Joint Bottom-Up Method for Probabilistic Forecasting of Hierarchical Time Series

Nicolò Bertani*[1], Shane T. Jensen[2], and Ville A. Satopää[3]

[1]Católica Lisbon School of Business & Economics, Lisbon, Portugal
[2]The Wharton School of the University of Pennsylvania, Philadelphia, PA, United States
[3]INSEAD, Fontainebleau, France

January 1, 2025

Abstract

Many domains involve a hierarchy of time series, where the granular bottom-level series sum to upper-level series based on geography, product category, temporal granularity, or other features. Decision making in these domains requires forecasts that are accurate, probabilistic, and coherent in the sense of respecting the summing structure. In this paper, we first show that accurate and coherent probabilistic forecasts for all series in the hierarchy can be obtained by focusing on a joint model of the bottom-level series. Based on this result, we devise a Bayesian method that models the bottom-level series jointly, takes into account their contemporaneous and lagged dependence, and outputs a coherent probabilistic forecast of all series in the hierarchy. For empirical validation, we compare our method against many state-of-the-art techniques on data on Australian domestic tourism and product sales at Walmart. On each dataset, our method outperforms its competition in terms of prediction accuracy. To conclude, we demonstrate how our method can support decisions in inventory management of multiple Walmart products.

Keywords: Bayesian statistics; Dimensionality reduction; Multivariate autoregressive models; Probabilistic forecasting; Spike-and-slab

*Corresponding author. E-mail: nbertani@ucp.pt

1 Introduction

In a wide range of domains, such as product sales (Pennings and van Dalen, 2017; Oliveira and Ramos, 2019), wind power (Jeon et al., 2019), electricity demand (Taieb et al., 2017a), performance metrics in an organization (Novak et al., 2017), exports of goods (Eckert et al., 2021), and risk analysis (Li and Tang, 2019), agents require accurate forecasts of a *hierarchy* of time series, where the granular *bottom-level series* sum to *upper-level series*. For instance, in inventory management, agents often need forecasts of future product sales at each store, each state, and the entire country. This forms a three-level hierarchy where store-level sales sum to state-level sales that, in turn, sum to the total sales in the country. In some cases, instead of summing *over time series*, the hierarchy is formed by summing the values of a time series *over time intervals*. For instance, the agents might require forecasts of monthly, quarterly, and annual sales of a product.

In this article, we introduce a new *hierarchical forecasting*[1] technique called the *joint bottom-up* (JBU) method. Our method can be applied to hierarchies that sum over time series or time intervals, and makes forecasts that are *accurate*, *probabilistic*, and *coherent*. These properties are essential to optimal decision making. In fact, only probabilistic forecasts allow decision makers to assess uncertainty around the associated point predictions and to calculate expected profits or other metrics that can guide investment and policy decisions. Our method outputs a joint probabilistic forecast that describes both uncertainty and dependence among the series. This can become important if the decisions depend on the future values of multiple time series. For instance, a non-risk-neutral agent managing the inventory of multiple products must treat the products together, as a portfolio, and hence needs a joint probabilistic forecast to make optimal decisions around order quantities (e.g., Choi et al. 2011).

Forecasts of hierarchical time series should also be coherent in the sense that they respect the summing structure implied by the hierarchy. For instance, the forecast of the electricity demand in a city should equal the sum of the forecasts of the individual demands in its neighborhoods. Coherence is important because it aligns decision making among agents working at different levels of the hierarchy. In particular, in supply chain management, it can foster agreement about key decisions, such as optimal order quantities, at different points of the value chain (Oliveira and Ramos, 2019).

To meet these requirements, we develop the JBU method in several steps. First, we show under general conditions that a joint model of the bottom-level series is sufficient to optimize accuracy to

[1]Sometimes "hierarchical forecasting" is taken to refer exclusively to contexts where there is only one possible sequence of summations, leading from the bottom-level series to the upper-level series. "Grouped forecasting" then refers to contexts where different orders of summing are possible. For an illustration of a grouped forecasting context, see Figure 3. Given that our proposed method works the same in both hierarchical and grouped forecasting, we simplify the exposition and use the term "hierarchical forecasting" to refer to the general problem. See Di Fonzo and Girolimetto (2021) for an excellent summary of these different terms in the literature.

forecast all series in the hierarchy. This result offers novel perspective to the hierarchical forecasting literature, where the most recent developments have looked at different ways to incorporate the upper-level series and the structure of the hierarchy in the model (e.g., Pennings and van Dalen 2017; Wickramasuriya et al. 2019) instead of explicitly developing a joint model of the bottom-level series. A potential reason for this is that the estimation of a joint model of a large number of time series is notoriously challenging: It requires separating contemporaneous and lagged dependence, which, in turn, imposes the formally and computationally demanding task of controlling a potentially large number of lag terms both within and across time series. Such a high-dimensional environment makes model estimation vulnerable to spurious correlation, overfitting, and hence poor predictive accuracy.

Our JBU method estimates dependence among the bottom-level series using a Bayesian *base model*. This model takes into account the contemporaneous covariance, and allows lagged versions of each series to act as predictors of all other series, including themselves. To mitigate the risk of overfitting, the model performs automatic selection of the lag terms via adaptive regularization. Unfortunately, in some applications of hierarchical forecasting, there can be many (hundreds or more) bottom-level series, leading to a prohibitively large number (hundreds of thousands or millions) of potential predictors. To apply our JBU in such cases, we supplement the base model with a fast filter that induces sparsity. Specifically, our filter approximates the base model, ranks predictors in terms of their importance in predicting the target outcome, and selects a user-specified number of predictors for each of the bottom-level series. The filtered predictors are then input to our base model that estimates a joint model of the bottom-level series and performs regularization as needed.

Being a Bayesian approach operating at the most granular bottom level, our JBU method can produce probabilistic and coherent forecasts of all time series in the hierarchy by appropriately summing up the forecasts of the bottom-level series. By accounting for inter-series dependence in estimation, it can borrow strength across the bottom-level series and improve prediction accuracy over the previously proposed methods that model each series independently (e.g., Dunn et al. 1976). To the best of our knowledge, our joint bottom-up approach is the first hierarchical forecasting method that incorporates inter-series dependence and performs joint modeling of all time series.

To evaluate its potential to make accurate predictions, we compare JBU against the current state-of-the-art techniques in hierarchical forecasting. First, we show how our JBU outperforms its competition in a controlled simulation study with a small hierarchy. Then, we evaluate its performance on two real-world datasets. The first considers different types of domestic tourism at different geographic resolutions in Australia. This represents an important application that has been used at multiple stages to illustrate new methodology for hierarchical forecasting (e.g., Wick-

ramasuriya et al. 2019; Di Fonzo and Girolimetto 2021). Our second real-world data application turns to time series of Walmart's sales of hobby products over different time intervals. On these two datasets, we find that the probabilistic forecasts of JBU systematically improve the accuracy of the other methods. Compared to the classical bottom-up method (Dunn et al., 1976), that models each bottom-level series independently and neglects all inter-series dependence, the improvement is often particularly large. Overall, the improvements of JBU tend to be larger at the upper levels of the hierarchy, which may seem surprising given that our method only operates at the bottom level. This emphasizes the importance of modeling dependence of hierarchical time series.

In addition to these accuracy comparisons, we use the Walmart data to explore how JBU can support decision making in inventory management. Specifically, we estimate how well JBU can help a firm to control the risk of having many products out of stock simultaneously. Among all competing forecasting techniques, JBU leads to order quantities that are the most aligned with the firm's targeted service levels.

The rest of the article is structured as follows. Section 2 describes our notation and how the accuracy of probabilistic forecasts are evaluated. Section 3 presents our theoretical and methodological developments, namely the Bayesian base model and its filtering step. Section 4 reviews previously proposed methods for hierarchical forecasting. Sections 5 evaluates our JBU method on synthetic data. Sections 6 and 7 apply our JBU method to the Australian domestic tourism data and the Walmart sales data, respectively. Section 8 concludes with a summary and a brief discussion of limitations and other properties of the JBU method, including interpretability, broad applicability, and minimal requirements for user intervention. The Supplementary Material presents proofs of all propositions, the technical details required to estimate our Bayesian base model, and further performance comparisons under both synthetic and real-world data.

2 Preliminaries

2.1 Indexing

In this article, we denote all matrices and tensors (i.e., higher dimensional matrices) with bold upper-case letters, vectors with bold lower-case letters, and scalars with non-bold letters. All vectors are column vectors. To index subcomponents, we use subscripts for fixed coordinates and the dot $\cdot$ if the coordinate in that dimension is unspecified.

To make this specific, suppose there are M bottom-level series (i.e., series that cannot be formed by summing up some other series) with length T. We collect all these vectors into an $M \times T$ matrix $\boldsymbol{Y}$. Then, y_{mt} is the scalar value of the m-th bottom-level series at time t, and $\boldsymbol{y}_{\cdot t} = (y_{1t}, \ldots, y_{Mt})'$ is the M-dimensional vector of all values of the bottom-level series at time t. We denote the values of all upper-level series at time t with $\boldsymbol{u}_{\cdot t}$. Given that the upper-level

series are sums of the bottom-level series, there exists a binary matrix $\boldsymbol{S}$ such that $\boldsymbol{z}_{\cdot t} = \boldsymbol{S}\boldsymbol{y}_{\cdot t}$, where $\boldsymbol{z}_{\cdot t} = (\boldsymbol{u}'_{\cdot t}, \boldsymbol{y}'_{\cdot t})'$ denotes the values of all series in the hierarchy (i.e., both the bottom- and upper-level series) at time t.

For each bottom-level series at every time point, we consider P predictors. We collect all predictors for $\boldsymbol{Y}$ into a $M \times T \times P$ tensor $\boldsymbol{X}$. Then, by our convention, $\boldsymbol{x}_{mt\cdot}$ is the P-dimensional vector of all predictors of y_{mt}, and $\boldsymbol{X}_{m\cdot\cdot}$ is the $T \times P$ matrix of all P predictors of the m-th bottom-level series $\boldsymbol{y}_{m\cdot}$. Finally, we denote an M-dimensional vector of zeros with $\mathbf{0}_M$, an $M \times M$ matrix of zeros with $\mathbf{0}_{M\times M}$, and an $M \times M$ identity matrix with $\boldsymbol{I}_M$.

2.2 Scoring Rules

A probabilistic forecast takes the form of a predictive probability distribution over future quantities or events of interest (Gneiting and Katzfuss, 2014). In probabilistic hierarchical forecasting, the goal is to predict the h-step-ahead probability distribution of all series in the hierarchy conditional on their previous observations, namely $p(\boldsymbol{z}_{\cdot T+h}|\boldsymbol{z}_{\cdot 1}, \ldots, \boldsymbol{z}_{\cdot T})$. The quality of the prediction should be assessed using *strictly proper scoring rules*. In fact, using improper scoring rules can result in grossly misguided inference about forecasting accuracy (Gneiting and Raftery, 2007; Gneiting, 2011; Hilden and Gerds, 2014).

Definition 1 (Strictly Proper Scoring Rules)**.** Suppose the random vector $\mathbf{z}$ follows the distribution $p_{\mathbf{z}}$. The function $S(\cdot, \cdot)$ is a strictly proper scoring rule if $\mathbb{E}_{\mathbf{z}\sim\mathbf{p_z}}[S(q_{\mathbf{z}}, \mathbf{z})] > \mathbb{E}_{\mathbf{z}\sim\mathbf{p_z}}[S(p_{\mathbf{z}}, \mathbf{z})]$ for all distributions $q_{\mathbf{z}} \neq p_{\mathbf{z}}$.

