Temporal hierarchies with autocorrelation for load forecasting

Nystrup, Peter; Lindström, Erik; Pinson, Pierre; Madsen, Henrik

Published in:
European Journal of Operational Research

Link to article, DOI:
10.1016/j.ejor.2019.07.061

Publication date:
2020

Document Version
Peer reviewed version

Temporal hierarchies with autocorrelation for load forecasting

Peter Nystrup[ab]*, Erik Lindström[b], Pierre Pinson[c], and Henrik Madsen[a]

July 17, 2019

[a] *Department of Applied Mathematics and Computer Science, Technical University of Denmark*
[b] *Centre for Mathematical Sciences, Lund University, Sweden*
[c] *Department of Electrical Engineering, Technical University of Denmark*

Abstract

We propose four different estimators that take into account the autocorrelation structure when reconciling forecasts in a temporal hierarchy. Combining forecasts from multiple temporal aggregation levels exploits information differences and mitigates model uncertainty, while reconciliation ensures a unified prediction that supports aligned decisions at different horizons. In previous studies, weights assigned to the forecasts were given by the structure of the hierarchy or the forecast error variances without considering potential autocorrelation in the forecast errors. Our first estimator considers the autocovariance matrix within each aggregation level. Since this can be difficult to estimate, we propose a second estimator that blends autocorrelation and variance information, but only requires estimation of the first-order autocorrelation coefficient at each aggregation level. Our third and fourth estimators facilitate information sharing between aggregation levels using robust estimates of the cross-correlation matrix and its inverse. We compare the proposed estimators in a simulation study and demonstrate their usefulness through an application to short-term electricity load forecasting in four price areas in Sweden. We find that by taking account of auto- and cross-covariances when reconciling forecasts, accuracy can be significantly improved uniformly across all frequencies and areas.

Keywords: Forecasting; Forecast combination; Temporal Aggregation; Autocorrelation; Reconciliation.

1 Introduction

Temporal aggregation has been studied since the seminal work by Amemiya and Wu (1972) and Tiao (1972) (see Silvestrini and Veredas, 2008, for a literature review). Different temporal aggregations can reveal important information about the underlying data-generating process. When temporal aggregation is applied to a time series, it can strengthen or attenuate different features.

*Correspondence to: Peter Nystrup, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Asmussens Allé, Building 303B, 2800 Kgs. Lyngby, Denmark.
Email: pnys@dtu.dk

Nonoverlapping temporal aggregation is a filter of high-frequency components. At an aggregate view, low frequency components, such as trend and cycle, will dominate. The opposite is true for disaggregate data, where short-term seasonality may be visible. Hence, temporal aggregation can be seen as a tool to better understand and model the data at hand.

Kourentzes et al. (2014) and Petropoulos and Kourentzes (2015) showed that combining forecasts from multiple aggregation levels can lead to improvements in forecast accuracy, while overcoming the need to select a single optimal level. The greatest improvements typically occur at the highest level of a hierarchy, where information from lower levels—i.e., higher-resolution data—is aggregated up (Athanasopoulos et al., 2011; Rostami-Tabar et al., 2013; Kourentzes et al., 2014; Athanasopoulos et al., 2017).

Temporal hierarchies for forecasting, as introduced by Athanasopoulos et al. (2017), can be constructed for any time series by means of nonoverlapping temporal aggregation. Rather than attempting to build one complex model that captures all temporal attributes (see, e.g., Ghysels et al., 2004; Livera et al., 2011; Nystrup et al., 2017; Sedoc et al., 2018), using temporal hierarchies forecasts for different horizons can be made with different (simple) methods. Since these forecasts are produced by different approaches and are based on different information, they will most likely not be coherent. This incoherency can lead to decisions that are not aligned, or even conflicting. Optimal decision making requires coherent forecasts; thus, reconciliation is necessary. In the framework proposed by Athanasopoulos et al. (2017), forecasts constructed at different aggregation levels can be combined to yield temporally reconciled, accurate, and robust forecasts, independently of forecasting models.

1.1 Related work

Traditionally, either top-down or bottom-up approaches are used to produce forecasts for a hierarchy. According to the former, forecasts are generated for the time series at the top level and then disaggregated down all the way to the bottom level, while for the latter forecasts are generated at the very bottom level and then aggregated up. Such approaches ensure that forecasts add up across a hierarchy. The advantages and disadvantages of these traditional approaches are not complementary. The top-down approach requires forecasts for only one time series at the very aggregate level; however, aggregation implies a large loss of information, and it is challenging to disaggregate the forecasts down the hierarchy (see Gross and Sohl, 1990; Athanasopoulos et al., 2009, for a summary of top-down approaches). By contrast, bottom up implies no loss of information; but it requires many and possibly very noisy time series to be forecast.

Hyndman et al. (2011) formulated the forecast reconciliation problem for a structural hierarchy as a linear regression model. In order to circumvent the problem of estimating the covariance of the base forecasts, they proposed the use of the ordinary least-squares (OLS) estimator for computing reconciled forecasts. Hyndman et al. (2016) suggested using weighted least squares (WLS), taking account of the variances on the diagonal of the covariance matrix but ignoring the offdiagonal covariances. Later, Wickramasuriya et al. (2019) considered the generalized least-squares (GLS) estimator and found the incorporation of correlation information into the reconciliation procedure to be beneficial for forecast accuracy, when combined with a simple shrinkage estimator.

Van Erven and Cugliari (2015) proposed a game-theoretically optimal reconciliation method that guarantees that the total weighted quadratic error of the reconciled forecasts is never greater than that of the base forecasts. Their approach is fundamentally different in that they formulate the

forecast reconciliation problem as an optimization problem rather than a regression model. Their formulation included only a diagonal weight matrix and, thus, no information sharing between the base forecasts. Van Erven and Cugliari (2015) proved that the reconciled forecasts are guaranteed to be at least as good as the base forecasts for any loss function that is based on a Bregman divergence. Wickramasuriya et al. (2019) reiterated the proof for their reconciliation method.

A crucial reason for this improvement is that the implied combination mitigates model uncertainty. Forecast combination is widely regarded as beneficial, leading to a reduction of forecast error variance (see, e.g., Clemen, 1989; Timmermann, 2006; Hall and Mitchell, 2007). Ways to best combine forecasts have been widely investigated, resulting in various sophisticated weighting methods; yet, simple approaches, such as the unweighted average, are often found to perform as well as more sophisticated methods (Timmermann, 2006).

Taieb et al. (2017a,b) considered reconciliation of density forecasts, as opposed to point forecasts. Their algorithm does not require distributional assumptions and imposes dependencies between forecast distributions using samples from the empirical copulas. Following van Erven and Cugliari (2015), they used the reconciliation approach to produce load forecasts for individual electricity consumers at the bottom to the total grid at the top of a structural hierarchy. Gamakumara et al. (2018) showed in a simulation study how a projection derived from the estimator of Wickramasuriya et al. (2019) can be applied to reconcile probabilistic forecasts. Jeon et al. (2018) proposed a cross-validation approach for selecting the reconciliation weights applied to probabilistic wind-power forecasts.

Temporally aggregated time series can be represented as a hierarchical time series. As a consequence, Athanasopoulos et al. (2017) showed that it is possible to use the reconciliation framework proposed by Hyndman et al. (2011) to produce coherent forecasts. They considered three diagonal estimators as approximations to the sample covariance matrix. By construction, these estimators share the property that they ignore any autocorrelation in the forecast errors. In most cases the simplest of the three, which is based purely on the structure of the hierarchy and requires no estimation of forecast errors, performed as well as the more complicated ones.

The work of Athanasopoulos et al. (2017) was extended by Taieb (2017), who considered load forecasting for individual households. He introduced regularization terms in order to obtain sparse and smooth adjustments that satisfy the aggregation constraints and minimize forecast errors. Imposing sparsity means that some base forecasts remain unaffected by the adjustments. Smoothness provides additional regularization by exploiting the fact that adjustments are applied to consecutive observations of a time series. The reconciled forecasts can be found by solving a sparse fused LASSO (least absolute shrinkage and selection operator) problem (Tibshirani et al., 2005).

In two successive articles, Yang et al. (2017a,b) applied a structural and temporal hierarchy, respectively, for reconciling solar-power forecasts. In the structural case, they followed Wickramasuriya et al. (2019) by considering both diagonal and nondiagonal, regularized and nonregularized covariance estimators. The largest forecast accuracy improvements occurred when including correlations in the reconciliation process. In the temporal case, they followed Athanasopoulos et al. (2017) by considering only diagonal estimators, thus disregarding potential information in the autocorrelation structure. Zhang and Dong (2018) documented the benefit from taking into account correlations when reconciling short-term wind-power forecasts across several wind farms in a structural hierarchy. Finally, Kourentzes and Athanasopoulos (2019) proposed a cross-temporal reconciliation approach for generating coherent forecasts across both geographical divisions and planning horizons for tourist flows in Australia.

1.2 Contribution

Given the well-documented benefits to incorporating correlation information when reconciling forecasts in a structural hierarchy (Yang et al., 2017a; Zhang and Dong, 2018; Wickramasuriya et al., 2019), we propose an estimator that considers the full autocovariance matrix within each aggregation level when reconciling forecasts in a temporal hierarchy. With the purpose of temporal aggregation being to exploit important information in a time series at different frequencies, it does not make sense to disregard the potentially most important information, namely its autocorrelation structure; at least not when there is enough data available that it can be estimated with reasonable precision. This is often the case in high-frequency settings. Hence, it should be possible to improve accuracy by considering autocorrelation information when reconciling forecasts.

Even with high-frequency data available, it can be difficult to estimate the autocovariance matrix without assuming that it has some special form (see, e.g., Bien et al., 2016). Therefore, we propose a second estimator that blends autocorrelation and variance information, but only requires estimation of the first-order autocorrelation coefficient at each aggregation level. This estimator is based on decomposing the autocovariance matrix into two diagonal variance matrices and a block-diagonal autocorrelation matrix that imposes a first-order Markov process on the reconciliation errors.

In order to facilitate information sharing between aggregation levels, we propose a third estimator that uses the graphical LASSO (GLASSO) for estimating a sparse representation of the inverse cross-correlation matrix across aggregation levels (Banerjee et al., 2008; Friedman et al., 2008). This estimator overcomes the problem of inverting a potentially singular cross-covariance matrix and is robust to noise and high dimensionality due to the imposed sparsity.

Our fourth proposal is a Stein-type shrinkage estimator of the cross-correlation matrix (Ledoit and Wolf, 2004; Schäfer and Strimmer, 2005), similar to the estimator Wickramasuriya et al. (2019) proposed for a structural hierarchy. This estimator is simpler to implement and shares most of the beneficial properties of the GLASSO estimator.

We document the usefulness of the proposed estimators through an application to short-term electricity load forecasting in four price areas in Sweden. We do not consider advanced forecasting techniques, nor include explanatory variables such as weather data (see, e.g., Fan and Hyndman, 2012; Clements et al., 2016), rather we only use exponential-smoothing methods to generate base forecasts. We show that incorporating information about the auto- and cross-correlation structure significantly improves forecast accuracy, both compared to traditional approaches and compared to the diagonal estimators proposed by Athanasopoulos et al. (2017). Improvements in forecast accuracy are in several cases greatest out of sample where the accuracy of the base forecasts, on average, is lower. In other words, reconciliation increases robustness by increasing accuracy the most when needed the most.

To better understand the differences and advantages of the proposed estimators, we compare their performance in a simulation study. We simulate data from a model similar to those used for load forecasting. The simulation study shows that the best performing estimator depends on the length of the sample available for estimation. Even when very limited data is available compared to the dimension of the temporal hierarchy, forecast accuracy can be improved by considering autocorrelation and dependencies between aggregation levels.

