Ensembles and combinations:
using multiple models to improve forecasts

An increasing size of the toolbox of forecasting methods is available for decision makers. These methods, including statistical and econometric models, machine learning algorithms, and even judgemental forecasting (see an encyclopedic overview by Petropoulos et al., 2020), have their own specialities and are developed under different model specifications with assumptions on the Data Generation Process (DGP) or the associated error distributions. Given a pool of forecasting methods, how to best exploit information in the individual forecasts obtained from these methods?

Numerous studies in the forecasting literature are devoted to identifying a single 'best model' for a given time series. Given a family of models, information criteria, such as the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978), are commonly implemented for model selection (e.g., Qi and Zhang, 2001; Billah et al., 2005; Yang, 2005). More generally, cross-validation, in its various forms, such as the hold-out approach and the out-of-sample rolling scheme, has been used successfully to select the best forecast when multiple model families or model-free forecasts are considered (e.g., Kohavi, 1995; Poler and Mula, 2011; Fildes and Petropoulos, 2015; Inoue et al., 2017; Talagala et al., 2018). However, different criteria may lead to different results of forecast selections. Kourentzes et al. (2019) argued that model selection was a challenging task for two reasons: the sample, parameter and model uncertainty associated with identifying a single best forecast, and the ill-defined best forecast.

(Modified to use multiple models instead of combination ↓.)

Given these challenges, alternatively Bates and Granger (1969) have suggested combining multiple forecasts. The idea of combining forecasts is derived from the simple portfolio diversification argument (Timmermann, 2006), which is a risk management strategy with an obvious intuition: do not put all eggs into one basket. Even though slightly earlier articles have provided empirical justification of the superiority of forecast combinations over individual forecasts (e.g., Barnard, 1963; Crane and Crotty, 1967), the work by Bates and Granger (1969) is often considered to be the seminal article on forecast combinations as they developed a general analysis and further explored more possibilities for forecast combinations by extending a simple average to a weighted combination. Furthermore, the idea of combining forecasts is also widely used in machine learning, referred to as forecast ensembles. Similar to combination, the ensemble is a machine learning paradigm using multiple models to solve the same problem. It is difficult to trace the beginning of the history of ensemble forecasting. However, it is clear that ensemble techniques have become a hot topic in various fields, especially weather forecasting (see an overview by Leutbecher and Palmer, 2008), since the 1990s. Lewis (2005) provided a genealogy to depict the scientific roots of ensemble forecasting from several fundamental lines of research.

There are nearly five decades of empirical and theoretical investigations support that combining multiple forecasts often achieves improved forecasting performance on average than selecting a single individual forecast. Important early contributions in this area were summarized by Granger (1989), Clemen (1989), Palm and Zellner (1992), and Timmermann (2006). Clemen (1989) surveyed over two hundred statistical literature on forecast combinations and provided a primary conclusion that forecasting accuracy could be substantially improved by combining multiple forecasts. Timmermann (2006) attributed the superiority of forecast combinations over a single model to the fact that individual forecasts obtained based on heterogeneous information sets, may be very differently affected by structural breaks and subject to misspecification bias of unknown form. They further concluded that forecast combinations were beneficial due to diversification gains. More recently, Atiya (2020) illustrated graphically why forecast combinations were superior.

1 Different ways of using multiple models

- Combinations: A (usually linear) combination of forecasts from multiple models are used for one series. This includes combining point forecasts, quantile forecasts or full distributional forecasts. It covers simple averaging, weighted averaging, and sometimes combinations based on ML algorithms. e.g., FFORMA and related methods.

- Ensembles: Although "ensembles" has been used in different ways in different literatures, we will use "ensemble" to mean a mixture of the forecast distributions from multiple models. In many ways this is simpler than combinations as the relationship between the methods can be ignored. Need to discuss when they are equivalent.

- Boosting: Multiple models used for one series in sequence. Equivalent to hybrid forecasting where residuals from one method are modelled using a different method.

- Bagging: One or more models applied to multiple similar series, and then a combination or ensemble is taken. Bagging requires a method for generating multiple series. Some possibilities are STL-ETS and GRATIS.

- Stacking.

Simple example to illustrate differences. Suppose we have one series and two methods: an ARIMA model and a CNN.

- A combination would apply both to the same series and average the results. Unless we are only interested in point forecasting, the averaging would need to take account of the correlation between the forecast errors.
- An ensemble would apply both to the same series and generate forecast distributions from each. These would then be mixed (possibly with weighting) to form the final forecast distribution.
- Boosting would apply the ARIMA model to the series, and then apply the CNN to the residuals. The final forecasts would be the forecasts from the ARIMA model plus the forecasts from the CNN.
- Bagging would generate multiple series like the series of interest, and apply one of the methods to all the generated series. These could then be combined, or ensembled.

2 Point forecast combinations

2.1 Simple combinations

Considerable literature has accumulated over the years regarding the way in which individual forecasts are combined. A unanimous conclusion is that simple combination schemes are hard to beat (Clemen, 1989; Fischer and Harvey, 1999; Stock and Watson, 2004; Lichtendahl and Winkler, 2020). More specifically, simple combination rules which ignore past information regarding the precision of individual forecasts and correlations between forecast errors work reasonably well relative to more sophisticated combination schemes, as noted in Clemen's (1989) survey. Lichtendahl and Winkler (2020) attributed this phenomenon to a lower risk of simple combination methods resulting in bad forecasts than more refined combination methods. Timmermann (2006) concisely summarized the reasons for the success of simple combinations using the importance of parameter estimation error—simple combination schemes did not require estimating parameters such as combination weights based on forecast errors, thus avoiding parameter estimation error that often existed in weighted combinations.

The vast majority of studies on combining multiple models has dealt with point forecasting, even though point forecasts generally provide insufficient information for decision making. The simple average of forecasts based on equal weights stands out as the most popular and surprisingly robust combination rule (see Bunn, 1985; Clemen and Winkler, 1986; Stock and Watson, 2003; Genre et al., 2013). Makridakis et al. (1982) reported the results of M-competition, a forecasting competition involving $1,001$ economic time series, and found that the simple average outperformed the individual techniques. Clemen (1989) provided an extensive bibliographical review of the early work on the combination of forecasts, and then addressed the issue that the arithmetic means often dominated more refined forecast combinations. Makridakis and Winkler (1983) concluded empirically that the accuracy of combined forecasts was improved and the variability associated with the choice of methods was reduced, as the number of individual methods included in a simple average increased. Palm and Zellner (1992) concisely summarized the advantages of adopting a simple average into three points: (i) combination weights were equal and did not have to be estimated, (ii) a simple average significantly reduced variance and bias by averaging out individual bias in many cases, and (iii) a simple average should be considered when the uncertainty of weight estimation was taken into account. Additionally, Timmermann (2006) pointed out that the good average performance of the simple average depended strongly on model instability and the ratio of forecast error variances associated with different forecasting models.

