Forecast with Forecasts: Diversity Matters

Yanfei Kang[a], Wei Cao[a], Fotios Petropoulos[b], Feng Li[c,*]

[a]*School of Economics and Management, Beihang University, Beijing 100191, China.*
[b]*School of Management, University of Bath, Claverton Down, Bath BA2 7AY, UK.*
[c]*School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China.*

Abstract

Forecast combinations have been widely applied in the last few decades to improve forecasting. Estimating optimal weights that can outperform simple averages is not always an easy task. In recent years, the idea of using time series features for forecast combinations has flourished. Although this idea has been proved to be beneficial in several forecasting competitions, it may not be practical in many situations. For example, the task of selecting appropriate features to build forecasting models is often challenging. Even if there was an acceptable way to define the features, existing features are estimated based on the historical patterns, which are likely to change in the future. Other times, the estimation of the features is infeasible due to limited historical data. In this work, we suggest a change of focus from the historical data to the produced forecasts to extract features. We use out-of-sample forecasts to obtain weights for forecast combinations by amplifying the diversity of the pool of methods being combined. A rich set of time series is used to evaluate the performance of the proposed method. Experimental results show that our diversity-based forecast combination framework not only simplifies the modeling process but also achieves superior forecasting performance in terms of both point forecasts and prediction intervals. The value of our proposition lies on its simplicity, transparency, and computational efficiency, elements that are important from both an optimization and a decision analysis perspective.

Keywords: Forecasting, Forecast Combination, Forecast Diversity, Prediction Intervals, Empirical Evaluation

*Corresponding author
Email addresses: yanfeikang@buaa.edu.cn (Yanfei Kang), caowei08@buaa.edu.cn (Wei Cao), f.petropoulos@bath.ac.uk (Fotios Petropoulos), feng.li@cufe.edu.cn (Feng Li)
URL: https://orcid.org/0000-0001-8769-6650 (Yanfei Kang), https://orcid.org/0000-0003-3039-4955 (Fotios Petropoulos), https://orcid.org/0000-0002-4248-9778 (Feng Li)

1. Introduction

Many real-world problems are too complex for a single model that assumes a specific data generation process. Dating back to 1818, Laplace stated that "In combining the results of these two methods, one can obtain a result whose probability law of error will be more rapidly decreasing" (Clemen, 1989). The literature shows that model combinations improve overall performance in a variety of research areas, such as regression (e.g., Mendes-Moreira et al., 2012), classification (e.g., Rokach, 2010), anomaly detection (e.g., Perdisci et al., 2006), and time series forecasting (e.g., De Menezes et al., 2000). A recent overview of forecast combinations is provided in Section 2.5 of the encyclopedic review article by Petropoulos et al. (2021).

The motivation for forecast combinations has focused on finding the optimal weights of combining different *forecasts*, which are values at certain specific future times based on multiple forecasting methods. The seminal work of Bates and Granger (1969) in the area of combining forecasts suggests that forecast combinations can improve forecasting accuracy, provided that the sets of forecasts contain some independent information. The usefulness of forecast combinations has been demonstrated since then by numerous researchers using a variety of weighting methods (e.g., Winkler and Makridakis, 1983; Mostaghimi, 1996; Watson and Stock, 2004; Petropoulos and Kourentzes, 2015; Montero-Manso et al., 2020; Kang, Hyndman and Li, 2020).

Despite the large number of studies on forecast combinations, the "forecast combination puzzle" –the arithmetic mean performing better than more sophisticated combination methods in some applications– remains hard to tackle (Watson and Stock, 2004; Smith and Wallis, 2009; Claeskens et al., 2016; Petropoulos and Svetunkov, 2020). This situation is presumably related to that of equally weighted models in linear prediction (Dawes, 1979). We summarize the main reasons for the forecast combination puzzle as follows. On the one hand, the optimal weights estimated by forecast combination are often sensitive to historical data, and thus it is difficult to assemble robust forecasts that could consistently outperform a simple average. On the other hand, the merits of forecast combinations stem from independent information across multiple forecasts, which is further explained in Section 2. If all forecasts are close to identical, forecast combinations will be close to a simple average. Therefore, an optimal forecast combination depends to some degree on the *diversity* of the individual forecasts, which is nonetheless difficult to satisfy in reality (Thomson et al., 2019).

Because the relative performance of different forecast methods changes, depending on the nature of the time series (Reid, 1972), one way to forecast a time series involves feature-based forecasting. The majority of the studies in this line of work focus on developing rules or selecting the best forecast model or averaging the models according to the historical features

of the data (e,g, Collopy and Armstrong, 1992; Meade, 2000; Wang et al., 2009; Petropoulos et al., 2014). A recent implementation of feature-based forecasting was proposed by Talagala et al. (2018). They used 42 time series features to train a random forecast classifier to select the best forecasting method. Montero-Manso et al. (2020) used the same set of features to estimate optimal combination weights through an algorithm named eXtreme Gradient Boosting (XGBoost, Chen and Guestrin, 2016). Their approach, FFORMA (Feature-based FORecast Model Averaging), placed second in the M4 competition. Both of these approaches used meta-learning, meaning that a group (reference set) of series is used to model the links between the time series and the out-of-sample performance of the available forecasting models. Then, given a new series and its features, the most suitable model is selected, or a set of weights for a forecast combination is estimated.

Regardless of the task (model selection or model combination), a common challenge in feature-based forecasting is the choice and estimation of time series features. Time series features vary from tens to thousands (Fulcher and Jones, 2014; Hyndman et al., 2019), and choosing different sets of features will inevitably result in different forecasts and varied performance. Moreover, the studies reviewed above focus on extracting such features by using the historical, observed data of each time series. For example, two commonly used features are the strength of a trend and the strength of the seasonality. The estimation of these two features is not unique, because of the existence of several approaches that are typically based on different assumptions. However, even if there was one acceptable way to define them, it would become inadequate because existing features are estimated based on observed historical patterns that are likely to change over time. Moreover, the estimation of some of the features might not be feasible or robust in the case of a limited number of available past observations. Finally, when the chosen features involve large numbers, this might increase the computational time required to estimate them.

In this paper, we suggest a change of the focus in producing forecasts from extraction of time series features from historical data. Specifically, we suggest the use the out-of-sample forecasts from a pool of models and measure their *diversity*, a feature that has been identified as a crucial factor in improving the performance of forecast combinations (Thomson et al., 2019; Lichtendahl and Winkler, 2020). Through meta-learning, we use a group of series to model the diversity of their forecasts and the optimal combination weights by minimizing the total forecasting loss. Once the model has been trained, and for any new series that needs to be forecast, we can calculate the combination weights based on the diversity of their forecasts produced by the models in the pool and produce both point forecasts and prediction intervals.

We empirically show that a single feature, the *diversity* of the forecasts, is sufficient to achieve levels of postsample performance similar to those of large set of features derived from estimates based on historical data.

Our study is in line with other studies that exploit information from the forecasts without utilization of forecast diversity. For example, Petropoulos and Siemsen (2020) proposed forecast representativeness and derived a new selection criterion that works remarkably well in cases of low signal to noise ratios in comparison with other established selection criteria. The ability of the representativeness criterion to more often than not select the best (and avoid the worst) models leads to significant accuracy improvement both in selecting single models and combinations across models. Zhao and Feng (2020) also used postsample forecasts as the input in a machine learning model as a step toward enhancing the performance of standard statistical time series forecasting models. Although they also use forecasts as a feature, they do not explicitly focus on a specific aspect of the forecasts (such as diversity), making it difficult to obtain insights into why their approach performs well. They also did not discuss how prediction intervals may be estimated.

