Probabilistic Forecasts in Hierarchical Time Series

Puwasala Gamakumara

December 18, 2017

Supervised by: Prof. Rob J. Hyndman
Associate Prof. George Athanasopoulos,
Dr. Anastasios Panagiotelis

1 Introduction

Many research applications often involve a large collection of multiple time series some of which are aggregates of others. These collection of time series are called hierarchical time series. For example, electricity demand of a country can be disaggregated into a geographical hierarchy. One would be interested in forecasting electricity demand of the whole country along with the demand of states, cities and households. This is referred to as hierarchical forecasting. An important aspect of hierarchical forecasting is the aggregate consistency of forecasts across the hierarchy; that is, forecasts at lower levels should add up to forecasts at higher levels of aggregation so that these reflects the properties of the actual data.

Various approaches have been used in literature to produce aggregate consistent forecasts of hierarchical time series incorporating valuable, structural information of the hierarchy. The traditional approaches involve bottom-up, top-down and middle-out methods. In the bottom-up approach, forecasts of the lowest level are generated and these are simply aggregated to forecast upper levels of the hierarchy (Dunn, Williams, and DeChaine, 1976). In contrast, the top-down approach involves forecasting the most aggregated series first and then disaggregating these forecasts down the hierarchy. Usually, this disaggregation is done based on the proportions of observed data, which is referred to as historical proportions. Gross and Sohl (1990) provide a comprehensive summary on using historical proportions in this context. A compromise between these two approaches is the middle-out method which entails forecasting each series of a selected middle level in the hierarchy and then forecasting upper levels by the bottom-up method and lower levels by the top-down method.

Hyndman et al. (2011) proposed an alternative way to produce aggregate consistent forecasts, which they referred to as optimal reconciliation approach. In this approach, the independent forecasts of all series at all levels of the hierarchy are optimally combined using a regression model to get the reconciled forecasts. This approach was further improved by Wickramasuriya, Athanasopoulos, and Hyndman (2015), by incorporating the information from the full covariance matrix of forecast errors. The most important feature of their method that differs from other reconciliation methods is that it considers the correlation structure of the whole hierarchy.

The main purpose of time series forecasting is to provide reliable conclusions about the future quantities depending on the historical values. Recently, forecasters have moved

from point forecasts to probabilistic forecasts so that they could provide better forecasts with the associated uncertainty (Gneiting et al., 2008). There is a rich literature on probabilistic forecasts of univariate and multivariate time series. However, in hierarchical time series forecasting, so far the attention has only been given to point forecasting and no published literature can be found in the context of probabilistic forecasting.

When generating the probabilistic forecasts of hierarchical time series, particular attention should be given to preserving certain inherent properties of hierarchical nature. For example, they should be aggregate consistent and have an equivalent correlation structure as the realizations of the hierarchy. Therefore the main purpose of my research is to develop techniques that optimally estimate the probabilistic forecasts of hierarchical time series. This will be achieved in two frameworks, namely, in the parametric Gaussian framework and the non-parametric framework.

The rest of the report can be outlined as follows. The next section will review related literature on forecasting hierarchical time series. Section 3 will discuss the theory on existing techniques and methodologies to construct probabilistic forecasts of hierarchical time series in Gaussian and non-parametric frameworks. Section 4 will present a simulation study followed by discussion and future works in Section 5.

2 Literature Review

This chapter will review the necessary literature related to probabilistic forecasting in hierarchical time series. First, it will review the literature on point forecasts of hierarchical time series. Then it will discuss methodologies for obtaining univariate and multivariate probabilistic forecasts followed by the methods used to evaluate their predictive performances.

2.1 Point forecasts of hierarchical time series

In hierarchical forecasting, one can build individual models at all levels and forecast each time series independently. Yet it is extremely rare that these forecasts will be aggregate consistent. In most applications, it is a necessary requirement that the forecasts add up to their respective upper levels, reflecting the aggregation structure of the realizations. Hence there is a need to use some techniques to make these point forecasts be aggregate consistent, so that they reflect the properties of hierarchical nature.

The Bottom-up, Top-down and Middle-out approaches have been widely used in earlier studies on hierarchical literature to produce aggregate consistent point forecasts. Even though these methods are easy to construct, they have their own weaknesses. For example, bottom up approach provides accurate forecasts only if the bottom level series of the hierarchy are accurately forecast. However, if the bottom level series are highly volatile or too noisy, they are challenging to forecast. Then the bottom-up approach would produce inaccurate point forecasts. On the other hand, the top-down approach provides good forecasts only if the historical proportions are properly calculated. In this context, Athanasopoulos, Ahmed, and Hyndman (2009) introduced a new top-down approach which disaggregates the top level forecasts according to the forecast proportions rather than historical proportions. They found that their method outperforms the conventional top-down approach. Moreover, many studies in the literature have broadly discussed the relative advantages and disadvantages of bottom-up and top-down methods and situations in which they would provide reliable forecasts (Fliedner, 2001; Kahn, 1998; Lapide,

1998; Schwarzkopf, Tersine, and Morris, 1988).

Hyndman et al. (2011) proposed a novel approach where they optimally combine the independent forecasts of each series at all levels through a regression model. The resulting forecasts from this approach is referred to as reconciled forecasts. They first independently forecast each series, to generate what they refer to as "base forecasts". The base forecasts are then modeled as the sum of the expected values of future series and the error term. The error term in this regression model is referred to as "reconciliation error" which occurs due to the aggregate inconsistency of base forecasts. They have shown that if the covariance matrix of the reconciliation errors is known, the generalized least squares (GLS) solution gives the minimum variance unbiased estimate for the expected values of bottom level series of the hierarchy. This solution will be then used to revise the forecasts so that they become aggregate consistent. Further assuming that the reconciliation error also satisfies the aggregation structure, they showed that the GLS solution coincides with the ordinary least square (OLS) solution. Moreover, Athanasopoulos, Ahmed, and Hyndman (2009) and Hyndman et al. (2011) have shown that reconciling the forecasts through this OLS solution provides better forecasts than the other traditional forecasting techniques.

Another important feature of hierarchical time series is that they often consist of thousands or millions of individual series and this imposes computational challenges in implementing any forecasting solutions. The optimal reconciliation method was generalized to handle these constraints and to scale to large hierarchies by Hyndman, Lee, and Wang (2016). Further, they proposed to use a weighted least squares (WLS) solution to the regression model by taking the weights as the diagonal matrix of the variance of reconciliation errors. If these variances are unknown, they suggest to use the variance of base forecast errors.

Wickramasuriya, Athanasopoulos, and Hyndman (2015) recently improved the work of Hyndman et al. (2011) and Hyndman, Lee, and Wang (2016), proposing a new approach for obtaining optimal reconciled point forecasts for hierarchical time series. They first showed that the variance covariance matrix of the reconciliation error in the regression model introduced by Hyndman et al. (2011) is not identifiable. Therefore they showed that the variance covariance matrix of the reconciliation error is not possible to be estimated and hence the GLS solution is unattainable. In addition to that, they find a unique analytical solution for the reconciled point forecasts which is unbiased, by minimizing the sum of the variances of reconciled forecast errors. An important feature of this method is that it imposes the correlation structure of the whole hierarchy to produce reconciled point forecasts. Their simulation study illustrates that this approach outperforms all existing hierarchical forecasting methods. They named this as MinT solution and further improved it enabling to obtain a computationally feasible solution.

Almost all existing studies on forecasting hierarchical time series concentrate on point forecasts. Only a few studies can be found on interval forecasts of hierarchical time series. For example, Shang and Hyndman (2016), and Shang (2016) used a method based on bootstrap errors of base forecasts to obtain aggregate consistent interval forecasts of hierarchical time series. There is no published literature on forecasting the entire probability distribution of the hierarchical time series.

2.2 Probabilistic forecasts

Many fields such as economics, medicine, finance, social sciences, meteorological sciences, depend on the prediction of future behavior, in their day to day applications. For these, providing a single predictive value for the uncertain future is not sufficient for reliable

decision making. This requirement motivated researchers to find the entire probability distribution of the future values so that it provides a full description of the uncertainty associated with the prediction (Abramson and Clemen (1995);Tay and Wallis (2000); Rossi (2014)). Earlier studies on estimating probabilistic forecasts involved survey based methods where the uncertainty of point forecasts was obtained from the survey respondents. Density forecasts of macroeconomic and financial data by Tay and Wallis (2000) is a case in point.

2.2.1 Univariate and Multivariate Probabilistic forecasting

Reviewing techniques of univariate and multivariate probabilistic forecasting is of great importance to this study. Mainly there are parametric approaches and non-parametric approaches for estimating probabilistic forecasts of time series.

Parametric approaches for density forecasting involves assuming a known parametric distribution for the errors of the conditional mean model and characterizing the predictive density with the conditional mean and the distribution of errors. For example, the conditional mean may be modeled using a suitable time series model and the errors may be assumed to have Gaussian (Rossi, 2014; Wijaya, Sinn, and Chen, 2015) or some other parametric distribution such as Beta (Bludszuweit, Domínguez-Navarro, and Llombart, 2008). Then density forecasting entails estimating the parameters of these predictive densities characterized by the conditional mean and distribution of errors, using appropriate techniques, such as maximum likelihood estimation (MLE). Alternatively, some applications involve using Bayesian techniques to obtain parametric predictive densities. For example Panagiotelis and Smith (2008), Clark (2011) and Huber (2016).