Therefore, the expected value of a proper scoring rule is minimized by the actual or "true" distribution of the target outcome. Proper scoring rules jointly evaluate two of the most important properties of probabilistic forecasting, namely statistical consistency (i.e., calibration) and concentration (i.e., sharpness) of the forecast around the observed outcome (Gneiting and Katzfuss, 2014). Calibration is important because only a well-calibrated probabilistic forecast can be interpreted as a probability distribution in the frequentist sense and reliably used in the calculation of decision making criteria such as expected profits. A forecast that is both sharp and calibrated is maximally useful to the decision maker.

There are many strictly proper scoring rules. A well-known example for both univariate and multivariate distributions is the logarithmic scoring rule, which is simply the logarithm of the probabilistic forecast evaluated at the realized outcome. A univariate alternative is the cumulative ranked probability score (Matheson and Winkler, 1976) and its multivariate generalization is called the energy score (Gneiting and Raftery, 2007):

$$\mathrm{ES}(\mathbf{p}, \mathbf{z}) = \mathbb{E}_{\tilde{\mathbf{z}}_1\sim\mathbf{p}}||\tilde{\mathbf{z}}_1 - \mathbf{z}|| - \frac{1}{2}\mathbb{E}_{\tilde{\mathbf{z}}_1,\tilde{\mathbf{z}}_2\sim\mathbf{p}}||\tilde{\mathbf{z}}_1 - \tilde{\mathbf{z}}_2||, \tag{1}$$

where $||\cdot||$ is the Euclidian norm, $\mathbf{p}$ is the predicted probability, $\mathbf{z}$ is the realized multivariate outcome, and the vectors $\tilde{\mathbf{z}}_\mathbf{1}$ and $\tilde{\mathbf{z}}_\mathbf{2}$ are independent draws from the multivariate distribution $\mathbf{p}$.

In the empirical part of this paper, we choose to score probabilistic forecasts with the energy score (1) because it aligns well with hierarchical forecasting where the outcome is, by definition, multivariate. Unlike the logarithmic score, the energy score evaluates the alignment of the entire distribution with the outcome (i.e., it is sensitive to distance; see, e.g., Staël von Holstein 1970). The logarithmic score is also problematic in practice because it requires either a closed form expression or, when only a predictive sample is available, a kernel smoothed approximation of the predictive distribution to be evaluated at the observation. Unfortunately, a closed form expression is often unavailable and kernel smoothing involves a subjective choice of the kernel function. In contrast, the energy score only depends on two expected values that can be directly computed based on the observations and our posterior predictive samples.

3 Joint Bottom-Up Method

3.1 Sufficiency of the Bottom Level

The simplest approach to hierarchical forecasting is the *bottom-up* (BU) method (Dunn et al., 1976). This focuses on the bottom-level series and estimates the distribution $p(y_{mT+h}|\boldsymbol{y}_{m\cdot})$ independently for each $m = 1, \ldots, M$. In other words, BU first estimates the distribution of the future value of each bottom-level series based on its own history only. It then samples the bottom-level distributions and sums each draw according to the structure $\boldsymbol{S}$ to form a coherent predictive sample of all series in the hierarchy $\boldsymbol{z}_{\cdot T+h}$ at time $T+h$. If the bottom-level series are independent of each other, the classical bottom-up method uses all available information and can perform well.

In practice, however, time series are often dependent. For instance, tourism in one region can correlate with tourism in its neighboring regions due to bundled holiday packages (Cao et al., 2017). Similarly, sales of a given product can be associated with sales of related products due to complementarity effects. Failing to account for the inter-series dependence can lead to poor forecasting accuracy especially at the upper levels because the distribution of a sum of multiple quantities is determined by their mutual influence.[2]

This explains why we expect the hierarchical forecasts of the BU to improve if we allow it to learn and incorporate dependence among series. In fact, Proposition 1 explains that a bottom-up approach can be sufficient to minimize any strictly proper scoring rule over all series in the

[2]For instance, consider estimating variance that is often a key input to risk assessment. The variance of the sum of random variables increases in the total pairwise covariance of the individual series. Therefore, if the bottom-level series are positively (negatively) correlated, the classical bottom-up method can end up under-estimating (over-estimating, respectively) the variability and hence the risk in the upper-level series.

hierarchy as long as the bottom-level series are modeled jointly.[3]

Proposition 1. *The distribution of the bottom-level series $p(\boldsymbol{y}_{\cdot T+h}|\boldsymbol{Y})$ is sufficient to minimize the expectation of any strictly proper scoring rule over all series in the hierarchy.*

Proof. See Supplementary Material S1.1. □

Intuitively, the joint distribution of the bottom-level series fully characterizes the distribution of all series in the hierarchy because the upper-level series are mere linear combinations of the bottom-level series.[4] Therefore, an accurate estimate of $p(\boldsymbol{y}_{\cdot T+h}|\boldsymbol{Y})$ naturally propagates to an accurate estimate of $p(\boldsymbol{S}\boldsymbol{y}_{\cdot T+h}|\boldsymbol{Y})$.

Estimating the joint distribution of the bottom-level series $p(\boldsymbol{y}_{\cdot T+h}|\boldsymbol{Y})$ requires us to model their dependence. On the positive side, incorporating dependence into our model allows us to "borrow strength" and stabilize estimates of the bottom-level series, which can be noisy and hence more difficult to forecast separately. On the negative side, this dependence can be complicated and take on various forms. In particular, it can be contemporaneous (i.e., between concurrent values of different bottom-level series) or lagged (i.e., between the current and past values of the same or different bottom-level series). There are numerous ways such dependencies can arise, and without any further restrictions, all such possibilities cannot be explored in practice. Therefore, to bring the general result in Proposition 1 to practice, we need more structure. The next subsection considers an autoregressive model that allows both contemporaneous and lagged dependence while restricting their forms enough so that the model can be estimated from data.

3.2 Bayesian Base Model

To operationalize Proposition 1, we need a tractable model that can describe both contemporaneous and lagged dependence among the bottom-level series. Our approach, called the Joint Bottom-Up or simply the JBU method, derives from a Bayesian base model with three core assumptions. First, we assume that the contemporaneous dependence is stationary.

[3]We have stated that the study of all time series in the hierarchy can be limited, without informational loss, to the bottom-level series. More precisely, the retained subset of the series does not have to be the bottom-level series. Instead, it could be any non-degenerate subset of the series that allows the reconstruction of all series in the hierarchy. However, the set of the bottom-level series is the most natural candidate, as well as the choice considered in past literature (e.g., recall the classical bottom-up method). Hence other non-degenerate subsets are disregarded in this paper.

[4]This intuition extends to any properties of the upper-level series that can be constructed based on the same properties of the bottom-level series. For instance, similarly to probabilistic forecasts, point forecasts of a given functional $T(\cdot)$ are evaluated using consistent scoring functions for that functional: $\mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y})}[S(t,\boldsymbol{y})]$ is uniquely minimized by the functionals of the actual distribution $t \in T(p(\boldsymbol{y}))$. Clearly, here the functionals of the bottom-level series are sufficient only if they characterize the functionals of the entire hierarchy. Even though this is true, for instance, for the mean, minimum, and maximum, it is not generally true for quantiles and dispersion measures. By Proposition 1, however, all functionals can be estimated based on probabilistic forecasts of the bottom-level series. In this sense probabilistic forecasting offers a fully general solution to hierarchical forecasting.

Assumption 1 (Contemporaneous Dependence)**.** For all time points t, the bottom-level series respect the model $\boldsymbol{y}_{\cdot t} = \boldsymbol{\mu}_{\cdot t} + \boldsymbol{\epsilon}_{\cdot t}$, where $\boldsymbol{\mu}_{\cdot t} = (\mu_{1t}, \ldots, \mu_{Mt})' \in \mathbb{R}^M$ is any mean process and $\boldsymbol{\epsilon}_{\cdot t} = (\epsilon_{1t}, \ldots, \epsilon_{Mt})' \in \mathbb{R}^M$ is a noise vector such that $\mathbb{E}[\boldsymbol{\epsilon}_{\cdot t}] = \mathbf{0}_M$, $\mathbb{E}[\boldsymbol{\epsilon}_{\cdot t}\boldsymbol{\epsilon}_{\cdot t}'] = \boldsymbol{\Sigma}$, and $\mathbb{E}[\boldsymbol{\epsilon}_{\cdot t}\boldsymbol{\epsilon}_{\cdot u}'] = \mathbf{0}_{M\times M}$ for all $t \neq u$.

The error term $\boldsymbol{\epsilon}_{\cdot t}$ captures all contemporaneous dependence among the bottom-level series at time t. The exact form of this dependence is given by the contemporaneous covariance matrix (CCM) $\boldsymbol{\Sigma}$. No restrictions are placed upon $\boldsymbol{\Sigma}$, allowing it to represent a wide range of contemporaneous dependences among the series.

Second, to specify a likelihood, we assume that the errors are normally distributed.

Assumption 2 (Normality)**.** Let $\boldsymbol{\epsilon}_{\cdot t} \overset{iid}{\sim} \mathcal{N}_M(\mathbf{0}_M, \boldsymbol{\Sigma})$ for all time points t.

Any distributional assumption limits the generality of a model. Assuming normality is a natural starting point for model development because in practice it often (possibly after a transformation) offers a reasonable description of the data (e.g., Winkler 1981; Chen and Zhao 2023). In our context, normality yields a tractable model for which the estimation procedure can be expressed in closed-form (see Supplementary Material S2) and hence computed more efficiently. Naturally, the model can be extended to other domains by following the theory of generalized linear models (McCullagh and Nelder, 1983).[5]

Finally, we assume that the mean process μ_{mt} is a linear function of some P predictors.[6]

Assumption 3 (Lagged Dependence)**.** For all $m = 1, \ldots, M$ and all $t = 1, \ldots, T$, we have that $\mathbb{E}[y_{mt}|\boldsymbol{x}_{mt\cdot}] = \mu_{mt} = \boldsymbol{x}_{mt\cdot}'\boldsymbol{\beta}_{m\cdot}$ is a linear function of some P predictors $\boldsymbol{x}_{mt\cdot} = (x_{mt1}, \ldots, x_{mtP})' \in \mathbb{R}^P$ with coefficients $\boldsymbol{\beta}_{m\cdot} = (\beta_{m1}, \ldots, \beta_{mP})' \in \mathbb{R}^P$.

There are no restrictions on the nature of these predictors. In principle, they can include available leading indicators or even judgmental predictions made by human experts. Non-linear effects, if deemed necessary, could be captured with an appropriate basis function expansion of the predictors (e.g., Friedman et al. 2001). In this paper, we focus on modeling pure time series data and hence form each predictor vector $\boldsymbol{x}_{mt\cdot}$ only based on previous observations of the m-th bottom-level series itself (self-lags) and other bottom-level series (cross-lags). This way the mean process is modeled with an autoregressive linear model that can capture complex forms of lagged dependence. Given that the coefficients $\boldsymbol{\beta}_{m\cdot}$ are specific to the m-th bottom-level series $\boldsymbol{y}_{m\cdot}$, the structure of the lagged dependence can be different for each of these series.

[5]For binary or count data one can augment the problem with continuous latent variables that are modeled by our current Bayesian base model and then transformed to the desired scale using an appropriate link function. For instance, the latent variable construction of the probit model transforms a normally distributed variable to a binary variable via the threshold function (Albert and Chib, 1993). Such extensions are left for future work.

[6]Even though our method allows the bottom-level series to have different numbers of predictors, for the sake of notational clarity and readability, we do not consider such models in this article.

If there are many bottom-level series, considering even a small number of past observations of each series as potential predictors quickly multiplies into a high-dimensional vector of predictors $\boldsymbol{x}_{mt\cdot}$ for all $m = 1, \ldots, M$. This makes our method vulnerable to spurious correlation (Fan and Lv, 2008), overfitting, and poor forecasting accuracy. To mitigate such risks, we need to regularize the estimation of our model. Unfortunately, we are not aware of any existing methodology that regularizes the joint estimation of lagged and contemporaneous dependence among multiple time series. To make this possible, we use Bayesian statistics with carefully chosen prior distributions.