More attention has been given to other options, including the median and mode, as well as trimmed means (e.g., Chan et al., 1999; Stock and Watson, 2004; Genre et al., 2013; Jose et al., 2014; Grushka-Cockayne et al., 2017), due to their robustness in the sense of being less sensitive to extreme forecasts than a simple average (Lichtendahl and Winkler, 2020). There is little consensus in the literature as to whether the mean or the median of individual forecasts performs better in terms of point forecasting (Kolassa, 2011). Specifically, McNees (1992) found no significant difference between the mean and the median, while the results of Stock and Watson (2004) supported the mean and Agnew (1985) recommended the median. Jose and Winkler (2008) studied the forecasting performance of the mean and median, as well as the trimmed and winsorized means. Their results suggested that the trimmed and winsorized means were appealing, particularly when there was a high level of variability among the individual forecasts, because of their simplicity and robust performance. Kourentzes et al. (2014) compared empirically the mean, mode and median combination operators based on kernel density estimation, and found that the three operators dealt with outlying extreme values differently, with the mean being the most sensitive and the mode operator the least. Based on these experimental results, they recommended further investigation of the use of the mode and median operators, which had been largely overlooked in the relevant literature.

Compared to various refined combination approaches and advanced machine learning algorithms, simple combinations seem to be outdated and uncompetitive in the big data era. However, the results from the recent M4 competition (Makridakis, Spiliotis and Assimakopoulos, 2020*a*) showed that simple combinations could achieve fairly good forecasting performance and still be competitive. Specifically, a simple equal-weights combination achieved the third best performance for yearly time series (Shaub, 2019) and a median combination of four models achieved sixth place for the point forecasts (Petropoulos and Svetunkov, 2020). Genre et al. (2013) encompassed a variety of combination methods in the case of forecasting GDP growth and the unemployment rate. They found that the simple average set a high benchmark, with few of the combination schemes outperforming it. Therefore, simple combination rules have been consistently the choice of many researchers and provide a tough benchmark to measure the effectiveness of the newly proposed weight estimation algorithms (e.g., Makridakis and Hibon, 2000; Stock and Watson, 2004; Makridakis, Spiliotis and Assimakopoulos, 2020*a*; Montero-Manso et al., 2020; Kang, Hyndman and Li, 2020; Wang et al., 2021). They have a less computational burden and can be implemented quickly.

Despite the fact that simple combination schemes can be intuitively implemented, the success of combination still highly depends on the choice of the model pool. If all component models are established in a very similar way based on the same, or highly overlapping set of information, forecast combination makes no sense and is not likely to be beneficial for the improvement of forecasting accuracy. Mannes et al. (2014) and Lichtendahl and Winkler (2020) emphasized two important issues concerning the performance of simple combination rules: one for the level of accuracy (or expertise) in the method pool and another for diversity among component models. Including component models with low accuracy is not likely to improve the combined forecasts. In addition, a high degree of diversity among component models facilitates the achievement of the best possible forecasting accuracy from their simple combinations (Thomson et al., 2019). In conclusion, simple, easy-to-use combination rules can provide good, and robust forecasting performance, especially when considering issues such as accuracy, and diversity of the method pool used for combining.

2.2 Weighted combinations

Though the combined forecasts formed by simple combination rules are acceptable for illustrative and concise purposes, the accumulated evidence of the forecasting literature suggests assigning greater weights to the individual forecasts which contain lower errors. The issue to be addressed is how to best weight the different forecasts used for combination. The general point forecast combination problem can be defined as seeking a one-dimensional aggregator that reduces the information up to time t in an N-vector of h-step-ahead forecasts, $\hat{\mathbf{y}}_{t+h|t} = \left(\hat{y}_{t+h|t,1}, \hat{y}_{t+h|t,2}, \ldots, \hat{y}_{t+h|t,N}\right)'$, to a single combined h-step-ahead forecast $\tilde{y}_{t+h|t} = C\left(\hat{\mathbf{y}}_{t+h|t}; \boldsymbol{w}_{t+h|t}\right)$, where $\boldsymbol{w}_{t+h|t}$ is an N-vector of combination weights. The general class of combination methods represented by the mapping, C, from $\hat{\mathbf{y}}_{t+h|t}$ to y_{t+h}, comprises linear, nonlinear, and time-varying combinations. Below we discuss in detail the use of various weighting schemes to determine combination weights associated with each individual forecast.

2.2.1 Linear combinations

Typically, the combined forecast is commonly constructed as a linear combination of the individual forecasts, which can be written as

$$\tilde{y}_{t+h|t} = \boldsymbol{w}'_{t+h|t}\hat{\mathbf{y}}_{t+h|t}, \tag{1}$$

where $\boldsymbol{w}_{t+h|t} = \left(w_{t+h|t,1}, \ldots, w_{t+h|t,N}\right)'$ is an N-vector of linear combination weights assigned to N individual forecasts.

Optimal weights

The seminal work of Bates and Granger (1969) proposed a method to find 'optimal' weights by minimizing the variance of the combined forecast error, and discussed only the combination of pairs of forecasts. Newbold and Granger (1974) then extended the method to the combination of several forecasts. Specifically, assuming that individual forecasts are unbiased and their variances of errors are consistent over time, the combined forecast obtained by a linear combination will also be unbiased. Differentiating with respect to $\boldsymbol{w}_{t+h|t}$ and solving the first order condition, the variance of the combined forecast error is minimized by taking

$$\boldsymbol{w}_{t+h|t}^{\text{opt}} = \frac{\boldsymbol{\Sigma}_{t+h|t}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}_{t+h|t}^{-1}\mathbf{1}}, \tag{2}$$

where $\boldsymbol{\Sigma}_{t+h|t}$ is the $N \times N$ covariance matrix of the lead h forecast errors and $\mathbf{1}$ is the N-dimensional unit vector. Unfortunately, in practice, the elements of the covariance matrix $\boldsymbol{\Sigma}_{t+h|t}$ are usually unknown and required to be properly estimated.

It follows that if $\boldsymbol{w}_{t+h|t}$ is determined by Equation (2), one can identify a combined forecast $\tilde{y}_{t+h|t}$ with no greater error variance than the minimum error variance of all individual forecasts. The fact was further demonstrated in detail in Timmermann (2006) to illustrate the diversification gains offered by forecast combinations by simply considering the combination of two forecasts. Under mean squared error (MSE) loss, Timmermann (2006) characterized the general solution of the optimal linear combination weights given the joint Gaussian distribution of the outcome y_{t+h} and forecasts $\hat{\mathbf{y}}_{t+h|t}$.

The loss assumed in Bates and Granger (1969) and Newbold and Granger (1974) is quadratic and symmetric in the forecast error from the linear combination. Elliott and Timmermann (2004) examined forecast combinations under more general loss functions that account for asymmetries, and forecast error distributions with skew. They demonstrated that the optimal combination weights in a combination strongly depended on the degree of asymmetry in the loss function and skews in the underlying forecast error distribution. Subsequently, Patton and Timmermann (2007) demonstrated that the properties of optimal forecasts established under MSE loss were not generally robust under more general assumptions about the loss function. The properties of optimal forecasts were also generalized to consider asymmetric loss and nonlinear DGP.