If it is not appropriate to assume a parametric distribution, then forecasters have used non-parametric approaches to obtain the probabilistic forecasts of univariate and multivariate time series. In the non-parametric approach, an appropriate model will first be fitted to the variable of interest and the point forecasts will be obtained. Then the associated probabilistic forecasts are estimated by incorporating in-sample errors of the fitted model. McSharry, Bouwman, and Bloemhof (2005) used this method to estimate probabilistic forecasts for peak electricity demand in terms of magnitude and timing. A slightly different approach was proposed by Laio, Ridolfi, and Tamea (2007). In that study, a local polynomial regression model was used to obtain point forecasts and the associated probabilistic forecasts were estimated by incorporating the global forecast errors. The global forecast errors were quantified as the sum of uncertainty associated with the regression coefficient estimation and the propagation of uncertainty in prediction. Moreover, in some applications they use non-parametric bootstrap methods to estimate the probabilistic forecasts of time series. These methods involve estimating the distribution of innovations using bootstrapped in-sample errors. For example, Manzan and Zerom (2008) in univariate time series and Vilar and Vilar (2013) in multivariate time series.

Probabilistic forecasts can also be obtained by estimating the set of quantiles of the future distribution. This involves estimating a set of quantiles for different probabilities using different forecast models and then combining them using some interpolation method. Taieb et al. (2016) proposed using a boosting procedure to estimate an additive quantile regression model for the set of quantiles of the future distribution in application to smart meter data. Pinson et al. (2009) also used quantile forecasts to obtain probabilistic forecasts of wind power generation. They further proposed an interesting method to jointly forecast the short-term (up to 2-3 forecast horizons ahead) probability distribution by incorporating the interdependency of forecast horizons. In order to estimate the interdependency

between prediction errors of different forecast horizons, the vector of short term quantile forecasts was transformed into a Gaussian random vector under the assumption of Gaussianity of errors. The mean and the covariance matrix of the transformed variable were then estimated assuming that it reflects the interdependency between forecast horizons. Finally, the Gaussian random vector was transformed back to the original random vector through the inverse cumulative function of quantile forecasts. This seems very useful in estimating probabilistic forecasts of more than one forecast horizons such that they reflect the interdependency between forecast horizons.

Ensemble forecasting is another implementation in probabilistic forecasting which is often used in weather prediction(Gneiting and Raftery, 2005). Ensemble forecasts comprise multiple runs of a numerical prediction model with slightly different initial conditions. Ensemble forecasts are considered as a random sample from the probability distribution of the future value of target variable. For example, Gel, Raftery, and Gneiting (2004) proposed a method to obtain multivariate probabilistic forecasts based on ensembles for mesoscale weather quantities. Since ensemble forecasts on their own are often biased (Pinson, 2012), certain statistical post-processing techniques are used to obtain unbiased and calibrated ensemble forecasts (Gneiting and Katzfuss, 2014; Gneiting et al., 2005). Also, these statistical post-process techniques enable us to understand the likelihood of the ensemble forecasts. For statistical post-processing techniques in ensemble forecasting, see Gneiting et al. (2005), Raftery et al. (2005), Berrocal, Raftery, and Gneiting (2008), Thorarinsdottir and Gneiting (2010), Pinson (2012), Schuhen, Thorarinsdottir, and Gneiting (2012) and McLean Sloughter, Gneiting, and Raftery (2013). In addition to ensemble forecasting, meteorological and statistical expertise often use Regime-Switching space-time (RST) models introduced by Gneiting et al. (2006), to obtain probabilistic forecasts of wind resources. Moreover, see Zhang, Wang, and Wang (2014) for a detailed discussion on recent developments of probabilistic forecasts in application to the wind power generation.

2.2.2 Modeling dependency with Copula

The interdependency between variables is traditionally described using the families of multivariate distributions. For example, multivariate normal, Gamma, Lognormal, Extreme-value distribution. However, these families of distributions require that the marginal distributions of variables also to be characterized by the univariate distribution of the same multivariate family. In many applications, it is required to find the joint distribution of variables with different marginal distributions. A remarkable solution to this is the concept of the copula. A copula allows modeling the dependence structure of random variables with any univariate distributions (Genest and Favre, 2007)

To understand the concept let us consider the simplest bivariate case. This can be easily extended to the multivariate case as well. Suppose the two variables are denoted by X_1 and X_2 with cumulative distribution functions, $F_1(x_1) = P(X_1 \leq x_1)$ and $F_2(x_2) = P(X_2 \leq x_2)$. For each pair of (x_1, x_2), there are three associated functionals, $F(x_1)$, $F_2(x_2)$ and $H(x_1, x_2)$ where each lie in the interval $[0, 1]$. Further each pair (x_1, x_2) leads to a point $[F(x_1), F_2(x_2)]$ in the domain $[0, 1] \times [0, 1]$ and this ordered pair corresponds to $H(x_1, x_2)$ in $[0, 1]$. The function which creates this correspondence is called copula. Sklar (1959) represented this useful concept in a theorem and is the foundation for many applications of copulas in statistics.

Sklars Theorem:

Let $H(x_1, x_2)$ be the joint distribution function of any pair (X_1, X_2) with marginal distributions $F_1(x_1)$ and $F_2(x_2)$. Then there exist a copula $C : [0,1]^2 \mapsto [0,1]$ such that,

$$H(x_1, x_2) = C[F_1(x_1), F_2(x_2)], \quad x_1, x_2 \in \mathbb{R}. \tag{2.1}$$

The function C uniquely characterizes the dependency structure between X_1 and X_2.

The multivariate extension of the Sklar's theorem can be given as follows. For a random vector $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ with corresponding cumulative distribution functions $F_1(X_1) = U_1, F_2(X_2) = U_2, ..., F_n(X_n) = U_n$, the joint distribution function $H(x_1, x_2, ..., x_n)$ is given by,

$$H(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) = C(u_1, u_2, ..., u_n), \tag{2.2}$$

where $x_i \in \mathbb{R}$ and $u_i \in [0,1]$ for all $i = 1, ..., n$.

For continuous, strictly increasing margins, the copula function can be extracted as (Joe, 1997),

$$C(u_1, u_2, ..., u_n) = H(F_1^{-1}(x_1), F_2^{-1}(x_2), ..., F_n^{-1}(x_n)), \tag{2.3}$$

where F_i^{-1} denotes the quantile function of the margins.

The importance of Sklar's theorem in statistical applications is that it allows us to pick a copula and model different univariate marginals together to get the joint multivariate distribution.

The copula functions can be classified into different families according to their characteristics. A comprehensive description of all the copula families and their theoretical aspects can be found in Joe (1997) and Nelsen (1999). A few of them will be briefly discussed below.

Gaussian and t Copula

The dependency structures of the Gaussian copula and t copula are implied by the well known Gaussian distribution and the student t-distribution respectively using (2.3). The Gaussian copula can be used to capture the joint symmetric dependency with non-normal marginals. The t-copula is useful for capturing the dependency of extreme values, such as the variables from leptokurtic distributions. More properties of the t-copula are broadly discussed in Demarta and McNeil (2005).

Archimedean copulas

Any copula C in the form of,

$$C(u_1, u_2, ..., u_n) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2) + ... + \varphi(u_n)) \tag{2.4}$$

is called an Archimedean copula. Here φ is called as the generator which is a continuous decreasing function from $[0,1]$ to $[0,\infty]$ such that $\varphi(1) = 0$. φ^{-1} is the pseudo-inverse of φ where $Dom(\varphi^{-1}) = [0,\infty]$ and $Ran(\varphi^{-1}) = [0,1]$. A full description of various generating functions of Archimedean copulas is given in Nelsen (1999). Clayton and Gumbel copula are two asymmetric copulas and can capture the negative tail dependence and positive tail

dependence of variables respectively (Joe (1997), Nelsen (1999)). Genest and Rivest (1993) have broadly examined how to select the appropriate Archimedean copula to capture the dependency structure in bivariate case.

2.2.3 Multivariate Probabilistic forecasting using Copulas

An important aspect of multivariate probabilistic forecasting is that it should reflect the joint dependency between the variables. A few studies on probabilistic forecasting of multivariate quantities under the Gaussian framework have been discussed in Section 2.2.1. However, it is challenging to capture the joint dependency in the multivariate forecast densities with different marginal distributions. The copula modeling gives a reasonable solution to this end since it is capable to capture the dependency between variables with different univariate marginal distributions.

A few studies on multivariate probabilistic forecasting using copulas could be found in the literature. For example, Möller, Lenkoski, and Thorarinsdottir (2013) used the Gaussian copula to capture the dependency between marginal predictive distributions of weather variables. They used the Bayesian model averaging (BMA) to produce the marginal predictive distributions and then used a Gaussian copula to capture the dependency between marginals and produced joint probability forecasts. Further, Wytock and Kolter (2013) used sparse Gaussian conditional random fields (SGCRF) to produce the probabilistic forecasts of multivariate time series. They also used Gaussian copula to transform the non-Gaussian variables into Gaussian variables to apply SGCRF and then transformed back each variable using inverse copula transformation.

2.3 Assessing Probabilistic Forecasts

The necessary final step in time series forecasting is to make sure that we have accurate forecasts about the uncertain future. When computing probabilistic forecasts, forecasters prefer to maximize the sharpness of the predictive distribution subject to the calibration (Gneiting and Katzfuss, 2014). Therefore the probabilistic forecasts should be evaluated with respect to these two properties.

Calibration refers to the statistical compatibility between probabilistic forecasts and realizations. In other words, random draws from well calibrated predictive distribution should be equivalent to the realizations. The uniformity in Probability Integral Transformation (PIT) of predictive densities evaluated at realizations implies well calibrated probabilistic forecasts. Therefore, assessing the uniformity in PIT of predictive densities, by inspecting their histograms and correlograms can be used as a simple tool for evaluating the calibration of probabilistic forecasts. In earlier studies, Diebold, Gunther, and Tay (1998) have used the PIT to evaluate univariate density forecasts whereas Diebold, Hahn, and Tay (1999), Clements and Smith (2002), Ko and Park (2013) used to evaluate multivariate density forecast. Alternatively, Gneiting et al. (2008) proposed a method based on Box density ordinate transform (BOT) to check the calibration of multivariate density forecasts.