Forecasting is vital for the efficient operation of supply chains (Tliche et al., 2020; Ali et al., 2017) as well as other operations-related decisions. To support the efficacy of our proposition for decision-making purposes, we offer not only large-scale empirical evaluations for the mean (point) forecasts but also for the uncertainty around this mean, in terms of quantile forecasts. We also provide trade-off curves based on upper coverage levels versus upper prediction intervals that approximate utility functions related to inventory forecasting.

The rest of this paper is organised as follows. In Section 2, we describe the calculation of forecast diversity for forecast combinations and present a framework of forecasts with forecasts. We demonstrates the superiority of the proposed approach via extensive experiments in Section 3. Section 5 provides our discussions and Section 6 concludes the paper.

2. Forecast combination: diversity matters

2.1. Diversity of forecasts

Ambiguity Decomposition (Krogh and Vedelsby, 1994) indicates in the literature on machine learning that accuracy and diversity are two main factors that should be taken into consideration when designing ensembles. In the forecasting community, many studies have emphasized the importance of the forecasting method's diversity pool when constructing forecast combinations (Bates and Granger, 1969; Batchelor and Dua, 1995; Thomson et al., 2019). Lichtendahl and Winkler (2020), in exploring why some combinations performed better than others in the recent

M4 competition (Makridakis et al., 2018), also identified diversity as an important factor for efficient forecast combinations, along with the robustness of the individual models.

Ambiguity Decomposition can be easily applied to the forecast combination task. For a given time series $\{y_t, t = 1, 2, \cdots, T\}$, we denote the h-th step forecast produced by the i-th individual method as f_{ih}, where $i = 1, 2, \cdots, M$ and $h = 1, 2, \cdots, H$. Furthermore, M and H are the number of algorithms in the forecast pools and the forecast horizon, respectively. Let f_{ch} be the h-th step combined forecast given by $\sum_{i=1}^{M} w_i f_{ih}$, where w_i is the combination weight for the i-th method. The overall mean squared error of a weighted forecast combination model MSE_{comb} over the whole forecast horizon H can be written as follows.

$$\begin{aligned} MSE_{comb} &= \frac{1}{H}\sum_{i=1}^{H}\left(\sum_{i=1}^{M} w_i f_{ih} - y_{T+h}\right)^2 \\ &= \frac{1}{H}\sum_{i=1}^{H}\left[\sum_{i=1}^{M} w_i (f_{ih} - y_{T+h})^2 - \sum_{i=1}^{M} w_i (f_{ih} - f_{ch})^2\right] \\ &= \frac{1}{H}\sum_{i=1}^{H}\left[\sum_{i=1}^{M} w_i (f_{ih} - y_{T+h})^2 - \sum_{i=1}^{M-1}\sum_{j>i}^{M} w_i w_j (f_{ih} - f_{jh})^2\right] \\ &= \sum_{i=1}^{M} w_i MSE_i - \sum_{i=1}^{M-1}\sum_{j>i}^{M} w_i w_j Div_{i,j}, \end{aligned} \tag{1}$$

where MSE_i represents the mean squared error for the i-th method. $Div_{i,j}$ denotes the degree of diversity between the i-th and j-th method in the forecast method pool, which is defined as follows.

$$Div_{i,j} = \frac{1}{H}\sum_{i=1}^{H}(f_{ih} - f_{jh})^2. \tag{2}$$

Equation (1) says that the mean squared error of the combined forecast is guaranteed to be less than or equal to the weighted mean squared error of the individual forecasts. The second term in the last line of Equation (1) tells us how diverse the individual forecasts are. Out of two combination methods with identical weighted mean squared error, the one with greater diversity will have a lower overall squared error. That is, the more diversity existing in the forecast method pool leads to overall better forecasting accuracy.

2.2. Diversity for forecast combination

How can we exploit the diversity information among different forecasting methods for forecast combination? As Montero-Manso et al. (2020) and Kang, Hyndman and Li (2020) point out, we can use a group of series to estimate the forecast combination weights by linking a set of time series features with the forecasting performances of the individual methods. The key point is to find a set of features that can represent the information affecting forecasting performance.

We propose to use the pairwise diversity measures as a proper set of features to represent the forecast diversity among different methods.

To make the diversity comparable between time series with different scales, we can scale the diversity measure in Equation (2) by averaging across all pairs of methods. We use the scaled diversity in Equation (3) for all the experiments in the following sections.

$$sDiv_{i,j} = \frac{\sum_{h=1}^{H}(f_{ih} - f_{jh})^2}{\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\left[\sum_{h=1}^{H}(f_{ih} - f_{jh})^2\right]}. \tag{3}$$

Figure 1 shows the diversity extraction procedure in the context of point forecasting. Given a time series data set $\{y_t^{(n)}\}_{n=1}^{N}$, for each time series $y_t^{(n)}$, its h-th step point forecast produced by the m-th method is denoted as $f_{mh}^{(n)}$, where $m = 1, 2, \cdots, M$ and $h = 1, 2, \cdots, H$. Therefore, we can get an $M \times H$ matrix for each time series to forecast, thus, representing the forecasts produced by the M methods for the entire forecasting horizon. Then the pairwise forecast diversity among the M methods can be calculated by using Equation (3). Thus, for each time series $y_t^{(n)}$, we get an $M \times M$ symmetric matrix. We then concatenate the diversity measures in the lower diagonal into a vector. This can be used as a feature vector for time series $y_t^{(n)}$ when estimating the corresponding forecast combination weights based on the feature-based forecasting framework. For a forecasting pool containing M methods, we can construct $M(M-1)/2$ pairwise diversity measures to extract the independent information in the pool.

The same procedure can be extended to the context of interval forecasting. For a forecasting pool containing M methods, we can construct $M(M-1)/2$ pairwise diversity measures for both the upper and lower prediction intervals.

Assuming that we need to forecast N time series, we obtain $N \times (M(M-1)/2)$ matrices (for the point forecasts and the upper and lower prediction intervals), where each row can be viewed as a diversity (feature) vector for the corresponding series.

2.3. Forecast with forecasts: the framework

To construct a forecast model using diversity, we tailor the state-of-the-art feature-based forecast model averaging (FFORMA) framework proposed by Montero-Manso et al. (2020) to allow for the diversity of the forecasts as the inputs. Next, we estimate the combination weights based on the diversity information of the out-of-sample forecasts produced by the pool of methods being combined.

In the original FFORMA framework, estimations of forecast combination weights is implemented by finding a function to assign weights to each forecasting method. To forecast with

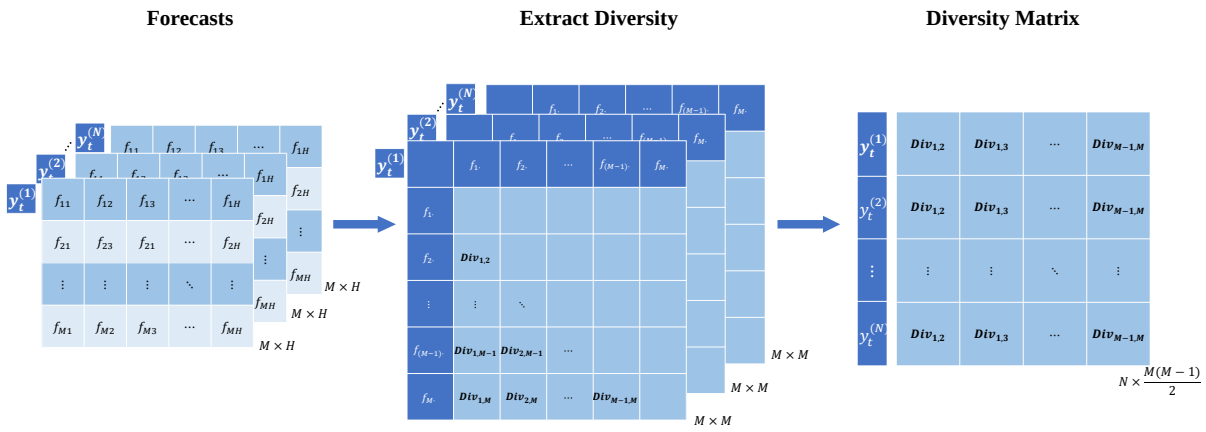

Figure 1: Diversity extraction from forecasts.