Regression approach

The seminal work by Granger and Ramanathan (1984) provided an important impetus for approximating the optimal weights under a linear regression framework. They recommended the strategy that the combination weights can be estimated by Ordinary Least Squares (OLS) in regression models having the vector of past observations as the response variable and the matrix of past forecasts as the explanatory variables. Three alternative approaches involving various possible restrictions are considered

$$y_{t+h} = \boldsymbol{w}_h' \hat{\mathbf{y}}_{t+h|t} + \varepsilon_{t+h}, \quad s.t. \quad \boldsymbol{w}_h' \mathbf{1} = 1, \tag{3}$$

$$y_{t+h} = \boldsymbol{w}' \hat{\mathbf{y}}_{t+h|t} + \varepsilon_{t+h}, \tag{4}$$

$$y_{t+h} = \omega_{0h} + \boldsymbol{w}' \hat{\mathbf{y}}_{t+h|t} + \varepsilon_{t+h}. \tag{5}$$

The constrained OLS estimation of the regression (3) in which the constant is omitted and the weights are constrained to sum to one yields results identical to the optimal weights proposed by Bates and Granger (1969). Furthermore, Granger and Ramanathan (1984) suggested the unrestricted OLS regression (5) which allowed for a constant term and did not impose the weights sum to one was superior to the popular optimal method regardless of whether the constituent forecasts were biased. However, De Menezes et al. (2000) put forward some consideration required when using the unrestricted regression, including the stationarity of the series being forecast, the possible presence of serial correlation in forecast errors (see also Diebold, 1988; Edward Coulson and Robins, 1993), and the issue of multicollinearity.

More generalizations of the combination regressions have been considered in a large body of literature. Diebold (1988) exploited the serial correlation in least squares framework by characterizing the combined forecast errors as the AutoRegressive Moving Average (ARMA) processes, leading to improved combined forecasts. Gunter (1992) and Aksu and Gunter (1992) provided an empirical analysis to compare the performance of various combination strategies, including the simple average, the unrestricted OLS regression, the restricted OLS regression where the weights were restricted to sum to unity, and the nonnegativity restricted OLS regression where the weights were constrained to be nonnegative. The results revealed that constraining weights to be nonnegative was at least as robust and accurate as the simple average and yielded superiority over other combinations based on regression framework. Conflitti et al. (2015) addressed the problem of determining the optimal weights by imposing two restrictions that the weights should be nonnegative and sum to one, which turned out to be a special case of a lasso regression. Edward Coulson and Robins (1993) found that allowing a lagged dependent variable in forecast combination regressions could achieve improved performance. Instead of using the quadratic loss function, Nowotarski et al. (2014) applied the absolute loss function in the unrestricted regression to yield the least absolute deviation regression which was more robust to outliers than OLS combinations.

The forecast combinations using changing weights are developed in the relevant literature to solve various types of structural changes in the constituent forecasts. For instance, Diebold and Pauly (1987) explored the possibilities for time-varying parameters in regression-based combination approaches. Both deterministic and stochastic time-varying parameters were considered in the linear regression framework. Specifically, the combination weights were described as deterministic nonlinear (polynomial) functions of time or allowed to involve random variation. Deutsch et al. (1994) allowed the combination weights to evolve immediately or smoothly using switching regression models and smooth transition regression models.

Researchers have worked on dealing with a large number of forecasts in the regression framework to take advantages of many different models. Chan et al. (1999) examined a wide range of combination methods in a Monte Carlo experiment and a real-world dataset. Their results investigated the poor performance of OLS combinations when the number of forecasts to be combined was large and suggested alternative weight estimation methods, such as ridge regression and Principal Component Regression (PCR). Stock and Watson (2004) offered the details of principal component forecast combination, which entailed forming a regression having the actual value as the response variable and the first few principal components reduced from several forecasts as the explanatory variables. This method reduced the number of weights that must be estimated in a regression framework and frequently served as a way to solve the multicollinearity problem which was likely to lead to unstable behavior in the estimated weights. The superiority of the PCR involving dimension reduction techniques over OLS combinations was also supported in Rapach and Strauss (2008) and Poncela et al. (2011). Aiolfi and Timmermann (2006) argued in favour of a step of clustering forecasting models using the k-means clustering algorithm based on their historical performance. For each cluster, a pooled (average) forecast was then computed, which preceded the calculation of combination weights for the constructed clusters. Both the clustering strategy and PCR have good merits such as allowing for a large number of individual forecasting models, improving computational efficiency, and reducing parameter estimation error.

Performance-based weights

Estimation errors in the optimal weights and a diverse set of regression-based weights tend to be particularly large due to difficulties in properly estimating the entire covariance matrix $\mathbf{\Sigma}_{t+h|t}$, especially in situations with a large number of forecasts at hand. Instead, Bates and Granger (1969) suggested weighting the constituent forecasts in inverse proportion to their historical performance, ignoring correlations across forecast errors. In follow-up studies, Newbold and Granger (1974) and Winkler and Makridakis (1983) generalized the issue in the sense of considering more time series, more forecasting models, and multiple forecast horizons. Their extensive results demonstrated that combinations that took account of correlations performed poorly, and consequently reconfirmed Bates and Granger's (1969) argument that correlations can be poorly estimated in practice and should be ignored in calculating combination weights.

Let $\mathbf{e}_{t+h|t} = \mathbf{1}y_{t+h} - \hat{\mathbf{y}}_{t+h|t}$ be the N-vector of h-period forecast errors from the individual models, the five procedures suggested in Bates and Granger (1969) for estimating the combination weights when $\boldsymbol{\Sigma}_{t+h|t}$ is unknown, extended to the general case are as follows:

$$w^{\mathrm{bg1}}_{t+h|t,i} = \frac{\left(\sum_{\tau=t-\nu+1}^{t} e^2_{\tau|\tau-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=t-\nu+1}^{t} e^2_{\tau|\tau-h,j}\right)^{-1}}. \tag{6}$$

$$\boldsymbol{w}^{\mathrm{bg2}}_{t+h|t} = \frac{\hat{\boldsymbol{\Sigma}}^{-1}_{t+h|t}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}_{t+h|t}\mathbf{1}}, \quad \text{where} \quad (\hat{\boldsymbol{\Sigma}}_{t+h|t})_{i,j} = \nu^{-1}\sum_{\tau=t-\nu+1}^{t} e_{\tau|\tau-h,i}e_{\tau|\tau-h,j}. \tag{7}$$

$$w^{\mathrm{bg3}}_{t+h|t,i} = \alpha\hat{w}_{t+h-1|t-1,i} + (1-\alpha)\frac{\left(\sum_{\tau=t-\nu+1}^{t} e^2_{\tau|\tau-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=t-\nu+1}^{t} e^2_{\tau|\tau-h,j}\right)^{-1}}, \quad 0 < \alpha < 1. \tag{8}$$

$$w^{\mathrm{bg4}}_{t+h|t,i} = \frac{\left(\sum_{\tau=1}^{t} \gamma^{\tau} e^2_{\tau|\tau-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=1}^{t} \gamma^{\tau} e^2_{\tau|\tau-h,j}\right)^{-1}}, \quad \gamma \geq 1. \tag{9}$$

$$\boldsymbol{w}^{\mathrm{bg5}}_{t+h|t} = \frac{\hat{\boldsymbol{\Sigma}}^{-1}_{t+h|t}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}_{t+h|t}\mathbf{1}}, \quad \text{where} \quad (\hat{\boldsymbol{\Sigma}}_{t+h|t})_{i,j} = \frac{\sum_{\tau=1}^{t} \gamma^{\tau} e_{\tau|\tau-h,i}e_{\tau|\tau-h,j}}{\sum_{\tau=1}^{t}\gamma^{\tau}} \quad \text{and} \quad \gamma \geq 1. \tag{10}$$

These weighting schemes differ in the factors, as well as the choice of the parameters, ν, α, and γ. Correlations across forecast errors are either ignored by treating the covariance matrix $\boldsymbol{\Sigma}_{t+h|t}$ as a diagonal matrix or estimated using sample data points which, however, may lead to quite unstable estimates of $\boldsymbol{\Sigma}_{t+h|t}$ given highly correlated forecast errors. Some estimation schemes suggest computing or updating the relative performance of different models over rolling windows of the most recent ν observations, while others base the weights on exponential discounting with higher values of γ giving larger weights to recent observations. In consequence, these weighting schemes are well adapted to allow the non-stationary relationship between the individual forecasting procedures over time (Newbold and Granger, 1974), which, however, tends to increase the variance of the parameter estimates and works quite poorly provided that the DGP is truly covariance stationary (Timmermann, 2006).

