	–21.1	–22.9	–21.8	–9.1	–7.9	–6.7	–4.6	–3.6	–9.0	–10.5	–7.6	7.7	3.7	3.0	3.4	–1.9	–3.3	–3.9	–3.1
EMinT	–60.4	–14.0	1.4	–29.9	–6.0	12.0	10.7	–6.7	16.7	–0.9	–12.4	**–21.0**	23.3	17.2	16.7	10.1	7.7	10.8	9.0	–3.7
Elasso	–4.2	–3.3	**–22.3**	–8.0	**–19.7**	**–9.9**	**–19.9**	**–25.3**	**–24.6**	**–24.3**	**–22.6**	–14.6	**–10.8**	**–3.8**	**–0.2**	**–4.9**	**–15.7**	**–9.3**	**–11.4**	**–13.2**

Note: The Base row shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, the proposed methods are indicated with a gray background, and methods that outperform the benchmark method are marked in bold.

We provide results based on the final test set spanning from August 2022 to July 2023 in Table 7. This is the latest available data, enabling us to use more data for model training and explore the post-COVID pattern. All Subset methods (using various estimators of $\boldsymbol{W}$) produce improved or comparable reconciled forecasts compared to their benchmarks. The accuracy improvements become more noticeable for longer forecast horizons. Similar to the average results in Table 6, the Intuitive and Lasso methods yield results identical to the benchmark methods due to their tendency to offer dense estimates. Surprisingly, when relaxing the unbiasedness constraint, Elasso ranks the best and demonstrates significant improvement over EMinT, and outperforms other methods across almost all levels except for the top level.

Table 8 presents the number of series selected at each level and the optimal tuning parameter values obtained using different proposed methods. We only show results from the Subset and Elasso methods, as they had the best RMSE results. The variation in the scale of the optimal parameters for different methods

Table 8: *Number of time series selected using different proposed methods and the optimal parameter values identified in the labour application, considering a single test set (from August 2022 to July 2023). The None row shows the original number of series in the structure.*

	Number of time series retained					Optimal parameters		
	Top	Duration	STT	Duration x STT	Total	λ	λ_0	λ_2
None	1	6	8	48	63	-	-	-
OLS-subset	0	5	1	48	54	-	4.16	1.00
WLSs-subset	0	5	1	46	52	-	0.38	0.10
WLSv-subset	1	5	7	48	61	-	0.51	1.00
MinTs-subset	0	1	1	47	49	-	0.03	0.01
Elasso	1	5	2	3	11	213.59	-	-

comes from the difference in the scales of the objective function. Table 8 shows that all Subset methods exclude some series. Remarkably, the Elasso method consistently outperforms the others overall, even though it uses only 11 series for forecast reconciliation. Most of the series at the STT level are removed, while the majority of series at the Duration level are retained. This aligns with our data description, highlighting that the seasonal patterns in the Duration level series are more consistent and potentially easier to forecast compared to those at the STT level.

5.2 Forecasting Australian domestic tourism

Australian domestic tourism flows are measured as the number of overnight trips Australians spend away from home. The data are sourced from the National Visitor Survey and collected through computer-assisted telephone interviews involving approximately $120,000$ Australian residents aged 15 years and older. The data follow a geographic structure, with national total tourism flows in the top-level aggregation, then disaggregated into seven states and territories (referred to as *State* level hereafter), further dividing into 27 zones, and finally, into 76 regions. Thus, $n_b = 76$ and $n = 111$. Each series in the hierarchy spans the period from January 1998 to December 2017, with a total of 240 observations.

Figure 4 shows the aggregate tourism flows for Australia as well as individual states, revealing pronounced seasonal patterns across the national total and states, albeit with varying seasonal patterns among the series. Notably, there was significant growth starting from around 2010 for the national total flow and some states such as NSW, VIC, QLD, and WA. While flows are relatively flat for SA, TAS, and NT. Moreover, the

time plot displays that a large decrease in tourism flows for WA occurred in 2016.

Our objective is to forecast tourism flows for each series in the geographic hierarchy while ensuring coherence across all levels. We use a rolling forecast origin to evaluate the forecast accuracy of different methods. We start with a training set of 216 months for each series, and compute base forecasts from optimal ETS models. We then roll the forecast origin forward, month by month, until December 2016. The base forecasts are reconciled using our proposed methods and some existing reconciliation methods.

Table 9 reports the average RMSE values for base forecasts generated by ETS models, along with the percentage relative improvements obtained by each reconciliation method. Similar to Section 5.1, the MinT method and the respective proposed methods are not considered due to their poor performance. The