The article is organized as follows. In Section 2, we outline the forecast reconciliation problem and its relation to ordinary, weighted, and generalized least-squares estimation. In Section 3, we

propose four different estimators for taking account of the autocorrelation structure. Results from the application to short-term load forecasting are presented in Section 4. The simulation study is presented in Section 5. We discuss the results and their importance to operational research in Section 6 before concluding in Section 7.

2 Forecast reconciliation

2.1 Temporal hierarchies

Given n individual *base* forecasts stacked in a column vector $\hat{y} \in \mathbb{R}^n$, where $\hat{y}_1$ is a forecast of the aggregate, we want to find *reconciled* forecasts $\tilde{y} \in \mathbb{R}^n$, which are coherent, so that $\sum_{i=2}^{n} \tilde{y}_i = \tilde{y}_1$. For example, $\hat{y}_2, \ldots, \hat{y}_n$ could be sales forecasts for $n-1$ individual stores and $\hat{y}_1$ the aggregate sales forecast for the entire chain of stores. This is an example of a *structural* hierarchy.

The framework can easily be extended to multiple aggregation levels. In a temporal hierarchy, for example, quarterly forecasts should reconcile to half-year forecasts, which should reconcile to annual forecasts, as illustrated in Figure 1. This is most easily done by introducing a summation matrix S, which for the hierarchy illustrated in Figure 1 would be

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

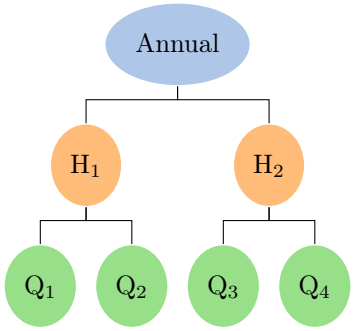


Figure 1: Temporal hierarchy for quarterly series.

It is even possible to have several different forecasts for each quarter, year, etc.

In general, there are $k \in \{k_1, \ldots, k_K\}$ aggregation levels, where k is a factor of m, with $k_1 = m$, $k_K = 1$, and m/k is the number of observations at aggregation level k. The summation matrix is given by

$$S = \begin{bmatrix} I_{m/k_1} \otimes 1_{k_1} \\ \vdots \\ I_{m/k_K} \otimes 1_{k_K} \end{bmatrix}, \tag{1}$$

where $\otimes$ denotes the Kronecker product, $I_{m/k}$ is an identity matrix of order m/k, and 1_k is an k-vector of ones. For example, the hierarchy illustrated in Figure 1 has aggregation levels $k_1 = 4$, $k_2 = 2$, and $k_3 = 1$ with $m = 4$ and $n = 7$.

By introducing a matrix

$$G = \left[0_{m \times (n-m)} \middle| I_m\right] \tag{2}$$

of order $m \times n$ that extracts the m bottom-level forecasts, the reconciliation constraint(s) can be written as

$$\tilde{y} = SG\tilde{y}. \tag{3}$$

Reconciliation is needed when base forecasts $\hat{y}$ do not satisfy this constraint.

2.2 Optimal reconciliation

Van Erven and Cugliari (2015) proposed to formulate the forecast reconciliation problem as

$$\begin{array}{ll} \text{minimize} & \left\| W^{1/2} \left(\tilde{y} - \hat{y} \right) \right\|_2^2 \\ \text{subject to} & \tilde{y} = SG\tilde{y}, \end{array} \tag{4}$$

where $\tilde{y} \in \mathbb{R}^n$ is the variable; the parameter $W \in \mathbb{R}^{n \times n}_{++}$ is a diagonal matrix with weights w_i along its diagonal; $S \in \mathbb{R}^{n \times m}$ and $G \in \mathbb{R}^{m \times n}$ are constant matrices given by the structure of the hierarchy; and $\|y\|_2 = \left(\sum_{i=1}^n y_i^2 \right)^{1/2}$ denotes the ℓ_2 norm of a vector $y \in \mathbb{R}^n$ of dimension n.

The squared error is the most common choice of loss function. The reconciled forecasts are optimal in that the base forecasts are adjusted by the least amount (in the sense of least squares) so that these become *coherent*. Formulation (4) is a convex optimization problem that can readily be solved (Boyd and Vandenberghe, 2004).

2.3 Relation to generalized least squares

Hyndman et al. (2011) and Athanasopoulos et al. (2017) formulated the structural and temporal reconciliation problems, respectively, as linear regression models. The reconciled forecasts can be found using the *generalized* least-squares estimate:

$$\begin{array}{ll} \text{minimize} & \left(\tilde{y} - \hat{y} \right)^T \Sigma^{-1} \left(\tilde{y} - \hat{y} \right) \\ \text{subject to} & \tilde{y} = SG\tilde{y}, \end{array} \tag{5}$$

where $\tilde{y} \in \mathbb{R}^n$ is the variable and the parameter $\Sigma \in \mathbb{R}^{n \times n}_{++}$ is the covariance matrix for the *coherency errors* $\varepsilon = \tilde{y} - \hat{y}$, which are assumed to be multivariate Gaussian and unbiased, i.e., have zero mean.

If Σ were known, the solution to (5) would be given by the GLS estimator

$$\tilde{y} = S \left(S^T \Sigma^{-1} S \right)^{-1} S^T \Sigma^{-1} \hat{y}. \tag{6}$$

When the summation matrix S is very large, it can be faster to solve the optimization problem (5) than to evaluate the closed-form solution (6) (Boyd and Vandenberghe, 2004). Hyndman et al. (2016) showed how the computations required to evaluate (6) can be handled efficiently by exploiting the sparse structure of the summation matrix. Wickramasuriya et al. (2019) derived an alternative representation that is significantly less demanding in terms of computation.

The close correspondence between (4) and (5) is evident, as the two coincide when $\Sigma^{-1} = W$. This provides a benchmark for selecting the weights. In formulation (5), the precision matrix Σ^{-1} is used to scale deviations from the base forecasts; hence, it is often referred to as a weight matrix. It is more expensive to adjust base forecasts with a higher precision. Another option is to select weights based on prior knowledge about the precision or importance of the base forecasts. The machine-learning approach would be to run a number of tests to find the combination of weights that gives the best (in-sample) result.

Hyndman et al. (2011) and van Erven and Cugliari (2015) argued for selecting uniform weights to increase the importance of forecasting the aggregate. The unweighted case $\Sigma = I$ corresponds

to *ordinary* least-squares estimation. Unweighted implies that each aggregation level is assigned the same total weight; however, due to the different scales of the various levels, this means that OLS emphasizes the highest aggregation levels.

Wickramasuriya et al. (2019) showed that, in general, Σ is not known and is not identifiable. By minimizing the variances of the reconciled forecast errors, they proposed an estimator which results in unbiased reconciled forecasts given by the GLS estimator (6), but with a different covariance matrix. To distinguish it from the GLS estimator, they referred to this as minimum trace reconciliation. As a proxy for the unidentifiable covariance matrix for the coherency errors, they provided theoretical justification for using the covariance matrix for the reconciled *forecast errors* $e = y - \tilde{y}$. Although it does not suffer from a lack of identifiability, it can still be challenging to estimate. Wickramasuriya et al. (2019) proposed different estimators based on the in-sample base forecast errors.

2.4 Weighted least-squares estimators

Athanasopoulos et al. (2017) proposed three diagonal estimators Λ of increasing simplicity that approximate Σ. For the temporal hierarchy illustrated in Figure 1, the three estimators are

$$\begin{aligned}
\Lambda_{\text{struc}} &= \text{diag}\left(4,2,2,1,1,1,1\right), \\
\Lambda_{\text{svar}} &= \text{diag}\left(\sigma^2_{\text{A}},\sigma^2_{\text{H}},\sigma^2_{\text{H}},\sigma^2_{\text{Q}},\sigma^2_{\text{Q}},\sigma^2_{\text{Q}},\sigma^2_{\text{Q}}\right), \\
\Lambda_{\text{hvar}} &= \text{diag}\left(\sigma^2_{\text{A}},\sigma^2_{\text{H}_1},\sigma^2_{\text{H}_2},\sigma^2_{\text{Q}_1},\sigma^2_{\text{Q}_2},\sigma^2_{\text{Q}_3},\sigma^2_{\text{Q}_4}\right).
\end{aligned}$$

By definition, these ignore correlations across and within aggregation levels and lead to alternative *weighted* least-squares estimators. They referred to the simplest of the three estimators as *structural* scaling. As base forecast errors at each level of a temporal hierarchy are associated with a single time series, they argued that it is reasonable to assume that the variances at each level are approximately equal. Assuming that the variance of each bottom-level base forecast error is σ^2 and that the errors are uncorrelated between nodes, they set $\Sigma = \sigma^2\Lambda_{\text{struc}}$, where Λ_{struc} is a diagonal matrix with each element containing the number of forecasts errors contributing to that aggregation level:

$$\Lambda_{\text{struc}} = \text{diag}\left(S1_m\right).$$

This estimator has several desirable properties. First, it depends only on the seasonal period m of the most disaggregated observations and is independent of both data and forecasting model. Second, it permits forecasts which originate from any forecasting method or even predictions from human experts that are not described by a formal model, since no estimation of the variance of the forecast errors is needed. In their empirical evaluation, Athanasopoulos et al. (2017) found that structural scaling often performed at least as well as the more complicated estimators.

The second estimator proposed by Athanasopoulos et al. (2017) is referred to as *series variance* scaling. This estimator includes separate variance estimates for each aggregation level. That is, it assumes homogeneous error variance within a level, but not across levels. They argued that this is a reasonable assumption given that base forecast errors within the same aggregation level are for the same time series. Their proposal was to use the variances of the one-step-ahead forecast errors for each aggregation level, which are easily calculated. This tends to underestimate the variances of the lower levels, where forecasts are made multiple steps ahead. Generally, the further out in time a forecast is made, the more uncertain it is (see, e.g., Hyndman et al., 2008, Chapter 6.2). Our

proposal is to use pooled estimates of the multi-step-ahead forecast error variances. For example, σ^2_{Q} is a pooled estimate of $\sigma^2_{\mathrm{Q}_1}$, $\sigma^2_{\mathrm{Q}_2}$, $\sigma^2_{\mathrm{Q}_3}$, and $\sigma^2_{\mathrm{Q}_4}$.

The third estimator proposed by Athanasopoulos et al. (2017) includes separate variance estimates for each base forecast. This estimator is referred to as *hierarchy variance* scaling. For the same reason as before, we differ from Athanasopoulos et al. (2017) by arguing that $\sigma^2_{\mathrm{Q}_4}$ should be the variance of the errors when, at the end of each year, forecasting the fourth quarter next year, rather than the variance of one-step-ahead forecast errors for the fourth quarter.

3 Accounting for autocorrelation

3.1 Autocovariance scaling

In a temporal hierarchy, WLS corresponds to ignoring autocorrelation. As the purpose of temporal aggregation is to exploit important information about a time series at different frequencies, we argue that potential information in the autocorrelation structure should be included. Therefore, we extend the ideas of Wickramasuriya et al. (2019) to a temporal hierarchy by proposing four different estimators of the cross-covariance matrix based on the in-sample base forecast errors.