On the other hand, sharpness refers to the spread or the concentration of the prediction distribution and it is a property of forecasts only. The more concentrated the predictive distribution, the sharper the forecasts are (Gneiting, Balabdaoui, and Raftery, 2007). The sharpness of density forecasts can be assessed in terms of the prediction intervals. Sharper density forecasts will have shorter prediction intervals. However, independently assessing the calibration and sharpness will not help to properly evaluate the probabilistic forecasts. Therefore these properties should be assessed simultaneously in probabilistic forecasts.

2.3.1 Scoring rules

Scoring rules are summary measures based on the relationship between predictive distribution and the realizations which allow us to jointly assess the calibration and sharpness. In some studies, researchers take the scoring rules to be positively oriented which they would wish to maximize, for example, Gneiting and Raftery (2007). However, in many other studies, scoring rules were defined to be negatively oriented which forecasters wish to minimize. We also consider these negatively oriented scoring rules to evaluate probabilistic forecasts in hierarchical time series.

Let $\boldsymbol{F}$ denotes the predictive distribution and $\boldsymbol{y} \in \mathbb{R}^{\mathrm{d}}$ denote d dimension vector of realizations. Scoring rule is a numerical value $S(\boldsymbol{F}, \boldsymbol{y})$ assign to each pair $(\boldsymbol{F}, \boldsymbol{y})$. Suppose the true distribution of $\boldsymbol{y}$ is given by $\boldsymbol{G}$. Then the proper scoring rule is defined as,

$$S(\boldsymbol{G}, \boldsymbol{G}) \leq S(\boldsymbol{F}, \boldsymbol{G}),$$

where $S(\boldsymbol{F}, \boldsymbol{G}) = E_{\boldsymbol{G}}[S(\boldsymbol{F}, \boldsymbol{y})]$ is the expected score under the true distribution $\boldsymbol{G}$ (Gneiting and Katzfuss, 2014; Gneiting et al., 2008). Following are few scoring rules which have been widely used to assess probabilistic forecasts in literature.

Continuous Ranked Probability Score (CRPS)

CRPS is defined in terms of predictive cumulative distribution function (CDF) for evaluating univariate probabilistic forecasts and given as,

$$CRPS(F, y) = E_F \left| X - y \right| - \frac{1}{2} E_F \left| X - X' \right|, \tag{2.5}$$

where $y \in \mathbb{R}$ and X and X' are independent random variables from the forecast distribution F with finite first moment (Gneiting and Raftery, 2007). It reduces to the absolute error if F is a point forecast and therefore CRPS is meaningful to use to compare probabilistic forecasts and point forecasts (Gneiting and Katzfuss, 2014).

Energy score

The multivariate generalization of CRPS is the energy score proposed by Gneiting et al. (2008) and is given by,

$$ES(\boldsymbol{F}, \boldsymbol{y}) = E_{\boldsymbol{F}} ||\boldsymbol{X} - \boldsymbol{y}|| - \frac{1}{2} E_{\boldsymbol{F}} ||\boldsymbol{X} - \boldsymbol{X'}||, \tag{2.6}$$

where $\boldsymbol{y} \in \mathbb{R}^{\mathrm{d}}$ is the vector of realizations, $\boldsymbol{X}$ and $\boldsymbol{X'}$ are independent d dimension random vectors from the multivariate forecast distribution $\boldsymbol{F}$ and $||.||$ denotes the Euclidean norm. In many cases it is difficult to find the closed form expression for $ES(\boldsymbol{F}, \boldsymbol{y})$ and hence the Monte Carlo methods will be employed. Gneiting et al. (2008) has further given the Monte Carlo approximation to the equation (2.6) as,

$$\hat{ES}(\boldsymbol{F}, \boldsymbol{y}) = \frac{1}{k} \sum_{j=1}^{k} ||\boldsymbol{x}_i - \boldsymbol{y}|| - \frac{1}{2(k-1)} \sum_{j=1}^{k} ||\boldsymbol{x}_j - \boldsymbol{x}_{j+1}||, \tag{2.7}$$

where $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$ is a simple random sample of size k (possibly large) from the predictive density $\boldsymbol{F}$.
However, Pinson and Tastu (2013) has shown that energy score has a very low discrimination ability for incorrectly specified covariances even though it discriminates well in

misspecified means.

David-Sebastiani score (DSS)

DSS is another proper scoring rule proposed by Dawid and Sebastiani (1999) which depends on the forecast density only through its first and second moments. For univariate forecast densities DSS is defined as,

$$DSS(F, y) = \frac{(y - \mu_F)^2}{\sigma_F^2} + 2log\sigma_F, \tag{2.8}$$

where μ_F and σ_F^2 are the mean and the variance of forecast densities. For the multivariate predictive densities it is defined as,

$$DSS(\boldsymbol{F}, \boldsymbol{y}) = log|\boldsymbol{\Sigma_F}| + (\boldsymbol{y} - \boldsymbol{\mu_F})'\boldsymbol{\Sigma_F^{-1}}(\boldsymbol{y} - \boldsymbol{\mu_F}), \tag{2.9}$$

where $|.|$ denotes the determinant of the matrix, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma_F}$ denotes the mean and the variance covariance matrix of the predictive density respectively.

The log score, quadratic score and the spherical score are some other widely used scoring rules to evaluate probabilistic forecasts. They are defined as,

$$LogS(\boldsymbol{F}, \boldsymbol{y}) = -log\boldsymbol{f}(\boldsymbol{y}), \tag{2.10}$$

$$QS(\boldsymbol{F}, \boldsymbol{y}) = -2\boldsymbol{f}(\boldsymbol{y}) + ||\boldsymbol{f}||^2, \tag{2.11}$$

$$SphS(\boldsymbol{F}, \boldsymbol{y}) = -\frac{\boldsymbol{f}(\boldsymbol{y})}{||\boldsymbol{f}||}, \tag{2.12}$$

where $\boldsymbol{f}(\boldsymbol{y})$ is the value of the density function $\boldsymbol{f}$ evaluated at $\boldsymbol{y}$, and $||\boldsymbol{f}||^2 = \int(\boldsymbol{f}(\boldsymbol{z}))^2 d\boldsymbol{z}$. Further note that DSS and Log score for evaluating Gaussian predictive distribution are equivalent.

In practice any scoring function is presented as the average over comparable sets of probabilistic forecasts (Gneiting et al., 2008). That is,

$$S_m(\boldsymbol{F}, \boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^{m} S(\boldsymbol{F}_i, \boldsymbol{y_i}), \tag{2.13}$$

where $S(\boldsymbol{F}_i, \boldsymbol{y_i})$ refers to the scoring measure of the i^{th} comparable set. Further, the mean score of different scoring rules can be used to evaluate competitive probabilistic forecasts. Diebold and Mariano (1995) introduced a statistical test which can be used to test the equivalent performance of two competitive probabilistic forecasts.

2.3.2 A statistical test for the equivalence of two competitive probabilistic forecasting methods

Suppose $\boldsymbol{F}$ and $\boldsymbol{G}$ are two competitive probabilistic forecasting approaches. A test of equivalent predictive performance will be based on the test statistic,

$$t_m = \sqrt{m}\frac{S_m(\boldsymbol{F}, \boldsymbol{y}) - S_m(\boldsymbol{G}, \boldsymbol{y})}{\hat{\sigma}_m}, \tag{2.14}$$

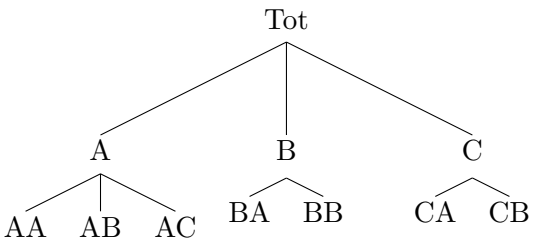

Figure 1: Two level hierarchical diagram

where $\hat{\sigma}_m = \frac{1}{m}\sum_{i=1}^{m}(S(\boldsymbol{F}_i, \boldsymbol{y}_i) - S(\boldsymbol{G}_i, \boldsymbol{y}_i))^2$ is the variance of the score differences and $S_m(\boldsymbol{F}, \boldsymbol{y})$ and $S_m(\boldsymbol{G}, \boldsymbol{y})$ follows from (2.14) with independent forecast cases. Diebold and Mariano (1995) showed that t_m follows a standard normal distribution under the null hypothesis of $\boldsymbol{F}$ and $\boldsymbol{G}$ having equivalent predictive performance. If t_m is significantly different from zero at 5% level of significance and have negative/(positive) value, then the probabilistic forecasts from $\boldsymbol{F}/(\boldsymbol{G})$ have better predictive performance than $\boldsymbol{G}/(\boldsymbol{F})$.

This test can be directly used to test the predictive performances of probabilistic forecasts in a single forecast horizon. If the interest is to evaluate the predictive performance in multiple forecast horizons, the same test statistic can be used with slightly modified variance component that allows having serial correlation between forecast horizons.

In conclusion, we have seen that univariate and multivariate probabilistic forecasting has a wide range of application in many fields. However, no published literature on probabilistic forecasting in hierarchical time series. Hence this study attempts to fill this gap by developing coherent techniques to produce probabilistic forecasts of hierarchical time series giving more attention to preserving aggregation consistency and the correlation structure across the hierarchy.