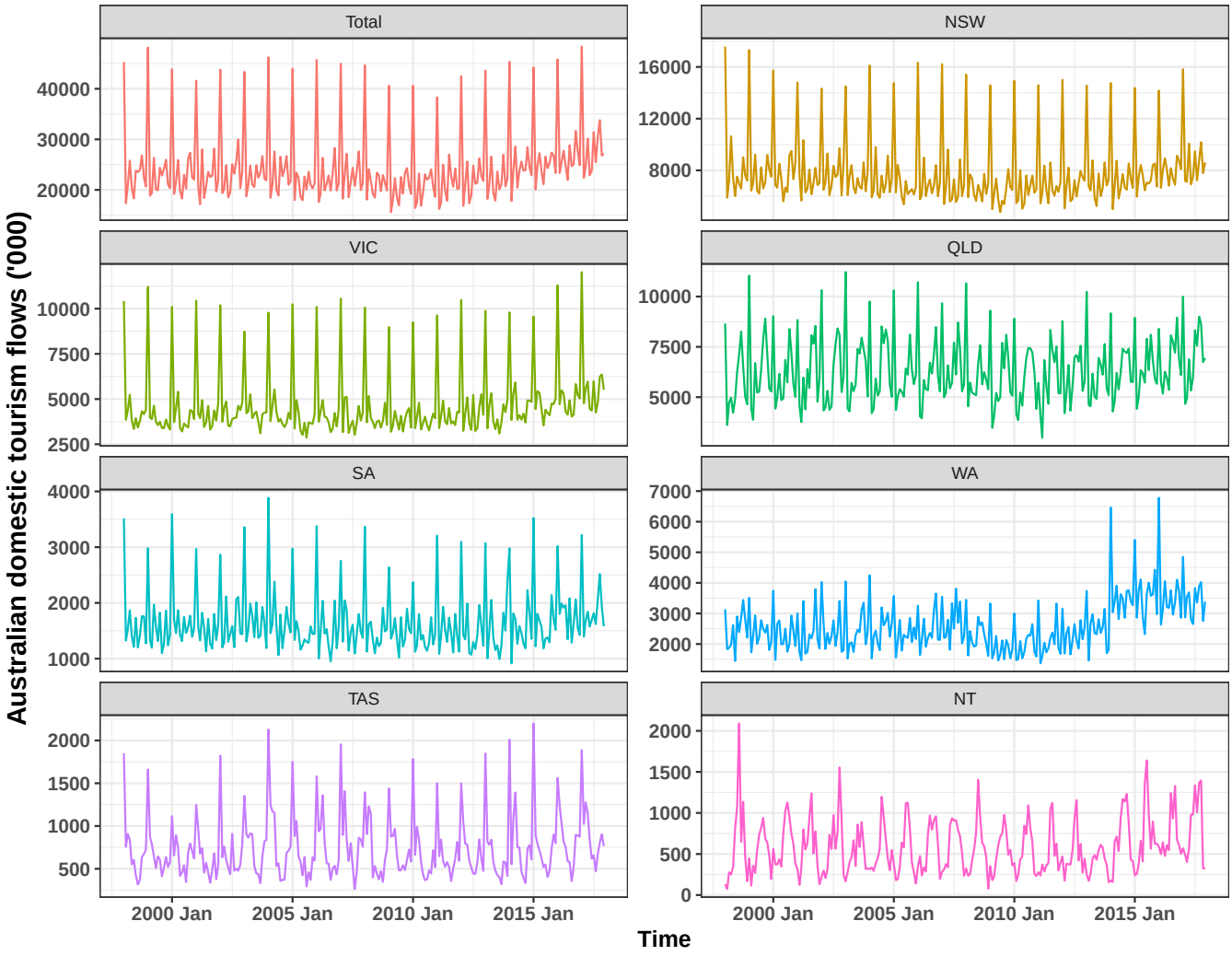

Figure 4: *Domestic tourism flows from January 1998 to December 2017 for the whole of Australia as well as the states.*

results show that the OLS method outperforms other benchmark methods like WLSs, WLSv and MinTs, despite the fact that WLSv and MinTs account for the in-sample covariance of base forecast errors. This highlights the effectiveness of the OLS method despite its simplicity.

Overall, the Subset methods outperform their respective benchmark methods, especially for aggregation levels and longer forecast horizons. The only exception is the OLS-subset method, which slightly reduces overall accuracy while still improving top-level forecasts. Moreover, the Intuitive and Lasso methods produce results almost identical to the corresponding benchmark methods, which is not surprising as ETS models typically do not yield extremely poor forecasts, making them challenging to be selected out using methods that tend to return dense estimates. When we relax the unbiasedness constraint, EMinT consistently performs the worst across all levels due to the evident lack of joint weak stationarity among the series in the hierarchy. The Elasso method presents significant improvement compared to the EMinT method, and it also outperforms other methods across almost all levels except for the bottom level.

Table 9: *Average out-of-sample forecast results for Australian domestic tourism data.*