First, consider the prior for the coefficients $\boldsymbol{\beta}_{m\cdot}$. Absolutely continuous Normal-Mixture Inverse Gamma (NMIG) spike-and-slab priors for $\boldsymbol{\beta}_{m\cdot}$ have useful properties to cope with high-dimensionality: They induce adaptive shrinkage in β_{mp} in proportion to the estimated predictive ability of $\boldsymbol{x}_{m\cdot p}$ and possess oracle-like risk performance (Ishwaran and Rao, 2005, 2011, 2014). In this article, we extend the previously univariate NMIG spike-and-slab regression to a multivariate model with a general CCM. The resulting model can be interpreted as a fully probabilistic (Bayesian) multivariate generalized ridge regression with a general contemporaneous covariance matrix.

Next, consider the prior distribution for the CCM $\boldsymbol{\Sigma}$. Past literature proposes two weakly informative priors for covariance matrices: the Jeffreys' prior $\pi(\boldsymbol{\Sigma}) \propto \boldsymbol{\Sigma}^{-\frac{M+1}{2}}$ (e.g., Kadiyala and Karlsson, 1997) and the RATS prior $\pi(\boldsymbol{\Sigma}) \propto \boldsymbol{\Sigma}^{-\frac{(P+1)M}{2}-1}$. The RATS prior is a modified version of the Jeffreys' prior and is used in the RATS (Regression Analysis of Time Series, Enders 1996) software package and many published works in empirical macroeconomics (e.g., Ni and Sun, 2005; Sims and Zha, 2006). However, neither of these priors guarantee a non-singular covariance matrix in high-dimensional settings, which are common in hierarchical forecasting applications. Consequently, we consider priors that are more (but still weakly) informative. The most common informative prior is the conjugate inverse-Wishart distribution $\mathcal{W}^{-1}(\nu_0, \boldsymbol{\Psi}_0)$ with two hyperparameters: degrees of freedom $\nu_0 > M - 1$ and a positive-definite scale matrix $\boldsymbol{\Psi}_0 \in \mathbb{R}_{\succ}^{M \times M}$. The first prior we consider is $\nu_0 = M + 1$ and $\boldsymbol{\Psi}_0 = \boldsymbol{I}_M$. Given that this aligns with Jeffreys' prior in terms of the number of pseudo-observations, we call this the Jeffreys-inspired prior. It is non-informative in the sense that it coincides with the marginal uniform distribution for all correlations (Barnard et al., 2000). By a similar logic, we also consider a RATS-inspired prior with $\nu_0 = (P + 1)M - 2$ and $\boldsymbol{\Psi}_0 = \boldsymbol{I}_M$. Comparing these two priors, our RATS-inspired prior shrinks more heavily towards the diagonal and hence imposes stronger regularization.

To bring this all together, we concatenate all coefficients into a single vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1\cdot}', \ldots, \boldsymbol{\beta}_{M\cdot}')' \in \mathbb{R}^{MP}$ and collect all corresponding predictors at time t into a block-diagonal matrix $\boldsymbol{D}_t = \text{diag}(\boldsymbol{x}_{1t\cdot}', \ldots, \boldsymbol{x}_{Mt\cdot}') \in \mathbb{R}^{M \times MP}$. Then, by assumptions 1, 2, and 3 and our chosen prior distributions, we have the follow-

ing Bayesian base model:

$$
\begin{aligned}
\text{For } t = 1, \ldots, T, \quad & \boldsymbol{y}_{.t} | \boldsymbol{D}_t, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}_M(\boldsymbol{D}_t \boldsymbol{\beta}, \boldsymbol{\Sigma}) \\
& \boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\nu_0, \boldsymbol{\Psi}_0) \\
\text{For } m = 1, \ldots, M, \;\; p = 1, \ldots, P, \quad & \beta_{mp} | \gamma_{mp}, \tau_{mp} \sim \mathcal{N}(0, \, \gamma_{mp} \tau_{mp}^2) \\
& \gamma_{mp} | \omega = \begin{cases} 1 & \text{with probability } \omega \\ \varepsilon > 0 & \text{otherwise} \end{cases} \\
& \omega \sim \mathcal{U}(0, 1) \\
& \tau_{mp}^2 | a_1, a_2 \sim \Gamma^{-1}(a_1, a_2),
\end{aligned}
\tag{2}
$$

where $\mathcal{W}^{-1}(\nu_0, \, \boldsymbol{\Psi}_0)$ is the (Jeffreys- or RATS-inspired) inverse-Wishart distribution, $\mathcal{U}(0, 1)$ is the continuous uniform distribution over the unit interval, and $\Gamma^{-1}(a_1, a_2)$ is the inverse gamma distribution with shape a_1 and scale a_2. The values of the hyperparameters ε, a_1, and a_2 determine the scale of the distribution of β_{mp} and hence control the level of regularization (i.e., shrinkage towards no effect $\beta_{mp} = 0$ for all m and p). By default, we follow Ishwaran and Rao (2005) and set $\varepsilon = 0.05$, $a_1 = 5$, and $a_2 = 50$. However, as is often the case in statistics or machine learning, performance can be improved by fine-tuning the parameters to the context at hand (Vandeput, 2021). Finally, following the typical recommendations for ridge regression (e.g., Friedman et al., 2001), training data and predictors are standardized to have mean zero and unit variance. To avoid data leakage, test data are standardized using the same means and variances of the associated training data.

In Supplementary Material S2, we derive a Gibbs sampler (Geman and Geman, 1984) to estimate the base model (2) by producing samples from the posterior distribution of the parameters. A blocked Gibbs formulation of the update step of $\boldsymbol{\beta}$ improves the computational efficiency and allows us to estimate models with up to thousands of predictors reasonably quickly. For instance, in Section 5 we have $M = 4$ bottom-level series each with $P = 4$ predictors, and the sampler takes around 0.002 seconds per iteration.[7] This runtime naturally increases in the number of bottom-level series M and the number of predictors per series P. For instance, in Section 7, with $M = 149$ and $P = 12$, each iteration takes around 2.26 seconds. In Section 6, with $M = 308$ and $P = 12$, the runtime reaches 36.8 seconds per iteration. In our applications, the full estimation of our model requires 700 iterations that in these three cases takes about 1.4 seconds, 26 minutes, and 7.2 hours. For how to handle much larger problems, see our discussion in Section 8.

[7] Runtimes were recorded on the corresponding author's laptop (MacBook Pro 16-inch, 2021, 2.4 GHz 8-Core Intel Core i9). The software of our implementation is available on the corresponding author's Github page.

3.3 Filtering a Large Number of Predictors

Even though the estimation of our base model can handle thousands of predictors, applications of hierarchical forecasting can have hundreds or thousands of bottom-level series, which leads to a prohibitively high number of potential predictors. For instance, in our application to Australian domestic tourism (Section 6), there are $M = 308$ bottom-level series. The data are monthly and exhibit yearly seasonality, which suggests that no fewer than 12 lags per series should be considered. A joint analysis of all (self- and cross-) lags then entails considering $308^2 \times 12 = 1,138,368$ predictors simultaneously. Given that our model estimation procedure must repeatedly invert a matrix of this dimension, applying our JBU method with such a large number of lag terms is infeasible.

In such contexts with an extremely high number of predictors, it is unlikely that all predictors are necessary for making accurate forecasts. A simple way forward is then to pair a well-performing, lower-dimensional method with a variable selection technique that pre-selects predictors and reduces dimensionality to a level where joint estimation is feasible. For similar reasons, Ishwaran and Rao (2003) pair the Normal-Mixture Inverse Gamma model with the Zcut procedure that excludes coefficients of $\boldsymbol{\beta}$ by comparing their approximate posterior means against zero. Unfortunately, this is computationally infeasible at our scale because it requires an approximate posterior sample for $\boldsymbol{\beta}$.

To align the filtering step with our base model (2), we introduce a scalable alternative that performs dimensionality reduction by approximating a Dirac spike-and-slab model (Mitchell and Beauchamp, 1988). Compared to our base model, the Dirac spike-and-slab model slightly changes the prior for the coefficients by using a mixture of a point mass at zero (spike) and a Gaussian density with mean zero and unknown variance (slab). This strategy induces sparsity – not adaptive regularization – by identifying predictors that ought to have exactly zero impact in the model. Specifically, focusing on the m-th bottom-level series, the Dirac spike-and-slab modifies the base model (2) to the following:

$$
\begin{aligned}
\boldsymbol{y}_{m\cdot}|\boldsymbol{X}_{m\cdot\cdot}, \boldsymbol{\beta}_{m\cdot}, \sigma_m^2 &\sim \mathcal{N}_T(\boldsymbol{X}_{m\cdot\cdot}\boldsymbol{\beta}_{m\cdot}, \sigma_m^2 \boldsymbol{I}_T) \\
\text{For } p = 1, \ldots, P, \qquad \beta_{mp}|\alpha_{mp}, \sigma_m^2 &\sim (1-\alpha_{mp})\,\delta_{\beta_{mp}}(0) + \alpha_{mp}\,\mathcal{N}(0,\, c\,\sigma_m^2) \qquad (3) \\
\alpha_{mp}|\psi &\sim \text{Br}(\psi),
\end{aligned}
$$

where $\alpha_{mp} \in \{0, 1\}$ is the inclusion parameter indicating whether the p-th predictor of the m-th bottom-level series $\boldsymbol{x}_{m\cdot p}$ is included in the final model, $\delta_{\beta_{mp}}(0)$ is the Dirac delta measure placing point mass at $\{\beta_{mp} = 0\}$, $\text{Br}(\psi)$ is the Bernoulli distribution with success probability ψ, and c is the reciprocal of the ridge penalty coefficient multiplied by the variance σ_m^2 to offset scale. The variance term σ_m^2 is assigned the default prior $p(\sigma_m^2) \propto \sigma_m^{-2}$. The remaining parameters $c \in \mathbb{R}_+$

and $\psi \in [0,1]$ will not affect our ultimate selection criterion and hence are kept as unspecified constants (for details, see Supplementary Material S1.2).

Unfortunately, direct identification of (3) or even its approximation through sampling (Malsiner-Walli and Wagner, 2018) are infeasible at our scale. Therefore we approximate Dirac selection with the marginalized conditional posterior inclusion probabilities of the predictors. Specifically, let $\boldsymbol{\alpha}_{-mp}$ denote the vector of all inclusion parameters for the m-th bottom-level series except α_{mp}. The posterior inclusion probability of $\boldsymbol{x}_{m\cdot p}$, conditional on $\boldsymbol{\alpha}_{-mp} = \mathbf{0}_{P-1}$ and marginalized over $\boldsymbol{\beta}_{m\cdot}$ and σ_m^2 (Malsiner-Walli and Wagner, 2018), takes the following form:

$$\mathbb{P}(\alpha_{mp} = 1 | \boldsymbol{y}_{m\cdot}, \boldsymbol{\alpha}_{-mp} = \mathbf{0}_{P-1}) \propto \pi(\boldsymbol{y}_{m\cdot} | \alpha_{mp} = 1, \boldsymbol{\alpha}_{-mp} = \mathbf{0}_{P-1}) \mathbb{P}(\alpha_{mp} = 1 | \psi),$$

where

$$\pi(\boldsymbol{y}_{m\cdot} | \alpha_{mp} = 1, \boldsymbol{\alpha}_{-mp} = \mathbf{0}_{P-1}) \propto \frac{(\boldsymbol{x}'_{m\cdot p}\boldsymbol{x}_{m\cdot p} + 1/c)^{-1/2}}{\left[\boldsymbol{y}'_{m\cdot}\boldsymbol{y}_{m\cdot} - (\boldsymbol{x}'_{m\cdot p}\boldsymbol{x}_{m\cdot p} + 1/c)^{-1}(\boldsymbol{x}'_{m\cdot p}\boldsymbol{y}_{m\cdot})^2\right]^{\frac{T-1}{2}}}. \tag{4}$$

Assuming that a priori each predictor is equally likely to be selected, i.e., that $\mathbb{P}(\alpha_{mp} = 1|\psi) = \mathbb{P}(\alpha_{mq} = 1|\psi)$ for all $p \neq q$, our Dirac filter selects predictor p over predictor q if

$$\pi(\boldsymbol{y}_{m\cdot} | \alpha_{mp} = 1, \boldsymbol{\alpha}_{-mp} = \mathbf{0}_{P-1}) > \pi(\boldsymbol{y}_{m\cdot} | \alpha_{mq} = 1, \boldsymbol{\alpha}_{-mq} = \mathbf{0}_{P-1}). \tag{5}$$

Following this ordering among the predictors, our Dirac filter selects a user-defined K predictors per bottom-level series.[8]

In addition to aligning with our base model, our filtering procedure establishes a strong connection between Bayesian variable selection based on the spike-and-slab priors and Sure Independence Screening.