3 Probabilistic forecasts for hierarchical time series

3.1 Notation

We start with introducing notation on hierarchical forecasting, consistent with the notations used in past studies such as Hyndman et al. (2011) and Wickramasuriya, Athanasopoulos, and Hyndman (2015). Suppose $\boldsymbol{y}_t$ is a n-dimensional vector comprising all observations of the whole hierarchy at time t and $\boldsymbol{b}_t$ is a m-dimensional vector comprising only the bottom level observations at time t. Then due to the aggregation nature of the hierarchy we have,

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t, \tag{3.1}$$

where $\boldsymbol{S}$ is a $n \times m$ constant matrix. $\boldsymbol{S}$ is called as the "summing" matrix since its role is to aggregate the observations of the bottom level to the corresponding upper levels. To understand the notations clearly, consider the hierarchy given in Figure 1.

In any hierarchy, the most aggregated level is termed as level 0, the second most aggregated level is termed as level 1 and so on. This example consists of two levels. At a particular time t, let $y_{T,t}$ denote the observation at level 0; $y_{A,t}, y_{B,t}, y_{C,t}$ denote observations at level 1; and $y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t}, y_{CA,t}, y_{CB,t}$ denote observations at level 2. Then $\boldsymbol{y}_t = [y_{T,t}, y_{A,t}, y_{B,t}, y_{C,t}, y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t}, y_{CA,t}, y_{CB,t}]^T$,

$\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t}, y_{CA,t}, y_{CB,t}]^T$, $m = 7$, n=11, and

$$\boldsymbol{S} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ & & & I_m & & & \end{pmatrix},$$

where I_m is a m-dimension identity matrix.

3.2 Existing point forecast methods

This subsection will briefly discuss the existing hierarchical point forecast methods. Suppose we fit time series models for each series independently, based on data up to time T and obtain h-step ahead forecasts. These forecasts will be then formed in a vector $\hat{\boldsymbol{y}}_{T+h}$ by stacking the forecasts at each node in the same order as $\boldsymbol{y}_t$. This is termed as the vector of base forecasts. Further let $\tilde{\boldsymbol{y}}_{T+h}$ denote h-step-ahead aggregate consistent point forecasts. Then all existing point forecast methods can be expressed as,

$$\tilde{\boldsymbol{y}}_{T+h} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h}, \tag{3.2}$$

where $\boldsymbol{P}$ that projects the base forecasts to the bottom level which are then summed by using $\boldsymbol{S}$ to obtain aggregate consistent h-step ahead forecasts, $\tilde{\boldsymbol{y}}_{t+h}$. Hyndman et al. (2011) showed that any $\boldsymbol{P}$ such that $\boldsymbol{SPS} = \boldsymbol{S}$ produces unbiased aggregate consistent point forecasts through (3.2).

The hierarchical forecasting methods differ from one another by the different choice of matrix $\boldsymbol{P}$. The most traditional, bottom up and top down approaches have,

$$\boldsymbol{P} = [0_{m\times(n-m)}|I_m] \tag{3.3}$$

and

$$\boldsymbol{P} = [\boldsymbol{p}|0_{m\times(n-1)}] \tag{3.4}$$

respectively, where $\boldsymbol{p} = [\boldsymbol{p}_1, ...\boldsymbol{p}_m]^T$ is a vector of proportions that sums to one. Hyndman et al. (2011) showed that in (3.4), $\boldsymbol{SPS} \neq \boldsymbol{S}$ for any choice of $\boldsymbol{p}$ and hence top-down method produces aggregate consistent forecasts that are biased even if the base forecasts are unbiased.

The optimal reconciliation method introduced by Hyndman et al. (2011) is based on the following regression model,

$$\hat{\boldsymbol{y}}_{T+h} = \boldsymbol{S}\boldsymbol{\beta}_{T+h} + \boldsymbol{\varepsilon}_{T+h}, \tag{3.5}$$

where $\boldsymbol{\beta}_{T+h} = E[\boldsymbol{y}_{T+h}|\boldsymbol{y}_1, ..., \boldsymbol{y}_T]$ is the unknown mean of the bottom level series at time $T + h$ and $\boldsymbol{\varepsilon}_{T+h}$ is the reconciliation error with mean zero and variance $\boldsymbol{V}$. Hyndman et al. (2011) further showed that ordinary least squares (OLS) solution of (3.5) yields

$$\boldsymbol{P} = (\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T, \tag{3.6}$$

and if $\boldsymbol{V}$ is known, the generalized least squares (GLS) solution to (3.5) yields

$$\boldsymbol{P} = (\boldsymbol{S}^T\boldsymbol{V}^\dagger\boldsymbol{S})^{-1}\boldsymbol{S}^T\boldsymbol{V}^\dagger, \tag{3.7}$$

where $\boldsymbol{V}^\dagger$ is the Moore-Penrose generalized inverse of $\boldsymbol{V}$. However Wickramasuriya, Athanasopoulos, and Hyndman (2015) showed that $\boldsymbol{V}$ is not identifiable and hence the

GLS solution in (3.7) is unattainable. Further they propose to minimize the sum of the variances of reconciled forecast errors to obtain an optimal solution to $\boldsymbol{P}$.

Suppose variance of h-step ahead base forecast errors is denoted by, $Var(\boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h}) = \boldsymbol{\Sigma}_{T+h}$. Wickramasuriya, Athanasopoulos, and Hyndman (2015) first showed that the variance of the reconciled forecast errors, i.e $Var(\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_{T+h}) = \tilde{\boldsymbol{\Sigma}}_{T+h}$ is given by,

$$\tilde{\boldsymbol{\Sigma}}_{T+h} = \boldsymbol{S}\boldsymbol{P}\boldsymbol{\Sigma}_{\boldsymbol{T+h}}\boldsymbol{P}^T\boldsymbol{S}^T \tag{3.8}$$

for any choice of $\boldsymbol{P}$. Then they minimize the trace of $\tilde{\boldsymbol{\Sigma}}_{T+h}$ with respect to $\boldsymbol{SPS} = \boldsymbol{S}$ to obtain optimal reconciliation matrix $\boldsymbol{P}$. They found a closed form solution to this minimization problem is given by,

$$\boldsymbol{P} = (\boldsymbol{S}^T\boldsymbol{\Sigma}_{T+h}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}^T\boldsymbol{\Sigma}_{T+h}^{-1} \tag{3.9}$$

and they named this the MinT approach. This implies the reconciled point forecasts are,

$$\tilde{\boldsymbol{y}}_{T+h} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h} = \boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{\Sigma}_{T+h}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}^T\boldsymbol{\Sigma}_{T+h}^{-1}\hat{\boldsymbol{y}}_{T+h}. \tag{3.10}$$

Since the variance covariance matrix of 1-step ahead base forecasts errors are approximately proportional to that of h-step ahead base forecasts errors, we have $\boldsymbol{\Sigma}_{T+h} = \alpha_{T+h}\boldsymbol{\Sigma}_{T+1}$ where $\alpha_{T+h} > 0$. Wickramasuriya, Athanasopoulos, and Hyndman (2015), discussed alternative ways to estimate $\boldsymbol{\Sigma}_{T+1}$ and how these estimates lead to different $\boldsymbol{P}$ matrices. In particular, if $\boldsymbol{\Sigma}_{T+1}$ is approximated by an identity matrix, then $\boldsymbol{P}$ in (3.9) collapses to OLS $\boldsymbol{P}$ in (3.6) introduced by Hyndman et al (2011). If $\boldsymbol{\Sigma}_{T+1}$ is approximated by a diagonal matrix with diagonal elements being the variances of base forecast errors, then (3.9) collapses to the weighted least square (WLS) solution introduced by Hyndman, Lee, and Wang (2016).

Alternatively, the unbiased sample variance covariance matrix of 1-step ahead base forecast errors can be used as an estimator for $\boldsymbol{\Sigma}_{T+1}$. They refer to this as MinT(Sample). Also they propose a shrinkage estimator for $\boldsymbol{\Sigma}_{T+1}$ which they referred to as MinT(Shrink). The shrinkage estimator they used is given by,

$$\hat{\boldsymbol{\Sigma}}_{T+1}^{shr} = \tau\hat{\boldsymbol{\Sigma}}_{T+1}^{D} + (1-\tau)\hat{\boldsymbol{\Sigma}}_{T+1}, \tag{3.11}$$

where $\hat{\boldsymbol{\Sigma}}_{T+1}^{D}$ is the diagonal matrix comprising diagonal entries of $\hat{\boldsymbol{\Sigma}}_{T+1}$ and

$$\tau = \frac{\sum_{i\neq j}\hat{Var}(\hat{r}_{ij})}{\sum_{i\neq j}\hat{r}_{ij}^2}$$

is the shrinkage parameter proposed by Schäfer and Strimmer (2005). $\hat{r}_{ij}$ is the ij-th element of sample correlation matrix. In this estimation, the off-diagonal elements of 1-step ahead sample covariance matrix will be shrunk to zero depending on the sparsity. Their simulation study shows that estimating $\boldsymbol{\Sigma}_{T+1}$ through a shrinkage procedure outperforms all existing point forecast methods in hierarchical time series.

3.3 Probabilistic forecasts in the Gaussian framework

Suppose we are interested in estimating the probabilistic forecasts of hierarchical time series, where all historical data are assumed to have a multivariate Gaussian distribution. Then the predictive distribution of the hierarchy will also follow a multivariate Gaussian

distribution. That is, let $\boldsymbol{Y}_{t+h}$ be a random vector from the predictive density of the hierarchy. Then,

$$\boldsymbol{Y}_{t+h} = \mathcal{N}(\tilde{\boldsymbol{y}}_{t+h}, \tilde{\boldsymbol{\Sigma}}_{t+h}), \tag{3.12}$$

where $\tilde{\boldsymbol{y}}_{t+h} = E(\boldsymbol{Y}_{t+h})$ and $\tilde{\boldsymbol{\Sigma}}_{t+h} = Var(\boldsymbol{Y}_{t+h})$. In other words, $\tilde{\boldsymbol{y}}_{t+h}$ and $\tilde{\boldsymbol{\Sigma}}_{t+h}$ are the first two moments of the Gaussian predictive distribution. In order to reflect the inherent properties of the hierarchical time series, the predictive distribution should be reconciled so that they become aggregate consistent. Since the first two moments will uniquely characterize the Gaussian distribution, it is sufficient to reconcile the first two moments of the future quantities, to obtain the reconciled predictive multivariate Gaussian distribution of the whole hierarchy.