FFORMA, one first needs to extract 42 features from the original time series and calculate the overall weighted average (OWA) error of each forecast algorithm in the forecast pool. Then, to obtain the optimal combination weights, features are linked with the OWA errors using the XGBoost algorithm, which is an ensemble machine learning algorithm and can deal with regression or classification problems by integrating plenty of decision tree models. We find that FFORMA has the following drawbacks: (1) a manual selection of features is required, with specific features more appropriate than others in some applications and contextss, and (2) because FFORMA extracts features based on historical data, it is not applicable to time series with an inadequate length of historical data; in such cases the estimations of features may be unreliable.

To address the above problems, we propose a diversity-based forecast combination framework. It consists of two phases, namely model-training and forecasting, as shown in Figure 2. In the model-training phase, we take all the series in a given data set (for which forecasts are required) as the reference data. Each time series in the reference data is split into training and testing periods. The length of the testing period for each series is the same as its forecasting horizon. We apply the forecasting methods in the pool by using the training periods, and extract the diversity matrix following Figure 1 from the forecasts produced by different forecasting methods on the testing periods. We then calculate the forecasting errors of each method and summarize them using an error metric. Finally, a forecast combination model is trained, by minimizing the total forecasting loss, to estimate the combination weights for each series as a function of its forecast diversity. Once the model has been trained, weights can be produced for any target series given the diversity of its forecasts produced by the method pool.

We present a detailed procedure forf constructing diversity-based forecasting combinations in Algorithm 1. Using the XGBoost algorithm, the following optimization problem is solved to

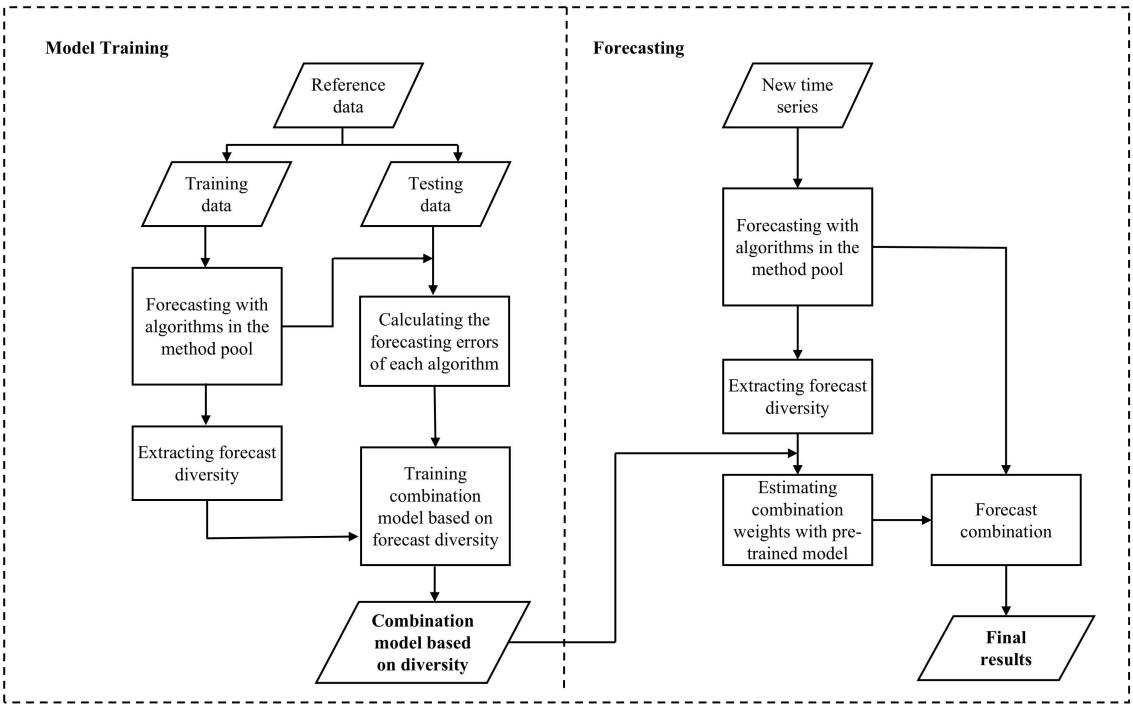

Figure 2: The framework of forecast with forecasts.

obtain the combination weights:

$$\arg\min_{w} \sum_{n=1}^{N} \sum_{i=1}^{M} w(Div_n)_i \times \mathrm{Err}_{ni}, \tag{4}$$

where Div_n indicates the forecast diversity of the n-th time series, $w(Div_n)_i$ is the combination weight assigned to method i for the n-th time series based on the diversity, and Err_{ni} is the error produced by method i for the n-th time series. The combination weight $w(Div_n)_i$ is obtained by the output of XGBoost model after a soft-max transformation:

$$w(Div_n)_i = \frac{\exp\{y(Div_n)_i\}}{\sum_{i=1}^{M} \exp\{y(Div_n)_i\}},$$

where $y(div_n)_i$ is the regression results from XGBoost. In other words, the key point of the combination model based on diversity is to search for the optimal weights that can minimize the weighted error. After obtaining the combination weights, we can calculate the point and interval forecasts as below.

$$\begin{aligned} f_n &= \frac{1}{M} \sum_{i=1}^{M} w_{ni} f_{ni}, \\ f_n^u &= \frac{1}{M} \sum_{i=1}^{M} w_{ni} f_{ni}^u, \\ f_n^l &= \frac{1}{M} \sum_{i=1}^{M} w_{ni} f_{ni}^l, \end{aligned} \tag{5}$$

where w_{ni} is the estimated weight for the n-th time series and the i-th method. And f_{ni}, f_{ni}^{u}, f_{ni}^{l} are the point, upper, and lower forecast values of the i-th algorithm for h-th forecast step, respectively.

Algorithm 1 The framework of forecasts with forecasts.

Phase 1: Model training

Input: y_{ref}: a time series reference set; a forecasting pool consisting of M methods.

Output: Forecast combination model based on diversity.

1: **for** $y_t^{(n)} \in y_{\text{ref}}$ **do**

2: Split time series $y_t^{(n)}$ into training and testing periods.

3: Produce the forecasts using the M methods.

4: Extract the diversity vector: Div_n (see Figure 1).

5: Calculate the forecasting errors of each method in the pool on the testing data.

6: **end for**

7: Estimate the combination model based on diversity with XGBoost, by minimizing the weighted errors:

$$\arg\min_{w} \sum_{n=1}^{N} \sum_{i=1}^{M} w(Div_n)_i \times \text{Err}_{ni}.$$

// Finish building combination model.

Phase 2: Forecasting

Input: Pretrained model; y_{new}: a time series data set to be forecast.

Output: Final forecasts of new time series y_{new}.

8: **for** $y_t^{(m)} \in y_{\text{new}}$ **do**

9: Produce forecasts using the methods in the forecasting method pool.

10: Extract the diversity vector: Div_m (see Figure 1).

11: Use the pretrained model to produce the optimal weight $w(Div_m)_i$ for method i.

12: Combine the individual forecasts using $w(Div_m)_i$ and obtain the final forecasts.

13: **end for**// *Obtain final results.*

The merits of using diversity for forecast combinations are twofold. First, the process of extracting diversity is straightforward and interpretable. The algorithm of measuring the diversity between different methods involves a simple calculation, and hence, it can reduce the computational complexity when extracting features. Meanwhile, diversity-based forecast combinations require only the forecasts from individual methods, avoiding exploiting information from other models as in Montero-Manso et al. (2020). Secondly, although traditional methods of time series feature extraction (Fulcher and Jones, 2014; Hyndman et al., 2019; Christ et al.,

2018) usually depend on the manual choice of an appropriate set of features, our approach can be applied automatically without the need for expert knowledge and human interaction.