A broader set of combination weights based on the relative performance of individual forecasting techniques is developed and examined in a series of studies. For example, Stock and Watson (1998) generalized the rolling window scheme (6) in the sense that the weights on the individual forecasts were inversely proportional to the kth power of their MSE. The weights with $k = 0$ correspond to assigning equal weights to all forecasts, while more weights are placed on the best performing models by considering $k \geq 1$. Other forms of forecast error measures, such as the Root Mean Squared Error (RMSE) and the symmetric Mean Absolute Percentage Error (sMAPE), are also considered to develop the performance-based combination weights (e.g., Nowotarski et al., 2014; Pawlikowski and Chorowska, 2020). Besides, a weighting scheme with the weights depending inversely on the exponentially discounted errors is proposed by Stock and Watson (2004) as an upgraded version of the scheme (9), and is encompassed in the sequent studies (e.g., Clark and McCracken, 2010; Genre et al., 2013) to achieve gains from combining forecasts. The pseudo out-of-sample performance used in these weighting schemes is commonly computed based on rolling or recursive (expanding) windows (e.g., Stock and Watson, 1998; Clark and McCracken, 2010; Genre et al., 2013). It is natural to adopt rolling windows in estimating the weights to deal with the structural change. But the window length should not be too short without the estimates of the weights becoming too noisy (Baumeister and Kilian, 2015).

Compared to constructing the weights directly using historical forecast errors, a new form of combination that is more robust and less sensitive to outliers is introduced based on the 'ranking' of models. Again this combination ignores correlations across forecast errors. The simplest and most commonly used method in the class is to use the median forecast as the output. Aiolfi and Timmermann (2006) constructed the weights proportional to the inverse of performance ranks (sorted according to increasing order of forecast errors), which were later employed by Andrawis et al. (2011) for tourism demand forecasting. Another weighting scheme that attaches a weight proportional to $\exp(\beta(N+1-i))$ to the ith ordered constituent model is adopted in Yao and Islam (2008) and Donate et al. (2013) to combine Artificial Neural Networks (ANNs), where β is a scaling factor. However, as mentioned by Andrawis et al. (2011), this class of combination method still comes with the drawback of the discrete nature because it limits the weight to only a few possible levels.

Combinations based on information criteria

Information criteria, such as the Akaike Information Criterion (AIC, Akaike, 1974), the corrected Akaike Information Criterion (AICc, Sugiura, 1978), and the Bayesian Information Criterion (BIC, Schwarz, 1978), are often advised to deal with model selection in forecasting. However, choosing a single model out of the candidate model pool may be misleading because of the loss of information gleaned from alternative models. An alternative way proposed by Burnham and Anderson (2002) is to combine different models based on information criteria to mitigate the risk of selecting a single model.

One such common approach is using Akaike weights. Specifically, in light of the fact that AIC estimates the Kullback-Leibler distance (Kullback and Leibler, 1951) between a model and the true DGP, differences in the AIC can be considered to weigh different models, providing a measure of the evidence for each model relative to other constituent models. Given N individual models, the Akaike weight of model i can be derived by the following steps:

$$\begin{aligned} w_i^{\text{aic}} &= \frac{\exp(-0.5\Delta\text{AIC}_i)}{\sum_{k=1}^{N}\exp\left(-0.5\Delta\text{AIC}_k\right)}, \\ \Delta\text{AIC}_i &= \text{AIC}_i - \min_{k\in\{1,2,\cdots,N\}} \text{AIC}(k). \end{aligned} \tag{11}$$

Akaike weights calculated in this manner can be interpreted as the probability that a given model performs best at approximating the unknown DGP, given the model set and data (Kolassa, 2011). Similar weights from AICc, BIC, and other variants with different penalties, can be derived analogously to Equation (11).

The outstanding performance of weighted combinations based on information criteria has been supported in some studies. For instance, Kolassa (2011) used weights derived from AIC, AICc and BIC to combine exponential smoothing forecasts, and resulted in superior accuracy over selection using these information criteria. A similar strategy was adopted by Petropoulos, Hyndman and Bergmeir (2018) to separately explore the benefits of bagging for time series forecasting. Additionally, an empirical study by Petropoulos, Kourentzes, Nikolopoulos and Siemsen (2018) showed that a weighted combination based on AIC improved the performance of the statistical benchmark they used.

Bayesian approach

Some effort has been directed toward the use of Bayesian approaches to updating forecast combination weights in face of new information from various sources. Recall that obtaining reliable estimates of the covariance matrix $\boldsymbol{\Sigma}$ (the time and horizon subscripts are dropped for simplicity) of forecast errors with the correlation being ignored or not, is a major challenge in the general case. With this in mind, Bunn (1975) suggested the idea of a Bayesian combination on the basis of the probability of respective forecasting model performing the best on any given occasion. Considering the beta and the Dirichlet distributions arising as the conjugate priors for the binomial and multinomial processes respectively, the suggested non-parametric method performs well when there is relatively little past data by means of attaching prior subjective probabilities to individual forecasts (Bunn, 1985; De Menezes et al., 2000). Öller (1978) presented another approach to involving subjective probability in a Bayesian updating scheme based on the self-scoring weights proportional to the evaluation of the expert's forecasting ability.

A different theme of research has also advocated the incorporation of prior information into the estimation of combination weights, but with the weights being shrunk toward some prior mean under a regression-based combination framework (Newbold and Harvey, 2002). Assuming that the vector of forecast errors was normally distributed, Clemen and Winkler (1986) developed a Bayesian approach with the conjugate prior for $\boldsymbol{\Sigma}$, represented by an inverted Wishart distribution with covariance matrix $\boldsymbol{\Sigma}_0$ and scalar degrees of freedom ν_0. Again we drop time and horizon subscripts for simplicity. If the last n observations are used to estimate $\boldsymbol{\Sigma}$, the combination weights derived from the posterior distribution for $\boldsymbol{\Sigma}$ are

$$\boldsymbol{w}^{\text{cw}} = \frac{\boldsymbol{\Sigma}^{*-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{*-1}\mathbf{1}}, \tag{12}$$

where $\boldsymbol{\Sigma}^* = \left[\left(\nu_0\boldsymbol{\Sigma}_0^{-1} + n\hat{\boldsymbol{\Sigma}}^{-1}\right)/(\nu_0+n)\right]^{-1}$ and $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix. Compared to estimating $\boldsymbol{\Sigma}$ using past data or treating it as a diagonal matrix, the proposed approach yields superiority, in terms of providing a relatively stable estimation and allowing correlations by specifying the prior estimate $\boldsymbol{\Sigma}_0$. The subsequent work by Diebold and Pauly (1990) allowed the incorporation of the standard normal-gamma conjugate prior by considering a normal regression-based combination

$$\mathbf{y} = \hat{\mathbf{Y}}\boldsymbol{w} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}\right), \tag{13}$$

where $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are T-dimensional vectors, and $\hat{\mathbf{Y}}$ is the $T \times N$ matrix of constituent forecasts. The approach results in estimated combination weights which can be viewed as a matrix weighted average of those for the two polar cases, least squares and prior weights. It can provide a rational transition between subjective and data-based estimation of the combination weights. In light of the fact that Bayesian approaches have been mostly employed to construct combinations of probability forecasts, we will elaborate on other newly developed methods of determining combination weights in a Bayesian context in the following Section 3.