Our first proposal is to estimate the full autocovariance matrix within each aggregation level, while ignoring correlations between aggregation levels. We refer to this estimator as *autocovariance* scaling. For example, for the temporal hierarchy illustrated in Figure 1, the estimator is

$$\Sigma_{\text{acov}} = \begin{bmatrix} \sigma^2_{\mathrm{A}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{\mathrm{H}_1} & \sigma^2_{\mathrm{H}_1,\mathrm{H}_2} & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{\mathrm{H}_1,\mathrm{H}_2} & \sigma^2_{\mathrm{H}_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{\mathrm{Q}_1} & \sigma^2_{\mathrm{Q}_1,\mathrm{Q}_2} & \sigma^2_{\mathrm{Q}_1,\mathrm{Q}_3} & \sigma^2_{\mathrm{Q}_1,\mathrm{Q}_4} \\ 0 & 0 & 0 & \sigma^2_{\mathrm{Q}_1,\mathrm{Q}_2} & \sigma^2_{\mathrm{Q}_2} & \sigma^2_{\mathrm{Q}_2,\mathrm{Q}_3} & \sigma^2_{\mathrm{Q}_2,\mathrm{Q}_4} \\ 0 & 0 & 0 & \sigma^2_{\mathrm{Q}_1,\mathrm{Q}_3} & \sigma^2_{\mathrm{Q}_2,\mathrm{Q}_3} & \sigma^2_{\mathrm{Q}_3} & \sigma^2_{\mathrm{Q}_3,\mathrm{Q}_4} \\ 0 & 0 & 0 & \sigma^2_{\mathrm{Q}_1,\mathrm{Q}_4} & \sigma^2_{\mathrm{Q}_2,\mathrm{Q}_4} & \sigma^2_{\mathrm{Q}_3,\mathrm{Q}_4} & \sigma^2_{\mathrm{Q}_4} \end{bmatrix}.$$

3.2 Markov scaling

In high-dimensional hierarchies, even with high-frequency data available, it can be difficult to estimate the autocovariance matrix within each aggregation level without assuming that it has some special form. Therefore, we also propose an estimator that blends autocorrelation and variance information, but only requires estimation of the first-order autocorrelation coefficient at each aggregation level. The estimator is based on decomposing the autocovariance matrix into two diagonal variance matrices Λ and a block-diagonal autocorrelation matrix that imposes a first-order autoregressive structure on the reconciliation errors.

We refer to this estimator as either *structural*, *series*, or *hierarchy Markov* scaling, depending on which diagonal variance matrix is used. For example, for the temporal hierarchy illustrated in Figure 1, the series Markov estimator is

$$\Sigma_{\text{seMarkov}} = \Lambda_{\text{svar}}^{1/2} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & \rho_{\text{H}} & 0 & 0 & 0 & 0 \\ 0 & \rho_{\text{H}} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho_{\text{Q}} & \rho_{\text{Q}}^2 & \rho_{\text{Q}}^3 \\ 0 & 0 & 0 & \rho_{\text{Q}} & 1 & \rho_{\text{Q}} & \rho_{\text{Q}}^2 \\ 0 & 0 & 0 & \rho_{\text{Q}}^2 & \rho_{\text{Q}} & 1 & \rho_{\text{Q}} \\ 0 & 0 & 0 & \rho_{\text{Q}}^3 & \rho_{\text{Q}}^2 & \rho_{\text{Q}} & 1 \end{bmatrix} \Lambda_{\text{svar}}^{1/2}.$$

If the models used to generate the base forecasts were perfect, then there would not be any residual autocorrelation. This is seldom the case in practical applications when forecasting multiple steps ahead. Yet, if the residual autocorrelation structure is complex and, for example, features seasonality, then it is probably worthwhile revisiting the model before considering forecast reconciliation. Thus, it seems reasonable to assume that the reconciliation errors can be approximated by a first-order Markov process.

3.3 Inverse cross-correlation estimation using GLASSO

A higher-order autoregressive structure can be imposed by considering the inverse of the autocorrelation matrix—i.e., the partial autocorrelation—which has a very simple structure for autoregressive processes (Madsen, 2008, Chapter 5). After all, we only need to estimate the inverse in order to solve (5). Our third proposal is to estimate a sparse representation of the inverse cross-correlation matrix using the graphical LASSO (Banerjee et al., 2008; Friedman et al., 2008). Because the LASSO penalty is sensitive to the scale of variables, we estimate the inverse cross-correlation rather than the inverse cross-covariance matrix to avoid problems caused by the inherent heteroscedasticity.

We refer to this estimator as either *series* or *hierarchy GLASSO* scaling, depending on which diagonal variance matrix is used. It will become clear from the discussion in Section 4.4 why we do not introduce a structural version of this estimator. With the series variances the estimator is

$$\Sigma_{\text{sGLASSO}}^{-1} = \Lambda_{\text{svar}}^{-1/2} \Theta \Lambda_{\text{svar}}^{-1/2},$$

where the inverse cross-correlation matrix Θ is found by maximizing the penalized Gaussian log-likelihood

$$\log\det\Theta - \operatorname{tr}(R\Theta) - \lambda \sum_{i \neq j} |\Theta_{ij}|. \tag{7}$$

We follow Yuan and Lin (2007) by omitting the diagonal elements from the penalty. The parameter R is the empirical cross-correlation matrix and $\lambda \geq 0$ is a regularization parameter that controls the degree of sparsity in the solution.

For large values of λ, GLASSO scaling is equivalent to scaling by the diagonal variance matrix. When $\lambda = 0$ and $\Lambda = \Lambda_{\text{hvar}}$, it is equivalent to scaling by the sample cross-covariance matrix, which can be difficult to estimate, depending on the number of observations available. There exist several specialized algorithms for finding the solution to the maximization problem (7); however, the problem is convex and can be solved using standard software for convex optimization, such as CVXPY (Diamond and Boyd, 2016; Akshay Agrawal and Boyd, 2018).

The GLASSO estimator has several beneficial properties. First, it overcomes the problem of inverting a potentially singular cross-covariance matrix. Second, it is robust to noise and high

dimensionality due to the imposed sparsity. Third, contrary to autocovariance and Markov scaling, this estimator allows for *information sharing* between aggregation levels by accounting for correlations between forecast errors from different aggregation levels. For example, the forecast errors for Q_1 and Q_2 should reasonably be correlated with the forecast error for H_1.

3.4 Cross-correlation shrinkage

Wickramasuriya et al. (2019) obtained good results by using a Stein-type shrinkage estimator of the sample covariance matrix (Ledoit and Wolf, 2004; Schäfer and Strimmer, 2005) when reconciling forecasts in a structural hierarchy. In a similar vein, our final proposal is to consider a shrinkage estimator. We consider a shrinkage estimator of the cross-correlation rather than the cross-covariance matrix to avoid problems with heteroscedasticity.

The estimator is based on decomposing the cross-covariance matrix into two diagonal variance matrices Λ and a cross-correlation matrix R_{shrink}. We refer to this estimator as either *series* or *hierarchy shrinkage* scaling, depending on which diagonal variance matrix is used. With the hierarchy variances the estimator is

$$\Sigma_{\text{hshrink}} = \Lambda_{\text{hvar}}^{1/2} R_{\text{shrink}} \Lambda_{\text{hvar}}^{1/2},$$

where the shrinkage estimator of the cross-correlation matrix is

$$R_{\text{shrink}} = (1 - \lambda)\, R + \lambda I_n. \tag{8}$$

R is the empirical cross-correlation matrix and $0 \leq \lambda \leq 1$ is a regularization parameter that controls the degree of shrinkage towards the identity matrix.

When $\lambda = 1$, shrinkage scaling is equivalent to scaling by the diagonal variance matrix. When $\lambda = 0$ and $\Lambda = \Lambda_{\text{hvar}}$, it is equivalent to scaling by the sample cross-covariance matrix.

The shrinkage estimator has the same beneficial properties as the GLASSO estimator. The difference is that the GLASSO estimator tries to achieve sparsity in the partial cross-correlation structure, whereas the shrinkage estimator shrinks all cross-correlations uniformly toward zero. The shrinkage estimator is simpler to implement; but if the partial cross-correlation structure is in fact sparse, then the GLASSO estimator should produce better results.

4 Empirical results

4.1 Load forecasting

Automated short-term load forecasting is needed for efficient operation of power systems and to support transactions by participants in deregulated electricity markets (Hahn et al., 2009). For day-ahead prediction, a weather-based model is typically used. Weather-based models are less important for intraday horizons, as weather variables tend to change relatively smoothly over short intervals of time. Moreover, weather forecasts are sometimes not available or only available with a delay. This prompts consideration of modeling approaches that use only historical load data (Taylor, 2010, 2012).

In short-term load forecasting, the seasonal Holt–Winters exponential smoothing is a common choice for modeling seasonality (Taylor, 2003; Gould et al., 2008; Taylor, 2010; Livera et al., 2011;

Taylor, 2012). Holt–Winters exponential smoothing was extended by Taylor (2003) to accommodate intraday and intraweek cycles in intraday data. It is well-suited for electricity-demand forecasting, as demand has both daily and weekly seasonalities.

Letting p_1 and p_2 denote the periods of the two seasons, the additive double-seasonal exponential smoothing method from Taylor (2012) can be presented as the following state-space model:

$$y_t = l_{t-1} + s^{(1)}_{t-p_1} + s^{(2)}_{t-p_2} + \phi e_{t-1} + \varepsilon_t, \tag{9a}$$

$$e_t = y_t - \left(l_{t-1} + s^{(1)}_{t-p_1} + s^{(2)}_{t-p_2}\right), \tag{9b}$$

$$l_t = l_{t-1} + \alpha e_t, \tag{9c}$$

$$s^{(1)}_t = s^{(1)}_{t-p_1} + \gamma_1 e_t, \tag{9d}$$

$$s^{(2)}_t = s^{(2)}_{t-p_2} + \gamma_2 e_t, \tag{9e}$$

where $\varepsilon_t \sim N\left(0, \sigma^2\right)$ and σ^2 is a constant variance; y_t is the load; l_t and $s^{(1)}_t$ are the state variables for the level and intraday cycle, respectively; $s^{(2)}_t$ is the state variable for the intraweek cycle remaining after $s^{(1)}_t$ is removed; α, γ_1, and γ_2 are smoothing parameters; and the term involving ϕ is an autoregressive adjustment for first-order residual autocorrelation. Taylor (2010) found that the inclusion of this term greatly improves forecast accuracy. The method assumes the same intraday cycle for all days of the week and does not include a trend component, as short-term electricity demand mostly does not have a trend.

4.2 Data

We consider hourly load data for 2016 and 2017 from Nord Pool.[1] Nord Pool Spot is the power market for Sweden, Norway, Denmark, Finland, Estonia, Latvia, and Lithuania. The day-ahead market is an auction, where power is traded for delivery each hour of the next day. The Nord Pool markets are divided into several bidding areas, as shown in Figure 2. We consider the load in GWh in the four Swedish areas SE1–SE4. We do not apply a log transformation to the load data, as it is custom to do, because it would ruin the summation property of the hierarchy. We use data from 2016 for in-sample training, with the first two weeks being used for initialization, and data from 2017 for out-of-sample testing.

Figure 2: Map showing the situation in the Nordic region. Sweden is divided into four electricity price areas: Malmö (SE4), Stockholm (SE3), Sundsvall (SE2) and Luleå (SE1). Norway currently has five price areas. Denmark is split into Eastern and Western Denmark. Finland, Estonia, Lithuania, and Latvia are undivided.

Figure 3 shows the autocorrelation functions (ACFs) for three different frequencies of the load data for the Swedish areas in 2016. Across all frequencies and areas, the load data is strongly autocorrelated. In Figure 3a, the

[1] https://www.nordpoolgroup.com/historical-market-data/

ACFs for the daily load reveal a clear weekly pattern, which is much stronger for SE4 than SE1. In Figure 3b, the ACFs for the six-hourly load display both a daily and weekly variation. Once again, the magnitude of the seasonal variation follows the order of the areas with both seasons being significantly more pronounced in SE4 compared to SE1. The daily and weekly cycles are also evident from Figure 3c, although the weekly cycle is less pronounced at the hourly frequency.