3. Empirical evaluation

3.1. Data

We used the M4 data set (Makridakis et al., 2020) to evaluate the forecasting performance of the proposed diversity-based forecast combination method in terms of both point and interval forecasting. The M4 data contains 100,000 time series with different seasonal periods from different domains such as demographics, finance, and industries. The lengths of the yearly, quarterly, monthly, weekly, daily, and hourly data lie in the ranges of [13, 835], [16, 866], [42, 2794], [80, 2597], [93, 9919] and [700, 960], respectively. The corresponding forecasting horizons are 6, 8, 18, 13, 14, and 48. The data set is publicly available in the `M4comp2018` R package (Montero-Manso et al., 2018). We optimized the combination weights separately for each frequency by using the respective M4 series to form the reference data (see Section 2.3 for more details).

3.2. The forecasting pool of methods

Our forecasting pool consists of eight individual forecast methods; they are described in Table 1. Note that compared with the commonly used nine individual methods in recent forecast combination studies (Talagala et al., 2021; Montero-Manso et al., 2020; Li et al., 2020; Kang, Hyndman and Li, 2020), we do not include the neural network time series forecasting method (nnetar) because it does not produce prediction intervals. The eight forecasting methods in our pool are implemented in the `forecast` package in R (Hyndman et al., 2020).

3.3. Diversity extraction

Since we are producing both point and interval forecasts, we calculate the diversity features based on the upper and lower prediction intervals. Considering the eight individual forecasting methods in the pool, we have 28 diversity features for upper prediction intervals, and 28 features for lower intervals. Therefore, in total, 56 diversity features are calculated for each time series. Note here we are not using the diversity of point forecasts, which are the midpoints of the prediction intervals for all the eight methods and do not contain more information than the upper and lower intervals.

Table 1: The methods used for forecast combination. All these methods are implemented using the `forecast` **R** package.

Forecasting Method	**R** implementation
auto_arima: the ARIMA family of models (Hyndman and Khandakar, 2008).	`auto.arima()`
ets: the exponential smoothing state space family of models (Hyndman et al., 2002).	`ets()`
tbats: the exponential smoothing state space model with a Box-Cox transformation, ARMA errors, trend and seasonal components (De Livera et al., 2011).	`tbats()`
stlm_ar: seasonal and trend decomposition using Loess with AR modeling of the seasonally adjusted series.	`stlm()` with `modelf = ar`
rw_drift: random walk with drift.	`rwf()` with `drift=TRUE`
thetaf: the theta method (Assimakopoulos and Nikolopoulos, 2000).	`thetaf()`
naïve: the naïve method.	`naive()`
snaïve: the seasonal naïve method.	`snaive()`

3.4. Forecasting evaluation metrics

We use the mean absolute scaled error (MASE, Hyndman and Koehler, 2006) to evaluate the point forecasts produced by our proposed combination model. MASE compares the forecast accuracy between a specific forecast algorithm and the naïve method. It is defined as

$$\text{MASE} = \frac{1}{H} \frac{\sum_{h=1}^{H} |f_h - y_{T+h}|}{\frac{1}{T-m}\sum_{t=m+1}^{T} |y_t - y_{t-m}|},$$

where H is the forecasting horizon, T is the length of the historical data, m is the frequency of the data, y_{T+h} is the actual value of the time series at time $T+h$, and f_h is the forecast value at the h-th step.

To assess the performances of the generated prediction intervals, we use the mean scaled interval score (MSIS, Gneiting and Raftery, 2007), as used in the M4 competition and other recent studies on forecast uncertainty estimation (e.g., Spiliotis et al., 2020; Kang, Spiliotis, Petropoulos, Athiniotis, Li and Assimakopoulos, 2020). The definition of MSIS is as follows.

$$\text{MSIS} = \frac{1}{H} \frac{\sum_{h=1}^{H} \left\{(U_h - L_h) + \frac{2}{\alpha}(L_h - y_{T+h})\mathbb{1}\left\{y_{T_h} < L_h\right\} + \frac{2}{\alpha}(y_{T+h} - U_h)\mathbb{1}\left\{y_{T+h} > U_h\right\}\right\}}{\frac{1}{T-m}\sum_{t=m+1}^{T} |y_t - y_{t-m}|},$$

where $[L_h, U_h]$ are the generated $(1-\alpha)100\%$ prediction intervals at the h-th step, and $\mathbb{1}$ is the

indicator function, which equals to 1 when y_{T+h} is within the postulated interval and returns 0 otherwise.

In the training process of Algorithm 1, we apply a new cost function that takes both point and interval forecasting into consideration when optimizing the combination weights. It is defined as

$$\text{Err} = \frac{1}{2}\left(\frac{\text{MASE}}{\text{MASE}_{\text{naive2}}} + \frac{\text{MSIS}}{\text{MSIS}_{\text{naive2}}}\right), \tag{6}$$

where $\text{MASE}_{\text{naive2}}$ and $\text{MSIS}_{\text{naive2}}$ are the MASE and MSIS values of the naive2 forecasting method, derived from the naïve method on the seasonally adjusted data (Makridakis et al., 2018). In this way, we are using the same group of features for both point and interval forecasts (see Section 3.3) and the same cost function for estimating the combination weights. The point and interval forecasts are then calculated following Equation (5).

3.5. Point and interval forecasting performance

We compared the performance of point and interval forecasts of the proposed diversity-based forecast combination approach as shown in Algorithm 1 against the FFORMA approach (Montero-Manso et al., 2020) that uses XGBoost to link 42 statistical time series features with forecast errors. We also benchmarked against a simple average (SA) approach, where the forecasts from all methods in the forecasting pool are combined with equal weights. Table 2 depicts the mean of the forecast errors across series from each frequency. Entries in bold highlight that our method outperforms the FFORMA approach. We can see that both our proposed method (Diversity) and FFORMA outperform SA. Overall, the mean values of MASE and MSIS of the proposed approach are 18.71% and 19.92% lower, respectively, compared with SA. In other words, the unequal weights produced by diversity-based modeling are effective and can help tackle the "forecast combination puzzle." More importantly, Diversity, without extracting and selecting sophisticated times series features, outperforms FFORMA when we focus on the mean MASE and MSIS values of the overall M4 data. Diversity outperforms FFORMA in most frequencies of data, apart from the weekly and daily data.

To further verify the performance of Diversity, we consider an approach that combines the features from Diversity and FFORMA. We combine the 42 statistical features and the 56 diversity features into a single set of time series features and use them as the inputs into XGBoost to obtain the optimal weights for the nine forecasting methods. The results for this approach, FD (FFORMA + Diversity), are also represented in Table 2. We observe that FD performs similarly with Diversity.

Note that in Table 2, FFORMA, Diversity and FD are using the same pool of forecasting methods and cost functions as described in Section 3.2 and Section 3.4 to make them comparable.

Table 2: Comparison of the mean MASE and MSIS values from our diversity-based forecast combination method (Diversity), FFORMA and forecast combination with simple averaging (SA). Entries in bold highlight that our method outperforms the FFORMA approach.