Proposition 2. *The Dirac filter based on the criterion* (5) *selects predictors equivalently to Sure Independence Screening of Fan and Lv (2008).*

Proof. See Supplementary Material S1.2. □

Therefore, our Dirac filter satisfies the *sure screening property*: For sufficiently large K and predictors that are neither strongly collinear nor both jointly correlated and marginally uncorrelated with the response variable, all important variables will be selected with probability tending to one (Fan and Lv, 2008).

[8]The filter can be used to select a different number of predictors per bottom-level series and, as mentioned earlier, these predictors can then be used in the Bayesian base model. For brevity, we do not explicitly consider different numbers of predictors per series in the current article.

Importantly, our Dirac filter only requires MP evaluations of the marginal likelihood (4) and hence can be applied efficiently even to a large number of predictors. For instance, on the corresponding author's laptop, the filter can process more than a million predictors in a matter of seconds. Our filter does not produce posterior distributions or even point estimates of the posterior inclusion probabilities. Indeed, its purpose is not model estimation. Instead, its goal is to induce sparsity in a manner that is feasible at scale and consistent with our base model. The selected variables are then input to the base model that is responsible for joint estimation and adaptive regularization.

4 Hierarchical Forecasting Methods

The rest of this article compares our joint bottom-up method against alternative hierarchical forecasting methods on synthetic (Section 5) and real-world data with hierarchies summing both over time series (Section 6) and over time intervals (Section 7). The set of *competing methods* includes variants of our JBU method and previously proposed methods (PPMs). The variants of our JBU method and their respective acronyms are as follows.

- $\text{JBU}_{\text{J, K}}$ and $\text{JBU}_{\text{R, K}}$: This is our JBU method with the Jeffreys- ($\text{JBU}_{\text{J, K}}$) and RATS-inspired ($\text{JBU}_{\text{R, K}}$) priors for the CCM. If the Dirac filter is applied (see Section 3.3), the number of filtered predictors per bottom-level series is emphasized with the subscript K. When no filtering is performed, subscript K is omitted. Probabilistic forecasts are obtained by sampling the joint posterior predictive distribution of the bottom-level series and then summing each draw according to the structure of the hierarchy.

- IND_{K}: This is like our JBU method in the previous point but under the assumption that the CCM is proportional to the identity matrix, i.e., the bottom-level series are assumed to be contemporaneously independent.

- VAR_{K}: Vector AutoRegression (VAR) is a classical way to model multiple time series (Hamilton, 1994; Madsen, 2007). Similarly to our JBU, VAR predicts each bottom-level series based on the lagged observations of all bottom-level series, including itself. Contrary to JBU, however, VAR coefficients are obtained via maximum-likelihood estimation that includes no (adaptive) regularization nor joint estimation of the CCM. If the dimension becomes too large for a direct application of VAR, we use our Dirac filter (see Section 3.3) to select K relevant predictors per bottom-level series. When no filtering is performed, subscript K is omitted. Probabilistic forecasts are obtained by first sampling the bottom-level series from a multivariate normal distribution, where the mean is given by the point predictions and the covariance matrix is given by the estimated CCM. The draws are then summed according to the structure of the hierarchy.

Our JBU is compared against the following *previously proposed methods (PPMs)* for probabilistic hierarchical forecasting. These methods require a separate time series model, typically from the ARIMA or exponential smoothing (ETS) family, for making base forecasts that are then transformed differently depending on the PPM. The exact base models used in the empirical applications of Sections 5, 6, and 7 are described separately in those sections.

- BU: This is the classical bottom-up method described in the beginning of Section 3.1.

- Reconciliation techniques for hierarchies summing over time series. Many previous techniques for hierarchical forecasting rely on *forecast reconciliation*. This proceeds in two steps. First, every series in the hierarchy (i.e., every value in $\mathbf{z}_{\cdot T+h}$) is forecast independently using a base time series model. Second, given that these *base forecasts* are unlikely to be coherent, they are transformed or "reconciled" to respect coherency. There are many ways to perform reconciliation. Our comparison includes the following approaches.

  - MinT_{shr}: In the forecast reconciliation literature, the MinT shrink method of Wickramasuriya et al. (2019) is typically regarded as the reference approach. This reconciles the base forecasts using the covariance matrix of the errors of 1-step-ahead in-sample point predictions, shrunk towards the diagonal according to Schäfer and Strimmer (2005). Probabilistic forecasts can be obtained by first sampling a multivariate normal distribution, where the mean is given by the reconciled point predictions of the bottom-level series and the covariance matrix is given by the covariance of 1-step ahead errors of these point forecasts, and then summing each draw according to the structure of the hierarchy (see, e.g., Gamakumara et al. 2018 and Athanasopoulos et al. 2020).
  - Copula: Taieb et al. (2021) introduce a hybrid method that uses MinT_{shr} for point forecasts and a copula-based procedure for probabilistic forecasts. First, the method independently constructs predictive samples for each bottom-level series and re-centers them based on the point predictions of MinT_{shr}. Next, the re-centered samples are permuted according to the empirical copula strategy of Gijbels and Herrmann (2014). Finally, each draw of the re-centered and permuted samples is summed according to the structure of the hierarchy.

- Reconciliation techniques for hierarchies summing over time intervals. Jeon et al. (2019) propose a modular approach for probabilistic prediction of hierarchical time series over different time intervals. This proceeds in three steps: First, a probabilistic forecast is made independently for every series in the hierarchy and sampled. These samples are then all either left as they are (stacked), sorted in increasing order (ranked), or randomly permuted

(shuffled). Finally, each draw of the resulting samples is made coherent using one of the following three techniques.

- Str: Structural reconciliation (Athanasopoulos et al., 2017) follows a procedure similar to MinT_{shr} but based on a diagonal covariance matrix where the diagonal entries represent the numbers of bottom-level series needed to construct each series in the hierarchy. For instance, if the hierarchy consists of monthly, quarterly, and annual values, then the diagonal entries are 1, 4, and 12, respectively.
- Average: This first uses the predictions of all series in the hierarchy to form implied predictions of the bottom-level series, averages those implied predictions, and then sums these averages according to the structure of the hierarchy. For instance, with monthly, quarterly, and annual values, the revised forecast is the average of the monthly forecast, the associated quarter's forecast divided by four, and the associated year's forecast divided by 12.
- CV: This is like Average in the previous point but, instead of averaging, it uses a general linear combination. The linear coefficients are obtained via cross-validation and can be unconstrained, constrained to be positive, or constrained to respect a convex combination.

Given that the focus of this article is on probabilistic forecasting, the main text does not include PPMs that have been previously used only for point forecasting of hierarchical time series. For a discussion of such methods and how our JBU compares against them in terms of point forecasting accuracy, see Supplementary Materials S3-S6.

5 Simulation Study

5.1 Data Generation

In this section, we compare the competing methods (see Section 4) on simulated data that satisfy our Assumptions 1-3. The purpose is to present a controlled environment in which we can illustrate the correctness of our implementation, further explore the theory presented in Section 3, and understand how cross- and contemporaneous dependance affects our model and its alternatives. The following Sections 6 and 7 then complement this evaluation with empirical tests of forecast accuracy in two real-world applications where our modeling assumptions are unlikely to represent the data generating process exactly.

In this section our data generating process introduces heterogeneity in the dependence structure of the bottom-level series and allows us to discuss the behavior of the competing methods in a variety of contexts that can appear in practice. Specifically, the data are organized into the small

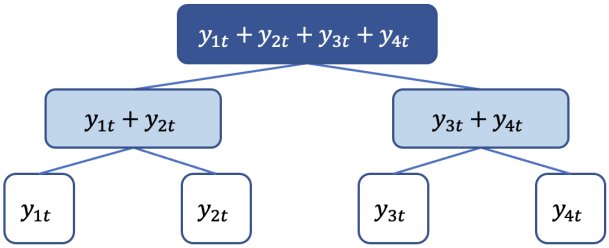

Figure 1: The hierarchy used in our simulation study. There are four bottom-level series (y_{1t}, y_{2t}, y_{3t}, and y_{4t}) that are summed by pairs into two series ($y_{1t} + y_{2t}$ and $y_{3t} + y_{4t}$) that, in turn, are summed into the grand total ($y_{1t} + y_{2t} + y_{3t} + y_{4t}$).

but non-trivial hierarchy illustrated in Figure 1. This hierarchy sums over time series: It has four bottom-level series (y_{1t}, y_{2t}, y_{3t}, and y_{4t}) that are summed first by pairs into two series ($y_{1t} + y_{2t}$ and $y_{3t} + y_{4t}$) and then into a single grand total ($y_{1t} + y_{2t} + y_{3t} + y_{4t}$). Similarly to Wickramasuriya et al. (2019), we generate the bottom-level series using a vector autoregression process with coefficients 0.4. This coefficient value preserves stationarity and introduces a moderate level of lagged dependence. The bottom-level series differ in their configurations of self- and cross-lags:

$$
\begin{aligned}
y_{1t} &= 0.4\, y_{1(t-1)} + 0.4\, y_{1(t-4)} + \epsilon_{1t}, \\
y_{2t} &= 0.4\, y_{2(t-1)} + 0.4\, y_{1(t-4)} + \epsilon_{2t}, \\
y_{3t} &= 0.4\, y_{3(t-1)} + 0.4\, y_{1(t-4)} + \epsilon_{3t}, \text{ and} \\
y_{4t} &= 0.4\, y_{3(t-1)} + 0.4\, y_{1(t-4)} + \epsilon_{4t}.
\end{aligned}
$$

The series y_{1t} depends on self-lags only and hence represents self-lagged dependence. The series y_{2t} and y_{3t} depend on one self-lag and one common cross-lag and hence represent mixed-lagged dependence. Finally, the series y_{4t} depends on cross-lags only and hence represents dependence that is only cross-lagged.

These series' contemporaneous dependence is given by the errors $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \epsilon_{2t}, \epsilon_{3t}, \epsilon_{4t})'$ that follow a multivariate normal distribution with mean zero and contemporaneous covariance matrix (CCM) $\boldsymbol{\Sigma}$. The diagonal values of the CCM are chosen uniformly at random from the set $\{0.6, 0.8, 1.0, 1.2, 1.4\}$. The off-diagonal entries are fixed to yield correlation 0.3 between any pair of error terms. As a result, the covariances range between 0.18 and 0.42. The value 0.3 is chosen because it introduces non-negligible contemporaneous dependence, while remaining a plausible real-world value. For comparison, the time series in our Australian domestic tourism (Section 6) and product sales at Walmart (Section 7) datasets have (absolute) average correlations of around 0.10 and 0.36, respectively.