The point forecasts in hierarchical time series in the Gaussian framework produce the mean of the predictive distribution, that is $\tilde{\boldsymbol{y}}_{t+h}$. Therefore from (3.2) it follows that, $\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{SP}\hat{\boldsymbol{y}}_{t+h}$. Further,

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_{t+h} &= Var(\boldsymbol{Y}_{t+h}), \\
&= E[\boldsymbol{Y}_{t+h} - E(\boldsymbol{Y}_{t+h})]^2, \\
&= Var[\boldsymbol{Y}_{t+h} - E(\boldsymbol{Y}_{t+h})] + (E[\boldsymbol{Y}_{t+h} - E(\boldsymbol{Y}_{t+h})])^2, \\
&= Var[\boldsymbol{Y}_{t+h} - E(\boldsymbol{Y}_{t+h})] = Var[\boldsymbol{Y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h})].
\end{aligned}$$

Therefore from (3.8) it follows that, $\tilde{\boldsymbol{\Sigma}}_{t+h} = \boldsymbol{SP}\boldsymbol{\Sigma}_{t+h}\boldsymbol{P}^T\boldsymbol{S}^T$. The role of $\boldsymbol{P}$ in $\tilde{\boldsymbol{\Sigma}}_{t+h}$ is to project the variances and covariances of base forecast errors to the bottom level and then aggregate them through the summing matrix to obtain the aggregate consistent variance of the future values.

Now that we have the reconciled means and the variances of future values of the hierarchical time series, we already have estimated the entire predictive distribution under the Gaussian framework. It is worth to note that $\tilde{\boldsymbol{\Sigma}}_{t+h}$ is a singular matrix due to the aggregation nature of the hierarchy and hence the predictive Gaussian distribution of the hierarchy will be a singular multivariate Gaussian distribution.

Different choices of $\boldsymbol{P}$ matrices lead to different estimations of means and variances of future quantities and hence different predictive Gaussian distributions. In Section 4 we compare the predictive performance of these Gaussian distributions with respect to different choices of forecasting approaches, namely, Bottom up (3.3), OLS (3.6), MinT(Sample) and MinT(Shrink) (3.9).

The MinT solution introduced by Wickramasuriya, Athanasopoulos, and Hyndman (2015) is known to produce best unbiased point forecasts in hierarchical time series. However, it is not theoretically proven yet as an optimal reconciliation matrix that produces the best probabilistic forecasts of hierarchical time series. Since the scoring functions measure the predictive performance of probabilistic forecasts, finding a $\boldsymbol{P}$ matrix that minimizes an appropriately chosen scoring function would yield the best probabilistic forecasts of hierarchical time series. I expect to address this problem under the Gaussian framework in my future work.

3.4 Probabilistic forecasts using non-parametric bootstrap procedure

In many applications, it may not be reasonable to assume that the forecast densities follow a Gaussian distribution or any other parametric distribution. Non-parametric approaches

have been used in the literature to estimate the probabilistic forecasts in such situations. I propose a non-parametric bootstrap procedure to obtain aggregate consistent probabilistic forecasts of hierarchical time series in this section.

Suppose we fit univariate time series models for the series at each node in the hierarchy by using past observations up to time T. Then we generate future sample paths at each node from fitted univariate models, that are conditional on past observations. The error series for generating these sample paths will be obtained by bootstrapping in-sample errors from the fitted model. That is, let, $\boldsymbol{R}_{(T\times n)} = (\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_T)$ denote the in-sample residual matrix where, $\boldsymbol{e}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ is a vector that consists of residuals in each node at time t and stacked in the same order as $\boldsymbol{y}_t$. Then we block bootstrap a sample of size h from $\boldsymbol{R}$ and these bootstrapped errors will be incorporated as the error series of simulating future paths.

The future sample paths that were simulated will be then formed in a vector $\boldsymbol{y}^b_{T+h}$ by stacking the sample paths at each node in the same order as $\boldsymbol{y}_t$. Then these sample paths will be revised to make them aggregate consistent by,

$$\tilde{\boldsymbol{y}}^b_{T+h} = \boldsymbol{SP}\boldsymbol{y}^b_{T+h}, \tag{3.13}$$

where, $\tilde{\boldsymbol{y}}^b_{T+h}$ denote h-step-ahead aggregate consistent future paths. Thousands of such bootstrapped reconciled future paths can be considered as a possible sample from the predictive distribution of the future values of the hierarchy. The predictive performance of these probabilistic forecasts will experiment in a simulation set up in section 4.2.

The bootstrap procedure explained above does not explicitly model the dependency across the hierarchy. In future work, I expect to modify this approach by incorporating the dependency structure of the hierarchy. Copula modeling might be useful in this case since it allows to model the dependency between variables with different marginals.

Further it is worth to mention about the comparison of two approaches. Gaussianity is a reasonable assumption for hierarchies with small dimensions. Therefore, it is meaningful to compare the probabilistic forecasts from Gaussian densities with that of non-parametric approach to see how much we loose by not assuming Gaussianity in the problem. On the other hand, where there is clearly no Gaussianity in the problem, we can see how much we loose by the wrong assumption of Gaussianity.

4 Monte Carlo Simulations

This section presents the results of the Monte Carlo simulation, that was carried out to experiment with the two probabilistic forecasting approaches discussed in the previous section. Subsection 4.1 will analyze the predictive performance of the probabilistic forecasts in the Gaussian framework discussed in 3.3, whereas section 4.2 will analyze the predictive performance of the probabilistic forecasts based on non-parametric bootstrap procedure discussed in 3.4. Both simulation studies were based on the two hierarchical structures presented in Figure 1 and Figure 2. In each case, the data for the bottom level series were generated and then aggregated up along the tree to obtain the data for the top levels. The data generation process in two structures will be explained below.

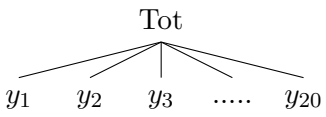

Figure 2: Single level hierarchical diagram with 20 bottom level series

Data generation process for the hierarchy 1

The hierarchy in Figure 1 consists of seven bottom level series. Each of the seven series was generated from a univariate ARMA (1,1) process. For each generated series, parameters for AR and MA components were randomly and uniformly chosen from the parameter space defined as $[0.4, 0.7]$ satisfying the stationarity and invertibility conditions of the coefficients.

Two correlation structures for the contemporaneous errors of ARMA processes were considered.
They are,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0.8 & 0.6 & 0.5 & 0.4 & 0.2 & -0.2 \\ 0.8 & 1 & 0.7 & 0.4 & 0.5 & 0.35 & -0.1 \\ 0.6 & 0.7 & 1 & 0.2 & 0.3 & 0.15 & -0.2 \\ 0.5 & 0.4 & 0.2 & 1 & 0.75 & 0.4 & 0.3 \\ 0.4 & 0.5 & 0.3 & 0.75 & 1 & 0.55 & 0.4 \\ 0.2 & 0.35 & 0.15 & 0.4 & 0.55 & 1 & 0.7 \\ -0.2 & -0.1 & -0.2 & 0.3 & 0.4 & 0.7 & 1 \end{pmatrix}, \boldsymbol{B} = \begin{pmatrix} 1 & .0.8 & 0.6 & 0 & 0 & 0 & 0 \\ 0.8 & 1 & 0.7 & 0 & 0 & 0 & 0 \\ 0.6 & 0.7 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.75 & 0 & 0 \\ 0 & 0 & 0 & 0.75 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}.$$

The correlation structure $\boldsymbol{A}$ will allow us to have a strong positive correlation among the series within the same parent node and moderate correlation (positive/negative) among the series between different parent nodes. The results that follow from this correlation structure will be referred to as "Hierarchy 1 - Case A".

Alternatively, correlation structure $\boldsymbol{B}$ allows the series within the same parent node to be positively correlated but series with different parent nodes to be independent. The results that follow from this correlation structure are referred to as "Hierarchy 1 - Case B".

Data generation process for the hierarchy 2

The twenty bottom level series of the hierarchical structure in Figure 2 were also generated from univariate ARMA(1,1) processes. Parameters were chosen in a similar manner as before, from the parameter space $[0.2, 0.8]$. The correlation structure considered for the contemporaneous errors have a *moderate correlation* between all the series. The results that follow from this hierarchical structure will refer as "Hierarchy 2".

4.1 Analyzing the predictive performance of probabilistic forecasts in the Gaussian framework

Simulation setup:
Data were generated with sample size T = 501 for each hierarchical structures discussed above. To ensure the data follows a multivariate Gaussian distribution, the contemporaneous errors of the ARMA processes were randomly generated from a multivariate Gaussian distribution with mean zero and variance covariance matrix described as before.

In each hierarchical structure, univariate ARIMA models were fitted using the first 500 observations, for each series independently, and 1-step ahead base forecasts were generated. The *forecast* package in R (Hyndman, 2015) was used for model fitting and for generating base forecasts. Then means and variances of 1-step ahead future distribution were estimated using the different methods discussed in Sections 3.2 and 3.3. When estimating the reconciled variance covariance matrix $\tilde{\boldsymbol{\Sigma}}_{t+h} = \boldsymbol{SP\Sigma_{t+h}P}^T\boldsymbol{S}^T$, $\boldsymbol{\Sigma}_{t+h}$ will be estimated using the shrinkage method given in (3.11), which was used in MinT(Shrink) estimation by Wickramasuriya, Athanasopoulos, and Hyndman (2015). This process was replicated using 1000 different data sets from the same data generating process (DGP).