Method	Overall	Yearly	Quarterly	Monthly	Weekly	Daily	Hourly
				MASE			
SA	1.9040	3.6907	1.2432	0.9813	6.3826	5.8921	3.3319
FFORMA	1.5586	3.0842	1.1220	0.8980	**2.2309**	3.2464	0.8822
Diversity	**1.5478**	**3.0670**	**1.1095**	**0.8915**	2.2744	**3.2296**	**0.8540**
FD	**1.5507**	**3.0615**	**1.1096**	0.8997	2.2639	**3.2345**	**0.8574**
				MSIS			
SA	17.5077	42.0776	9.9248	8.3012	22.4778	31.5910	11.4214
FFORMA	14.5934	32.0185	9.2388	7.8189	**16.0496**	**27.7694**	6.6161
Diversity	**14.0197**	**30.3312**	**8.7805**	**7.6385**	16.4015	28.0220	**6.3587**
FD	**14.0254**	**30.3980**	**8.7995**	**7.6248**	16.0936	27.8723	**6.3145**

Therefore, the results for FFORMA are slightly different from those in Montero-Manso et al. (2020) where a different pool (including NNETAR) and cost functions (OWA) are used.

To investigate the statistical significance of the performance differences, we performed Multiple Comparisons from the Best (MCB) test (Koning et al., 2005) on each data frequency separately but also over all the M4 series. The aim was to test whether the average ranks of each forecasting method are significantly different from the others. The MCB test was applied based on the MASE errors as shown in Figure 3. One can read the results as follows. Lower average ranks are better, althouogh the performance differences between any two methods are not significant if their confidence intervals overlap.

According to Figure 3, we observe:

- Although Diversity outperforms FFORMA on average, their differences are not significant (see the top panel, "Overall"). However, Diversity is more intuitively appealing and easier to compute.
- Overall, FD performs significantly better than Diversity, FFORMA, and SA. In other words, the diversity features, focusing on future forecasts, could bring significant improvements to combination forecasts based on traditional time-series statistical features, which only represent information from time series historical values. That suggests that the two groups of features could complement each other when used for estimating weights for forecast combinations.

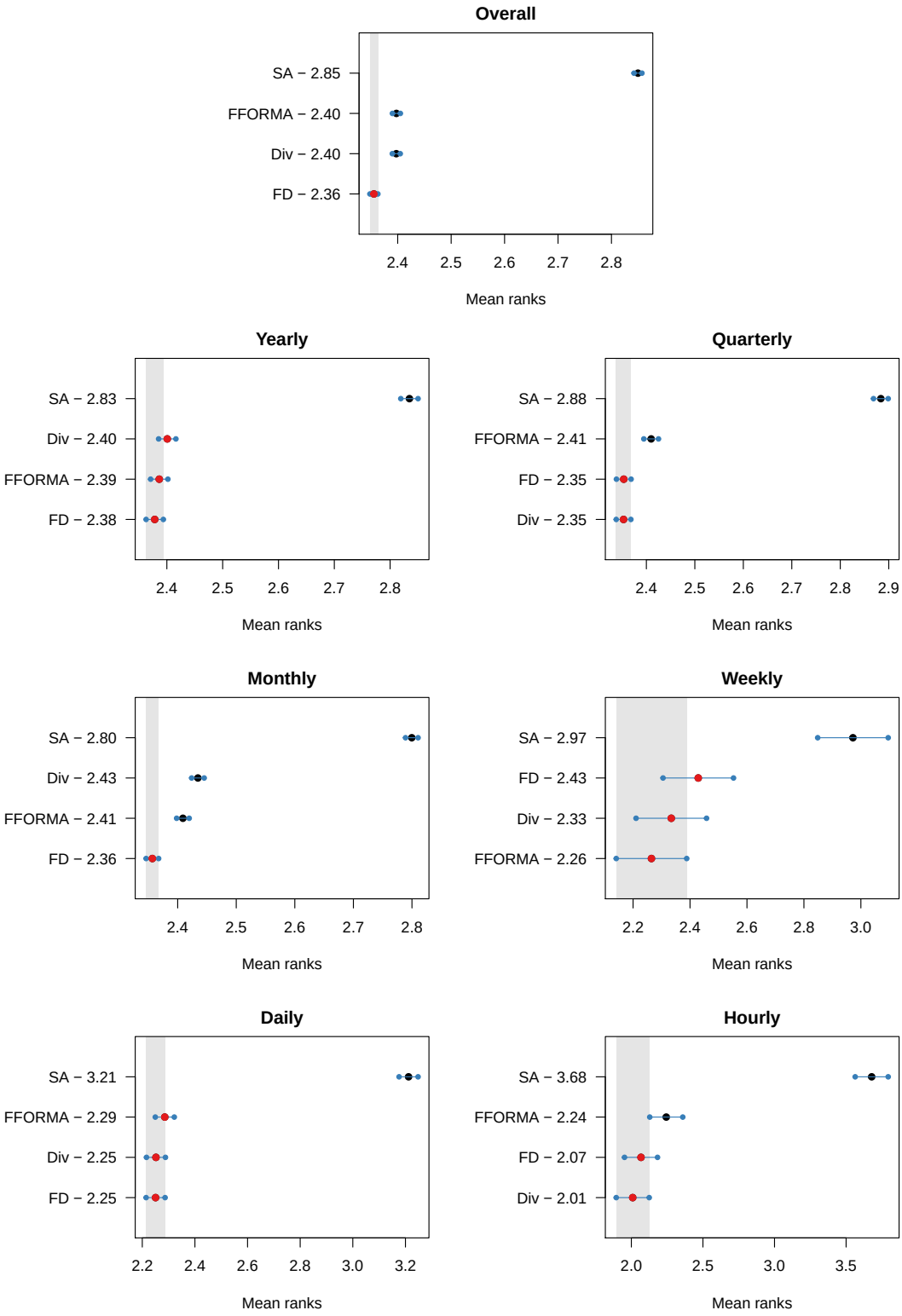

Figure 3: MCB tests on the ranks of the MASE errors of SA, FFORMA, Diversity and FD for each data frequency separately and across all frequencies (Overall).

- Looking at the MCB results for different frequencies, in most cases FD or Diversity outperforms FFORMA. SA is always significantly outperformed by the other three approaches.

3.6. Trade-off curves

We approximate the performance of the proposed Diversity approach on inventory forecasting by assuming that the prediction intervals produced are directly used for inventory-related decisions. We produce prediction intervals for various confidence levels: 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% and 99%. We then consider two quantities. First, we measure the upper coverage level (the percentage of times where the actual outcome is below the corresponding upper prediction interval, i.e., the 97.5% percentile for prediction intervals produced at a 5% confidence level). The average upper coverage level is a proxy of the average achieved service level, as it effectively shows the percentage of times that a demanded product was in stock (see also Svetunkov and Petropoulos, 2018) if the prediction intervals of the forecasts were directly used for inventory decisions. Second, we calculate the average upper prediction interval across time series after scaling with the mean value of the historical data for each series. This is a proxy of the holding cost that would be required to achieve the respective service level (see also Svetunkov and Petropoulos, 2018; Petropoulos and Siemsen, 2020). Rendering the upper prediction interval scale-independent is required as the different series in our data set refer to different quantities (hundreds versus thousands versus tens of thousands units) and we are interested to explore the average effect across all series.

The trade-off curves for these two quantities (scaled upper prediction interval versus upper coverage), for Diversity, SA and the eight individual forecasting methods in the pool, are presented in Figure 4. The results for each data frequency are presented in a different panel. One can read these graphs as follows (Petropoulos et al., 2019). Assuming a vertical line (similar upper prediction interval, i.e., holding cost), the methods that achieve higher upper coverage levels (higher achieved service levels) should be preferred. Similarly, assuming a horizontal straight line (similar upper coverage levels), the methods with lower scaled upper prediction intervals (lower costs) are better. It can be seen that the Diversity approach offers a very competitive trade-off between the two quantities considered that outperfroms all other methods for all frequencies. The only exception is the yearly frequency where Theta method performs well in terms of upper prediction interval values but cannot reach high upper coverage levels.