Following this process, we generate $10,000$ synthetic datasets. For each dataset, we train all

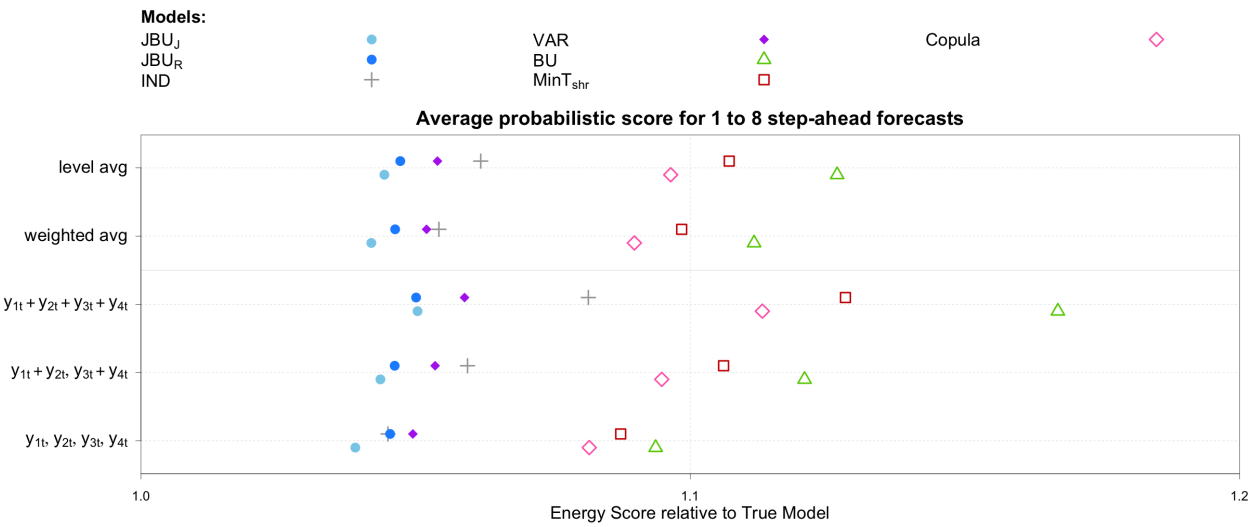

Figure 2: Forecast accuracy on simulated data. Scores are plotted relative to the score of an oracle that uses the true data generating process. Slight vertical jitter has been added to make overlaps visible.

competing methods on the first 100 training observations and then evaluate their (out-of-sample) forecasts up to 8 steps ahead. All competing methods are assumed to know that the true data generating process is autoregressive with maximum lag of order 4. Methods without cross-lags, namely previously proposed methods BU, MinT$_{shr}$, and Copula, model each individual time series with the AR(4) (i.e., autoregression with four self-lag terms) process. JBU, IND, and VAR model the time series using all self- and cross-lags up to order 4, yielding 16 predictors per bottom-level series. Given that the dimensionality is small, no Dirac filtering is performed. This allows us to focus on the performance of our Bayesian base model.

5.2 Accuracy Comparisons

Figure 2 summarizes probabilistic accuracy of all competing methods. The scores show relative improvement over an oracle that knows the true data generating process. For instance, a relative score of 1.1 means that the accuracy of the method is 10% worse than that of the oracle. The bottom three rows represent the probabilistic scores at each of the three levels of the hierarchy in Figure 1. Using the geometric mean, the relative scores are averaged over all simulated datasets and prediction horizons. Given that the relative accuracy of the competing methods is consistent across horizons, we do not present scores separately for each horizon. The top two rows present summary accuracy metrics. The top row *level avg* is the geometric mean of the probabilistic scores at each level of the hierarchy, i.e. of the bottom three rows. This gives equal weight to each level, regardless of the number of time series in it. The second row *weighted avg* weights each level-specific score by the number of time series in that level. This places more weight on the lower

levels with more individual series.

Among all competing methods, our JBU achieves the highest accuracy at each row in Figure 2. The Jeffreys-inspired prior on the CCM leads to slightly better accuracy than the RATS-inspired prior, but this difference is small. The second and third most accurate methods are VAR and IND, respectively. These are followed rather far behind by the PPMs. Given that only JBU, IND and VAR can incorporate cross lags, this large performance gap illustrates the benefits of modeling lagged dependence. The benefits of modeling contemporaneous dependence are in turn illustrated by the performance gap between JBU and IND. This gap increases at the upper levels of the hierarchy, highlighting the importance of CCM in probabilistic forecasting of hierarchical time series. Interestingly, MinT_{shr} and Copula improve over BU even at the bottom level, suggesting that these methods manage to indirectly integrate some part of the lagged or contemporaneous dependence in these series through their sums. The extent to which they manage to do so, however, is unclear and reduced compared to JBU, which considers both lagged and contemporaneous dependence explicitly. Finally, even in this small problem, both JBU and VAR already involve $16 \times 4 = 64$ parameters for the coefficients and 10 for the CCM. Estimating such models on the relatively low number of 100 training observations benefits from regularization, as is illustrated by the performance gap between JBU and VAR. We expect this gap to widen as the dimensionality of the problem increases and regularization becomes even more important.

In Supplementary Material S4, we repeat this study over a larger 2-by-2 design that additionally considers a diagonal CCM and a configuration where y_{3t} also depends only on self-lags. We also compare JBU against PPMs with base models based on exponential smoothing instead of autoregression. In short, the JBU continues to outperform its competition, and the results therein remain qualitatively similar to the ones presented in this section. Importantly, even if the PPMs are correctly specified for 50% of the series, failing to control for cross-lags continues to significantly impair their predictive accuracy. This relative loss in accuracy worsens at upper levels of the hierarchy, even if the CCM is diagonal.

6 Summing Over Time Series: Domestic Tourism in Australia

6.1 Data Description

In this section, we evaluate the potential of our JBU method to forecast future domestic tourism in Australia.[9] The data consist of monthly observations from January 1998 to December 2019 and have become a benchmark in hierarchical forecasting (Hyndman et al., 2011; Wickramasuriya et al., 2019; Hollyman et al., 2021). These data were collected by Tourism Research Australia

[9]All data analyzed in this paper are available from the corresponding author upon request. For more information about the data, see Tourism Research Australia (2015).

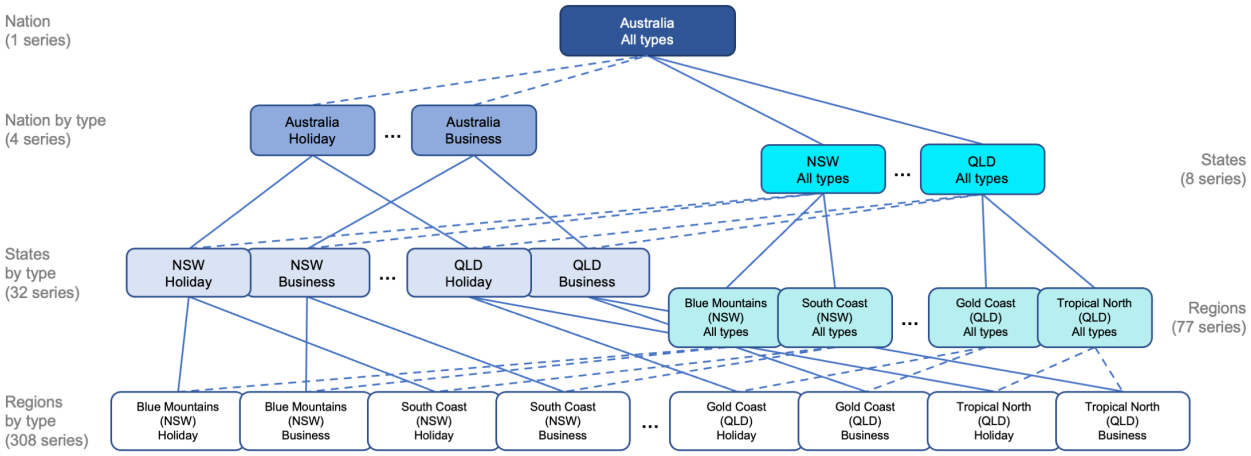

Figure 3: Illustration of the hierarchy in the Australian domestic tourism application. For clarity, the illustration only explicitly displays two types of tourism for two states each with two regions. Depending on at what level in the hierarchy summation is done over type of tourism, there are three paths of summation from the bottom level to the top level. Full (dashed) lines indicate summation by geography (type). The hierarchy has six levels to be predicted. Abbreviations: NSW = New South Wales, QLD = Queensland.

(TRA) and measure tourism in terms of "visitor nights", i.e., the nights spent away from home by a visitor. TRA splits the data based on a geographical partition of Australia with the *nation* divided into 8 *states* or *territories* which are further divided into a total of 77 *regions*. TRA also groups these data by four *types* of tourism: business, holiday, visiting family and friends, and other. This all means that, at the bottom level, a time series represents a type of tourism in a region (regions-by-type). There are 308 such series, each with 264 observations. The bottom-level time series can then be summed by geography or by type. In particular, by geography, we can sum the 308 bottom-level region-by-type series within their respective states to yield 32 states-by-type series that can then be, in turn, summed to 4 nation-by-type series. Any of these geographical levels can be collapsed over type, leading to series of total tourism of any type at the regional (77 series), state-wide (8 series), and national (1 series) levels. This yields a hierarchy where there are six different levels and multiple ways to sum the bottom-level series to the grant total on top. Figure 3 illustrates the hierarchy and Figure 4 illustrates the temporal dynamics of our data with a few selected time series at each unique level of the hierarchy.

6.2 Accuracy Comparisons

We compare different variants of our JBU against PPMs, namely BU, MinT_{shr}, and Copula (see Section 4), designed for hierarchies summing over time series. All competing methods are trained on a rolling window of 96 observations (8 years of monthly observations) and asked to forecast out-of-sample up to 12 months ahead. With a total of 264 time points, we have 157 complete rolling windows. Given that there are 308 bottom-level series, it is not feasible to consider the

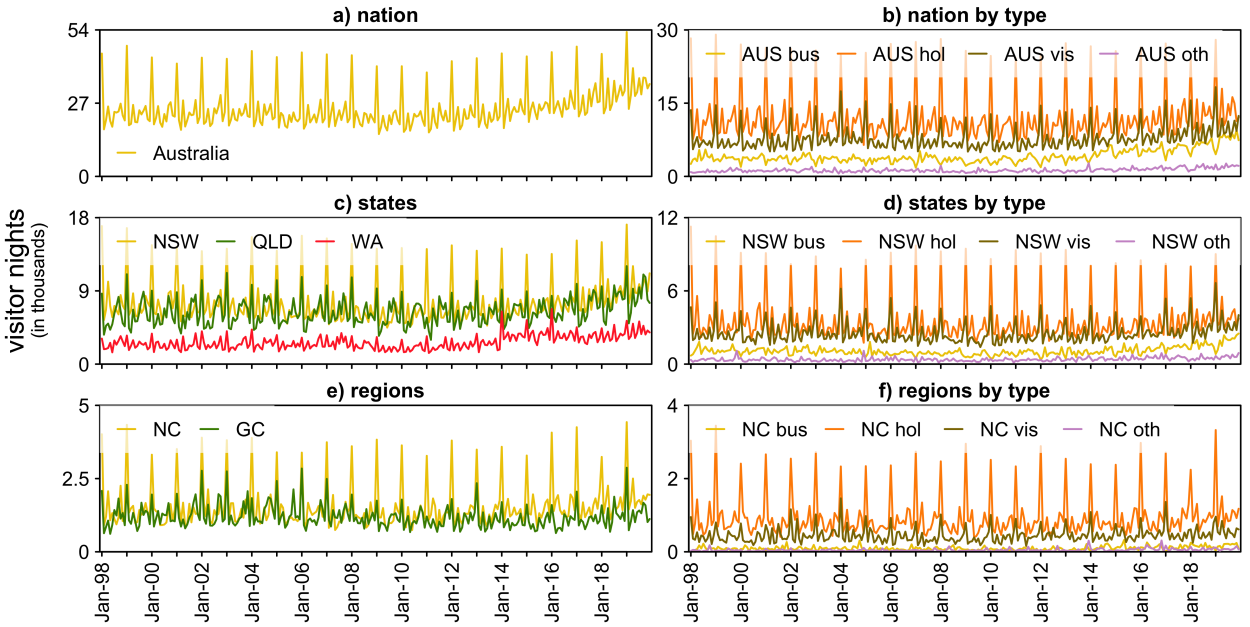

Figure 4: Evolution of Australian domestic tourism from January 1998 to December 2019, summed by geography and type of tourism. Abbreviations: bus = business, hol = holiday, vis = visiting family and friends, oth = other, AUS = Australia, NSW = New South Wales, QLD = Queensland, WA = Western Australia, NC = North Coast of NSW, and GC = Gold Coast of Queensland.

entire lagged history in JBU, IND, and VAR. Instead, we initially consider self- and cross-lags up to order 12 and then use our Dirac filter to select the most promising $K = 12$ or $K = 24$ terms for each bottom-level series. The base forecasts for the PPMs are given by three models: a 12-lag autoregression AR(12) aligning with the annual seasonality of the data (see Figure 4), ARIMA, and exponential smoothing (ETS). The specifications of ARIMA and ETS are found with the automatic selection procedure of Hyndman and Khandakar (2008).