	Top				State				Zone				Region				Average			
Method	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12
Base	1565.8	1520.2	1548.3	1773.1	366.0	406.1	421.4	442.0	142.5	170.8	178.5	185.3	72.1	86.4	90.9	94.4	121.2	140.0	146.2	153.6
BU	14.3	38.8	42.0	38.8	4.6	10.3	13.8	15.7	–0.1	0.9	1.3	1.6	0.0	0.0	0.0	0.0	2.5	6.0	6.9	7.4
OLS	–0.6	1.1	1.8	1.9	–1.2	–1.0	–1.0	–1.3	–2.8	–4.0	–4.8	–5.6	–0.1	–0.8	–1.6	–2.4	–1.1	–1.6	–2.1	–2.7
OLS-subset	–0.6	**–1.9**	**–4.9**	**–3.0**	–1.2	**–1.2**	1.5	0.9	–2.6	–0.5	–0.2	–1.1	0.2	2.1	1.7	0.6	–1.0	0.3	0.5	–0.2
OLS-intuitive	–0.6	1.1	1.8	1.9	–1.2	–1.0	–1.0	–1.3	–2.8	–4.0	–4.8	–5.6	–0.1	–0.8	–1.6	–2.4	–1.1	–1.6	–2.1	–2.7
OLS-lasso	**–0.8**	2.1	2.8	2.9	**–1.3**	–0.4	0.5	0.3	–2.3	–3.5	–4.2	–4.9	0.0	**–0.9**	–1.5	–2.2	–1.0	–1.3	–1.5	–2.0
WLSs	4.1	16.5	19.0	18.1	0.6	2.0	4.1	5.2	–2.7	–3.1	–3.3	–3.4	–0.5	–1.0	–1.4	–1.8	–0.4	0.7	1.0	1.1
WLSs-subset	4.1	**6.9**	**8.9**	**10.2**	0.6	**1.7**	**4.0**	**4.3**	–2.7	–3.0	–2.1	–2.1	–0.5	–0.1	–0.1	–0.5	–0.4	**0.0**	**0.9**	**1.0**
WLSs-intuitive	4.1	16.5	19.0	18.1	0.6	2.0	4.1	5.2	–2.7	–3.1	–3.3	–3.4	–0.5	–1.0	–1.4	–1.8	–0.4	0.7	1.0	1.1
WLSs-lasso	**3.6**	17.1	19.5	18.5	**0.3**	2.1	4.3	5.5	–2.7	–2.9	–3.2	–3.3	**–0.6**	–1.0	–1.4	–1.7	**–0.5**	0.8	1.1	1.2
WLSv	6.2	22.3	25.0	23.6	1.1	3.7	6.3	7.9	–2.6	–2.4	–2.4	–2.2	–1.6	–1.6	–1.8	–1.9	–0.5	1.5	2.1	2.4
WLSv-subset	**0.9**	**4.5**	**4.9**	**5.7**	**–1.2**	**–0.3**	**–0.1**	**0.7**	**–3.2**	**–3.8**	**–4.5**	**–4.9**	–1.3	–1.2	–1.7	**–2.3**	**–1.6**	**–1.2**	**–1.6**	**–1.7**
WLSv-intuitive	6.2	22.3	25.0	23.6	1.1	3.7	6.3	7.9	–2.6	–2.4	–2.4	–2.2	–1.6	–1.6	–1.8	–1.9	–0.5	1.5	2.1	2.4
WLSv-lasso	6.2	22.3	25.0	23.6	1.1	3.7	6.3	7.9	–2.6	–2.4	–2.4	–2.2	–1.6	–1.6	–1.8	–1.9	–0.5	1.5	2.1	2.4
MinTs	5.1	17.2	19.7	18.6	0.1	1.9	4.4	5.7	–3.5	–3.3	–3.4	–3.4	**–1.9**	**–2.0**	**–2.4**	**–2.7**	–1.2	0.2	0.7	0.9
MinTs-subset	**1.8**	**1.3**	**2.0**	**3.2**	**–2.2**	**–2.1**	**–1.3**	**–0.7**	**–4.2**	**–4.5**	**–4.9**	**–5.4**	–1.5	–1.3	–1.9	–2.5	**–2.0**	**–2.2**	**–2.3**	**–2.5**
MinTs-intuitive	5.1	17.2	19.7	18.6	0.1	1.9	4.4	5.7	–3.5	–3.3	–3.4	–3.4	**–1.9**	**–2.0**	**–2.4**	**–2.7**	–1.2	0.2	0.7	0.9
MinTs-lasso	5.1	17.2	19.7	18.6	0.1	1.9	4.4	5.7	–3.5	–3.3	–3.4	–3.4	**–1.9**	**–2.0**	**–2.4**	**–2.7**	–1.2	0.2	0.7	0.9
EMinT	–2.3	24.3	58.8	59.7	36.9	56.0	68.4	70.4	51.4	64.6	75.8	81.4	65.9	72.3	81.9	85.9	48.3	62.3	75.4	79.0
Elasso	**–17.0**	**–19.4**	**–19.8**	**–18.7**	**–21.6**	**–17.3**	**–19.3**	**–19.6**	**–6.5**	**–9.4**	**–11.5**	**–12.6**	**2.2**	**0.4**	**–1.0**	**–1.8**	**–7.0**	**–7.7**	**–9.2**	**–9.9**

Note: The Base row shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, the proposed methods are indicated with a gray background, and methods that outperform the benchmark method are marked in bold.

We also present the results based on the last one training set spanning from January 2017 to December 2017 in Table 10. The reconciliation errors across each of the 111 series and across the four levels in the hierarchy are displayed in Figure 5. The results show a similar performance to the average results described above, indicating relatively high-quality forecasts from the Subset and Elasso methods.

Table 10: *Out-of-sample forecast results on a single test set (from January 2017 to December 2017) for Australian domestic tourism data.*