4. Case study: forecasting fast moving consumer goods

To highlight the usefulness of diversity in practice, we considered the forecasting of sales of fast-moving consumer goods (FMCG) from a major North American food manufacturer. We focused on the sales of stock keeping units (SKUs) in two countries, the USA and Canada. The sales are recorded according to monthly frequency and consist of 51 periods, from April 2013 to

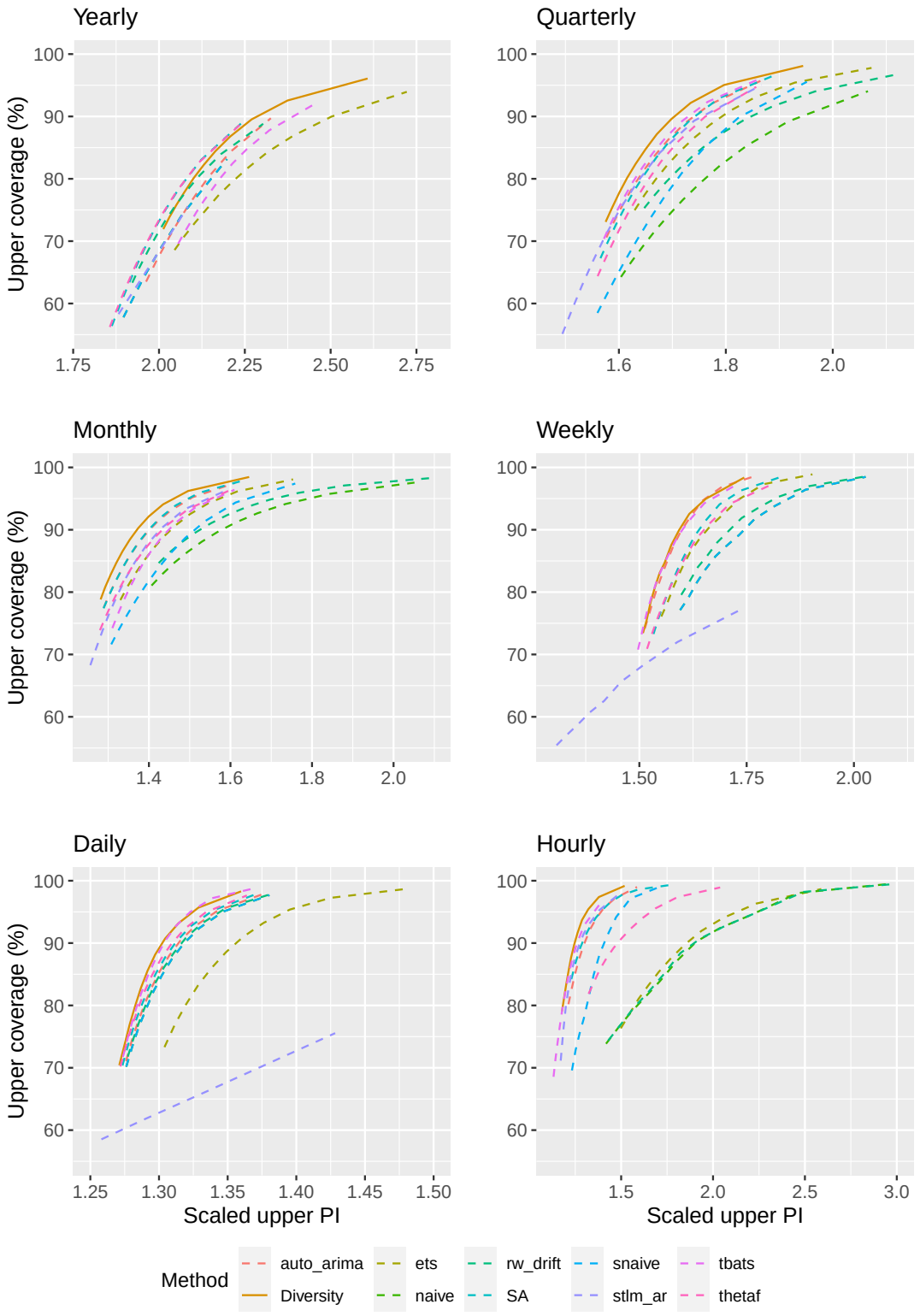

Figure 4: The upper coverage versus the scaled upper prediction intervals across different confidence levels for Diversity, SA and the eight individual methods, for each data frequency separately. Diversity is shown in solid lines while the other methods are shown in dashed lines.

June 2017. We set the forecast horizon to 12 and split each series into a training (27 periods), validation, and test sets (12 periods each). Some time series started with zero sales values,

which we trimmed. In those cases in which this trimming process resulted in a training set of fewer than two full seasonal cycles (less that 24 months), that time series was dropped from the set because it would not be feasible to use to estimate any seasonal patterns. The above process resulted in 955 unique combinations of SKU $\times$ location.

The model training part of our framework (see also Figure 2) was completed using the forecasts corresponding to the observations in the validation set (observations 28 to 39). The method pool, cost function, and the diversity extraction procedure used in the training process were the same as those used in Section 3. Once we estimated the combination model based on diversity, we then applied it to the forecasts for the test set (periods 40 to 51), where we also used MASE and MSIS to measure the out-of-sample performance. The same approach was used for the FFORMA and FD approaches, similar to the results in Section 3.5.

The results from our case study are presented in Table 3. Much like our main empirical results, we observed that the proposed combination approaches based on the diversity of the forecasts alone (Diversity) or on diversity combined with other times series features (FD) outperformed FFORMA and SA, in both terms of point forecast accuracy and estimation of the prediction intervals. It is important to highlight that the results of this case study were based solely on the 955 FMCG series for both the model training and the forecasting phases of our proposed framework (see also Algorithm 1), thus showcasing that Diversity does not require massive reference data sets to estimate a diversity-based combination model.

Table 3: Comparison of the mean MASE and MSIS values from our diversity-based forecast combination method (Diversity), FFORMA and forecast combination with simple averaging (SA) on the FMCG data. Entries in bold highlight that our method outperforms the FFORMA approach.

Method	SA	FFORMA	Diversity	FD
MASE	0.9555	0.9599	**0.9365**	**0.9367**
MSIS	8.5085	8.1254	**8.0066**	**7.9189**

5. Discussion

Feature-based forecast model selection and combinations face the challenge of selecting an appropriate set of time series features that vary according to different domains and forecasters (Fulcher and Jones, 2014; Wang et al., 2021; Kang, Hyndman and Li, 2020). More importantly, features' estimation is unreliable when historical data is limited (e.g., for fast-moving products), or even unavailable when there is no history at all (e.g., for new products). This study proposes to forecast with forecasts without manually choosing time series features, yielding comparable

performance with top contestants in the M4 competition data with regard to both point forecasts and prediction intervals.

The performance of forecast combinations is potentially highly related to the degree of diversity among the individual forecasts (Armstrong, 2001; Thomson et al., 2019). This study explores how to further improve forecast combinations and attempts to tackle the forecast combination puzzle by exploiting the forecast diversity in the forecasting method pool. The proposed diversity-based forecast combination can automatically control the combination via measuring the pairwise diversity between forecasts from different sources and linking them, via a meta-learner, to the accuracy of a test set. If the pool of available forecasts has low diversity, then our approach will approximate an equal-weight combination approach, which is an appropriate strategy with a lack of additional information. However, if the diversity across the candidate forecasts is high, then the combination weights will be modeled using two important factors–diversity and accuracy–in arriving at efficient ensembles (Lichtendahl and Winkler, 2020).

Our empirical results show that the diversity information among individual forecasts used for combination is informative in allocating the combination weights. In fact, using forecast diversity as the sole time-series feature results in performance that is equivalent to using an array of time-series features calculated on historical (in-sample) data. Our approach is not only faster to compute, but also simpler and more straightforward because it does not involve decisions related to which features to include and how to compute them. As such, it is in-line with the simplicity argument of Green and Armstrong (2015). In addition, forecast diversity can be used in conjunction with other established time-series features to boost forecasting performance.