All these combinations result in a total of 17 competing methods. Instead of presenting the scores of all methods at once, we improve the clarity of our discussion by presenting only the scores of the better performing methods in the main text and defer the rest to Supplementary Material S5. Specifically, JBU_{J} and VAR_{24} are omitted because they are slightly outperformed by JBU_{R} and VAR_{12}, respectively. In addition, we only present PPMs based on AR(12) because this base model choice gave the PPMs with the best average performance.

Figure 5 presents the probabilistic accuracy scores averaged over all forecast horizons and rolling windows. The structure of Figure 5 is similar to that of Figure 2 with the exception that now we have six sets of time series and that the scores have been normalized to the scores of IND_{12}. A relative score of, e.g., 0.9 or 1.1 then means that the average accuracy of the method is 10% better or worse, respectively, than that of IND_{12}.

The best score is achieved by a variant of our JBU method across all levels in Figure 5. Our 12-predictor version $\text{JBU}_{\text{R},12}$ marginally outperforms our 24-predictor version $\text{JBU}_{\text{R},24}$ at the bottom

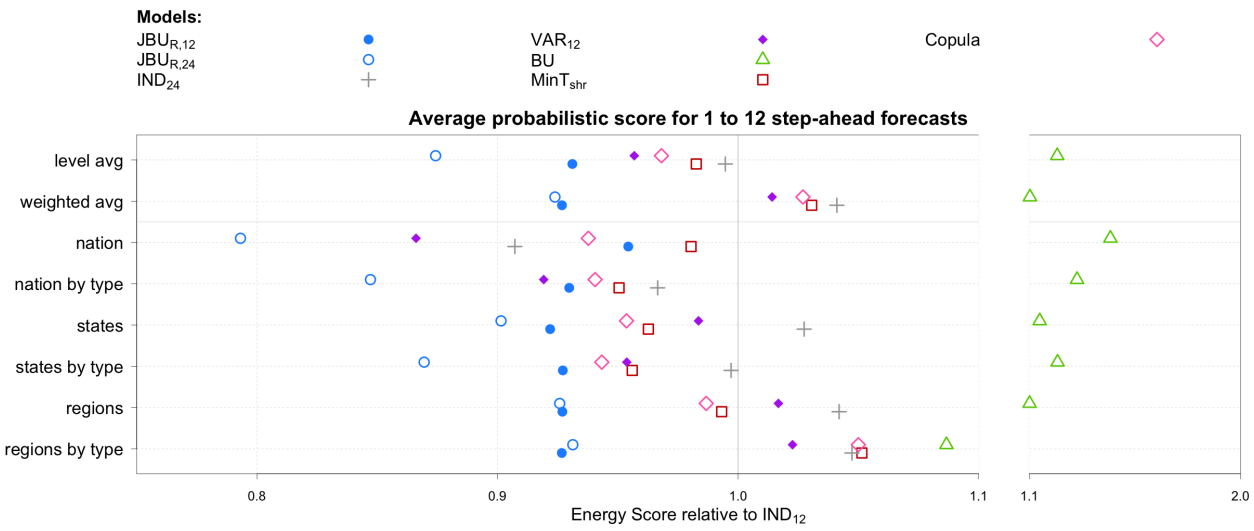

Figure 5: Probabilistic forecasting accuracy of Australian domestic tourism. Scores are shown relative to the score of the IND_{12} model. Slight vertical jitter has been added to make overlaps visible.

level but produces worse forecasts everywhere else. In fact, their accuracy scores cross and diverge at upper levels. This offers a compelling insight on the balance between model complexity in terms of number of predictors and the accuracy to capture inter-series dependence. Specifically, a model with fewer predictors is likely to be more robust against over-fitting and hence perform better in forecasting the bottom-level series that are often more volatile. However, controlling for a very small number of lag terms may not adequately address lagged correlation in the data. This can hinder the estimation of the joint distribution of the bottom-level series that is needed to propagate the bottom-level distributions to the upper levels (Proposition 1). Conversely, a model with an excessive number of predictors may be exposed to over-fitting. In JBU the potential for overfitting is mitigated, to some degree, by adaptive regularization. In practice, the user should find the right balance by comparing models with different numbers of predictors per bottom-level series K. In our case, we considered two possible values, and $K = 24$ appears preferable to $K = 12$.

Our $\text{JBU}_{\text{R, 24}}$ improves the accuracy of IND_{24} and VAR_{12} at all levels and typically by around 10%. In comparison to our simulation study in Section 5, the benefits of adaptive regularization and joint estimation of the CCM are now larger. The best and worst performing PPMs are Copula and BU, respectively. In Figure 5, our $\text{JBU}_{\text{R, 24}}$ improves the accuracy of PPMs by at least 6% at any level and by at least 10% in terms of *level avg*. In the case of the classical BU, the improvement at any level is at least 17%.

Further analysis can be found in the supplementary material. In particular, Supplementary Material S5 presents complete probabilistic results, including the accuracy scores of many competing methods that were omitted in the current section. Considering all combinations of PPMs and base

models, $\text{JBU}_{\text{R, 24}}$ improves the *level avg* of the PPMs by 32% on average. Supplementary Material S5.2 analyzes the competing methods in terms of point forecasting accuracy. The results remain qualitatively similar to their probabilistic counterparts presented in the current section. Finally, Supplementary Material S5.3 considers a hybrid approach that applies JBU at every level of the hierarchy and then reconciles the predictions of level-specific JBU models using MinT_{shr}. The results suggest that reconciling JBU tends to deteriorate its probabilistic accuracy but improve its point accuracy.

7 Summing Over Time Intervals: Product Sales at Walmart

7.1 Data Description

In this section, we use the M5 competition (Makridakis et al., 2022) data[10] to evaluate the potential of JBU to make temporally coherent forecasts. These data report daily sales of items in 7 product departments and 3 product categories, recorded at 10 Walmart stores in 3 states of the United States between January 2011 and June 2016. Similarly to our analysis of domestic tourism in Section 6, we let the bottom-level series be monthly values of per-day sales averaged over the days of the month and all stores.[11] This gives us 65 monthly observations per bottom-level series. To manage the computational burden of our analysis, we focus on the smallest department, *Hobbies 2*, with 149 products. This allows us to include many more variants of our JBU and the PPMs in the accuracy comparison. Figure 6 illustrates how the hierarchy sums each individual product's sales over different time intervals, and Figure 7 charts the sales of nine individual products at the different levels of the hierarchy.

7.2 Accuracy Comparisons

We compare different variants of our JBU against PPMs, namely BU, Average, Str, and CV (see Section 4), designed for hierarchies summing over time intervals. All competing methods are first trained on a rolling window of 48 observations (4 years of monthly observations) and then asked to forecast out-of-sample up to 6 months ahead. At this horizon, we can evaluate the accuracy of predictions of monthly, bimonthly, quarterly, and semesterly sales. With a total of 65 time points in the data, we have 12 complete rolling windows.

Similarly to Section 6, the base forecasts for the PPMs are produced either by AR(12), ARIMA,

[10]The competition was hosted by Kaggle. The data are publicly available and can be downloaded at `https://www.kaggle.com/competitions/m5-forecasting-accuracy/overview`.

[11]Using monthly data maintains a consistent level of temporal granularity across the two applications. Notably, the choice of temporal granularity is not dictated by computational constraints and we could have worked with any frequency (including daily). This is because the computational burden of the JBU is almost completely dictated by the number of predictors passed to the Bayesian base model. Therefore, so long as the number of filtered predictors remains unchanged, increased temporal granularity only increases the computational cost of the filtering step. This cost, however, grows linearly in the number of parameters and has a negligible effect on total runtime of our method.

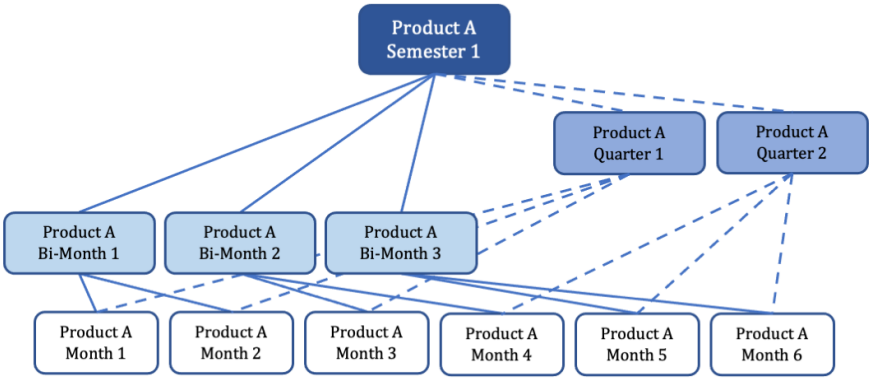

Figure 6: Illustration of the hierarchy for a generic "Product A" in the Walmart sales application. There are 149 products, and the hierarchy sums each product's monthly average per-day sales to form average per-day sales at the bi-monthly, quarterly, and semester level.