I used two matrix norms, namely, Frobenius norm and Spectral norm to measure how close the estimated variance covariance matrix of probabilistic forecasts is to the true variance covariance matrix.

Recall that the Frobenius norm of a matrix $\boldsymbol{M}$ is defined as,

$$\|\boldsymbol{M}\|_F = \sqrt{Trace(\boldsymbol{M}^T\boldsymbol{M})} \tag{4.1}$$

and the Spectral norm of a matrix $\boldsymbol{M}$ is defined as,

$$\|\boldsymbol{M}\|_S = \sqrt{\lambda_{max}(\boldsymbol{M})}, \tag{4.2}$$

where $\lambda_{max}(\boldsymbol{M})$ is the maximum eigenvalue of $\boldsymbol{M}$ and $\boldsymbol{M}$ is a positive semidefinite matrix.

The Frobenius norm and Spectral norm differences between two matrices $\boldsymbol{M}$ and $\boldsymbol{N}$ can be respectively calculated as,

$$\|\boldsymbol{M}-\boldsymbol{N}\|_F = \sqrt{Trace(\boldsymbol{M}^T\boldsymbol{M}) + Trace(\boldsymbol{N}^T\boldsymbol{N}) - 2Trace(\boldsymbol{M}^T\boldsymbol{N})}, \tag{4.3}$$

and

$$\|\boldsymbol{M}-\boldsymbol{N}\|_S = \sqrt{\lambda_{max}(\boldsymbol{M}-\boldsymbol{N})}. \tag{4.4}$$

Average norm differences over 1000 simulations are presented in Table 1. The smaller the average norm difference, the closer the estimated variance covariance matrix to the true covariance matrix is.

Note that the true variance covariance matrix for the whole hierarchy is given by $\boldsymbol{S\Omega S}^T$, where $\boldsymbol{\Omega}$ is the variance covariance matrix of the data generating process of the bottom levels series.

The predictive performance of the probabilistic forecasts was evaluated using Energy scores and Log scores using equation (2.7) and (2.10) respectively. Average scores over 1000 replications are presented in Table 2. The smaller the average scores, the better the probabilistic forecasts are. Energy scores were calculated based on a random sample drawn from the predictive multivariate Gaussian density.

Since the resulting reconciled Gaussian predictive distribution are singular, the log scores of these were calculated using the pseudo determinant and Moore-Penrose inverse of variance covariance matrix of the singular multivariate Gaussian density function. That is,

$$LogS(\boldsymbol{F},\boldsymbol{y}) = log(|\boldsymbol{\Sigma}|_+^{-1/2}) + (\boldsymbol{y}-\boldsymbol{\mu_F})'\boldsymbol{\Sigma}_{\boldsymbol{F}}^{\dagger}(\boldsymbol{y}-\boldsymbol{\mu_F}), \tag{4.5}$$

where, $|\boldsymbol{\Sigma}|_+$ is the product of only the positive eigenvalues of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{\boldsymbol{F}}^{\dagger}$ is the Moore-Penrose inverse of $\boldsymbol{\Sigma}$ and $\boldsymbol{F}$ is a singular Gaussian distribution with mean $\boldsymbol{\mu_F}$ and positive semidefinite variance covariance matrix $\boldsymbol{\Sigma}$.

However, it would not be sensible to compare the log score from a singular density function for the reconciled case to that of a non-singular density function for the base forecasts. Therefore the log scores for the base probabilistic forecasts are not presented in Table 2. Further, the Diebold-Mariano test described in Section 2.3.2 was used to test the equal predictive performance of probabilistic forecasts from the different methods.

Results:

As it can be seen from the results in Table 1, the variance covariance matrices of reconciled predictive Gaussian distributions are much closer to the true variance covariance matrix than that of the base predictive density in terms of Frobenius norm.

As it is evident from the Diebold-Mariano test results presented in Table 2, the MinT(Shrink) and OLS approaches significantly improve on the base predictive densities in all three hierarchical structures, in terms of the predictive performance measured by Energy score. The Bottom up and MinT(Sample) approaches improve the base forecast densities only for "Hierarchy 1-Case A" and "Hierarchy 2".

The Diebold-Mariano test results for log scores presented in Table 3 shows that, the MinT(Shrink) approach outperforms Bottom up and OLS for "Hierarchy 1 - Case A" whereas MinT(Shrink) outperforms predictive densities from all other methods for "Hierarchy 1 - Case B". On the other hand, both Bottom up and MinT(Shrink) outperform OLS and MinT(Sample) for "Hierarchy 2". Further the Bottom up and MinT(Shrink) approaches estimate densities under Gaussianity with equivalent predictive performance with respect to both Energy and Log scores in "Hierarchy 2".

In conclusion, the simulation results imply that the hierarchical approaches provide improved probabilistic forecasts compared to the base probabilistic forecasts. Further, Gaussian densities from the Bottom up and MinT(Shrink) approaches have better predictive performances than the other methods. Moreover, they both have equivalent predictive performance in hierarchies with much larger dimensions.

Table 1: *Comparison of covariance matrices of predictive distributions over different hierarchical forecasting methods under Gaussian framework*

Forecasting method	Hierarchy 1 - Case A		Hierarchy 1 - Case B		Hierarchy 2	
	F-norm	S-norm	F-norm	S-norm	F-norm	S-norm
Base	11.59	8.38	5.86	3.49	14.61	7.51
Bottom up	11.19	8.42	5.62	3.42	12.83	10.09
OLS	11.17	8.37	5.67	3.49	10.95	7.18
MinT(Sample)	11.24	8.52	5.64	3.48	9.40	4.15
MinT(Shrink)	11.25	8.56	5.65	3.49	10.29	7.17

Note: F-norm and S-norm denote the Frobenius norm difference and Spectral norm difference between estimated covariance matrix and the true covariance matrix respectively.

Table 2: *Energy score and log score of the probabilistic forecasts followed from different hierarchical forecasting methods under Gaussian framework*

Forecasting method	Hierarchy 1 - Case A		Hierarchy 1 - Case B		Hierarchy 2	
	Energy score	Log score	Energy score	Log score	Energy score	Log score
Base	9.26		6.65		9.76	
Bottom up	9.19**	9.06	6.63	8.35	9.57**	27.89
OLS	9.22**	9.07	6.63**	8.44	9.75**	27.94
MinT(Sample)	9.20*	9.04	6.66	8.44	9.58**	28.00
MinT(Shrink)	9.18**	9.03	6.62**	8.41	9.60**	27.90

Note: Each entry represent scores. If the Diebold-Mariano test rejects the null hypothesis of equivalent predictive performance of base probabilistic forecasts and probabilistic forecasts from other methods at 5% and 1% level of significance, then it is denoted by "$*$" and "$**$" respectively.

Table 3: *Diebold-Mariano test results for comparing probabilistic forecasts from different hierarchical forecasting methods under Gaussian framework (with respect to Energy score and Log score)*

Forecasting method	Hierarchy 1 - Case A		Hierarchy 1 - Case B		Hierarchy 2	
	Energy score	Log score	Energy score	Log score	Energy score	Log score
BU Vs OLS	−1.76	−2.24*	−0.09	0.00	−3.86**	−3.39**
BU Vs MinT(Sample)	−0.65	2.25*	−2.06*	0.19	−0.39	−7.98**
BU Vs MinT(Shrink)	0.18	4.81**	1.24	4.09**	−1.18	−0.68
OLS Vs MinT(Sample)	0.92	2.55*	−1.69	0.18	4.36**	−3.77**
OLS Vs MinT(Shrink)	2.01*	6.07**	1.24	3.84**	6.55**	4.86**
MinT(Shrink) Vs MinT(Sample)	−1.12	−1.78	−3.99**	−3.42**	1.34	−7.83**

Note: Each entry represent the Diebold-Mariano test statistic for testing the null hypothesis of equivalent predictive performance of two different forecasting methods. "$*$" and "$**$" denote if the test statistic rejects null hypothesis at 5% and 1% level of significance respectively. Further, if it has compared probabilistic forecasts F vs. G and given a negative(positive) significant test statistic, then it implies F(G) outperforms G(F).

4.2 Analyzing the predictive performance of reconciled probabilistic forecasts from non-parametric bootstrap procedure

Simulation setup:
This section will examine the performance of probabilistic forecasts based on non-parametric bootstrap procedure discussed in Section 3.4 in a simulation setup. Data were generated in the same manner as described before for the two hierarchical structures given in Figure 1 and 2.

After the data generation of each hierarchical structure, univariate ARIMA models were fitted for all series independently and the in-sample errors were constructed as described in Section 3.4. Then 1-step ahead future paths were generated for each variable in the hierarchy, based on the fitted model and conditioning on the historical observations. *forecast* package in R software (Hyndman, 2015) was used in model fitting and future path generation. Bootstrapped in-sample errors were added in generating future paths to incorporate the uncertainty in future values. Then these sample paths were revised such that they become aggregate consistent, using existing methods, namely, OLS, Mint(Sample) and MinT(Shrink) as described in equation (3.13). 5000 of such 1-step ahead aggregate

consistent future paths were constructed for the whole hierarchy. This process was replicated in 500 different datasets from the same DGP.

Frobenius norm and Spectral norm differences between the covariance matrix of future sample paths and true variance covariance matrix of the whole hierarchy were calculated using (4.3) and (4.4) respectively, and the average norm difference over 500 replications are presented in Table 4. This will provide some information on how close the variance covariance matrix of estimated probabilistic forecasts to the true variance covariance matrix.