	Top				State				Zone				Region				Average			
Method	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12	h=1	1–4	1–8	1–12
Base	1158.2	716.6	1279.5	1907.6	452.7	323.3	349.9	424.8	165.5	163.6	160.7	179.7	100.8	89.4	88.2	94.1	148.3	127.9	133.1	152.1
BU	89.1	132.8	53.4	42.0	–4.6	10.3	17.0	19.7	1.1	–2.4	0.4	1.0	0.0	0.0	0.0	0.0	5.7	7.6	7.6	8.5
OLS	–4.7	–0.4	0.5	1.4	–3.0	–3.9	–1.6	–1.5	–2.1	–4.2	–5.6	–7.5	1.0	–0.4	–1.9	–3.2	–1.0	–2.1	–2.7	–3.6
OLS-subset	–4.7	8.0	**–1.4**	**–14.1**	–3.0	5.5	0.3	**–7.9**	–2.1	–1.5	–3.7	**–8.7**	1.0	1.7	–0.1	–2.3	–1.0	1.7	–1.2	**–6.5**
OLS-intuitive	–4.7	–0.4	0.5	1.4	–3.0	–3.9	–1.6	–1.5	–2.1	–4.2	–5.6	–7.5	1.0	–0.4	–1.9	–3.2	–1.0	–2.1	–2.7	–3.6
OLS-lasso	–4.7	–0.4	0.5	1.4	–3.0	–3.9	–1.6	–1.5	–2.1	–4.2	–5.6	–7.5	1.0	–0.4	–1.9	–3.2	–1.0	–2.1	–2.7	–3.6
WLSs	25.1	55.2	20.8	19.1	–15.8	–5.0	3.5	6.2	–5.9	–5.4	–4.7	–5.0	–0.2	–0.8	–1.6	–2.2	–3.0	–0.1	0.3	0.9
WLSs-subset	25.1	**18.7**	**0.8**	**–7.8**	–15.8	–2.7	**–2.1**	**–6.2**	–5.9	–4.1	**–4.8**	**–8.5**	–0.2	0.3	–1.0	**–2.5**	–3.0	**–0.6**	**–2.1**	**–5.5**
WLSs-intuitive	25.1	55.2	20.8	19.1	–15.8	–5.0	3.5	6.2	–5.9	–5.4	–4.7	–5.0	–0.2	–0.8	–1.6	–2.2	–3.0	–0.1	0.3	0.9
WLSs-lasso	25.1	55.2	20.8	19.1	–15.8	–5.0	3.5	6.2	–5.9	–5.4	–4.7	–5.0	–0.2	–0.8	–1.6	–2.2	–3.0	–0.1	0.3	0.9
WLSv	38.2	76.2	29.6	25.6	–17.4	–3.1	7.0	9.9	–5.0	–4.3	–3.1	–3.2	–4.2	–1.6	–1.8	–2.1	–3.9	1.3	2.0	2.8
WLSv-subset	38.2	**34.5**	**10.7**	**8.5**	–17.4	**–8.8**	**–0.8**	**1.4**	–5.0	**–5.5**	**–5.3**	**–6.7**	–4.1	**–2.0**	**–2.6**	**–3.4**	–3.9	**–2.3**	**–2.0**	**–2.2**
WLSv-intuitive	38.2	76.2	29.6	25.6	–17.4	–3.1	7.0	9.9	–5.0	–4.3	–3.1	–3.2	–4.2	–1.6	–1.8	–2.1	–3.9	1.3	2.0	2.8
WLSv-lasso	38.2	76.2	29.6	25.6	–17.4	–3.1	7.0	9.9	–5.0	–4.3	–3.1	–3.2	–4.2	–1.6	–1.8	–2.1	–3.9	1.3	2.0	2.8
MinTs	20.6	53.6	21.6	19.0	**–22.2**	–7.2	3.5	6.3	**–12.1**	–6.6	–5.1	–5.3	**–5.3**	–2.6	–2.8	–3.1	–8.6	–1.8	–0.3	0.4
MinTs-subset	20.6	**20.0**	**6.4**	**5.6**	**–22.2**	**–11.3**	**–2.5**	**–0.1**	**–12.1**	**–7.5**	**–6.4**	**–7.8**	**–5.3**	**–2.9**	**–3.2**	**–3.9**	–8.6	**–4.5**	**–3.2**	**–3.3**
MinTs-intuitive	20.6	53.6	21.6	19.0	**–22.2**	–7.2	3.5	6.3	**–12.1**	–6.6	–5.1	–5.3	**–5.3**	–2.6	–2.8	–3.1	–8.6	–1.8	–0.3	0.4
MinTs-lasso	20.6	53.6	21.6	19.0	**–22.2**	–7.2	3.5	6.3	**–12.1**	–6.6	–5.1	–5.3	**–5.3**	–2.6	–2.8	–3.1	–8.6	–1.8	–0.3	0.4
EMinT	116.5	97.8	–15.8	–13.7	149.4	114.5	63.5	47.5	108.4	68.4	60.6	54.2	122.1	103.1	90.2	78.2	123.2	93.9	67.9	55.5
Elasso	**–84.5**	**–50.4**	**–16.3**	**–16.4**	**–18.3**	**0.6**	**–9.0**	**–11.4**	**–7.8**	**–8.8**	**–7.5**	**–10.4**	**2.9**	**1.6**	**4.1**	**0.3**	**–10.2**	**–4.4**	**–3.2**	**–6.7**

Note: The Base row shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, the proposed methods are indicated with a gray background, and methods that outperform the benchmark method are marked in bold.

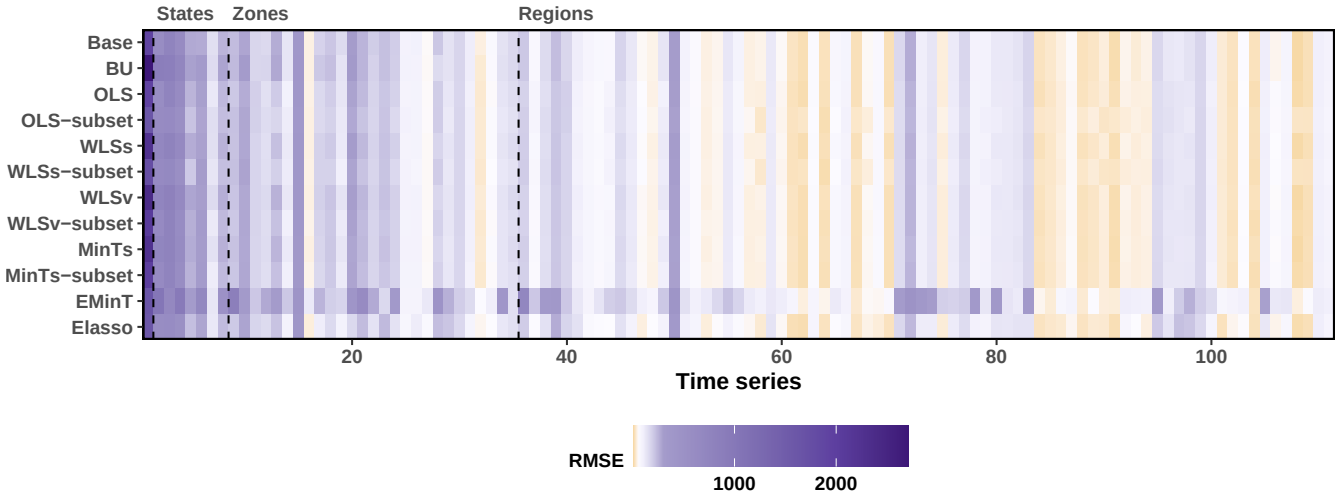

Figure 5: *Average out-of-sample forecasting performance, measured in terms of RMSE (from 1- to 12-step-ahead), for each series across different reconciliation methods. Time series are arranged along the horizontal axis.*