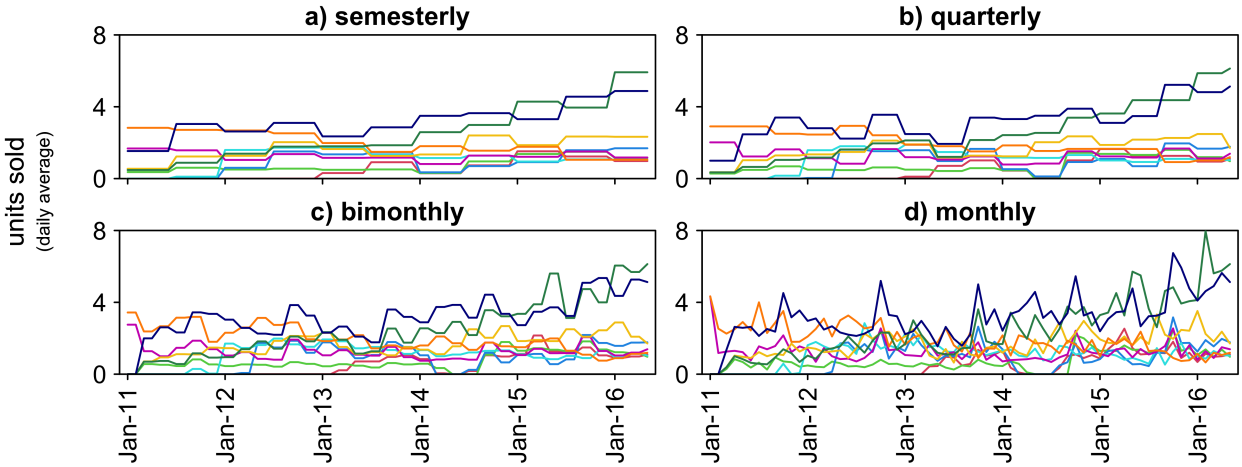

Figure 7: Evolution of the monthly average per-day unit sales of Hobbies 2 items at different time intervals. Only 9 series out of the total 149 are shown. These series were chosen because they correspond to the 10%, 20%, ..., 90% quantiles of the distribution of products by total units sold over the period of observation.

or ETS models. This results in a total of 48 PPMs.[12] The reconciliation techniques consider base predictions at the monthly, bi-monthly, quarterly, four-monthly, semesterly, and annual levels. The CV method uses the last 6 observations in each rolling window for validation. For JBU, IND, and VAR, we initially consider self- and cross-lags up to order 12 and then use our Dirac filter to select the most promising $K = 12$ or $K = 18$ terms for each bottom-level series. Once the lag terms are accounted for, these methods are trained on only 36 observations. This is less than half as much data as in our application to Australian domestic tourism (Section 6.2). Given that in Section 6.2 our JBU with the RATS-inspired prior performs better than with the Jeffreys-inspired prior, for

[12]Jeon et al. (2019) propose 3 sampling schemes and 3 techniques to enforce coherence (see Section 4). Given that the CV technique can use 3 different linear combinations, their method has a total of $3 \times 5 = 15$ different configurations. Adding BU to the list and considering 3 base models per method, we have $16 \times 3 = 48$ PPMs.

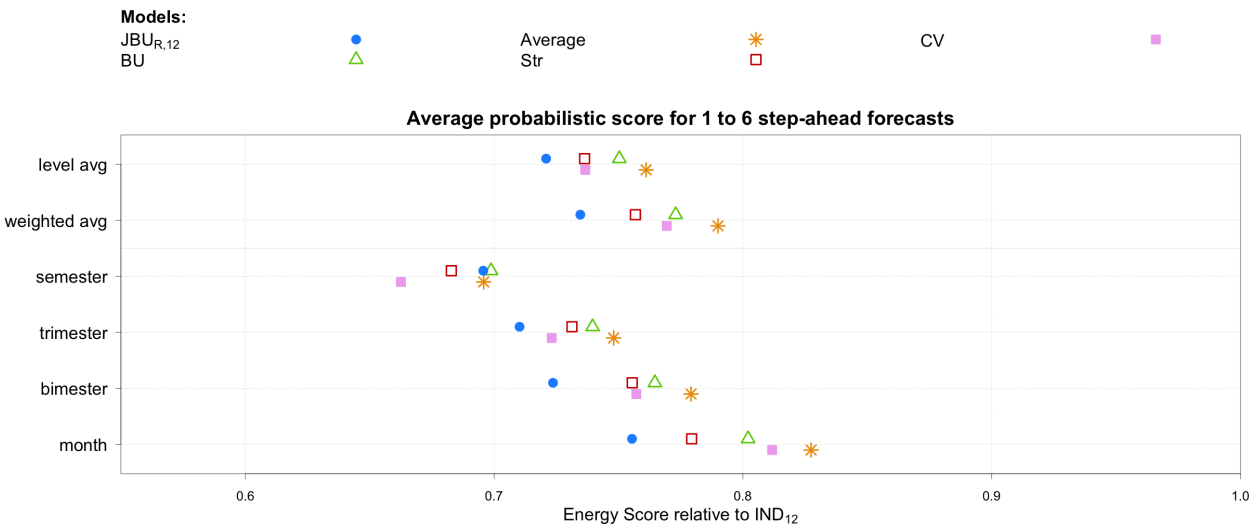

Figure 8: Probabilistic forecasting accuracy of product sales at Walmart. Scores are shown relative to the score of the IND$_{12}$ model. Slight vertical jitter has been added to make overlaps visible.

brevity we only consider the RATS-inspired prior in the current application.

Figure 8 presents accuracy scores at different time intevals. The scores are normalized to those of IND$_{12}$ and averaged over all forecast horizons and rolling windows. The labels are analogous to those in Figure 5. As before, we only display the scores of the best performing variants in the main text and defer the rest of the scores to Supplementary Material S6. Specifically, we only present here the PPMs with exponential smoothing (ETS) as the base model and stacking as the sample combination criterion. Among variants of our JBU, we only present JBU$_{R,12}$ and IND$_{12}$. Finally, VAR is omitted because it diverges and performs poorly on these data. Given that VAR is estimated based on the maximum likelihood principle, there may be too little data in this application to pin down parameter values based on the likelihood alone, but with the addition of proper priors in our Bayesian model the search space over parameters becomes sufficiently regularized to stabilize estimation and allow convergence (for a related discussion, see Gelman et al. 2008).

The results show that JBU produces the most accurate probabilistic forecasts at all levels, except at the top level (semester) where it is outperformed slightly by CV and Str. Recall that here JBU is compared against more than 50 competing methods. With so many competitors, it is likely that some of them outperform JBU by chance, especially at the top semester level where the out-of-sample scores are based on only 12 dependent observations (one per rolling window). However, the fact that JBU continues to perform well systematically at all levels is an indication of its robust performance. In comparison to our application to Australian domestic tourism, the improvements in Figure 8 are now more modest: at least 2% in terms of *level avg* and on average around 4% at any temporal level.

Interestingly, JBU improves the accuracy of IND$_{12}$ by 26%. This highlights the benefits of

separating contemporaneous and lagged dependence. Indeed, sales of multiple products are often subject to large contemporaneous shocks, and our current application is no exception: The absolute correlation is on average 0.36 and exceeds 0.5 for 30% of the pairs of time series. Given that IND only incorporates lagged dependence, it can end up explaining some contemporaneous variation falsely as lagged dependence. Such misinterpretation leads to over-estimation of lagged dependence and ultimately to poor centering of the probabilistic forecasts. This same problem is less visible in our applications to synthetic data (Section 5) and to Australian domestic tourism (Section 6) because there we have much more training data per bottom-level series or lower contemporaneous dependence, respectively.

As before, further analysis can be found in the supplementary material. In particular, Supplementary Material S6.1 presents the complete results including all the competing methods that were omitted in the current section. Considering all variants of PPMs, $\text{JBU}_{\text{R, 12}}$ improves *level avg* by 10% on average. Supplementary Material S6.2 considers point forecasting accuracy. In short, the PPMs tend to be more competitive in point forecasting, but their performance varies markedly depending on the chosen base model.

7.3 Joint Service Level

As mentioned in Section 1, probabilistic forecasts are often necessary in optimal decision making. To illustrate this, consider a firm that must choose the order quantities of multiple products with uncertain future demand. This is a classical problem in inventory management and is often solved with the multi-product newsvendor model that can help the firm to choose optimal order quantities. Before doing this, however, the firm must construct a probabilistic forecast of the future demand. If the firm operates under a risk-neutral performance measure, such as maximizing the expected future profits, the newsvendor problem can be solved based on independent probabilistic forecasts of each product. Firms, however, operate in finite time and hence must worry about the near-future variability of profits instead of focusing exclusively on long-term expected levels. In fact, managers can be risk-averse, especially for high-value products (Schweitzer and Cachon, 2000). If the performance measure is not risk-neutral, the newsvendor model must consider the products simultaneously as a portfolio (e.g., Choi et al. 2011). This requires a joint probabilistic forecast of the future demands of the products. Managing the inventory of multiple products then often begins with the challenging statistical task of estimating the joint distribution of the product demands.

This is precisely the problem that our JBU method aims to solve. To illustrate how our JBU method can support such multi-product inventory management decisions, we posit a firm that seeks to control its *joint service level* for the products considered in Section 7.2.[13] Specifically, the firm

[13]Demand can exceed sales when there are stockouts. In this study, however, we follow many previous authors (e.g. Osadchiy et al. 2016; Spiliotis et al. 2021) and treat sales as a direct proxy for demand.

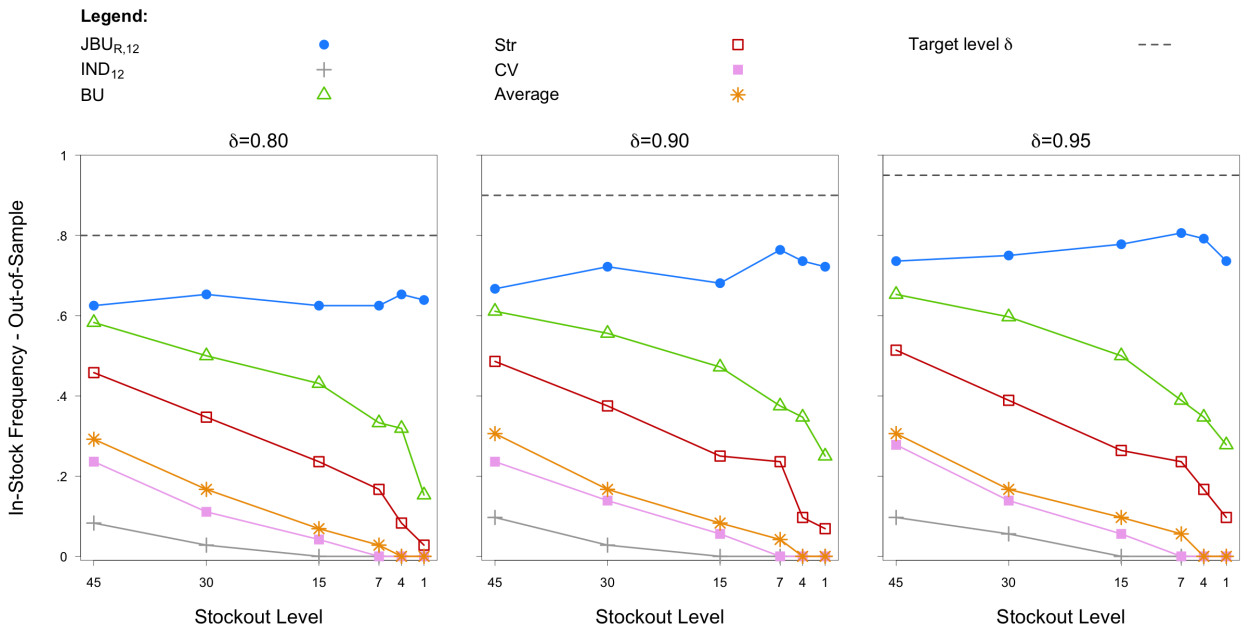

Figure 9: Setting joint service levels. The goal is to observe at most L stockouts with confidence δ. Any point below (above) the nominal confidence level indicates that the number of stockouts exceeds L too often (rarely, respectively).

wants to set order quantities such that, with a given confidence level $\delta \in [0, 1]$, the number of products simultaneously out of stock will not exceed some stockout level L. This way the firm seeks to control the risk of a widespread stock-out. Given a joint predictive distribution of product demand, the optimal order quantities are the smallest δ^*-quantiles of the marginal predictive distributions such that no more than L products are predicted to be simultaneously out of stock with probability larger than $1 - \delta$. Our analysis uses the same competing methods, time-series of the sales of the 149 products, train-to-test split of the dataset, rolling window approach, forecast horizon, and probabilistic forecasts as in Section 7.2.