2.2.2 Nonlinear combinations

So far our attention has been placed on linear combination schemes, including simple and weighted methods. Linear combination approaches implicitly assume a linear dependence between constituent forecasting models and the variable being forecasted (Donaldson and Kamstra, 1996; Freitas and Rodrigues, 2006), which may result in the optimal combination at a particular point but not the best forecast (Shi et al., 1999). Therefore, the trustworthiness of linear combinations may be quite questionable if the individual forecasts come from nonlinear models or if the true relationship between combination members and the best forecast is characterized by nonlinear systems (Babikir and Mwambi, 2016). In such cases, it follows naturally to relax the linearity assumption and consider nonlinear combination schemes which, however, have received very limited research attention so far.

As classified by Timmermann (2006), two types of non-linearities can be considered in forecast combinations. The first type involves nonlinear functions of the individual forecasts, but with the unknown parameters of combination weights being in linear form. While the second method allows a more general combination where non-linearities are considered in the combination parameters. However, considerable estimation errors may be produced in such cases and thus may require additional numerical efforts.

Shi and Liu (1993) suggested the use of the neural networks for nonlinear combination by demonstrating that it served as an effective method to solve the problem of forecasting the price of

IBM stock. Since then, a stream of studies has been devoted to conducting nonlinear combination of forecasts using the neural networks. Donaldson and Kamstra (1996) used ANNs to obtain the combined forecasts $\tilde{y}_{t+h|t}$ by the following form

$$
\begin{aligned}
\tilde{y}_{t+h|t} &= \beta_0 + \sum_{j=1}^{k} \beta_j \hat{y}_{t+h|t,j} + \sum_{i=1}^{p} \delta_i g\left(\mathbf{z}_{t+h|t}\gamma_i\right), \qquad (14)\\
g\left(\mathbf{z}_{t+h|t}\gamma_i\right) &= \left(1 + \exp\left(-\left(\gamma_{0,i} + \sum_{j=1}^{N} \gamma_{1,j} z_{t+h|t,j}\right)\right)\right)^{-1}, \\
z_{t+h|t,j} &= \left(\hat{y}_{t+h|t,j} - \bar{y}_{t+h|t}\right)/\hat{\sigma}_{yt+h|t}, \\
k &\in \{0, N\} \quad p \in \{0, 1, 2, 3\},
\end{aligned}
$$

where $\bar{y}_{t+h|t}$ and $\hat{\sigma}_{yt+h|t}$ were the in-sample mean and standard deviation across the forecasting models respectively using data up to time t. This scheme permitted special cases of both purely linear combination ($k = N, p = 0$) and nonlinear combination ($k = 0, p \neq 0$). Following them came an extension work by Harrald and Kamstra (1997) to evolve ANNs and demonstrate its utility when only forecasts to combine were given. Krasnopolsky and Lin (2012) and Babikir and Mwambi (2016) employed neural network approaches with various activation functions to approximate the nonlinear dependence of individual forecasts and achieve the nonlinear mapping, which were essentially variants of Equation (14). The empirical results of nonlinear combination from these studies offered forecasts which generally dominated forecasts from traditional linear combining procedures, such as simple average, OLS weights, and performance-based weights. The superiority arose possibly due to the neural networks' outstanding learning ability and flexibility to uncover hidden nonlinear relationships not easily captured by traditional linear combinations (Donaldson and Kamstra, 1996; Babikir and Mwambi, 2016).

However, these empirical evidence for the gains from incorporating nonlinear combinations was given based on less than ten time series (mainly economic or financial data), and thus was not statistically significant and probably produced by the dodgy selection of data. Besides, these nonlinear combination methods suffered from other drawbacks, including the neglect of correlation in forecast errors, the instability of parameter estimation, and the multicollinearity issue caused by the overlap in the information sets used to produce the individual forecasts. The superiority of nonlinear combinations over linear combining schemes, thus, need to be further investigated.

Other researchers sought to construct nonlinear combinations via including a nonlinear term to cope with the case where the forecast errors from constituent models were correlated. This combination mechanism can be generalized to the following form

$$\tilde{y}_{t+h|t} = \beta_0 + \sum_{j=1}^{N} \beta_j \hat{y}_{t+h|t,j} + \sum_{\substack{i,j=1 \\ i<j}}^{N} \pi_{ij} v_{ij}. \tag{15}$$

By this way, the usual framework for linear combinations is extended. The involved nonlinear terms take various definitions. For example, Freitas and Rodrigues (2006) intuitively defined v_{ij} as the product of individual forecasts from different models, $\hat{y}_{t+h|t,i} \cdot \hat{y}_{t+h|t,j}$, while Adhikari and Agrawal (2012) took into account the linear correlations among the forecast pairs by including the term, $(\hat{y}_{t+h|t,i} - \bar{y}_i)(\hat{y}_{t+h|t,j} - \bar{y}_j)/(\sigma_i \sigma_j)^2$, where $\bar{y}_i$ and σ_i were the mean and standard deviation of the ith model. Moreover, Adhikari (2015) defined the nonlinear term using $(\hat{z}_i - m_{ij}\hat{z}_j)(\hat{z}_j - m_{ji}\hat{z}_i)$, where the term m_{ij} was the degree of mutual dependency between the ith and jth forecasting model, and $\hat{z}_i$ was the standardized individual forecasts using the mean $\bar{y}_i$ and standard deviation σ_i. Five correlation measures were investigated for measuring the mutual associations between two different forecasts.

Clearly, the area of considering non-linearities in forecast combinations requires further research.

- On one hand, the forecasting performance of nonlinear combination schemes should be further investigated for a large, diverse collection of time series datasets as well as with some statistical inference.
- On the other hand, the high correlation across forecast errors as well as the multicollinearity issue has to be addressed carefully in the framework of nonlinear combinations. For example, the PCR model can be considered as an alternative approach to dealing with such concerns by projecting the individual forecasts onto a smaller subspace via principal component analysis.

2.2.3 Combining by learning

Stacking (stacked generalization, Wolpert, 1992) provides a strategy to adaptively combine the available forecasting models by learning from a single model. While stacking is frequently employed on a wide variety of classification tasks (Zhou, 2012), it introduces the concept of meta-learning in the context of time series forecasting with the purpose of boosting forecasting accuracy beyond the level achieved by any of the individual models. On one hand, stacking is a general framework that consists of at least two levels. Considering a stacking approach involving two levels. The first level entails the training of individual forecasting models using the original data, while the model, too called meta-model, is learned in the second level using the first-level forecasts as attributes to form a final set of forecasts. Note that the meta-model in a certain level must use the forecasts of models in the previous level as inputs. On the other hand, stacking can be regarded as a generic combination method that combines by learning. In this respect, we introduce the stacking approach here as a class of combination methods, which adaptively weighs individual forecasts using meta-learning processes.