4.3 Base forecasts

We model the entire temporal hierarchy from the daily frequency at the top to the hourly frequency at the bottom. With aggregation levels $k = 24, 12, 8, 6, 4, 3, 2, 1$, the total dimension of the hierarchy is 60.

Base forecasts at the daily frequency are generated based on exponential smoothing with a weekly seasonality using the automatic forecasting procedure implemented by Hyndman and Khandakar (2008). If multiple years of data were available for estimation, the models could be extended to accommodate yearly seasonality. Forecasts at all frequencies are generated using the additive double-seasonal exponential smoothing method from Taylor (2012). All model parameters are fitted by minimizing the mean squared error of the in-sample one-step-ahead forecasts using the data from 2016.

We focus on squared errors to be consistent with the objective function in (5). Table 1 shows the root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}, \tag{10}$$

the root mean squared scaled error

$$\text{RMSSE} = \frac{\text{RMSE}}{\sqrt{\frac{1}{T-m/k}\sum_{t=m/k+1}^{T}\left(y_t - y_{t-m/k}\right)^2}}, \tag{11}$$

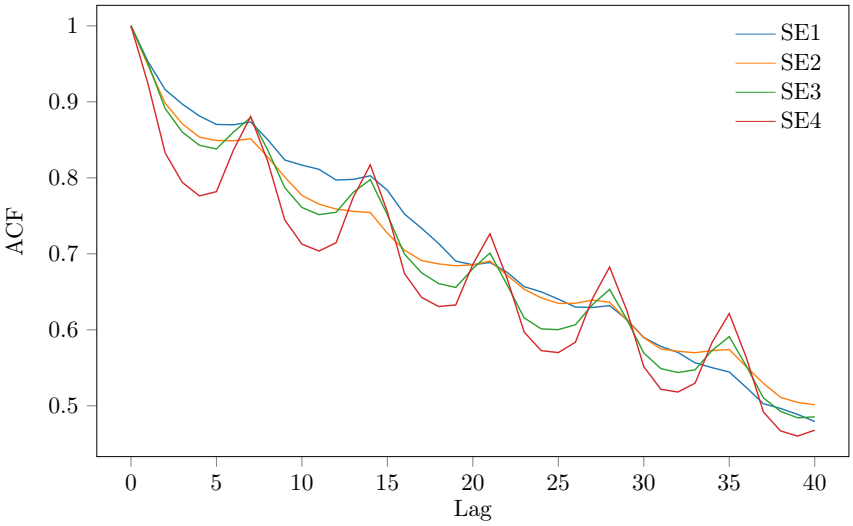
where m/k is the sampling frequency per day, and the root mean squared percentage error

$$\text{RMSPE} = 100\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t - \hat{y}_t}{y_t}\right)^2} \tag{12}$$

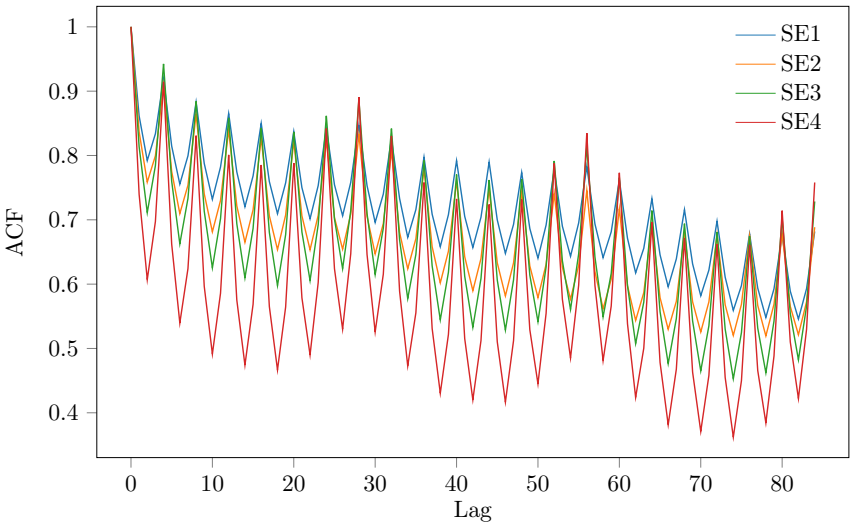
for base forecasts of one-day-ahead power consumption in the Swedish areas at the different data frequencies. At the daily frequency, this corresponds to one-step-ahead forecasts, whereas at the hourly frequency, it corresponds to 24-steps-ahead forecasts, respectively, once per day at the end of each day. For example, at the hourly frequency the RMSE is an average across the 24 steps each day across all days in the sample.

The RMSE is much larger for SE3 compared to the other areas, because consumption in this area is larger. The RMSSE and RMSPE are better suited for comparing the forecast accuracy for the different areas, since they have the advantage of being scale-independent. The RMSPE is easier to interpret than the RMSSE, but can only be applied to time series that do not overlap with zero. The RMSSE is significantly lower in SE3 and SE4 compared to SE1 and SE2.

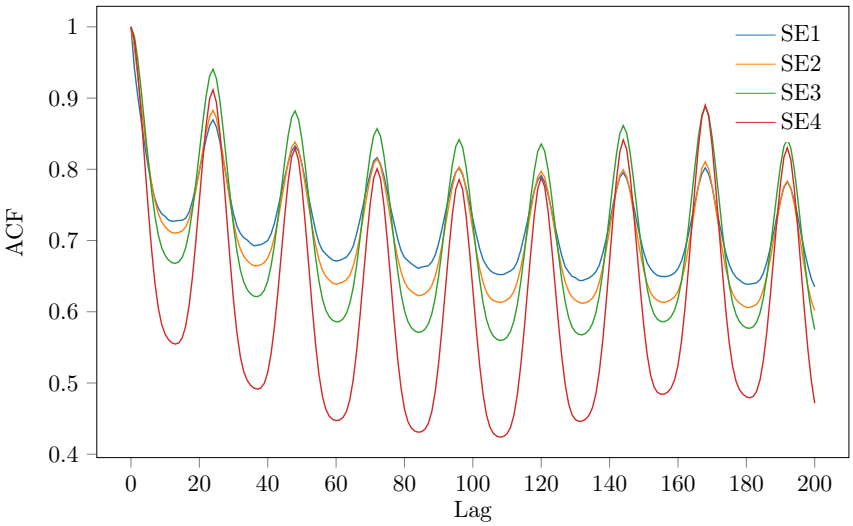
Looking at the RMSPE, it is evident that load forecasts for SE3 are more accurate, relatively speaking, compared to the other areas. This is true both in and out of sample. The RMSPE for SE3 is relatively similar across the different data frequencies, whereas it doubles, for example



(a) Daily



(b) Six-hourly



(c) Hourly

Figure 3: Autocorrelation functions for three frequencies of load data for Swedish areas in 2016.

	In sample (2016)				Out of sample (2017)			
	SE1	SE2	SE3	SE4	SE1	SE2	SE3	SE4
Daily								
RMSE	1.12	2.34	8.77	2.89	1.21	2.58	9.66	3.33
RMSSE	0.86	0.86	0.56	0.53	0.90	0.89	0.61	0.59
RMSPE	4.17	4.95	3.61	4.27	4.32	5.39	3.97	4.92
Twelve-hourly								
RMSE	0.56	1.17	4.02	1.40	0.56	1.21	3.90	1.46
RMSSE	0.72	0.70	0.48	0.48	0.70	0.73	0.46	0.48
RMSPE	4.32	4.99	3.28	4.14	4.16	5.07	3.22	4.44
Eight-hourly								
RMSE	0.44	0.88	3.02	1.05	0.44	0.98	3.15	1.13
RMSSE	0.78	0.73	0.52	0.51	0.76	0.83	0.54	0.53
RMSPE	5.05	5.75	3.72	4.68	4.85	6.06	3.83	5.03
Six-hourly								
RMSE	0.36	0.72	2.51	0.85	0.33	0.78	2.62	0.93
RMSSE	0.78	0.74	0.54	0.52	0.72	0.82	0.56	0.56
RMSPE	5.64	6.18	4.07	5.08	4.94	6.39	4.20	5.47
Four-hourly								
RMSE	0.25	0.51	1.88	0.62	0.24	0.54	1.92	0.68
RMSSE	0.76	0.76	0.61	0.57	0.73	0.82	0.62	0.61
RMSPE	5.85	6.56	4.49	5.48	5.34	6.62	4.56	5.92
Three-hourly								
RMSE	0.19	0.40	1.44	0.50	0.19	0.41	1.48	0.55
RMSSE	0.75	0.75	0.61	0.59	0.72	0.80	0.62	0.64
RMSPE	6.05	6.69	4.58	5.80	5.61	6.73	4.70	6.27
Two-hourly								
RMSE	0.14	0.27	0.93	0.33	0.14	0.27	0.96	0.36
RMSSE	0.77	0.75	0.59	0.57	0.72	0.77	0.60	0.62
RMSPE	6.66	6.98	4.48	5.74	6.11	6.94	4.62	6.24
Hourly								
RMSE	0.09	0.14	0.46	0.16	0.08	0.14	0.44	0.16
RMSSE	0.84	0.75	0.58	0.54	0.75	0.76	0.55	0.56
RMSPE	8.31	7.46	4.52	5.62	7.39	7.43	4.37	5.94

Table 1: In- and out-of-sample RMSE, RMSSE, and RMSPE for base forecasts of one-day-ahead power consumption in Swedish areas at different data frequencies.

	In sample (2016)					Out of sample (2017)				
	SE1	SE2	SE3	SE4	Average	SE1	SE2	SE3	SE4	Average
Bottom up	18	2	4	0	6	4	−4	−4	−4	−2
Identity	−3	−4	−8	−8	−6	−1	−4	−6	−6	−4
Structural	−5	−5	−7	−7	−6	−5	−5	−6	−6	−6
Series variance	−3	−3	−4	−5	−4	−5	−5	−5	−5	−5
Hierarchy variance	−3	−5	−9	−9	−6	−6	−7	−9	−9	−8
Structural Markov	−6	−7	−12	−11	−9	−5	−7	−10	−9	−8
Series Markov	−6	−6	−10	−10	−8	−6	−7	−10	−9	−8
Hierarchy Markov	−7	−9	−23	−20	−15	−8	−11	−21	−19	−15
Autocovariance	−11	−12	−26	−23	−18	−10	−14	−25	−22	−18
Series GLASSO	−31	−25	−28	−32	−29	−30	−26	−27	−30	−28
Series shrinkage	−30	−27	−31	−34	−30	−29	−28	−29	−33	−30
Cross-covariance	−39	−38	−57	−53	−47	−35	−41	−53	−48	−44

Table 2: In- and out-of-sample average percentage difference in RMSE between reconciled and base forecasts of one-day-ahead power consumption in Swedish areas across the aggregation levels. The last columns show the average across the price areas.

in SE1, when going from the daily to the hourly frequency. It is important to emphasize that this does not mean that the one-day-ahead forecasts based on hourly data are worse than those based on daily data, as it will become evident from the next tables. The RMSPE increases when moving down the hierarchy from the daily level toward the hourly level, with the exception of the twelve-hourly frequency, where base forecasts, in most cases, are more accurate than at the daily level, relatively speaking.

4.4 Reconciled forecasts

We follow Athanasopoulos et al. (2017) and the recommendation of Hyndman and Koehler (2006) by considering the relative root mean squared error

$$\text{RRMSE} = \frac{\text{RMSE}}{\text{RMSE}^{\text{base}}} - 1 \tag{13}$$

when comparing the accuracy of reconciled and base forecasts. A negative entry shows a percentage decrease in RMSE relative to the base forecast, i.e., improved accuracy.