Additionally, Table 11 presents a summary of the number of series selected using different proposed methods for each level as well as the optimal tuning parameter values identified. Here we only give the results of the Subset and Elasso methods since they are useful in the tourism application. Note that the variation in the scale of the optimal parameters for different methods comes from the difference in the scales of the objective. We observe that the OLS-subset and WLSs-subset methods exclude some series at the State and Zone levels for forecast reconciliation. In contrast, the WLSv and MinTs methods retain all

series, which is reasonable because they take into account the in-sample covariance, making themselves allow for larger adjustments made to series with large in-sample forecast error variances in forecast reconciliation. Nonetheless, the WLSv and MinTs methods can still enhance the quality of reconciled forecasts due to the inclusion of shrinkage through additional ridge regularization. It is surprising that Elasso performs exceptionally well despite using only 13 series for reconciliation.

Table 11: *Number of time series selected using different proposed methods and the optimal parameter values identified in the tourism application, considering a single test set (from January 2017 to December 2017). The None row shows the original number of series in the structure.*

	Number of time series retained					Optimal parameters		
	Top	State	Zone	Region	Total	λ	λ_0	λ_2
None	1	7	27	76	111	-	-	-
OLS-subset	1	2	13	76	92	-	27.98	10.00
WLSs-subset	1	1	15	76	93	-	18.73	10.00
WLSv-subset	1	7	27	76	111	-	0.03	0.01
MinTs-subset	1	7	27	76	111	-	0.05	0.01
Elasso	1	4	0	8	13	71759.21	-	-

6 Conclusion

In the existing literature on forecast reconciliation, we map all base forecasts into bottom-level disaggregated forecasts, which are then summed to yield coherent forecasts for the entire structure. The mapping step can be conceptually regarded as a forecast combination. It is common that the base forecasts for some time series perform poorly. This may reduce the overall effectiveness of forecast reconciliation methods. In this paper, we have addressed this issue by introducing a selection mechanism to forecast reconciliation; i.e., incorporating time series selection when reconciling forecasts, while ensuring the generation of coherent forecasts for all series.

Under the unbiasedness constraint, we developed three reconciliation methods with selection mechanisms to automatically remove some base forecasts when forming reconciled forecasts. These methods include group best-subset selection with ridge regularization (Subset), an intuitive method with L_0 regularization (Intuitive), and a group lasso method (Lasso). These methods use different penalty functions designed to penalize the columns of the weighting matrix, $\boldsymbol{G}$, towards zero. Additionally, we relaxed the unbiasedness

constraint and proposed the empirical group lasso method (Elasso) which selects series based on in-sample observations and fitted values.

Simulation experiments and two empirical applications have demonstrated the superiority of the proposed methods over existing reconciliation methods that do not involve series selection. When model misspecification was introduced for some series in the hierarchy, our proposed methods guaranteed coherent forecasts that outperformed or, at least, matched their respective benchmark methods in the minimum trace reconciliation framework. In both empirical applications, where no apparent model misspecification was present, the Subset and Elasso methods were always preferred, particularly for aggregation levels and longer forecast horizons, while the Intuitive and Lasso methods yield results identical to the corresponding benchmark methods, as they tend to provide dense estimates.

A feature of the proposed methods is their ability to reduce the disparities arising from using different estimates of the base forecast error covariance matrix, thereby mitigating the challenges associated with estimator selection, which is a prominent issue within the field of forecast reconciliation research.

As the number of series grows, solving these problems efficiently becomes challenging, and the exact computation of these estimators remains a hurdle. In our study, we have used Gurobi, one of the most widely used commercial solvers, to address NP-hard MIP problems. Despite various efforts to develop MIP-based approaches for solving L_0-regularized regression problems, extending these methods to incorporate additional constraints remains a challenge. We leave this aspect to be addressed in future research.

References

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N. & Panagiotelis, A. (2024), 'Forecast reconciliation: A review', *International J Forecasting* **forthcoming**.

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

Ben Taieb, S. & Koo, B. (2019), Regularized regression for hierarchical forecasting without unbiasedness conditions, *in* 'Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', KDD '19, Association for Computing Machinery, New York, NY, USA, pp. 1337–1347.

Bertsimas, D., King, A. & Mazumder, R. (2016), 'Best subset selection via a modern optimization lens', *The Annals of Statistics* **44**(2), 813–852.

Di Fonzo, T. & Girolimetto, D. (2023), 'Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives', *International J Forecasting* **39**(1), 39–57.

Hazimeh, H. & Mazumder, R. (2020), 'Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms', *Operations Research* **68**(5), 1517–1537.

Hazimeh, H., Mazumder, R. & Radchenko, P. (2023), 'Grouped variable selection with discrete optimization: Computational and statistical perspectives', *The Annals of Statistics* **51**(1), 1–32.

Hazimeh, H., Mazumder, R. & Saab, A. (2022), 'Sparse regression at scale: branch-and-bound rooted in first-order optimization', *Mathematical Programming* **196**(1), 347–388.

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.

Hyndman, R. J. & Athanasopoulos, G. (2021), *Forecasting: principles and practice*, 3rd edn, OTexts, Melbourne, Australia. https://OTexts.com/fpp3.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. & Yasmeen, F. (2023), *forecast: Forecasting functions for time series and linear models*. R package version 8.21.1. https://pkg.robjhyndman.com/forecast/.

Hyndman, R. J. & Khandakar, Y. (2008), 'Automatic time series forecasting: the forecast package for R', *J Statistical Software* **26**(3), 1–22.