There are many different ways to implement the stacking strategy. Its primary implementation is as a technique for combining individual models in a series-by-series fashion. Simply put, individual forecasting models in the method pool are trained using only data of the single series they are going to forecast, their forecast outputs are subsequently fed to a meta-model tailored for the target series to calculate the combined forecasts. This means that n meta-models are required for n different series. Unsurprisingly, OLS regression (e.g., Granger and Ramanathan, 1984; Gunter, 1992) falls into this category and can be viewed as the most simple, common learning algorithm used in stacking. Instead of applying multiple linear regression, Moon et al. (2020) suggested the PCR model as the meta-model predominantly due to its desirable characteristics such as dimensionality reduction and avoidance of multicollinearity between the input forecasts of individual models. Similarly, lasso regression, as well as machine learning techniques, such as ANN, Wavelet Neural Network (WNN), and Support Vector Regression (SVR) can therefore be conducted in a series-by-series fashion to combine constituent models (e.g., Donaldson and Kamstra, 1996; Conflitti et al., 2015; Ribeiro et al., 2019; Ribeiro and dos Santos Coelho, 2020). One could consider the use of the expanding or rolling window method to ensure that enough individual forecasts are generated for the training of meta-models. Time series cross-validation, also known as ‘evaluation on a rolling forecasting origin’, is also recommended in the training procedures for both individual models and meta-models to help with the determination of parameters. Nevertheless, stacking approaches implemented in a series-by-series fashion still suffer from some limitations such as time wasting, high requirements for time series length, and insufficiency of training data for meta-models.

An alternative way to perform the stacking strategy sheds some light on the potential of cross-learning. Specifically, the meta-model is trained using information derived from multiple series without relying only on a single series, thus various patterns can be captured along different series. The M4 competition (Makridakis, Spiliotis and Assimakopoulos, 2020*a*), comprising $100,000$ time series, recognized the benefits of cross-learning in the sense that the top three performing methods of the competition utilized the information across the whole dataset rather than a single series. Cross-learning can therefore be identified as a promising strategy to boost forecasting accuracy, at least when appropriate strategies for extracting information from large, diverse time series datasets are considered (Kang, Spiliotis, Petropoulos, Athiniotis, Li and Assimakopoulos, 2020; Semenoglou et al., 2020). Zhao and Feng (2020) trained a neural network model across the M4 competition dataset to learn how to combine different models in the method pool. They adopted the temporal holdout strategy to generate the training dataset and utilized only the out-of-sample forecasts produced by standard individual models as the input in the neural network model.

An increasing stream of studies has shown that time series features, additional inputs describing each series in a dataset, provide valuable information for forecast combination in a cross-learning fashion, leading to an extension of stacking. The pioneering work by Collopy and Armstrong (1992) developed a rule base consisting of 99 rules to combine forecasts from four statistical models using 18 time series features. Petropoulos et al. (2014) identified the main determinants of forecasting accuracy through an empirical study involving 14 forecasting models and seven time series features. The findings can provide useful information for forecast combination. More recently, Montero-Manso et al. (2020) introduced a Feature-based FORecast Model Averaging (FFORMA) approach available in the R package M4metalearning[1], which employed 42 statistical features (implemented using the R package tsfeatures, Hyndman et al., 2019) to estimate the optimal weights for combining nine different traditional models trained per series based on an XGBoost model. The method reported the second-best forecasting accuracy in M4 competition. Li et al. (2020) extracted time series features automatically with the idea of time series imaging, then these features were used for forecast combination. Gastinger et al. (2021) demonstrated the value of a collection of combination methods on a large and diverse amount of time series from M3, M4, M5 datasets (Makridakis and Hibon, 2000; Makridakis, Spiliotis and Assimakopoulos, 2020*a*,*b*) and Federal Reserve Economic Data (FRED) datasets[2]. In light of the finding that it was not clear which combination strategy should be selected, they introduced a meta-learning step to select a promising subset of combination methods for a newly given dataset based on its extracted features.

[1] The R package M4metalearning is available at `https://github.com/robjhyndman/M4metalearning`.

[2] The FRED dataset is available at `https://fred.stlouisfed.org`.

In addition to the time series features extracted from the historical data, it is crucial to look at the diversity of the individual model pool in the context of forecast combination (Batchelor and Dua, 1995; Thomson et al., 2019; Atiya, 2020; Lichtendahl and Winkler, 2020). An increase in diversity among forecasting models can improve the accuracy of their combination. In this respect, features describing the diversity of the method pool should be included in the feature pool to provide additional information possibly relevant to combining models. Lemke and Gabrys (2010) calculated six diversity features and created an extensive feature pool describing both the time series and the individual method pool. Three meta-learning algorithms were implemented to link knowledge on the performance of individual models to the features, and to improve forecasting performance. Kang, Cao, Petropoulos and Li (2020) utilized a group of features only measuring the diversity across the candidate forecasts to construct a forecast combination model mapping the diversity matrix to the forecast errors. The proposed approach yielded comparable forecasting performance with the top-performing methods in the M4 competition.

As expected, the implementations of stacking in a cross-learning manner also come with their own limitations. The first limitation is the requirement for a large, diverse time series dataset to enable meaningful training. This issue can be addressed by simulating series on the basis of the assumed DGPs (Talagala et al., 2018), which are exponential smoothing models and ARIMA models, and by generating time series with diverse and controllable characteristics (Kang, Hyndman and Li, 2020). Moreover, given considerable literature on feature identification and feature engineering (e.g., Wang et al., 2009; Kang et al., 2017; Lemke and Gabrys, 2010; Montero-Manso et al., 2020; Li et al., 2020), the feature-based forecast combination methods naturally raise the concern about how to design an appropriate feature pool in order to achieve the best out of such methods. Other major limitations lie with the design of the loss function for the meta-model and the requirement of significant training time.

2.3 Which forecasts should be combined?

The benefits of forecast combinations highly depend on how to best weight the individual forecasts being combined (Stock and Watson, 2004; Timmermann, 2006). As discussed previously, forecast combinations implemented by simple rules or alternative weighting schemes have both merits and limitations. For example, simple combination rules provide simple, robust forecasts while ignoring past information regarding the precision of individual forecasts and correlations among forecast errors. Weighting operators adaptively weigh individual forecasts according to their historical performance, while suffering from various uncertainties during the estimation of weights. Alternatively, the gains from forecast combinations are directly related to the quality of the selected individual forecasts that are combined (Batchelor and Dua, 1995; Geweke and Amisano, 2011). Intuitively, we prefer an ideal situation that the component forecasts fall on opposite sides of the truth (the realisation), so that these forecasts being combined 'bracket' the true value (Bates and Granger, 1969; Larrick and Soll, 2006). In such a manner, forecast errors tend to cancel each other out. Forecast combinations, thus, are likely to achieve the greatest gains in terms of forecasting accuracy. Unfortunately, though, this case rarely occurs in practice, as these forecasts may be based on a similar training process and overlapping information set. It is natural to highlight the question, which forecasts should be combined?

One important issue with respect to the forecasting models being combined is accuracy. Including models with extremely poor performance degrades the performance of the forecast combination. One prefers to exclude models that perform poorly and use top performers to combine. In judgemental forecasting, Mannes et al. (2014) highlighted the importance of the crowd's mean level of accuracy (expertise). They argued that the mean level of expertise set a floor on the performance of combining. The gains in accuracy from selecting top-performing models for combination have been investigated and confirmed by a stream of articles such as Budescu and Chen (2015), and Kourentzes et al. (2019). Lichtendahl and Winkler (2020) emphasized that the variance of accuracy across series, which provided an indication of the accuracy risk, exerted a great influence on the performance of combined forecasts. They suggested balancing the tradeoffs between the average accuracy and the variance of accuracy when choosing component models from a set of available models.