The results for all price areas and aggregation levels are shown in Appendix A, where the two-sided Diebold–Mariano test (Diebold and Mariano, 1995) with the modification suggested by Harvey et al. (1997) is used to compare the accuracy of the reconciled and base forecasts. Most of the differences in accuracy are statistically significant at the 0.01 level. In all areas, the best performing approach is to scale by the full cross-covariance matrix. In Appendix B, we compare forecasts using the relative root mean squared percentage error (RRMSPE) and obtain similar results. The difference between the RRMSE and the RRMSPE is never more than a few percentage points.

Table 2 summarizes the average percentage difference in RMSE between reconciled and base forecast for each price area and across price areas in and out of sample. The in- and out-of-sample results are very similar. The largest improvements in accuracy are achieved for SE3 and SE4.

Table 3 summarizes the out-of-sample average percentage difference in RMSE between reconciled and base forecast for each aggregation level and across aggregation levels. The daily forecasts improve the most, while the smallest improvements are at the hourly and twelve-hourly levels. At

	Out of sample (2017)								
	24	12	8	6	4	3	2	1	Average
Bottom up	−10	14	1	−2	−6	−7	−4	0	−2
Identity	−16	9	−2	−4	−7	−8	−5	0	−4
Structural	−16	8	−3	−6	−9	−10	−7	−2	−6
Series variance	−14	9	−3	−6	−9	−9	−7	−2	−5
Hierarchy variance	−17	5	−6	−8	−11	−12	−9	−4	−8
Structural Markov	−19	5	−6	−8	−10	−11	−8	−4	−8
Series Markov	−18	5	−6	−8	−11	−11	−9	−4	−8
Hierarchy Markov	−28	−5	−14	−15	−17	−17	−15	−10	−15
Autocovariance	−30	−8	−16	−18	−20	−20	−17	−13	−18
Series GLASSO	−46	−20	−28	−28	−29	−29	−26	−20	−28
Series shrinkage	−47	−22	−29	−30	−31	−30	−27	−21	−30
Cross-covariance	−61	−38	−45	−44	−45	−44	−41	−35	−44

Table 3: Average out-of-sample percentage difference in RMSE between reconciled and base forecasts of one-day-ahead power consumption in Swedish areas for each aggregation level. The last column shows the average across the aggregation levels.

the twelve-hourly level, reconciled forecasts are only more accurate than the base forecasts when accounting for both variance differences and autocorrelation.

Bottom up The bottom-up approach, where forecasts made at the hourly frequency are aggregated to all other levels, in general, is not very successful, though it improves out-of-sample forecasts in some areas at certain frequencies. We did also consider the middle-out and top-down approaches, but decided to omit the results in order to save space since they are worse than those using bottom up.

Identity scaling The OLS approach leads to significant improvements of the daily base forecasts in all areas, but mixed results at the other aggregation levels. This makes sense given that Hyndman et al. (2011) and van Erven and Cugliari (2015) argued for selecting uniform weights to increase the importance of forecasting the aggregate.

Structural scaling Structural scaling, where the variance is assumed to be proportional to the number of forecast errors contributing to each aggregation level, leads to slightly better results than identity scaling.

Variance scaling Scaling by the estimated pooled multi-step-ahead variance at each aggregation level of the series in most cases yields slightly worse results than structural scaling. The results for hierarchy variance scaling—i.e., scaling by the nonpooled variances at each aggregation level of the hierarchy—are slightly better than series variance and structural scaling.

Markov scaling Hierarchy Markov scaling improves accuracy compared to hierarchy variance scaling in nearly all cases, which is also reflected in the average improvement in RMSE of 15%. In most cases it is better than structural and series Markov scaling, especially for SE3 and SE4, and it is never worse. On average, structural and series Markov scaling is only slightly better than OLS and WLS. Besides its simplicity, the advantage of structural scaling compared to series and hierarchy variance scaling is that it does not misestimate the aggregated forecast error variance by failing to account for autocorrelation. When accounting explicitly for autocorrelation, it is

important to adequately capture the ratio between the forecast error variances within and across aggregation levels.

Autocovariance scaling Autocovariance scaling improves forecast accuracy uniformly across all frequencies and areas, and compared to all of the previous approaches. The simpler approaches have all struggled to improve accuracy at the twelve-hourly level. The average improvement in RMSE of 18% both in and out of sample is only slightly better than that of hierarchy Markov scaling. The large improvements that can be obtained using hierarchy Markov scaling or autocovariance scaling compared to OLS and WLS show the value of accounting for autocorrelation when reconciling forecasts.

GLASSO scaling The regularization parameter in (7) is found by doing a grid search over the values $\lambda = 0$, 10^{-5}, 10^{-4}, $\ldots$, 10^{-1} and comparing the average in-sample improvement in RMSE. With the series variances the optimal value is $\lambda = 0.01$, while for the hierarchy variances the optimal value is $\lambda = 0$. This indicates that a year of hourly observations is sufficient to estimate the cross-covariance matrix without any need for regularization, as will be further discussed in the following sections. In Table 5, results are shown for series GLASSO scaling, while hierarchy GLASSO scaling is equivalent to cross-covariance scaling. Even though it is suboptimal to use the same regularization parameter for all price areas, this increases robustness of the results. The 28% average improvement in out-of-sample RMSE is a substantial improvement over autocovariance scaling.

Shrinkage scaling The regularization parameter in (8) is found by doing a grid search over the values $\lambda = 0$, 0.01, 0.02, $\ldots$, 1 and comparing the average in-sample improvement in RMSE. With the series variances the optimal value is $\lambda = 0.05$, while for the hierarchy variances the optimal value is $\lambda = 0$, i.e., no regularization. In Table 5, results are shown for series shrinkage scaling, while hierarchy shrinkage scaling is equivalent to cross-covariance scaling. The results using series shrinkage are very similar to those using series GLASSO scaling.

Cross-covariance scaling Scaling by the full cross-covariance matrix is the best performing approach in all areas and across all aggregation levels. The greatest improvements in RMSE are at the daily level, where they range from 52% to 68%. The average improvement in RMSE is 44% out of sample, with the greatest improvements occurring in SE3 and SE4. In several cases improvements in forecast accuracy are greater out of sample compared to in sample, which shows the robustness of the reconciliation approach. The significant improvement that is achieved using cross-covariance scaling compared to autocovariance scaling emphasizes the importance of capturing dependencies between forecast errors from different aggregation levels.

5 Simulation study

5.1 Setup

To better understand the differences and advantages of the proposed estimators, we compare their performance in a simulation study. We simulate data from the same double-seasonal exponential smoothing model (9) that was used for load forecasting with constant variance $\sigma^2 = 0.004$;

	Length of training samples in weeks			
	2	4	13	52
Bottom up	16 [−21; 83]	−13 [−33; 22]	−15 [−33; 11]	−7 [−25; 20]
Identity	−12 [−27; 14]	−6 [−21; 15]	−9 [−21; 5]	−13 [−24; 0]
Structural	−18 [−27; 1]	−20 [−27; −10]	−18 [−25; −8]	−16 [−22; −4]
Series variance	−14 [−31; 19]	−24 [−35; −10]	−20 [−30; −4]	−13 [−24; 3]
Hierarchy variance	−17 [−32; 11]	−26 [−36; −14]	−22 [−30; −9]	−15 [−23; 0]
Structural Markov	−23 [−37; −6]	−18 [−29; −2]	−17 [−26; −6]	−18 [−25; −8]
Series Markov	−26 [−38; −7]	−26 [−34; −15]	−22 [−30; −12]	−19 [−26; −6]
Hierarchy Markov	−26 [−40; −8]	−26 [−36; −14]	−24 [−32; −14]	−21 [−28; −10]
Autocovariance	13×10^3 $[100; 11 \times 10^4]$	−29 [−41; −17]	−26 [−35; −15]	−23 [−31; −13]
Series GLASSO	−19 [−44; 29]	−42 [−57; −22]	−50 [−61; −33]	−49 [−63; −31]
Hierarchy GLASSO	−19 [−44; 21]	−42 [−59; −17]	−52 [−65; −34]	−55 [−67; −36]
Series shrinkage	−20 [−45; 24]	−43 [−57; −24]	−51 [−61; −36]	−51 [−63; −33]
Hierarchy shrinkage	−21 [−44; 17]	−41 [−59; −17]	−52 [−65; −35]	−55 [−67; −38]
Cross-covariance	19×10^2 $[34; 59 \times 10^2]$	11×10^2 $[8; 64 \times 10^2]$	−53 [−69; −29]	−62 [−74; −41]

Table 4: Average percentage difference in RMSSE between reconciled and base forecasts across aggregation levels and across the seven-day test period. The numbers in square brackets are the 2.5% and 97.5% quantiles, respectively, for the 1,000 repetitions for each sample length.

smoothing parameters $\alpha = 0.14$, $\gamma_1 = 0.03$, and $\gamma_2 = 0.08$; and autoregressive parameter $\phi = 0.5$. These are the parameters values that were estimated for SE1.

For each series, after a burn-in period of 500 observations, we simulate 3, 5, 14, or 53 weeks of hourly observations using the initial values that were estimated for SE1. We discard the first 500 samples to avoid dependency on the initial values. In all cases, the last seven days are withheld as test set. Using the remaining samples as training data, we compare the performance of the estimators when forecasting one-day ahead for seven days. We repeat this process 1,000 times for each sample length.

5.2 Results

Table 4 shows the average percentage difference in RMSSE between reconciled and one-day-ahead base forecasts across aggregation levels and across the seven-day test period. We consider RMSSE rather than RMSE or RMSPE, because the scale of the simulated series varies and in some cases the simulated series overlap with zero. The numbers in square brackets are the 2.5% and 97.5% quantiles, respectively. For GLASSO scaling results are shown for $\lambda = 0.01$ and for shrinkage scaling results are shown for $\lambda = 0.05$. These are the same values that were used in Section 4.4. Although it is suboptimal to use the same regularization parameters for all sample lengths, this increases robustness of the results.

With only two weeks of data available for estimation, autocovariance scaling leads to reconciled forecasts that are 130 times worse than the base forecasts. The high figure itself is meaningless and only serves as evidence of the sample size being insufficient for estimating the autocovariances. With the total dimension of the temporal hierarchy being 60, a minimum of 455 observations, equivalent to 19 days of hourly measurements, are needed in order to estimate the autocovariance matrix within each aggregation level.[2] In comparison, hierarchy Markov scaling only requires 67 observations, which explains why this estimator yields the largest improvements in RMSSE when only two weeks of data is available. The GLASSO and shrinkage estimators perform somewhere in between Markov scaling and the WLS estimators, but the confidence intervals are quite wide.

[2]The number of parameters in a covariance matrix of dimension n is $\frac{n(n+1)}{2}$.

In some cases the shrinkage estimators do very well, while in other cases they do much worse than the simpler estimators.

With four weeks of data, the autocovariance matrix within each aggregation level can be estimated with reasonable precision, which leads to slightly better results than Markov scaling. Four weeks of hourly data is not enough to estimate the full cross-covariance matrix, which requires at least 1830 observations, corresponding to about 11 weeks of hourly measurements. Yet, the GLASSO and shrinkage estimators are able to significantly outperform autocovariance scaling by facilitating information sharing between aggregation levels. This clearly shows their advantage compared to the sample estimator.

With 13 weeks of hourly data available for estimation, cross-covariance scaling yields a similar average improvement as the GLASSO and shrinkage estimators, although with a wider confidence interval. This is enough data that forecast accuracy can be improved by more than 50% by considering both auto- and cross-correlations. The advantage of information sharing leads to improvements in RMSSE that are twice as large compared to only considering correlations within aggregation levels.