Figure 9 compares the chosen order quantities to the realized (out-of-sample) outcomes and shows how frequently each competing method manages to keep the number of stockouts below the desired level L. The stockout level L varies over the x-axis and takes on values $45, 30, 15, 7, 4$, and 1 that correspond roughly to $30, 20, 10, 5, 2.5$, and 1 percent of the 149 products in our dataset, respectively. From left to right, the three panels represents nominal confidence levels δ of 0.80, 0.90, and 0.95, as indicated by the dashed horizontal lines. The nominal values represent the company's targeted in-stock frequencies. The closer the points of a method are to the horizontal lines, the better the method is in capturing the targeted joint service levels.

At the extreme case with $L = 1$ and $\delta = 0.95$, our joint service criterion allows more than one product out of stock only 5% of the time. To offer such high levels of service with such high level of confidence, the firm must act conservatively and increase its order sizes. Indeed, as L decreases

and δ increases, the chosen marginal quantile δ^* and the associated order quantities increase and shift further out in the right tail of the probabilistic forecast. For instance, for $L = 45$ ($L = 15$, $L = 1$) and $\delta = 0.8$ ($\delta = 0.9$, $\delta = 0.95$), δ^* is 0.726 (0.925, 0.998) averaged over all rolling windows and competing methods.

Estimating joint tail behavior is a notoriously difficult problem. This can explain why, in our study, all methods satisfy the joint service level criterion too infrequently, resulting in curves below the nominal levels. Among the competing methods, our JBU performs the best at all points in Figure 9. It satisfies the joint service criterion with frequency reasonably close to the nominal level at all values of L and δ. In contrast, all PPMs perform much worse, especially as L approaches 1. In fact, at small values of L, most of the PPMs completely fail to maintain the desired joint service level. They lead to too many stock outs, suggesting that their probabilistic forecasts are over-confident and their order quantities are often too low.

8 Discussion

Hierarchical time series emerge in a wide range of domains: tourism, electricity demand, product sales, crime occurrences, and many other areas. Forecasts of such series must be accurate, probabilistic, and coherent. Our joint bottom-up (JBU) method offers a generic tool for these applications. It is inspired by the result (Proposition 1) that the bottom-level series and their dependence structure are sufficient for accurately estimating a joint coherent model of all series in the hierarchy. This offers new perspective to the hierarchical forecasting literature, where most recent developments have either neglected inter-series dependence or have focused on incorporating the upper-level series and the hierarchy in the model. These methods typically involve two steps that first make predictions under a misspecified model and then use an appropriate transformation to ensure that the final predictions respect any coherence requirements. From a modeling perspective, we believe that a more direct procedure that models all series jointly within the natural constraints of the problem, i.e., coherence with respect to the summing structure, should perform better. Our results on a synthetic and two real-world datasets support this reasoning: JBU leverages dependence among the bottom-level series and improves the forecasting accuracy of the current state-of-the-art methods. Furthermore, given that our JBU method is a multivariate Bayesian approach that operates at the most granular bottom level, its forecasts are inherently both probabilistic and coherent.

Several empirical studies in the literature have illustrated how reconciliation can improve the accuracy of BU by incorporating the upper-level series and the structure of the hierarchy. This suggests that there is value in modeling the upper-level series, which appears to be at odds with our theory saying that hierarchical forecasting can safely focus on joint modeling of the bottom-level series. How can we rationalize this seeming contradiction? Indeed, the literature often explains

that reconciliation improves upon BU by incorporating additional "information in the hierarchy" (e.g., Athanasopoulos et al. 2017; Jeon et al. 2019; Wickramasuriya et al. 2019; Hollyman et al. 2021). To explain this statement, we must first recognize that the behavior of a sum can only be characterized fully based on the joint behavior of its components. BU models each series independently and does not capture their joint behavior, but reconciliation may do so – at least partially – based on the sums. If the benefits of reconciliation truly derive from this implicit incorporation of dependence, then reconciliation is unlikely to use any information beyond what is already used by JBU that models the very same dependence explicitly. This intuition aligns with our results in Supplementary Material S5.3, where we apply reconciliation to JBU and find that it can improve our point forecasts but not necessarily our probabilistic forecasts. A future research direction could explore this more rigorously by developing theory to explain when reconciliation helps, why, and how, and, based on that, possibly uncover the mechanism by which reconciliation can contribute beyond a joint model of the bottom-level series.

A key strength of the popular reconciliation techniques is that they operate on base forecasts that can stem from different sources. For instance, some base forecasts can be made by ARIMA models, others by AR models, and some may be judgments made by human experts. Our JBU can also leverage such external forecasts. We just need to input them as predictors in the model. We can even complement those predictors with the (default) lagged (self- and cross) terms. JBU will then perform adaptive regularization and selection among all such predictors. In this way, our JBU can coalesce diverse sources of information, including lagged terms, external forecasts, and any other information that the user can input as predictors, determine their predictive relevance, and use them to generate inherently coherent probabilistic forecasts of all series in the hierarchy.

In addition to being accurate, coherent, and probabilistic, the predictions of our JBU model are also highly interpretable which can facilitate its widespread adoption among different decision makers in the same industry (Ribeiro et al., 2016). Unfortunately, in statistical forecasting and machine learning, the most accurate models are often complicated "black boxes" that lack transparency and are difficult to interpret (Lipton, 2018). Even understanding the origins of the reconciled predictions can be challenging as it requires a certain level of familiarity with projections. Our JBU method, however, is accurate and can be interpreted as easily as a collection of linear models. For instance, in our application to Australian domestic tourism, the current tourism count is, in expectation, equal to a linear combination of past tourism counts with known temporal lags and locations. In this sense, the JBU method offers a decision making tool that is accessible even to the less technical audience.

The benefits of JBU, however, do not come without a cost. Compared to alternative approaches to hierarchical forecasting, joint estimation of the bottom-level models with self- and cross-lags is computationally more demanding. For instance, in our applications, many of the PPMs can be es-

timated within minutes, whereas estimating our JBU can take hours. Even though the filtering step can reduce the computational burden as necessary, it can be challenging to apply our JBU method in some of the larger hierarchical forecasting problems. For instance, in supply chain management there can be thousands of individual stock keeping units (Oliveira and Ramos, 2019), and in the analysis of the national power grid there can be tens of thousands of individual households (Taieb et al., 2017b). In such cases, partitioning the bottom-level series, assuming independence between series in different parts, and applying the joint bottom-up method separately but in parallel to each part can offer a partial solution.

To conclude, our work seeks to bring fresh perspective to the hierarchical forecasting literature that has largely focused on "vertical" improvements to BU based on the series at the upper levels of the hierarchy. In this article we show that BU can also benefit from principled "horizontal" improvements based on cross-series interactions at the most granular bottom level. Our JBU method is broadly applicable, easy to interpret, involves minimal user intervention, and can produce accurate joint probabilistic forecasts of all series in the hierarchy. We hope that these properties encourage practitioners to use JBU and that our work can pave the way for new research in hierarchical forecasting methodology.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R. J., and Affan, M. (2020). Hierarchical forecasting. In Fuleky, P., editor, *Macroeconomic Forecasting in the Era of Big Data*, pages 689–719. Springer International Publishing.

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

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10:1281–1311.

Cao, Z., Li, G., and Song, H. (2017). Modelling the interdependence of tourism demand: The global vector autoregressive approach. *Annals of Tourism Research*, 67:1–13.

Chen, Z. and Zhao, L. (2023). Constructing quantiles via forecast errors: Theory and empirical evidence. *Available at SSRN 4371538*. February 27, 2023.

Choi, S., Ruszczyński, A., and Zhao, Y. (2011). A multiproduct risk-averse newsvendor with law-invariant coherent measures of risk. *Operations Research*, 59(2):346–364.

Di Fonzo, T. and Girolimetto, D. (2021). Forecast combination based forecast reconciliation: insights and extensions. *arXiv preprint arXiv:2106.05653*.

Dunn, D., Williams, W., and DeChaine, T. (1976). Aggregate versus subaggregate models in local area forecasting. *Journal of the American Statistical Association*, 71(353):68–71.

Eckert, F., Hyndman, R. J., and Panagiotelis, A. (2021). Forecasting Swiss exports using Bayesian forecast reconciliation. *European Journal of Operational Research*, 291(2):693–710.

Enders, W. (1996). *RATS Handbook: Handbook for Econometric Time Series*. Wiley.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(5):849.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gamakumara, P., Panagiotelis, A., Athanasopoulos, G., Hyndman, R. J., et al. (2018). Probabilistic forecasts in hierarchical time series. Technical report, Monash University, Department of Econometrics and Business Statistics.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gijbels, I. and Herrmann, K. (2014). On the distribution of sums of random variables with copula-induced dependence. *Insurance: Mathematics and Economics*, 59:27–44.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hamilton, J. D. (1994). *Time series analysis*. Princeton university press.

Hilden, J. and Gerds, T. A. (2014). A note on the evaluation of novel biomarkers: Do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine*, 33(19):3405–3414.

Hollyman, R., Petropoulos, F., and Tipping, M. E. (2021). Understanding forecast reconciliation. *European Journal of Operational Research*, 294(1):149–160.

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

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589.

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association*, 98(462):438–455.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Ishwaran, H. and Rao, J. S. (2011). Consistency of spike and slab regression. *Statistics & Probability Letters*, 81(12):1920–1928.

Ishwaran, H. and Rao, J. S. (2014). Geometry and properties of generalized ridge regression in high dimensions. *Contemporary Mathematics*, 622:81–93.

Jeon, J., Panagiotelis, A., and Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research*, 279(2):364–379.

Kadiyala, K. R. and Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, 12(2):99–132.

Li, H. and Tang, Q. (2019). Analyzing mortality bond indexes via hierarchical forecast reconciliation. *ASTIN Bulletin: The Journal of the IAA*, 49(3):823–846.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Madsen, H. (2007). *Time series analysis*. CRC Press.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022). The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4):1325–1336.

Malsiner-Walli, G. and Wagner, H. (2018). Comparing spike and slab priors for Bayesian variable selection. arXiv preprint arXiv:1812.07259.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models*. Routledge.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Ni, S. and Sun, D. (2005). Bayesian estimates for vector autoregressive models. *Journal of Business & Economic Statistics*, 23(1):105–117.

Novak, J., McGarvie, S., and Garcia, B. E. (2017). A Bayesian model for forecasting hierarchically structured time series. arXiv preprint arXiv:1711.04738.

Oliveira, J. M. and Ramos, P. (2019). Assessing the performance of hierarchical forecasting methods on the retail sector. *Entropy*, 21(4):436.

Osadchiy, N., Gaur, V., and Seshadri, S. (2016). Systematic risk in supply chain networks. *Management Science*, 62(6):1755–1777.

Pennings, C. L. and van Dalen, J. (2017). Integrated hierarchical forecasting. *European Journal of Operational Research*, 263(2):412–418.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):1–30.

Schweitzer, M. E. and Cachon, G. P. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management science*, 46(3):404–420.

Sims, C. A. and Zha, T. (2006). Does monetary policy generate recessions? *Macroeconomic Dynamics*, 10(2):231–272.

Spiliotis, E., Makridakis, S., Kaltsounis, A., and Assimakopoulos, V. (2021). Product sales probabilistic forecasting: An empirical evaluation using the m5 competition data. *International Journal of Production Economics*, 240:108237.

Staël von Holstein, C.-A. S. (1970). A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology and Climatology*, 9(3):360–364.

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

Taieb, S. B., Taylor, J. W., and Hyndman, R. J. (2021). Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*, 116(533):27–43.

Taieb, S. B., Yu, J., Barreto, M., and Rajagopal, R. (2017b). Regularization in hierarchical time series forecasting with application to electricity smart meter data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1):4474–4480.

Tourism Research Australia (2015). Tourism forecasts. Technical report, Australian Goverment, Canberra, Australia.

Vandeput, N. (2021). *Data science for supply chain forecasting*. Walter de Gruyter GmbH & Co KG.

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

Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science*, 27(4):479–488.