The other key issue is diversity. Diversity among the individual models is often recognized as one of the elements required for accurate forecasting using a combination (Batchelor and Dua, 1995; Brown et al., 2005; Thomson et al., 2019). Atiya (2020) utilized the bias-variance decomposition of MSE to study the effects of forecast combinations and confirmed the finding that an increase in diversity among the individual models was responsible for the error reduction displayed in combined forecasts. Diversity among individual models is frequently measured in terms of correlations among their forecasting errors, with lower correlations indicating a higher degree of diversity. The distance of top-performing clusters introduced by Lemke and Gabrys (2010), where k-means clustering algorithm is applied to construct clusters, and a measure of coherence proposed by Thomson et al. (2019) are also considered as other measures to reflect the degree of diversity among forecasts.

Researchers attempt to choose independent forecasts to amplify the gains of diversity in forecasting accuracy when forming a combination. However, individual forecasts available are often produced based on similar training, similar models and overlapping information set, leading to highly positively correlated forecast errors. Including pairs of forecasts that have highly correlated forecast errors in a combination creates redundancy, which is likely to contribute to multicollinearity problems in some combination methods, especially in the class of regression-based combinations (Granger and Ramanathan, 1984). In this respect, different types of forecasting models (i.e., statistical, machine learning, and judgemental) or different sources of information (i.e., exogenous variables) are often recommended to achieve diversity (Atiya, 2020). The results of the M4 competition reconfirmed the benefits of a combination with both statistical and machine learning models (Makridakis, Spiliotis and Assimakopoulos, 2020*a*).

It is often suggested to involve an appropriate number of individual forecasts rather than the full set of forecasts in a combination, as there are decreasing returns to adding additional forecasts (Armstrong, 2001; Zhou et al., 2002; Geweke and Amisano, 2011; Lichtendahl and Winkler, 2020). Simply put, *many could be better than all*. In this regard, given a method pool with a large number of forecasting models available, we can consider an additional step ahead of combining: forecast pooling. Instead of using all available forecasts in a combination, pooling aims to eliminate some forecasts from the combination and select only a subset of the available forecasts.

The most common technique of pooling is using the top quantile to form a model pool, discarding the worst-performing models (e.g., Granger and Jeon, 2004). Mannes et al. (2014) investigated the gains in accuracy from the *select-crowd* strategy, which selected top-performing models based on the historical forecasting accuracy. However, the use of top quantiles can be criticized for using arbitrary cut-off points of how many quantiles to use. Kourentzes et al. (2019) proposed a heuristic, which is identical to top quantiles, to automatically stop component forecasts with a sharp drop in performance from entering the model pool using the outlier detection methods in boxplots. Their empirical results over four diverse datasets showed that forecast pooling outperformed selecting a single forecast or combining all of them. Nonetheless, the approach suffers the limitation of not considering diversity when formulating appropriate pools. More recently, Lichtendahl and Winkler (2020) developed a pooling approach comprising two screens: one screen for removing individual models that performed poorly than the Naive2 benchmark and another for excluding pairs of models with highly correlated forecast errors. In this way, both accuracy and diversity issues are addressed when forming a combination.

Similarly to the PCR method (Stock and Watson, 2004) and the clustering strategy (Aiolfi and Timmermann, 2006), pooling techniques take advantages of allowing the large number of forecasts to be combined, reducing weight estimation errors, and improving computational efficiency. However, instead of focusing on dealing with combination using a large number of forecasts, forecast pooling deliberates on the trimming of individual models when developing a model pool. Forecast pooling has received scant attention in the context of forecast combination, and it is mainly focused on trimming based on the principles of expertise. Therefore, automatic pooling techniques considering both expertise and diversity merit further attention and development.

2.4 Forecast combination puzzle

Despite the explosion of a variety of popular and sophisticated combination methods, empirical evidence and extensive simulations repeatedly show that the simple average with equal weights often outperforms more complicated weighting schemes. This somewhat surprising result has occupied a very large literature, including the early studies by Stock and Watson (1998, 2003, 2004), the series of Makridakis competitions (Makridakis et al., 1982; Makridakis and Hibon, 2000; Makridakis, Spiliotis and Assimakopoulos, 2020*a*), and also the more recent articles by Blanc and Setzer (2016), etc. Clemen (1989) surveyed the early combination studies and raised a variety of issues remain to be addressed, one of which was ‘What is the explanation for the robustness of the simple average of forecasts?’. In a recent study, Gastinger et al. (2021) investigated the forecasting performance of a collection of combination methods on a large amount of time series from diverse sources and found that the winning combination methods differed for the different data sources, while the simple average strategies showed, on average, more gains at improving accuracy than other complex methods. Stock and Watson (2004) coined the term ‘forecast combination puzzle’ for the phenomenon—theoretically sophisticated weighting schemes should provide more benefits than the simple average from forecast combination, while empirically the simple average has been continuously found to dominate more complicated approaches to combining forecasts.

Some explanations of why the simple average might dominate the optimal combination in practice have centred on estimation error—the error on the estimation of the target optimal weights. Intuitions arise from the fact that the combination weights must be estimated in sophisticated weighting schemes, while no parameters are estimated at all in the simple average. Smith and Wallis (2009) demonstrated that the simple average was expected to overshadow the weighted average in a situation where the weights were theoretically equivalent. The results from simulations and an empirical study showed the estimation cost of weighted averages when the optimal weights were close to equality, thus providing an empirical explanation of the puzzle. Following them came an extension work by Claeskens et al. (2016) to provide a theoretical explanation for the empirical results. Taking the estimation of optimal weights into account, Claeskens et al. (2016) considered random weights rather than fixed weights during the optimality derivation and showed that, in this case, the forecast combination may introduce biases in combinations of unbiased component forecasts and the variance of the forecast combination may be larger than in the fixed-weight case, such as the simple average. More recently, Chan and Pauwels (2018) proposed a framework to study the theoretical properties of forecast combination. The proposed framework verified the estimation error explanation of the forecast combination puzzle and, more crucially, provided an additional insight into the puzzle. Specifically, the Mean Squared Forecast Error (MSFE) can be considered as a variance estimator of the forecast errors which may not be consistent, leading to biased results with different weighting schemes based on a simple comparison of MSFE values.

Explaining the puzzle using estimation error requires a hypothesis that potential gains from the optimal combination are not too large so that estimation error overwhelms the gains. Special cases, such as the covariance matrix of the forecast errors has all variances equal to each other and all covariances equal to a constant, are illustrated by Timmermann (2006) and Hsiao and Wan (2014) to arrive at equivalence between the simple average and the optimal combination. Elliott (2011) characterized the potential bounds on the size of gains from the optimal weights over the equal weights and illustrated that these gains were often too small to balance estimation error, providing a supplementary explanation of the puzzle for the large estimation error explanation.

The examination and explanation of the forecast combination puzzle can provide decision makers with some guidelines to identify which combination method to choose in specific forecasting problems.