When one year of hourly data is available for training, the results are fairly similar to those presented in Section 4.4. Cross-covariance scaling yields the largest improvement in RMSSE, with the GLASSO and shrinkage estimators not far behind. The amount of data available offsets the advantages of the GLASSO and shrinkage estimators. Across the different lengths of training samples, there does not appear to be any significant difference between GLASSO and shrinkage scaling.

6 Discussion

Many real-life decision problems naturally involve multiple time horizons. For example, in energy planning, production has to meet expected demand over the next hours, the next day, the next week, and even longer horizons. It is often beneficial to take into account longer-term forecasts when making short-term decisions to ensure that decisions made now do not have a negative impact on future possibilities. Managers who have had to make decisions based on forecasts that were not coherent will know the issues that this can cause. Optimal decision making requires coherent forecasts, which is why reconciliation is needed.

As a specific example, consider a combined heat and power plant (CHP) that uses biomass along with solar thermal collectors and heat pumps. Decisions related to the purchase of biomass are established one–two years in advance; these contracts are decided based on long-term forecasts of the heat load and solar thermal production. One–two months before the biomass is actually consumed, its transportation has to be arranged—often from remote areas. Most CHP plants have an accumulator tank that can store heat for up to four days. In the Nord Pool areas, the day-ahead price of electricity depends on, i.a., the wind-power production relative to the load. If the electricity price is low, it is beneficial to use heat pumps instead of biomass. On an even shorter time scale, say with a 3–12-hour horizon, the optimal temperature level for the heat supply has to be decided.

In the example of the CHP plant, if the sum of the short-term forecasts is not coherent with the longer-term forecast, decisions related to the purchase of biomass become a choice between excess supply or inability to meet demand. The benefit of using temporal hierarchies for reconciling forecasts is the significant improvement in forecast accuracy that can be achieved. If forecast

errors are correlated, it is important to describe the dependencies correctly in the reconciliation process, since this can be of great importance to the outcome. The resulting forecasts, which are both coherent and more accurate, are in a format that is already used by decision makers. Substantially improved forecasts provide a possibility for making better decisions.

Our initial idea was to explore a cross-temporal hierarchy—similar to Kourentzes and Athanasopoulos (2019)—which, in addition to coherency across the various forecast frequencies, would require the sum of forecasts for the four price areas to be coherent with aggregate forecasts. The intention was to exploit information in the correlation between price areas to improve forecasts in each of the areas. In our case of load forecasting, however, we found that autocorrelation was much more important than including information about the structural correlation.

The improvement in forecast accuracy that can be achieved by considering autocorrelation depends on the data-generating process. There has to be autocorrelation in the forecast errors, which is often the case in practical applications when forecasting multiple steps ahead. In their simulation study, Athanasopoulos et al. (2017) considered seasonal, first-order, integrated moving-average processes. As the memory of a moving-average process is bounded by its order, this is a case where there is no autocorrelation in the forecast errors.

The temporal resolution of the data is another factor to consider. Athanasopoulos et al. (2017) applied hierarchies going from annual data at the top to monthly data at the bottom with a dimension of 28. This means that at least 406 observations, corresponding to 34 years of monthly observations, are needed in order to estimate all cross-correlations. It is certainly not always the case that this much data is available for estimation.

In our case, only one quarter of hourly observations was needed in order to estimate the cross-correlation matrix. However, even with hourly data, if forecasts are made more than one-day ahead in time, the sample cross-correlation matrix quickly becomes infeasible to estimate. In settings where limited data is available—or relevant due to stationarity concerns—compared to the dimension of the temporal hierarchy, the Markov, GLASSO, and shrinkage estimators come into their own, as evidenced by the simulation study in Section 5.

7 Conclusion

We proposed to consider forecast error autocorrelation when reconciling forecasts in a temporal hierarchy. Our first proposal was to consider the autocovariance matrix within each aggregation level. Even with high-frequency data available, these can be difficult to estimate. Thus, we proposed a second estimator that blends autocorrelation and variance information in a Markov structure that only requires estimation of the first-order autocorrelation coefficient at each aggregation level. In order to facilitate information sharing between aggregation levels, we proposed a third estimator that uses the GLASSO for estimating a sparse representation of the inverse cross-correlation matrix across aggregation levels. Our fourth and final proposal was a shrinkage estimator of the cross-correlation matrix.

We demonstrated the usefulness of the proposed estimators in a simulation study and through an application to short-term load forecasting in four price areas in Sweden. The simulation study showed that the best performing estimator depends on the length of the sample available for estimation. Even in settings with very limited data compared to the dimension of the temporal hierarchy, forecast accuracy could be improved by considering both auto- and cross-correlation.

By taking account of the autocorrelation structure when reconciling load forecasts, accuracy

could be improved uniformly across all frequencies and areas. The most accurate forecasts were obtained when capturing dependencies in forecast errors across aggregation levels, with improvements in RMSE from 23% to 68% out of sample. We argued that the resulting coherent and significantly more accurate forecasts are likely to yield better multi-horizon decisions.

The forecast reconciliation problem is fundamentally an optimization problem. Considering a general and flexible formulation based on convex optimization would provide the flexibility necessary to explore other loss functions than the squared error, to add regularization penalties, to combine structural and temporal aggregation constraints in a cross-temporal hierarchy, or to relax the aggregation constraints. All of which are matters that should be explored in future research, but that it is hardly possible to do in the framework of linear regression.

Acknowledgments

The authors are thankful for the comments and suggestions from two anonymous referees, in particular for suggesting a shrinkage estimator and simulation study. The work of Peter Nystrup, Pierre Pinson, and Henrik Madsen was supported by the Centre for IT-Intelligent Energy Systems (CITIES) project funded in part by Innovation Fund Denmark under Grant No. 1305-00027B.

References

Akshay Agrawal, S. D., Robin Verschueren and S. Boyd. "A rewriting system for convex optimization problems." *Journal of Control and Decision*, vol. 5, no. 1 (2018), pp. 42–60.

Amemiya, T. and R. Y. Wu. "The effect of aggregation on prediction in the autoregressive model." *Journal of the American Statistical Association*, vol. 67, no. 339 (1972), pp. 628–632.

Athanasopoulos, G., R. A. Ahmed, and R. J. Hyndman. "Hierarchical forecasts for australian domestic tourism." *International Journal of Forecasting*, vol. 25, no. 1 (2009), pp. 146–166.

Athanasopoulos, G., R. J. Hyndman, N. Kourentzes, and F. Petropoulos. "Forecasting with temporal hierarchies." *European Journal of Operational Research*, vol. 262, no. 1 (2017), pp. 60–74.

Athanasopoulos, G., R. J. Hyndman, H. Song, and D. C. Wu. "The tourism forecasting competition." *International Journal of Forecasting*, vol. 27, no. 3 (2011), pp. 822–844.

Banerjee, O., L. E. Ghaoui, and A. d'Aspremont. "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data." *Journal of Machine Learning Research*, vol. 9 (2008), pp. 485–516.

Bien, J., F. Bunea, and L. Xiao. "Convex banding of the covariance matrix." *Journal of the American Statistical Association*, vol. 111, no. 514 (2016), pp. 834–845.

Boyd, S. and L. Vandenberghe. *Convex Optimization*. Cambridge University Press: New York (2004).

Clemen, R. T. "Combining forecasts: A review and annotated bibliography." *International Journal of Forecasting*, vol. 5, no. 4 (1989), pp. 559–583.

Clements, A., A. Hurn, and Z. Li. "Forecasting day-ahead electricity load using a multiple equation time series approach." *European Journal of Operational Research*, vol. 251, no. 2 (2016), pp. 522–530.

Diamond, S. and S. Boyd. "CVXPY: A Python-embedded modeling language for convex optimization." *Journal of Machine Learning Research*, vol. 17, no. 83 (2016), pp. 1–5.

Diebold, F. X. and R. S. Mariano. "Comparing predictive accuracy." *Journal of Business & Economic Statistics*, vol. 13, no. 3 (1995), pp. 253–263.

Fan, S. and R. J. Hyndman. "Short-term load forecasting based on a semi-parametric additive model." *IEEE Transactions on Power Systems*, vol. 27, no. 1 (2012), pp. 134–141.

Friedman, J., T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, vol. 9, no. 3 (2008), pp. 432–441.

Gamakumara, P., A. Panagiotelis, G. Athanasopoulos, and R. J. Hyndman. "Probabilisitic forecasts in hierarchical time series." Working Paper 11/18, Monash University (2018).

Ghysels, E., P. Santa-Clara, and R. Valkanov. "The MIDAS touch: Mixed data sampling regression models." Working paper, UNC and UCLA (2004).

Gould, P. G., A. B. Koehler, J. K. Ord, R. D. Snyder, R. J. Hyndman, and F. Vahid-Araghi. "Forecasting time series with multiple seasonal patterns." *European Journal of Operational Research*, vol. 191, no. 1 (2008), pp. 207–222.

Gross, C. W. and J. E. Sohl. "Disaggregation methods to expedite product line forecasting." *Journal of Forecasting*, vol. 9, no. 3 (1990), pp. 233–254.

Hahn, H., S. Meyer-Nieberg, and S. Pickl. "Electric load forecasting methods: Tools for decision making." *European Journal of Operational Research*, vol. 199, no. 3 (2009), pp. 902–907.

Hall, S. G. and J. Mitchell. "Combining density forecasts." *International Journal of Forecasting*, vol. 23, no. 1 (2007), pp. 1–13.

Harvey, D., S. Leybourne, and P. Newbold. "Testing the equality of prediction mean squared errors." *International Journal of Forecasting*, vol. 13, no. 2 (1997), pp. 281–291.

Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. "Optimal combination forecasts for hierarchical time series." *Computational Statistics & Data Analysis*, vol. 55, no. 9 (2011), pp. 2579–2589.

Hyndman, R. J. and Y. Khandakar. "Automatic time series forecasting: The forecast package for R." *Journal of Statistical Software*, vol. 27, no. 3 (2008), pp. 1–22.

Hyndman, R. J. and A. B. Koehler. "Another look at measures of forecast accuracy." *International Journal of Forecasting*, vol. 22, no. 4 (2006), pp. 679–688.

Hyndman, R. J., A. B. Koehler, J. K. Ord, and R. D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer: Berlin (2008).

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

Jeon, J., A. Panagiotelis, and F. Petropoulos. "Reconciliation of probabilistic forecasts with an application to wind power." *arXiv:1808.02635* (2018).

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

Kourentzes, N., F. Petropoulos, and J. R. Trapero. "Improving forecasting by estimating time series structural components across multiple frequencies." *International Journal of Forecasting*, vol. 30, no. 2 (2014), pp. 291–302.

Ledoit, O. and M. Wolf. "A well-conditioned estimator for large-dimensional covariance matrices." *Journal of Multivariate Analysis*, vol. 88, no. 2 (2004), pp. 365–411.

Livera, A. M. D., R. J. Hyndman, and R. D. Snyder. "Forecasting time series with complex seasonal patterns using exponential smoothing." *Journal of the American Statistical Association*, vol. 106, no. 496 (2011), pp. 1513–1527.

Madsen, H. *Time Series Analysis*. Chapman & Hall: London (2008).

Nystrup, P., H. Madsen, and E. Lindström. "Long memory of financial time series and hidden Markov models with time-varying parameters." *Journal of Forecasting*, vol. 36, no. 8 (2017), pp. 989–1002.

Petropoulos, F. and N. Kourentzes. "Forecast combinations for intermittent demand." *Journal of the Operational Research Society*, vol. 66, no. 6 (2015), pp. 914–924.

Rostami-Tabar, B., M. Z. Babai, A. Syntetos, and Y. Ducq. "Demand forecasting by temporal aggregation." *Naval Research Logistics*, vol. 60, no. 6 (2013), pp. 479–498.