Moreover, the Energy score was used to measure the predictive performance of future paths generated from different hierarchical forecasting methods and Diebold-Mariano test was used to test the equivalence between them. Average energy score over 500 replications are presented in Table 5 and the Diebold-Mariano test statistics for comparing predictive performances are presented in Table 6.

Results:
Average norm difference of aggregate consistent future paths from different forecasting approaches is smaller than that of base future paths. This implies that aggregate consistent future paths can better capture the true correlation structure of the hierarchy.

From the results presented in Tables 5 and 6 it is evident that, all methods, except for MinT(Sample) outperform *base*, for "Hierarchy 1 - Case A" and "Hierarchy 1 - Case B" in terms of the predictive performance of future paths, whereas only OLS and MinT(Shrink) outperform *base* for "Hierarchy 2". Furthermore, future paths from the Bottom up approach have significantly better predictive performance than all other methods in "Hierarchy 1 - Case A" and "Hierarchy 1 - Case B". However in "Hierarchy 2", MinT(Shrink) outperforms all other methods including the Bottom up approach. This implies that MinT(Shrink) has better predictive performance than other existing methods in hierarchies with large dimensions. The significantly better performance in the Bottom up approach in comparatively smaller dimensions is perhaps due to simulation setup, where we obtain the data for the whole hierarchy by generating a well structured bottom level series.

Table 4: *Comparing covariance matrices of probabilistic forecasts over different hierarchical forecasting methods under non-parametric bootstrap procedure*

Forecasting method	Hierarchy 1 - Case A		Hierarchy 1 - Case B		Hierarchy 2	
	F-norm	S-norm	F-norm	S-norm	F-norm	S-norm
Base	12.35	9.19	11.85	8.85	12.25	8.65
Bottom up	11.90	8.95	11.38	8.59	8.38	3.42
OLS	12.11	9.19	11.59	8.83	11.84	8.59
MinT(Sample)	12.12	9.26	11.58	8.87	8.31	3.35
MinT(Shrink)	11.98	9.10	11.47	8.76	8.59	3.90

Note: F-norm and S-norm denote the Frobenius norm difference and Spectral norm difference between estimated covariance matrix and the true covariance matrix respectively.

Table 5: *Energy scores of probabilistic forecasts based on non-parametric bootstrap procedure*

Forecasting method	Energy score		
	Hierarchy 1 - Case A	Hierarchy 1 - Case B	Hierarchy 2
Base	14.54	12.44	13.59
Bottom up	13.87**	11.76**	13.77
OLS	14.17**	12.11**	13.53**
MinT(Sample)	15.12	12.98	13.61
MinT(Shrink)	14.15**	12.15**	13.37**

Note: each entry represent scores. If the Diebold-Mariano test rejects the null hypothesis of equivalent predictive performance of base probabilistic forecasts and probabilistic forecasts from other methods at 5% and 1% level of significance, then it is denoted by "∗" and "∗∗" respectively.

Table 6: *Diebold-Mariano test results for comparing probabilistic forecasts from different hierarchical forecasting methods under non-parametric bootstrap approach*

Forecasting methods	Diebold-Mariano test results with respect to Energy scores		
	Hierarchy 1 - Case A	Hierarchy 1 - Case B	Hierarchy 2
BU Vs OLS	−2.09*	−2.63**	1.46
BU Vs MinT(Sample)	−5.19**	−5.71**	4.46**
BU Vs MinT(Shrink)	−2.34*	−3.63**	4.01**
OLS Vs MinT(Sample)	−3.87**	−4.01**	−0.60
OLS Vs MinT(Shrink)	0.13	−0.35	2.03*
MinT(Shrink) Vs MinT(Sample)	−5.88**	−5.64**	−3.28**

Note: Each entry represent the Diebold-Mariano test statistic for testing the null hypothesis of equivalent predictive performance of two forecasting methods. "∗" and "∗∗" denote if the test statistic rejects the null hypothesis at 5% and 1% level of significance respectively. Further, if it has compared probabilistic forecasts F vs G and given a negative(positive) significant test statistic, then it implies F(G) outperforms G(F)

5 Discussion and future work

To the best of our knowledge, many of the studies on hierarchical forecasting have involved generating point forecasts, only a few have attempted in producing the interval forecasts and no study has focused on the probabilistic forecasts in hierarchical time series. My research work will attempt to contribute to the hierarchical literature by inventing methods that produce the entire predictive distribution for hierarchical time series.

Under the Gaussian framework, estimating the predictive means and the variance covariance matrix of future quantities will uniquely estimate the predictive Gaussian density. Estimating these predictive means and the variances of hierarchical time series are described in equations (3.2) and (3.8) respectively in Section 3.3. Different $\boldsymbol{P}$ matrices yield different estimates of means and variance covariance matrices and hence different predictive Gaussian densities. We attempt to obtain the optimal density estimate that gives the best predictive performance. In the present work I have compared the predictive

performance of the Gaussian densities that are estimated with respect to existing hierarchical forecasting methods, namely, Bottom up, OLS, MinT(Sample) and MinT(Shrink).

The simulation study provides evidence that the reconciled variance covariance matrix of the predictive Gaussian density is much closer to the true variance covariance matrix of the whole hierarchy than the base variance covariance matrix. In general, the hierarchical forecasting approaches yields Gaussian densities with improved predictive performance compared to the base Gaussian densities. It was further assessed the predictive performances of Gaussian densities followed from competing hierarchical forecasting approaches. These results imply that MinT approach outperforms Bottom up and OLS approaches in hierarchies with a small number of dimensions. Further, it was noticed that MinT(Shrink) outperforms MinT(Sample) when there is a sparsity in the correlation matrix of the hierarchy. This is perhaps due to the ability of shrinkage estimator in capturing the sparsity of the true covariance matrix. However, when it comes to the hierarchies with large dimensions, the Bottom up and MinT(Shrink) have equivalent predictive performance and they both outperform all other forecasting approaches.

Even though the simulation study favored the MinT(Shrink) approach for generating the predictive Gaussian distributions in hierarchical time series, there is still a lack of theoretical evidence for this. In my future work, I expect to find an optimal matrix $\boldsymbol{P}$ that combines the entire predictive Gaussian distribution, by minimizing an appropriately chosen scoring function.

I have also presented a non-parametric bootstrap procedure to obtain probabilistic forecasts for hierarchical time series where it relaxes the assumptions on a predictive parametric distribution for the hierarchy. This entails using a hierarchical forecasting approach to estimate the aggregate consistent future paths of the whole hierarchy. Thousands of such aggregate consistent future paths were generated conditioning on the historical observations and incorporating the bootstrapped in-sample errors.

It was observed from the simulation study that the variance covariance matrix of future paths generated from the hierarchical approaches is much closer to the true covariance matrix than base future paths. This implies that the correlation structure of the hierarchy may be improved by future paths from hierarchical approaches over base future paths. Further, it was evident that the future paths generated from MinT(Shrink), Bottom-up and OLS approaches outperform base future paths in hierarchies with comparatively small dimensions whereas only MinT(Shrink) and OLS outperform base future paths in large dimensions.

Next, it was evaluated the predictive performances of future paths generated from different types of hierarchical approaches. These results evident that the future paths generated using MinT(Shrink) and Bottom up provide significantly better predictive performance than OLS and MinT(Sample) approaches. However, for hierarchies with small dimensions, Bottom up outperform MinT(Shrink) whereas MinT(Shrink) outperform Bottom up in hierarchies with large dimensions. Perhaps, the Bottom up approach dominates other methods in small hierarchies due to the well structured simulation setup.

I expect to improve this approach by incorporating the information of the dependency structure of the hierarchy by using semi-parametric copula approach. Further, I expect to find an optimal forecasting method that can be used to generate probabilistic forecasts in the non-parametric framework by minimizing a suitable score function.

5.1 Structure of the thesis outline

The lack of attention on probabilistic forecasts for hierarchical time series has motivated this research work to invent new methods that estimate the entire probability distribution of the hierarchical time series. Centralizing on this main objective, the structure of the research can be outlined as follows.

Chapter 1: Introduction

Understand the problem and the background of the study

Chapter 2: Literature Review

Review on literature associated with the problem of interest.

Chapter 3: Probabilistic forecasts of hierarchical time series

In order to preserve the inherent properties of the hierarchical structure, probabilistic forecasts of hierarchical time series should be reconciled so that they produce aggregate consistent probabilistic forecasts. Chapter 1 will be centralized on this main objective following two subsections.

Section 1: Probabilistic forecasts of hierarchical time series in the Gaussian framework.

Objective 1: Comparing the performances of the predictive Gaussian distribution followed from existing hierarchical forecasting methods.

This objective has been achieved so far and the description of methods followed by the simulation study have presented in section 3.3 and 4.1 of this report.

Objective 2: Inventing a new technique that optimally combines the entire predictive multivariate Gaussian distribution by minimizing an appropriately chosen scoring function.

Section 2: Probabilistic forecasts of hierarchical time series in the non-Gaussian framework: A non-parametric bootstrap approach.

This method involves obtaining aggregate consistent future paths using a non-parametric bootstrap approach.

Objective 1: Comparing the predictive performances of future paths followed from different types of existing hierarchical forecasting methods.

This objective has been achieved so far and the description of methods followed by the simulation study have presented in section 3.4 and 4.2 of this report.

Objective 2: Improving the future paths generation by modeling the dependency structure of the hierarchy using Copulas.

Chapter 4: Minimum Mean Squared Error (MSE) reconciliation

In the previous works on point forecasts of hierarchical time series involves minimizing the variance of aggregate consistent forecasts under the assumption of unbiasedness. I expect to relax this assumption on unbiasedness and minimize the trace of MSE of forecasts to obtain reconciled means and variances of future quantities of the hierarchical times series.