Research on diversity-based regression/classifier ensembles in the machine learning literature is in line with our findings. Specifically, Liu and Yao (1999) used negative correlation learning to create an ensemble with negatively correlated networks and encourage their specialization and cooperation. Kuncheva and Whitaker (2003) improve the ensemble accuracy by measuring the diversity in classifier ensembles. Mendes-Moreira et al. (2012) reviewed the ensemble approaches for regression and emphasized the importance of regression diversity. Our study aligns with this line of research in the sense that it aims to improve forecast ensembles by exploiting forecast diversity in the ensemble.

The good performance of our proposition is a result of two factors. First, we built on the rich and established literature on forecast combinations and explicitly took into account one of the critical elements in building effective forecast combinations: the diversity of the forecasts.

Second, and in line with the arguments made in the study of Petropoulos and Siemsen (2020), we explicitly considered the output of the forecasting models, i.e., the out-of-sample forecasts, rather than simply focusing on information that relates to the features and characteristics of the in-sample data (like in Montero-Manso et al., 2020) or how well forecasting models fit the in-sample data. The use of the out-of-sample forecasts toward making forecast-related decisions is crucial. In Petropoulos and Siemsen (2020), the evaluation of the representatives of the out-of-sample forecasts to the actual situation enabled the acceptance or rejection of some forecasting models (i.e., judging models by their outputs). In our context, out-of-sample forecasts informed our algorithmic calculations for obtaining weights for forecast combinations by amplifying the diversity of the final pool of methods being combined.

Another advantage of the proposed approach is its nonreliance on specific families of models. In this study, we focused on linear statistical methods. However, nonlinear methods could be part of the pool of models in other contexts. Moreover, our approach can be applied equally to both statistical and judgemental forecasts or even to a combination of the two. Although in this study we focused on exploring the benefits of forecast diversity in the context of statistically produced forecasts, our approach could be extended toward combining forecasts from experts, where prior research suggests that the performance of equal weights is hard to beat (Genre et al., 2013). Our suggestion to test diversity-based forecast combinations also extends the research of Grushka-Cockayne et al. (2017), who showed that combinations based on trimmed means (excluding the top and bottom $x\%$ forecasts) could improve the accuracy of overfitted and overconfident forecasts.

Although the proposed method improves point and interval forecasting, a possible future research path is to extend it to probability density forecasting. Meta-learning can presumably be used to produce a weighted mixture of the forecast distributions from multiple models or to generate a weighted average of the forecast distributions. Diversity could be measured based on probability distribution distances (e.g., Kullback–Liebler divergence). The loss function in meta-learning could also be adapted to density forecasting principles (e.g., calibration and sharpness).

One limitation of the current study is that it uses all the available individual forecasts without pooling the most heterogeneous forecasts. Although, naturally, some combination weights will be close to zero, excluding per series models with poor performance could further improve the forecasting performance. Future research could examine how diversity-based forecast combinations can be extended to the whole spectrum of selection-pooling-combinations (Cang and Yu, 2014; Kourentzes et al., 2019). Besides, the impact of adding, removing, and selecting

different methods in the approach is also a valuable research topic.

Another limitation of our study is that our empirical results on the M4 data are, in theory, not directly comparable with any of the original contestants of the M4 competition. Access to the postsample data allows for experimentation, testing, and hyper-parametrization, none of which were available to the original participants. Although we do not directly use future data values to inform our forecasts, we still cannot claim that our approach would have performed better than FFORMA in the M4 competition because our approach was never submitted. Even if we had also applied the proposed approach to a real application, the same forecasting method pool is used with most of the methods being drawn from results of societal activities and being used with linear approaches of various types. As such, a final avenue for future research would be to compare our framework against the performance of other established benchmarks in more/new data sets, including data with nonlinear and intermittent patterns.

6. Conclusion

In this paper, we proposed to use forecasts to improve forecasting performance. In essence, we measure the pairwise diversity among the forecasts from the methods in a pool for each time series and use these measurements as a group of features linked with the out-of-sample forecasting performances of the individual forecasting methods. We show that our approach–simply using forecast diversity–achieves equivalent performance to the use of manually selected time series features calculated from historical data. Combining diversity and other statistical features (depicting future and historical information, respectively) can be further advantageous.

Our approach provides an automatic and flexible tool for forecasting practice. Our proposed approach has the following merits. First, forecasters do not need to tackle the issue of feature selection when carrying out feature-based forecasting. The calculation of diversity is straightforward, easy to implement, and interpretable. Second, forecasts from any method (including statistical methods, nonlinear techniques, and judgment) can be easily used within our approach. Finally, our proposition offers improved point forecast accuracy coupled with better performance for interval forecasts.

Acknowledgments

The authors are grateful to the editors and three anonymous reviewers for helpful comments that improved the contents of the paper. Yanfei Kang is supported the National Natural Science Foundation of China (No. 72171011, and No. 72021001) and the National Key Research and Development Program (No. 2019YFB1404600) and Feng Li is supported by the Emerging

Interdisciplinary Project of CUFE and the Beijing Universities Advanced Disciplines Initiative (No. GJJ2019163). This research is supported by the high-performance computing (HPC) resources at Beihang University.

References

Ali, M. M., Babai, M. Z., Boylan, J. E. and Syntetos, A. A. (2017), 'Supply chain forecasting when information is not shared', *European Journal of Operational Research* **260**(3), 984–994.

Armstrong, J. S. (2001), Combining forecasts, *in* J. S. Armstrong, ed., 'Principles of Forecasting: A Handbook for Researchers and Practitioners', Springer US, Boston, MA, pp. 417–439.

Assimakopoulos, V. and Nikolopoulos, K. (2000), 'The Theta model: a decomposition approach to forecasting', *International Journal of Forecasting* **16**(4), 521–530.

Batchelor, R. and Dua, P. (1995), 'Forecaster diversity and the benefits of combining forecasts', *Management Science* **41**(1), 68–75.

Bates, J. M. and Granger, C. W. (1969), 'The combination of forecasts', *Journal of the Operational Research Society* **20**(4), 451–468.

Cang, S. and Yu, H. (2014), 'A combination selection algorithm on forecasting', *European Journal of Operational Research* **234**(1), 127–139.

Chen, T. and Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 785–794.

Christ, M., Braun, N., Neuffer, J. and Kempa-Liehr, A. W. (2018), 'Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – a Python package)', *Neurocomputing* **307**, 72 – 77.

Claeskens, G., Magnus, J. R., Vasnev, A. L. and Wang, W. (2016), 'The forecast combination puzzle: A simple theoretical explanation', *International Journal of Forecasting* **32**(3), 754–762.

Clemen, R. T. (1989), 'Combining forecasts: A review and annotated bibliography', *International Journal of Forecasting* **5**(4), 559 – 583.

Collopy, F. and Armstrong, J. S. (1992), 'Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations', *Management Science* **38**(10), 1394–1414.

Dawes, R. M. (1979), 'The robust beauty of improper linear models in decision making.', *American psychologist* **34**(7), 571.

De Livera, A. M., Hyndman, R. J. and Snyder, R. D. (2011), 'Forecasting time series with

complex seasonal patterns using exponential smoothing', *Journal of the American Statistical Association* **106**(496), 1513–1527.

De Menezes, L. M., Bunn, D. W. and Taylor, J. W. (2000), 'Review of guidelines for the use of combined forecasts', *European Journal of Operational Research* **120**(1), 190–204.

Fulcher, B. D. and Jones, N. S. (2014), 'Highly comparative feature-based time-series classification', *IEEE Transactions on Knowledge and Data Engineering* **26**(12), 3026–3037.

Genre, V., Kenny, G., Meyler, A. and Timmermann, A. (2013), 'Combining expert forecasts: Can anything beat the simple average?', *International Journal of Forecasting* **29**(1), 108–121.