Schäfer, J. and K. Strimmer. "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics." *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1 (2005).

Sedoc, J., J. Rodu, D. Foster, and L. Ungar. "Multiscale hidden Markov models for covariance prediction." (2018).

Silvestrini, A. and D. Veredas. "Temporal aggregation of univariate and multivariate time series models: A survey." *Journal of Economic Surveys*, vol. 22, no. 3 (2008), pp. 458–497.

Taieb, S. B. "Sparse and smooth adjustments for coherent forecasts in temporal aggregation of time series." In *Proceedings of the Time Series Workshop at NIPS 2016*, edited by O. Anava, A. Khaleghi, M. Cuturi, V. Kuznetsov, and A. Rakhlin, vol. 55 of *Proceedings of Machine Learning Research* (2017), pp. 16–26.

Taieb, S. B., J. W. Taylor, and R. J. Hyndman. "Coherent probabilistic forecasts for hierarchical time series." In *Proceedings of the 34th International Conference on Machine Learning*, edited by D. Precup and Y. W. Teh, vol. 70 of *Proceedings of Machine Learning Research* (2017a), pp. 3348–3357.

———. "Hierarchical probabilistic forecasting of electricity demand with smart meter data." (2017b).

Taylor, J. W. "Short-term electricity demand forecasting using double seasonal exponential smoothing." *Journal of the Operational Research Society*, vol. 54, no. 8 (2003), pp. 799–805.

———. "Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles." *International Journal of Forecasting*, vol. 26, no. 4 (2010), pp. 627–646.

———. "Short-term load forecasting with exponentially weighted methods." *IEEE Transactions on Power Systems*, vol. 27, no. 1 (2012), pp. 458–464.

Tiao, G. C. "Asymptotic behaviour of temporal aggregates of time series." *Biometrika*, vol. 59, no. 3 (1972), pp. 525–531.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1 (2005), pp. 91–108.

Timmermann, A. "Forecast combinations." In *Handbook of Economic Forecasting*, edited by G. Elliott, C. W. J. Granger, and A. Timmermann, vol. 1, chap. 4. Elsevier: Amsterdam (2006), pp. 135–196.

van Erven, T. and J. Cugliari. "Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts." In *Modeling and Stochastic Learning for Forecasting in High Dimensions*, edited by A. Antoniadis, J.-M. Poggi, and X. Brossat, vol. 217 of *Lecture Notes in Statistics*. Springer: Cham (2015), pp. 297–317.

Wickramasuriya, S. L., G. Athanasopoulos, and R. J. Hyndman. "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization." *Journal of the American Statistical Association*, vol. 114, no. 526 (2019), pp. 804–819.

Yang, D., H. Quan, V. R. Disfani, and L. Liu. "Reconciling solar forecasts: Geographical hierarchy." *Solar Energy*, vol. 146 (2017a), pp. 276–286.

Yang, D., H. Quan, V. R. Disfani, and C. D. Rodríguez-Gallegos. "Reconciling solar forecasts: Temporal hierarchy." *Solar Energy*, vol. 158 (2017b), pp. 332–346.

Yuan, M. and Y. Lin. "Model selection and estimation in the Gaussian graphical model." *Biometrika*, vol. 94, no. 1 (2007), pp. 19–35.

Zhang, Y. and J. Dong. "Least squares-based optimal reconciliation method for hierarchical forecasts of wind power generation." *IEEE Transactions on Power Systems* (2018), pp. 1–1. URL `http://dx.doi.org/10.1109/tpwrs.2018.2868175`.

A RMSE results

	In sample (2016)				Out of sample (2017)			
	SE1	SE2	SE3	SE4	SE1	SE2	SE3	SE4
Daily								
Bottom up	18**	−4	2	−1	−9*	−9**	−11**	−10**
Identity	−15**	−16**	−14**	−14**	−18**	−14**	−17**	−15**
Structural	−14**	−14**	−10**	−10**	−21**	−13**	−15**	−13**
Series variance	−9**	−11*	−6*	−6	−20**	−11**	−13**	−12**
Hierarchy variance	−10**	−12*	−11**	−11**	−21**	−13**	−18**	−16**
Structural Markov	−17**	−18**	−17**	−16**	−22**	−16**	−20**	−19**
Series Markov	−15**	−16**	−14**	−13**	−22**	−15**	−19**	−17**
Hierarchy Markov	−17**	−19**	−30**	−28**	−24**	−20**	−34**	−31**
Autocovariance	−20**	−22**	−33**	−30**	−27**	−23**	−37**	−33**
Series GLASSO	−47**	−40**	−40**	−43**	−53**	−42**	−44**	−46**
Series shrinkage	−46**	−42**	−43**	−46**	−51**	−44**	−46**	−48**
Cross-covariance	−54**	−52**	−68**	−64**	−57**	−57**	−68**	−63**
Twelve-hourly								
Bottom up	31**	13*	22**	13**	13**	8**	20**	13**
Identity	3	3	6**	2	5**	6**	14**	9**
Structural	2	2	8**	4	1	5*	15**	10**
Series variance	5*	5	12**	7*	1	6*	17**	11**
Hierarchy variance	4	4	6*	2	0	4	11**	6*
Structural Markov	−1	−1	2	−1	0	3	10**	5
Series Markov	0	1	4	0	−1	4	11**	6*
Hierarchy Markov	1	−3	−12**	−13**	−3	−2	−5	−8**
Autocovariance	−5**	−5*	−15**	−16**	−6**	−5**	−9**	−11**
Series GLASSO	−28**	−19**	−19**	−27**	−30**	−17**	−12**	−21**
Series shrinkage	−27**	−22**	−23**	−30**	−29**	−20**	−16**	−25**
Cross-covariance	−35**	−32**	−50**	−48**	−33**	−36**	−43**	−40**
Eight-hourly								
Bottom up	19**	6	11**	6*	3	−6	3	3
Identity	−3**	−2	−2	−2*	−3*	−6**	0	0
Structural	−5**	−2	0	−1	−7**	−8**	0	1
Series variance	−2	0	3	1	−7**	−8**	1	2
Hierarchy variance	−3	−2	−2	−3*	−8**	−10**	−4*	−2*
Structural Markov	−6**	−4	−4**	−5**	−7**	−9**	−4**	−3**
Series Markov	−6**	−4	−3**	−4**	−8**	−10**	−4**	−2**
Hierarchy Markov	−8**	−7*	−17**	−16**	−10**	−15**	−16**	−13**
Autocovariance	−11**	−9**	−20**	−18**	−13**	−17**	−20**	−16**
Series GLASSO	−33**	−23**	−22**	−27**	−35**	−29**	−22**	−25**
Series shrinkage	−32**	−25**	−26**	−31**	−33**	−31**	−25**	−28**
Cross-covariance	−41**	−36**	−54**	−51**	−41**	−45**	−51**	−46**
Six-hourly								
Bottom up	14**	2	4	1	6**	−8*	−5	−3
Identity	−3*	−3	−7**	−5**	2*	−7**	−7**	−5**
Structural	−5**	−5**	−5**	−5**	−3**	−9**	−7**	−5**
Series variance	−3*	−3	−3	−3*	−3*	−9**	−6**	−4*
Hierarchy variance	−4*	−5*	−7**	−7**	−4**	−11**	−10**	−8**
Structural Markov	−6**	−6**	−10**	−9**	−3**	−10**	−10**	−8**
Series Markov	−6**	−6**	−9**	−8**	−4**	−11**	−10**	−7**
Hierarchy Markov	−8**	−9**	−22**	−18**	−6**	−15**	−21**	−18**
Autocovariance	−11**	−12**	−25**	−21**	−9**	−18**	−25**	−21**
Series GLASSO	−33**	−25**	−25**	−29**	−30**	−29**	−26**	−28**
Series shrinkage	−32**	−27**	−28**	−32**	−29**	−31**	−28**	−31**
Cross-covariance	−42**	−38**	−56**	−51**	−35**	−43**	−52**	−47**

Table 5: In- and out-of-sample percentage difference in RMSE between reconciled and base forecasts of one-day-ahead power consumption in Swedish areas for different data frequencies. Differences that are statistically significant at the 0.05 and 0.01 level, as evaluated by the Diebold–Mariano test (Diebold and Mariano, 1995), are marked with * and **, respectively.