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.

Mazumder, R., Radchenko, P. & Dedieu, A. (2022), 'Subset selection with shrinkage: Sparse linear modeling when the SNR is low', *Operations Research* .

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

Wickramasuriya, S. L. (2021), 'Properties of point forecast reconciliation approaches'. arXiv:2103.11129.

Wickramasuriya, S. L., Athanasopoulos, G. & Hyndman, R. J. (2019), 'Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization', *J American Statistical Association* **114**(526), 804–819.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *J Royal Statistical Society. Series B, Statistical Methodology* **68**(1), 49–67.

Appendix

The section provides additional results for the simulation data in Section 4.

Table A1: *Out-of-sample forecast results for the simulated data in Scenario B, Setup 1.*

	Top				Middle				Bottom				Average			
Method	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16
Base	9.6	10.7	12.6	15.6	12.1	14.4	15.3	17.0	4.2	4.9	5.9	7.5	7.2	8.5	9.6	11.4
BU	**–1.0**	0.4	0.6	0.7	**–47.7**	**–49.6**	**–43.6**	**–36.2**	**0.0**	**0.0**	**0.0**	**0.0**	**–23.0**	**–24.0**	**–19.8**	**–15.3**
OLS	8.5	13.9	10.4	7.6	–28.2	–29.4	–26.7	–23.1	22.9	23.9	17.0	11.3	–4.2	–3.8	–4.2	–4.1
OLS-subset	**–0.5**	**0.5**	**0.6**	**0.7**	**–46.3**	**–49.0**	**–43.2**	**–35.9**	**2.2**	**1.0**	**0.7**	**0.5**	**–21.5**	**–23.4**	**–19.4**	**–15.0**
OLS-intuitive	**–0.5**	**0.5**	**0.6**	**0.6**	**–46.5**	**–49.0**	**–43.2**	**–36.0**	**2.2**	**1.2**	**0.7**	**0.5**	**–21.6**	**–23.4**	**–19.4**	**–15.0**
OLS-lasso	**–0.2**	**1.5**	**1.4**	**1.3**	**–46.9**	**–48.9**	**–43.1**	**–35.8**	**0.9**	**0.8**	**0.5**	**0.3**	**–22.1**	**–23.3**	**–19.3**	**–14.9**
WLSs	12.1	18.6	14.0	10.2	–34.4	–35.1	–31.7	–26.9	15.6	17.0	12.0	8.0	–9.0	–8.0	–7.6	–6.5
WLSs-subset	**–0.1**	**1.2**	**1.1**	**1.1**	**–46.7**	**–48.8**	**–43.1**	**–35.8**	**1.5**	**1.1**	**0.8**	**0.6**	**–21.8**	**–23.2**	**–19.2**	**–14.8**
WLSs-intuitive	**0.0**	**1.2**	**1.0**	**0.9**	**–46.5**	**–48.8**	**–43.1**	**–35.9**	**1.7**	**1.3**	**0.9**	**0.6**	**–21.6**	**–23.1**	**–19.2**	**–14.9**
WLSs-lasso	**–0.1**	**1.5**	**1.5**	**1.3**	**–46.7**	**–48.9**	**–43.1**	**–35.8**	**0.9**	**0.8**	**0.5**	**0.3**	**–22.0**	**–23.2**	**–19.3**	**–14.9**
WLSv	–0.8	2.3	1.8	1.6	–46.3	–47.9	–42.3	–35.2	1.6	1.9	1.2	0.8	–21.7	–22.2	–18.6	–14.4
WLSv-subset	–0.7	**1.3**	**1.4**	**1.4**	**–46.9**	**–48.7**	**–42.9**	**–35.6**	**1.0**	**1.0**	**0.8**	**0.6**	**–22.2**	**–23.1**	**–19.1**	**–14.7**
WLSv-intuitive	–0.4	**1.5**	**1.4**	**1.2**	**–46.9**	**–48.6**	**–42.8**	**–35.6**	**0.9**	**1.2**	**0.9**	**0.7**	**–22.2**	**–23.0**	**–19.0**	**–14.7**
WLSv-lasso	–0.6	**1.3**	**1.3**	**1.3**	**–47.2**	**–48.9**	**–43.0**	**–35.7**	**0.6**	**0.8**	**0.5**	**0.4**	**–22.4**	**–23.3**	**–19.2**	**–14.8**
MinT	0.2	0.5	0.6	0.5	–47.5	–49.4	–43.5	–36.1	1.1	0.5	0.3	0.1	–22.3	–23.7	–19.6	**–15.3**
MinT-subset	**–0.1**	0.8	0.9	0.9	–46.9	–49.1	–43.3	–36.0	1.7	0.9	0.5	0.3	–21.9	–23.4	–19.4	–15.1
MinT-intuitive	0.2	0.5	0.6	0.5	–47.5	–49.4	–43.5	–36.1	1.1	0.5	0.3	0.1	–22.3	–23.7	–19.6	**–15.3**
MinT-lasso	**–0.3**	**0.3**	0.6	0.5	**–47.6**	–49.4	–43.5	–36.1	**0.8**	**0.3**	**0.2**	0.1	**–22.5**	**–23.9**	**–19.7**	**–15.3**
MinTs	–0.3	0.3	**0.4**	**0.4**	–47.6	–49.5	**–43.6**	**–36.2**	0.7	0.2	0.1	**0.0**	–22.6	–23.9	**–19.8**	**–15.3**
MinTs-subset	**–0.8**	0.5	0.8	0.8	–47.2	–49.2	–43.4	–36.0	1.0	0.7	0.4	0.3	–22.3	–23.6	–19.5	–15.1
MinTs-intuitive	–0.3	0.3	**0.4**	**0.4**	–47.6	–49.5	**–43.6**	**–36.2**	0.7	0.2	0.1	**0.0**	–22.6	–23.9	**–19.8**	**–15.3**
MinTs-lasso	**–0.9**	**0.2**	0.5	0.5	**–47.7**	–49.5	**–43.6**	**–36.2**	**0.5**	0.2	0.1	0.1	**–22.8**	**–24.0**	**–19.8**	**–15.3**
EMinT	2.2	2.9	2.5	1.7	–46.2	–48.1	–42.4	–35.3	3.6	2.9	2.0	1.1	–20.5	–21.9	–18.2	–14.3
Elasso	**1.4**	**2.7**	**2.4**	**1.6**	**–46.4**	**–48.2**	–42.4	**–35.4**	**3.1**	3.2	2.1	1.2	**–20.9**	–21.9	–18.2	–14.3