Chapter 5: Application

Chapter 6: Conclusions

5.2 Time plan of the Research

Thesis Chapter	Task description	Time duration	Progress
Introduction and Literature Review	Understanding the problem and review on literature associate with the problem	Feb/2016 - Nov/2016	Completed
Probabilistic forecasts of the hierarchical time series in the Gaussian framework	Comparing the performances of predictive Gaussian distributions followed from existing hierarchical forecasting methods	Nov/2016 - Feb/2017	Completed
	Inventing a new technique that optimally combines the entire predictive multivariate Gaussian distribution by minimizing a scoring function	Mar/2016 - Aug/2017	In-progress
Probabilistic forecasts of the hierarchical time series in the non - Gaussian framework	Comparing the predictive performance of future paths followed from different types of existing hierarchical forecasting methods	Nov/2016 - Feb/2017	Completed
	Improving the future paths generation by modeling the dependency structure of the hierarchy using copulas	Sep/2016 - Feb/2018	Incomplete
Minimum Mean Squared Error reconciliation		Feb/2018 - Aug/2018	Incomplete
Application		Sep/2018 - Feb/2019	Incomplete

References

Abramson, B. and Clemen, R. (1995). "Probability forecasting". In: *International Journal of Forecasting* 11.1, pp. 1–4.

Athanasopoulos, G., Ahmed, R. A., and Hyndman, R. J. (2009). "Hierarchical forecasts for Australian domestic tourism". In: *International Journal of Forecasting* 25.1, pp. 146–166.

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2008). "Probabilistic quantitative precipitation field forecasting using a two-stage spatial model". In: *The Annals of Applied Statistics*, pp. 1170–1193.

Bludszuweit, H., Domínguez-Navarro, J. A., and Llombart, A. (2008). "Statistical analysis of wind power forecast error". In: *IEEE Transactions on Power Systems* 23.3, pp. 983–991.

Clark, T. E. (2011). "Real-Time Density Forecasts From Bayesian Vector Autoregressions With Stochastic Volatility". In: *Journal of Business & Economic Statistics* 29.3, pp. 327–341.

Clements, M. P. and Smith, J. (2002). "Evaluating multivariate forecast densities: a comparison of two approaches". In: *International Journal of Forecasting* 18.3, pp. 397–407.

Dawid, A. P. and Sebastiani, P. (1999). "Coherent dispersion criteria for optimal experimental design". In: *Annals of Statistics*, pp. 65–81.

Demarta, S. and McNeil, A. J. (2005). "The t copula and related copulas". In: *International Statistical Review/Revue Internationale de Statistique*, pp. 111–129.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). "Evaluating Density Forecasts with Application to Financial Risk Management". In: *International Economic Review* 39.4, pp. 863–883.

Diebold, F. X., Hahn, J., and Tay, A. S. (1999). "Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange". In: *Review of Economics and Statistics* 81.4, pp. 661–673.

Diebold, F. X. and Mariano, R. S. (1995). "Comparing predictive accuracy". In: *Journal of Business & economic statistics* 13.3, pp. 253–263.

Dunn, D., Williams, W., and DeChaine, T. (1976). "Aggregate versus subaggregate models in local area forecasting". In: *Journal of the American Statistical Association* 71.353, pp. 68–71.

Fliedner, G. (2001). "Hierarchical forecasting: issues and use guidelines". In: *Industrial Management & Data Systems* 101.1, pp. 5–12.

Gel, Y., Raftery, A. E., and Gneiting, T. (2004). "Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method". In: *Journal of the American Statistical Association* 99.467, pp. 575–583.

Genest, C. and Favre, A.-C. (2007). "Everything you always wanted to know about copula modeling but were afraid to ask". In: *Journal of hydrologic engineering* 12.4, pp. 347–368.

Genest, C. and Rivest, L.-P. (1993). "Statistical inference procedures for bivariate Archimedean copulas". In: *Journal of the American Statistical Association* 88.423, pp. 1034–1043.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). "Probabilistic forecasts, calibration and sharpness". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2, pp. 243–268.

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

Gneiting, T. and Raftery, A. E. (2005). "Weather forecasting with ensemble methods". In: *Science* 310.5746, pp. 248–249.

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation". In: *Monthly Weather Review* 133.5, pp. 1098–1118.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006). "Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method". In: *Journal of the American Statistical Association* 101.475, pp. 968–979.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds". In: *Test* 17.2, p. 211.

Gross, C. W. and Sohl, J. E. (1990). "Disaggregation methods to expedite product line forecasting". In: *Journal of Forecasting* 9.3, p. 233.

Huber, F. (2016). "Density forecasting using Bayesian global vector autoregressions with stochastic volatility". In: *International Journal of Forecasting* 32.3, pp. 818–837.

Hyndman, R (2015). "Forecasting functions for time series and linear models, R package version 8.0". In: *URL: http://github.com/robjhyndman/forecast*.

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

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). "Optimal combination forecasts for hierarchical time series". In: *Computational Statistics & Data Analysis* 55.9, pp. 2579–2589.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Kahn, K. B. (1998). "Revisiting top-down versus bottom-up forecasting". In: *The Journal of Business Forecasting* 17.2, p. 14.

Ko, S. I. and Park, S. Y. (2013). "Multivariate density forecast evaluation: A modified approach". In: *International Journal of Forecasting* 29.3, pp. 431–441.

Laio, F., Ridolfi, L., and Tamea, S. (2007). "Probabilistic prediction of real-world time series: A local regression approach". In: *Geophysical research letters* 34.3.

Lapide, L (1998). "A Simple View of Top-Down Versus Buttom-Up Forecasting". In: *Journal of Business Forecasting Methods and Systems* 17, pp. 28–31.

Manzan, S. and Zerom, D. (2008). "A bootstrap-based non-parametric forecast density". In: *International Journal of Forecasting* 24.3, pp. 535–550.

McLean Sloughter, J, Gneiting, T., and Raftery, A. E. (2013). "Probabilistic wind vector forecasting using ensembles and Bayesian model averaging". In: *Monthly Weather Review* 141.6, pp. 2107–2119.

McSharry, P. E., Bouwman, S., and Bloemhof, G. (2005). "Probabilistic forecasts of the magnitude and timing of peak electricity demand". In: *IEEE Transactions on Power Systems* 20.2, pp. 1166–1172.

Möller, A., Lenkoski, A., and Thorarinsdottir, T. L. (2013). "Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas". In: *Quarterly Journal of the Royal Meteorological Society* 139.673, pp. 982–991.

Nelsen, R. B. (1999). *An introduction to copulas, volume 139 of Lecture Notes in Statistics*.

Panagiotelis, A. and Smith, M. (2008). "Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions". In: *International Journal of Forecasting* 24.4, pp. 710–727.

Pinson, P. (2012). "Adaptive calibration of (u, v)-wind ensemble forecasts". In: *Quarterly Journal of the Royal Meteorological Society* 138.666, pp. 1273–1284.

Pinson, P. and Tastu, J. (2013). *Discrimination ability of the Energy score*. Tech. rep. Technical University of Denmark.

Pinson, P., Madsen, H., Nielsen, H. A., Papaefthymiou, G., and Klöckl, B. (2009). "From probabilistic forecasts to statistical scenarios of short-term wind power production". In: *Wind energy* 12.1, pp. 51–62.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). "Using Bayesian model averaging to calibrate forecast ensembles". In: *Monthly Weather Review* 133.5, pp. 1155–1174.

Rossi, B. (2014). *Density Forecasts in Economics, Forecasting and Policymaking*.

Schäfer, J., Strimmer, K., et al. (2005). "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics". In: *Statistical applications in genetics and molecular biology* 4.1, p. 32.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T. (2012). "Ensemble model output statistics for wind vectors". In: *Monthly weather review* 140.10, pp. 3204–3219.

Schwarzkopf, A. B., Tersine, R. J., and Morris, J. S. (1988). "Top-down versus bottom-up forecasting strategies". In: *The International Journal of Production Research* 26.11, pp. 1833–1843.

Shang, H. L. (2016). "Reconciling forecasts of infant mortality rates at national and sub-national levels: Grouped time-series methods". In: *Population Research and Policy Review*, pp. 1–30.

Shang, H. L. and Hyndman, R. J. (2016). "Grouped functional time series forecasting: An application to age-specific mortality rates". In: *Journal of Computational and Graphical Statistics* just-accepted.

Sklar, M (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Taieb, S. B., Huser, R., Hyndman, R. J., and Genton, M. G. (2016). "Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression". In: *IEEE Transactions on Smart Grid* 7.5, pp. 2448–2455.

Tay, A. S., Wallis, K. F., et al. (2000). "Density forecasting: a survey". In: *Journal of forecasting* 19.4, pp. 235–254.

Thorarinsdottir, T. L. and Gneiting, T. (2010). "Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173.2, pp. 371–388.

Vilar, J. A., Vilar, J. M., et al. (2013). "Time series clustering based on nonparametric multidimensional forecast densities". In: *Electronic Journal of Statistics* 7, pp. 1019–1046.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2015). *Forecasting hierarchical and grouped time series through trace minimization*. Tech. rep. Department of Econometrics and Business Statistics, Monash University.

Wijaya, T. K., Sinn, M., and Chen, B. (2015). "Forecasting uncertainty in electricity demand". In: *AAAI-15 Workshop on Computational Sustainability*. EPFL-CONF-203769.

Wytock, M. and Kolter, J. Z. (2013). "Large-scale probabilistic forecasting in energy systems using sparse gaussian conditional random fields". In: *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*. IEEE, pp. 1019–1024.

Zhang, Y., Wang, J., and Wang, X. (2014). "Review on probabilistic forecasting of wind power generation". In: *Renewable and Sustainable Energy Reviews* 32, pp. 255–270.