	In sample (2016)				Out of sample (2017)			
	SE1	SE2	SE3	SE4	SE1	SE2	SE3	SE4
Four-hourly								
Bottom up	15^{**}	-1	-5	-5	5^{*}	-7	-11^{**}	-10^{*}
Identity	0	-5^{**}	-14^{**}	-11^{**}	2^{**}	-5^{*}	-13^{**}	-14^{**}
Structural	-3^{**}	-6^{**}	-14^{**}	-11^{**}	-2^{**}	-8^{**}	-13^{**}	-11^{**}
Series variance	-1	-5^{**}	-12^{**}	-9^{**}	-3^{**}	-8^{**}	-13^{**}	-10^{**}
Hierarchy variance	-2	-7^{**}	-15^{**}	-13^{**}	-3^{**}	-10^{**}	-16^{**}	-14^{**}
Structural Markov	-3^{**}	-8^{**}	-18^{**}	-14^{**}	-2^{**}	-9^{**}	-17^{**}	-14^{**}
Series Markov	-4	-8^{**}	-17^{**}	-13^{**}	-4^{**}	-9^{**}	-16^{**}	-13^{**}
Hierarchy Markov	-5^{**}	-11^{**}	-28^{**}	-23^{**}	-5^{**}	-14^{**}	-27^{**}	-22^{**}
Autocovariance	-9^{**}	-13^{**}	-31^{**}	-25^{**}	-8^{**}	-16^{**}	-30^{**}	-25^{**}
Series GLASSO	-28^{**}	-26^{**}	-32^{**}	-33^{**}	-27^{**}	-27^{**}	-31^{**}	-32^{**}
Series shrinkage	-27^{**}	-27^{**}	-35^{**}	-35^{**}	-26^{**}	-29^{**}	-33^{**}	-34^{**}
Cross-covariance	-38^{**}	-38^{**}	-59^{**}	-53^{**}	-34^{**}	-41^{**}	-55^{**}	-49^{**}
Three-hourly								
Bottom up	15^{**}	-1	-6	-10^{**}	5^{**}	-6	-13^{**}	-14^{**}
Identity	2	-4^{*}	-16^{**}	-15^{**}	2^{*}	-4^{*}	-15^{**}	-15^{**}
Structural	-1	-6^{**}	-15^{**}	-15^{**}	-2^{**}	-6^{**}	-15^{**}	-15^{**}
Series variance	0	-5^{**}	-12^{**}	-14^{**}	-2^{**}	-7^{**}	-14^{**}	-15^{**}
Hierarchy variance	0	-6^{**}	-16^{**}	-17^{**}	-3^{**}	-8^{**}	-18^{**}	-18^{**}
Structural Markov	-1	-7^{**}	-19^{**}	-18^{**}	-1	-7^{**}	-18^{**}	-18^{**}
Series Markov	-2^{*}	-7^{**}	-18^{**}	-17^{**}	-3^{**}	-7^{**}	-18^{**}	-17^{**}
Hierarchy Markov	-3^{**}	-9^{**}	-29^{**}	-26^{**}	-4^{**}	-11^{**}	-28^{**}	-26^{**}
Autocovariance	-7^{**}	-12^{**}	-31^{**}	-29^{**}	-7^{**}	-14^{**}	-31^{**}	-28^{**}
Series GLASSO	-26^{**}	-25^{**}	-32^{**}	-35^{**}	-25^{**}	-25^{**}	-32^{**}	-35^{**}
Series shrinkage	-24^{**}	-26^{**}	-35^{**}	-38^{**}	-24^{**}	-26^{**}	-34^{**}	-37^{**}
Cross-covariance	-36^{**}	-37^{**}	-59^{**}	-55^{**}	-32^{**}	-39^{**}	-55^{**}	-51^{**}
Two-hourly								
Bottom up	12^{**}	0	-1	-6^{*}	4^{**}	-2	-8^{**}	-10^{**}
Identity	0	-3	-10^{**}	-11^{**}	3^{*}	-1	-10^{**}	-11^{**}
Structural	-2^{**}	-5^{**}	-10^{**}	-11^{**}	-2^{*}	-3^{**}	-11^{**}	-11^{**}
Series variance	-1^{**}	-4^{**}	-8^{**}	-10^{**}	-2^{**}	-4^{**}	-10^{**}	-11^{**}
Hierarchy variance	-2^{**}	-5^{**}	-12^{**}	-13^{**}	-3^{**}	-5^{**}	-14^{**}	-14^{**}
Structural Markov	-3^{*}	-6^{*}	-14^{**}	-15^{**}	-1	-4^{**}	-14^{**}	-14^{**}
Series Markov	-3^{**}	-6^{**}	-13^{**}	-14^{**}	-3^{**}	-5^{**}	-14^{**}	-14^{**}
Hierarchy Markov	-4^{**}	-9^{**}	-25^{**}	-23^{**}	-4^{**}	-8^{**}	-24^{**}	-22^{**}
Autocovariance	-8^{**}	-11^{**}	-28^{**}	-25^{**}	-6^{**}	-11^{**}	-27^{**}	-25^{**}
Series GLASSO	-24^{**}	-23^{**}	-28^{**}	-32^{**}	-22^{**}	-21^{**}	-28^{**}	-31^{**}
Series shrinkage	-23^{**}	-24^{**}	-31^{**}	-34^{**}	-21^{**}	-23^{**}	-30^{**}	-34^{**}
Cross-covariance	-34^{**}	-36^{**}	-57^{**}	-52^{**}	-29^{**}	-35^{**}	-53^{**}	-48^{**}
Hourly								
Identity	-8^{**}	-2	-8^{**}	-4^{*}	-1	2	-1	0
Structural	-10^{**}	-4^{**}	-9^{**}	-5^{**}	-4^{**}	-1	-2	-1
Series variance	-10^{**}	-4^{**}	-7^{**}	-4^{**}	-4^{**}	-1	-2	-1
Hierarchy variance	-10^{**}	-4^{**}	-10^{**}	-8^{**}	-4^{**}	-2	-6^{**}	-6^{*}
Structural Markov	-10^{**}	-5^{*}	-13^{**}	-9^{**}	-4^{**}	-1	-6^{**}	-4^{*}
Series Markov	-11^{**}	-5^{**}	-12^{**}	-8^{**}	-5^{**}	-2	-6^{**}	-4
Hierarchy Markov	-11^{**}	-7^{**}	-23^{**}	-17^{**}	-5^{**}	-5^{*}	-17^{**}	-13^{**}
Autocovariance	-14^{**}	-10^{**}	-26^{**}	-20^{**}	-7^{**}	-7^{**}	-20^{**}	-16^{**}
Series GLASSO	-27^{**}	-21^{**}	-27^{**}	-27^{**}	-19^{**}	-18^{**}	-21^{**}	-23^{**}
Series shrinkage	-26^{**}	-22^{**}	-29^{**}	-29^{**}	-18^{**}	-19^{**}	-23^{**}	-25^{**}
Cross-covariance	-35^{**}	-33^{**}	-55^{**}	-48^{**}	-23^{**}	-30^{**}	-47^{**}	-41^{**}

Table 5: In- and out-of-sample percentage difference in RMSE between reconciled and base forecasts of one-day-ahead power consumption in Swedish areas for different data frequencies. Differences that are statistically significant at the 0.05 and 0.01 level, as evaluated by the Diebold–Mariano test (Diebold and Mariano, 1995), are marked with * and **, respectively.

B RMSPE results

	In sample (2016)				Out of sample (2017)			
	SE1	SE2	SE3	SE4	SE1	SE2	SE3	SE4
Daily								
Bottom up	21	−3	2	1	−5	−10	−10	−8
Identity	−15	−16	−14	−13	−17	−16	−17	−15
Structural	−14	−15	−10	−9	−20	−16	−16	−14
Series variance	−8	−11	−6	−5	−19	−14	−13	−11
Hierarchy variance	−9	−12	−11	−10	−20	−16	−18	−16
Structural Markov	−17	−18	−17	−15	−21	−19	−21	−19
Series Markov	−15	−16	−14	−13	−22	−18	−20	−17
Hierarchy Markov	−16	−20	−28	−25	−24	−23	−33	−30
Autocovariance	−20	−22	−32	−28	−26	−25	−36	−32
Series GLASSO	−45	−41	−41	−43	−50	−44	−44	−46
Series shrinkage	−44	−43	−44	−46	−49	−46	−46	−48
Cross-covariance	−52	−52	−68	−64	−54	−58	−67	−62
Twelve-hourly								
Bottom up	31	10	21	13	14	7	19	12
Identity	1	1	7	3	4	4	13	6
Structural	−1	0	8	4	−1	2	13	6
Series variance	3	2	12	7	0	3	15	8
Hierarchy variance	3	1	6	2	−1	1	9	3
Structural Markov	−3	−2	2	−2	−2	0	8	1
Series Markov	−2	−2	4	0	−3	0	8	2
Hierarchy Markov	−3	−5	−11	−12	−5	−6	−7	−10
Autocovariance	−7	−8	−15	−15	−8	−8	−11	−13
Series GLASSO	−27	−21	−22	−28	−31	−23	−17	−26
Series shrinkage	−27	−24	−25	−32	−30	−25	−20	−29
Cross-covariance	−34	−34	−51	−49	−33	−37	−45	−43
Eight-hourly								
Bottom up	19	2	10	5	5	−4	3	4
Identity	−3	−3	−1	−2	−2	−4	1	1
Structural	−6	−5	−1	−2	−7	−8	−1	0
Series variance	−3	−4	2	0	−7	−8	1	1
Hierarchy variance	−4	−3	−6	−5	−8	−10	−4	−4
Structural Markov	−7	−7	−5	−6	−7	−9	−4	−4
Series Markov	−7	−7	−5	−5	−8	−10	−5	−4
Hierarchy Markov	−9	−10	−18	−16	−10	−15	−17	−14
Autocovariance	−13	−13	−21	−19	−13	−17	−21	−17
Series GLASSO	−32	−25	−26	−30	−35	−31	−25	−27
Series shrinkage	−31	−27	−29	−33	−34	−32	−27	−30
Cross-covariance	−40	−37	−55	−52	−40	−44	−51	−46
Six-hourly								
Bottom up	14	−1	3	0	7	−5	−3	−2
Identity	−1	−3	−6	−4	2	−4	−5	−3
Structural	−5	−6	−6	−5	−3	−8	−6	−5
Series variance	−4	−6	−3	−4	−3	−9	−5	−4
Hierarchy variance	−5	−7	−8	−9	−4	−11	−10	−8
Structural Markov	−5	−7	−10	−9	−3	−9	−9	−7
Series Markov	−6	−8	−9	−9	−4	−10	−10	−8
Hierarchy Markov	−9	−11	−21	−19	−7	−16	−21	−18
Autocovariance	−13	−14	−25	−22	−10	−18	−25	−20
Series GLASSO	−33	−26	−27	−31	−31	−30	−27	−29
Series shrinkage	−32	−28	−29	−33	−30	−32	−29	−31
Cross-covariance	−41	−38	−56	−52	−35	−42	−52	−46

Table 6: In- and out-of-sample percentage difference in RMSPE between reconciled and base forecasts of one-day-ahead power consumption in Swedish areas for different data frequencies.

	In sample (2016)				Out of sample (2017)			
	SE1	SE2	SE3	SE4	SE1	SE2	SE3	SE4
Four-hourly								
Bottom up	15	−2	−4	−4	7	−4	−9	−6
Identity	2	−3	−11	−8	3	−1	−9	−7
Structural	−3	−6	−12	−9	−2	−6	−11	−9
Series variance	−1	−6	−10	−8	−2	−7	−10	−8
Hierarchy variance	−2	−7	−14	−12	−3	−9	−15	−12
Structural Markov	−2	−6	−16	−12	−2	−6	−14	−11
Series Markov	−3	−7	−15	−12	−3	−8	−15	−12
Hierarchy Markov	−5	−11	−26	−21	−5	−13	−25	−20
Autocovariance	−9	−13	−29	−24	−8	−15	−29	−23
Series GLASSO	−27	−25	−32	−32	−27	−27	−31	−31
Series shrinkage	−26	−26	−34	−35	−26	−29	−33	−33
Cross-covariance	−36	−37	−58	−53	−33	−39	−54	−48
Three-hourly								
Bottom up	16	0	−5	−8	6	−2	−10	−10
Identity	5	−1	−13	−11	3	1	−12	−10
Structural	1	−4	−13	−12	−1	−4	−13	−12
Series variance	1	−4	−11	−12	−2	−5	−12	−12
Hierarchy variance	0	−5	−15	−15	−2	−6	−16	−15
Structural Markov	1	−4	−16	−15	0	−3	−16	−14
Series Markov	0	−5	−16	−15	−2	−5	−16	−15
Hierarchy Markov	−2	−8	−26	−24	−3	−10	−26	−23
Autocovariance	−6	−11	−30	−27	−7	−13	−30	−26
Series GLASSO	−24	−23	−32	−35	−25	−25	−31	−33
Series shrinkage	−23	−24	−34	−37	−24	−26	−33	−35
Cross-covariance	−34	−35	−58	−54	−31	−37	−54	−49
Two-hourly								
Bottom up	13	0	−1	−4	5	0	−7	−7
Identity	3	0	−8	−8	3	2	−8	−7
Structural	−1	−3	−9	−9	−1	−2	−10	−10
Series variance	−1	−3	−7	−8	−2	−3	−9	−9
Hierarchy variance	−1	−5	−12	−13	−2	−5	−14	−13
Structural Markov	−1	−2	−13	−12	−1	−2	−13	−12
Series Markov	−2	−5	−12	−12	−2	−4	−13	−12
Hierarchy Markov	−3	−7	−23	−21	−3	−8	−23	−21
Autocovariance	−7	−10	−26	−24	−7	−11	−27	−23
Series GLASSO	−22	−21	−29	−32	−22	−22	−29	−30
Series shrinkage	−22	−22	−31	−34	−21	−23	−31	−33
Cross-covariance	−32	−34	−56	−52	−28	−34	−52	−47
Hourly								
Identity	−7	1	−7	−3	−1	3	−1	0
Structural	−10	−3	−8	−5	−4	−2	−3	−3
Series variance	−9	−3	−6	−4	−5	−2	−3	−3
Hierarchy variance	−10	−4	−10	−8	−5	−3	−7	−6
Structural Markov	−9	−2	−12	−8	−4	−1	−6	−5
Series Markov	−10	−4	−11	−8	−5	−3	−7	−6
Hierarchy Markov	−11	−6	−22	−17	−5	−6	−17	−14
Autocovariance	−14	−9	−25	−20	−8	−9	−21	−17
Series GLASSO	−26	−20	−27	−27	−19	−20	−23	−24
Series shrinkage	−25	−21	−29	−30	−19	−21	−25	−26
Cross-covariance	−34	−31	−54	−48	−23	−30	−47	−42

Table 6: In- and out-of-sample percentage difference in RMSPE between reconciled and base forecasts of one-day-ahead power consumption in Swedish areas for different data frequencies.