Note: The Base row shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, the proposed methods are indicated with a gray background, and methods that outperform the benchmark method are marked in bold.

Table A2: *Out-of-sample forecast results for the simulated data in Scenario C, Setup 1.*

	Top				Middle				Bottom				Average			
Method	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16	h=1	1–4	1–8	1–16
Base	25.0	30.3	30.9	32.3	6.3	7.3	8.6	10.8	4.2	4.9	5.9	7.5	7.8	9.2	10.3	12.0
BU	–62.0	**–64.4**	**–59.0**	–51.5	**–0.3**	**0.0**	**0.1**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**–28.5**	**–30.2**	**–25.3**	**–19.8**
OLS	–34.8	–35.5	–33.5	–30.1	45.3	50.6	37.7	25.1	27.7	29.9	21.2	13.7	3.1	3.8	1.6	–0.2
OLS-subset	**–35.3**	**–41.9**	**–39.2**	**–35.0**	**43.9**	**39.5**	**29.5**	**19.6**	**27.1**	**23.6**	**16.8**	**10.9**	**2.4**	**–3.5**	**–4.2**	**–4.5**
OLS-intuitive	**–41.2**	**–49.2**	**–45.5**	**–40.0**	**35.1**	**26.8**	**20.3**	**13.7**	**21.9**	**15.9**	**11.5**	**7.6**	**–4.0**	**–12.2**	**–10.9**	**–9.1**
OLS-lasso	**–61.8**	**–63.6**	**–58.1**	**–50.9**	**0.4**	**1.3**	**1.3**	**0.7**	**0.3**	**0.8**	**0.6**	**0.4**	**–28.2**	**–29.3**	**–24.5**	**–19.2**
WLSs	–50.9	–52.4	–48.7	–43.3	17.6	20.0	14.5	9.3	9.6	11.3	7.7	4.9	–16.3	–16.7	–14.9	–12.5
WLSs-subset	**–61.8**	**–63.6**	**–58.1**	**–50.7**	**0.3**	**1.4**	**1.4**	**0.9**	**0.3**	**0.9**	**0.7**	**0.6**	**–28.2**	**–29.3**	**–24.4**	**–19.0**
WLSs-intuitive	**–61.8**	**–63.8**	**–58.3**	**–50.9**	**0.0**	**1.0**	**1.0**	**0.7**	**0.3**	**0.7**	**0.6**	**0.5**	**–28.3**	**–29.5**	**–24.6**	**–19.2**
WLSs-lasso	**–61.7**	**–63.5**	**–58.0**	**–50.7**	**0.5**	**1.5**	**1.4**	**0.9**	**0.3**	**0.9**	**0.7**	**0.5**	**–28.1**	**–29.2**	**–24.4**	**–19.1**
WLSv	–61.1	–63.4	–58.1	–50.8	1.0	1.7	1.3	0.8	0.7	1.0	0.6	0.4	–27.6	–29.1	–24.5	–19.2
WLSv-subset	**–61.9**	**–63.6**	**–58.2**	**–50.9**	**0.2**	**1.3**	**1.2**	0.8	**0.1**	**0.8**	0.6	0.5	**–28.3**	**–29.3**	–24.5	–19.2
WLSv-intuitive	**–61.8**	**–63.8**	**–58.3**	**–51.0**	**0.0**	**1.1**	**1.1**	**0.6**	**0.1**	**0.6**	**0.5**	0.4	**–28.4**	**–29.5**	**–24.7**	**–19.3**
WLSv-lasso	**–61.8**	**–63.9**	**–58.4**	**–51.1**	**0.2**	**0.9**	**0.9**	**0.5**	**0.1**	**0.5**	**0.4**	**0.3**	**–28.3**	**–29.6**	**–24.8**	**–19.4**
MinT	–62.1	–64.3	–58.9	**–51.6**	–0.2	0.6	0.5	0.2	0.8	0.5	0.3	0.1	–28.3	–29.9	–25.1	**–19.8**
MinT-subset	–61.8	–63.7	–58.2	–50.9	0.4	1.2	1.3	0.8	0.8	1.0	0.7	0.5	–28.0	–29.3	–24.5	–19.2
MinT-intuitive	–62.1	–64.3	–58.9	**–51.6**	–0.2	0.6	0.5	0.2	0.8	0.5	0.3	0.1	–28.3	–29.9	–25.1	**–19.8**
MinT-lasso	–62.1	**–64.4**	–58.9	–51.5	**–0.3**	**0.3**	**0.4**	**0.1**	**0.6**	**0.3**	**0.1**	0.1	**–28.4**	**–30.1**	**–25.2**	**–19.8**
MinTs	**–62.2**	**–64.4**	**–59.0**	**–51.6**	**–0.3**	0.3	0.4	0.1	0.4	0.3	0.1	**0.0**	**–28.5**	–30.1	–25.2	**–19.8**
MinTs-subset	–62.0	–63.8	–58.4	–51.1	0.4	1.1	1.2	0.7	0.5	0.9	0.7	0.5	–28.2	–29.5	–24.6	–19.3
MinTs-intuitive	**–62.2**	**–64.4**	**–59.0**	**–51.6**	**–0.3**	0.3	0.4	0.1	0.4	0.3	0.1	**0.0**	**–28.5**	–30.1	–25.2	**–19.8**
MinTs-lasso	**–62.2**	**–64.4**	–58.9	–51.5	–0.2	0.3	0.4	0.1	**0.2**	**0.2**	0.1	**0.0**	**–28.5**	–30.1	–25.2	**–19.8**
EMinT	–60.7	–63.5	–58.2	–51.0	2.5	2.9	2.3	1.3	3.6	2.9	2.0	1.1	–26.2	–28.3	–23.8	–18.9
Elasso	**–60.9**	**–63.6**	–58.2	**–51.1**	**2.3**	**2.8**	2.3	1.3	**3.1**	3.1	2.1	1.2	**–26.5**	–28.3	–23.8	–18.9