Gneiting, T. and Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Green, K. C. and Armstrong, J. S. (2015), 'Simple versus complex forecasting: The evidence', *Journal of Business Research* **68**(8), 1678–1685.

Grushka-Cockayne, Y., Jose, V. R. R. and Lichtendahl, K. C. (2017), 'Ensembles of overfit and overconfident forecasts', *Management Science* **63**(4), 1110–1130.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. and Yasmeen, F. (2020), *forecast: Forecasting functions for time series and linear models*. R package version 8.12.
URL: *http://pkg.robjhyndman.com/forecast*

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

Hyndman, R. J. and Koehler, A. B. (2006), 'Another look at measures of forecast accuracy', *International Journal of Forecasting* **22**(4), 679–688.

Hyndman, R. J., Koehler, A. B., Snyder, R. D. and Grose, S. (2002), 'A state space framework for automatic forecasting using exponential smoothing methods', *International Journal of Forecasting* **18**(3), 439–454.

Hyndman, R., Kang, Y., Montero-Manso, P., Talagala, T., Wang, E., Yang, Y. and O'Hara-Wild, M. (2019), *tsfeatures: Time Series Feature Extraction*. R package version 1.0.1.
URL: *https://CRAN.R-project.org/package=tsfeatures*

Kang, Y., Hyndman, R. J. and Li, F. (2020), 'GRATIS: GeneRAting TIme Series with diverse and controllable characteristics', *Statistical Analysis and Data Mining* **13**(4), 354–376.

Kang, Y., Spiliotis, E., Petropoulos, F., Athiniotis, N., Li, F. and Assimakopoulos, V. (2020), 'Déjà vu: A data-centric forecasting approach through time series cross-similarity', *Journal of Business Research,* **132**, 719–731.

Koning, A. J., Franses, P. H., Hibon, M. and Stekler, H. O. (2005), 'The M3 competition:

Statistical tests of the results', *International Journal of Forecasting* **21**(3), 397–409.

Kourentzes, N., Barrow, D. and Petropoulos, F. (2019), 'Another look at forecast selection and combination: Evidence from forecast pooling', *International Journal of Production Economics* **209**, 226–235.

Krogh, A. and Vedelsby, J. (1994), Neural network ensembles, cross validation and active learning, *in* 'Proceedings of the 7th International Conference on Neural Information Processing Systems', NIPS'94, MIT Press, Cambridge, MA, USA, p. 231–238.

Kuncheva, L. I. and Whitaker, C. J. (2003), 'Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy', *Machine learning* **51**(2), 181–207.

Li, X., Kang, Y. and Li, F. (2020), 'Forecasting with time series imaging', *Expert System with Applications* **160**, 113680.

Lichtendahl, K. C. and Winkler, R. L. (2020), 'Why do some combinations perform better than others?', *International Journal of Forecasting* **36**(1), 142–149.

Liu, Y. and Yao, X. (1999), 'Ensemble learning via negative correlation', *Neural Networks* **12**(10), 1399–1404.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2018), 'The M4 competition: Results, findings, conclusion and way forward', *International Journal of Forecasting* **34**(4), 802–808.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2020), 'The M4 competition: 100,000 time series and 61 forecasting methods', *International Journal of Forecasting* **36**(1), 54–74.

Meade, N. (2000), 'Evidence for the selection of forecasting methods', *Journal of Forecasting* **19**(6), 515–535.

Mendes-Moreira, J., Soares, C., Jorge, A. M. and Sousa, J. F. D. (2012), 'Ensemble approaches for regression: A survey', *ACM computing surveys* **45**(1), 1–40.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J. and Talagala, T. S. (2020), 'FFORMA: Feature-based forecast model averaging', *International Journal of Forecasting* **36**(1), 86 – 92.

Montero-Manso, P., Netto, C. and Talagala, T. S. (2018), *M4comp2018: Data from the M4-Competition*. R package version: 0.1.0.

Mostaghimi, M. (1996), 'Combining ranked mean value forecasts', *European Journal of Operational Research* **94**(3), 505–516.

Perdisci, R., Gu, G. and Lee, W. (2006), Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems, *in* 'Sixth International Conference on Data Mining (ICDM'06)', IEEE, pp. 488–498.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb,

S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., De Baets, S., Dokumentov, A., Ellison, J., Fiszeder, P., Franses, P. H., Frazier, D. T., Gilliland, M., Sinan Gönül, M., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G. M., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., James Reade, J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Arenas, J. R. T., Wang, X., Winkler, R. L., Yusupova, A. and Ziel, F. (2021), 'Forecasting: theory and practice', *arXiv* **2012.03854**.

Petropoulos, F. and Kourentzes, N. (2015), 'Forecast combinations for intermittent demand', *Journal of the Operational Research Society* **66**(6), 914–924.

Petropoulos, F., Makridakis, S., Assimakopoulos, V. and Nikolopoulos, K. (2014), "Horses for Courses' in demand forecasting', *European Journal of Operational Research* **237**(1), 152–163.

Petropoulos, F. and Siemsen, E. (2020), 'Forecast selection and representativeness', *Working paper* .

Petropoulos, F. and Svetunkov, I. (2020), 'A simple combination of univariate models', *International Journal of Forecasting* **36**(1), 110–115.

Petropoulos, F., Wang, X. and Disney, S. M. (2019), 'The inventory performance of forecasting methods: Evidence from the M3 competition data', *International Journal of Forecasting* **35**(1), 251–265.

Reid, D. (1972), 'A comparison of forecasting techniques on economic time series', *Forecasting in Action. Operational Research Society and the Society for Long Range Planning* .

Rokach, L. (2010), 'Ensemble-based classifiers', *Artificial Intelligence Review* **33**(1-2), 1–39.

Smith, J. and Wallis, K. F. (2009), 'A simple explanation of the forecast combination puzzle', *Oxford Bulletin of Economics and Statistics* **71**(3), 331–355.

Spiliotis, E., Assimakopoulos, V. and Makridakis, S. (2020), 'Generalizing the theta method for automatic forecasting', *European Journal of Operational Research* **284**(2), 550–558.

Svetunkov, I. and Petropoulos, F. (2018), 'Old dog, new tricks: a modelling view of simple moving averages', *International Journal of Production Research* **56**(18), 6034–6047.

Talagala, T., Li, F. and Kang, Y. (2021), 'FFORMPP: Feature-based forecast model perfor-

mance prediction', *International Journal of Forecasting* (in press).

Talagala, T. S., Hyndman, R. J., Athanasopoulos, G. et al. (2018), 'Meta-learning how to forecast time series', *Monash Econometrics and Business Statistics Working Papers* **6**, 18.

Thomson, M. E., Pollock, A. C., Onkal, D. and Gonul, M. S. (2019), 'Combining forecasts: Performance and coherence', *International Journal of Forecasting* **35**(2), 474–484.

Tliche, Y., Taghipour, A. and Canel-Depitre, B. (2020), 'An improved forecasting approach to reduce inventory levels in decentralized supply chains', *European Journal of Operational Research* **287**(2), 511–527.

Wang, X., Kang, Y., Petropoulos, F. and Li, F. (2021), 'The uncertainty estimation of feature-based forecast combinations', *Journal of the Operational Research Society* (in press).

Wang, X., Smith-Miles, K. and Hyndman, R. J. (2009), 'Rule induction for forecasting method selection: meta-learning the characteristics of univariate time series', *Neurocomputing* **72**(10-12), 2581–2594.

Watson, M. W. and Stock, J. H. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**(6), 405–430.

Winkler, R. L. and Makridakis, S. (1983), 'The combination of forecasts', *Journal of the Royal Statistical Society: Series A (General)* **146**(2), 150–157.

Zhao, S. and Feng, Y. (2020), 'For2For: Learning to forecast from forecasts', *arXiv* **2001.04601**.