Note: The Base row shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, the proposed methods are indicated with a gray background, and methods that outperform the benchmark method are marked in bold.

Table A3: *Proportion of time series being selected after using the proposed reconciliation methods with selection in Scenario B, Setup 1.*

	Top	A	B	AA	AB	BA	BB	Summary
OLS-subset	0.55	0.04	0.41	0.74	0.78	0.79	0.83	
OLS-intuitive	0.61	0.04	0.52	0.75	0.69	0.69	0.83	
OLS-lasso	0.04	0.35	0.02	1.00	1.00	1.00	1.00	
WLSs-subset	0.45	0.06	0.36	0.81	0.84	0.81	0.87	
WLSs-intuitive	0.61	0.06	0.48	0.75	0.71	0.73	0.84	
WLSs-lasso	0.02	0.33	0.02	1.00	1.00	1.00	1.00	
WLSv-subset	0.54	0.29	0.46	0.91	0.94	0.86	0.89	
WLSv-intuitive	0.59	0.32	0.53	0.82	0.86	0.77	0.86	
WLSv-lasso	0.27	0.42	0.26	1.00	1.00	1.00	1.00	
MinT-subset	0.69	0.64	0.66	0.95	0.96	0.90	0.90	
MinT-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinT-lasso	0.82	0.74	0.83	1.00	0.99	0.97	0.97	
MinTs-subset	0.62	0.63	0.58	0.95	0.96	0.90	0.86	
MinTs-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinTs-lasso	0.68	0.75	0.68	1.00	1.00	1.00	1.00	
Elasso	0.78	0.95	0.68	1.00	1.00	1.00	1.00	

Note: the last column displays a stacked barplot for each method, based on the total number of selected series data from 500 simulation instances, with a darker sub-bar indicating a larger number.

Table A4: *Proportion of time series being selected after using the proposed reconciliation methods with selection in Scenario C, Setup 1.*

	Top	A	B	AA	AB	BA	BB	Summary
OLS-subset	0.75	0.45	0.44	0.82	0.79	0.83	0.80	
OLS-intuitive	0.47	0.70	0.69	0.86	0.92	0.90	0.89	
OLS-lasso	0.38	0.01	0.01	1.00	1.00	1.00	1.00	
WLSs-subset	0.08	0.42	0.41	0.87	0.85	0.84	0.89	
WLSs-intuitive	0.06	0.55	0.50	0.66	0.87	0.69	0.88	
WLSs-lasso	0.35	0.03	0.03	1.00	1.00	1.00	1.00	
WLSv-subset	0.31	0.67	0.65	0.88	0.90	0.91	0.90	
WLSv-intuitive	0.34	0.63	0.60	0.80	0.89	0.84	0.87	
WLSv-lasso	0.45	0.35	0.36	1.00	1.00	1.00	1.00	
MinT-subset	0.69	0.78	0.80	0.91	0.91	0.91	0.91	
MinT-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinT-lasso	0.75	0.89	0.86	0.97	0.97	0.97	0.97	
MinTs-subset	0.67	0.74	0.76	0.90	0.89	0.88	0.91	
MinTs-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinTs-lasso	0.77	0.72	0.73	1.00	1.00	1.00	1.00	
Elasso	0.95	0.64	0.64	1.00	1.00	1.00	1.00	

Note: the last column displays a stacked barplot for each method, based on the total number of selected series data from 500 simulation instances, with a darker sub-bar indicating a larger number.

Table A5: *Proportion of time series being selected after using the proposed reconciliation methods with selection in Setup 2, with the error correlation being 0.8.*

	Top	A	B	AA	AB	BA	BB	Summary
OLS-subset	0.33	0.52	0.96	0.95	0.98	0.96	0.78	
OLS-intuitive	0.54	0.77	0.93	0.89	0.97	0.83	0.85	
OLS-lasso	0.69	0.53	0.60	1.00	1.00	1.00	1.00	
WLSs-subset	0.29	0.60	1.00	1.00	1.00	0.98	0.86	
WLSs-intuitive	0.63	0.67	0.99	0.98	1.00	0.93	0.86	
WLSs-lasso	0.69	0.76	0.91	1.00	1.00	1.00	1.00	
WLSv-subset	0.32	0.55	1.00	1.00	1.00	0.99	0.76	
WLSv-intuitive	0.58	0.56	1.00	1.00	0.98	1.00	0.75	
WLSv-lasso	0.77	0.84	0.99	1.00	1.00	1.00	1.00	
MinT-subset	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinT-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinT-lasso	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinTs-subset	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinTs-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinTs-lasso	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Elasso	0.73	0.65	0.98	0.98	0.86	1.00	0.99	

Note: the last column displays a stacked barplot for each method, based on the total number of selected series data from 500 simulation instances, with a darker sub-bar indicating a larger number.