- Estimation errors are identified as 'finite-sample estimation effects' in Smith and Wallis (2009), which suggests that insufficiently small sample size may be unable to provide robust weight estimates. Thus, if one faces limited historical data, the simple average or estimated weights with covariances between forecast errors being neglected are recommended. In addition, alternative simple combination operators such as trimmed and winsorized means can be adopted to eliminate extreme forecasts, and thus, offer more robust estimates than the simple average.

- Structural changes which may cause different weight estimates in the training and evaluation samples tend to impact sophisticated combination approaches more than the simple average. This case makes the simple average the better choice. The forecast combinations using changing weights can also be considered as a means to cope with structural changes, as suggested in Diebold and Pauly (1987) and Deutsch et al. (1994).

- If one has access to a large number of component forecasts, the PCR and the clustering strategy (for details, see Subsection 2.2.1) might be useful to diminish estimation errors and solve the multicollinearity problem by reducing the number of parameters need to be estimated.

- Involving time series features and individual forecasts with some extent of diversity in the process of weight estimation can enlarge the gains of the forecast combination, providing a possible way to untangle the forecast combination puzzle.

In summary, forecasters are encouraged to analyze the data prior to identifying the combination strategy and choose combination rules tailored to specific forecasting problems.

3 Probabilistic forecast combinations

- Combining quantiles and prediction intervals.

- Combining distributions as in fable.

The focuses of research on forecast combination are initially put on the combination of point forecasts (for details, see Section 2) since the pioneering work of Bates and Granger (1969). However, increasing attention has been shifted towards reporting probabilistic forecasts to provide decision makers with more insights than a single forecast. For example, the recent Makridakis competitions, the M4 and the M5 Uncertainty competitions (Makridakis, Spiliotis, Assimakopoulos, Chen, Gaba, Tsetlin and Winkler, 2020), encouraged participants to provide probabilistic forecasts of different types as well as point forecasts. Probabilistic forecasts are appealing for (i) involving an associated probability related to the reported forecasts, (ii) enabling optimal decision making with an understanding of uncertainties and the resulting risks, and (iii) allowing an overall comparison of forecasts from different forecasting models. In consequence, there has been a growing interest in brining together probabilistic forecasting and forecast combination, leading to probabilistic forecast combinations.

Probabilistic forecasts involve three main forms, namely prediction intervals (PIs), quantiles, and probability forecasts (also known as density or distribution forecasts), with probability forecasts being the most complete form. PIs are often constructed using quantile forecasts and the endpoints can be interpreted as the specific quantiles of a forecast distribution. Specifically, the lower and upper endpoints of a central $(1-\alpha) \times 100\%$ PI can be defined by the quantiles at level $\alpha/2$ and $1-\alpha/2$. In addition, the quantile forecast obtained from a forecasting model is the inverse of the corresponding probability forecast denoted by the cumulative distribution function. Nevertheless, when individual forecasts from different models are combined, the combined quantile forecast and the combined probability forecast may not be equivalent. Simple examples of averaging quantiles and probabilities with equal weights are provided in Lichtendahl et al. (2013).

In combining probabilistic forecasts, it is important to pose issues such as calibration and sharpness when considering which combination strategy should be selected and evaluating the quality of combined forecasts. Thus, in this section, we first introduce the assessment of the performance of probabilistic forecasts. Then the literature on the combination of probabilistic forecasts is divided into two streams: combining quantile forecasts and combining probability forecasts.

3.1 Evaluation of probabilistic forecasts

Principles concerning calibration, sharpness (Gneiting and Raftery, 2007; Gneiting et al., 2007), scoring rules, and shape (Lichtendahl et al., 2013), etc.

- calibration
  - perfectly calibrated: $\Pr(F(y) \leq u) = u$ for $0 < u < 1$.
  - overconfident: $\Pr(F(y) \leq u) \geq u$ for $0 < u < 1/2$, $\Pr(F(y) \leq u) = u$ for $u = 1/2$, and

  $\Pr(F(y) > u) \leq u$ for $1/2 < u < 1$, with the inequalities strict for some u.
  - underconfident: $\Pr(F(y) \leq u) \leq u$ for $0 < u < 1/2$, $\Pr(F(y) \leq u) = u$ for $u = 1/2$, and $\Pr(F(y) > u) \geq u$ for $1/2 < u < 1$, with the inequalities strict for some u.

- sharpness
  - we say F is sharper than G if F's variance of x is less than or equal to G's.
- goal of probabilistic forecasting
  - 'maximize the sharpness of the predictive distributions subject to calibration' (Gneiting and Raftery, 2007)
- scoring rules.

3.2 Combining quantile forecasts

Include the literatures on combining intervals (e.g., Lichtendahl et al., 2013; Park and Budescu, 2015; Gaba et al., 2017; Grushka-Cockayne et al., 2017; Grushka-Cockayne and Jose, 2020)

3.2.1 Simple combination heuristics

- Six heuristics (Park and Budescu, 2015; Gaba et al., 2017; Grushka-Cockayne and Jose, 2020) proposed for combining intervals.
- Generalize these heuristics to the combination of quantile forecasts.
- Limitations and the selection of heuristics in different forecasting problems.

quantile crossing

3.3 Combining probability forecasts

(e.g., Genest and Zidek, 1986; Ranjan and Gneiting, 2010; Clements and Harvey, 2011)

3.3.1 Linear opinion pool

- linear opinion pool (Genest and McConway, 1990)—a convex combination
- trimmed opinion pools (Jose et al., 2014)
- generalized linear opinion pool
- logarithmic opinion pool (Genest and Zidek, 1986)

Diversity hurt the average probability forecast

3.3.2 Bayesian approach

- Bayesian model averaging (Raftery et al., 1997)

4 Probabilistic ensembles

- Meteorological ensembles.
- True ensembles in other areas (i.e., not papers that use the word ”ensemble” but papers that use mixtures when forecasting).
- When is an ensemble equivalent to combination?
- When do point forecasts from an ensemble equal point forecasts from a combination?

5 Boosting in forecasting

6 Bagging in forecasting

7 Stacking in forecasting

References

Adhikari, R. (2015), 'A mutual association based nonlinear ensemble mechanism for time series forecasting', *Applied Intelligence* .

Adhikari, R. and Agrawal, R. (2012), A novel weighted ensemble technique for time series forecasting, *in* 'Pacific-Asia Conference on Knowledge Discovery and Data Mining', Springer, pp. 38–49.

Agnew, C. E. (1985), 'Bayesian consensus forecasts of macroeconomic variables', *Journal of Forecasting* **4**(4), 363–376.

Aiolfi, M. and Timmermann, A. (2006), 'Persistence in forecasting performance and conditional combination strategies', *Journal of Econometrics* **135**(1), 31–53.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Aksu, C. and Gunter, S. I. (1992), 'An empirical analysis of the accuracy of SA, OLS, ERLS and NRLS combination forecasts', *International Journal of Forecasting* **8**(1), 27–43.

Andrawis, R. R., Atiya, A. F. and El-Shishiny, H. (2011), 'Combination of long term and short term forecasts, with application to tourism demand forecasting', *International Journal of Forecasting* **27**(3), 870–886.

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.

Atiya, A. F. (2020), 'Why does forecast combination work so well?', *International Journal of Forecasting* **36**(1), 197–200.

Babikir, A. and Mwambi, H. (2016), ‘Evaluating the combined forecasts of the dynamic factor model and the artificial neural network model using linear and nonlinear combining methods’, *Empirical Economics* .

Barnard, G. A. (1963), ‘New methods of quality control’, *Journal of the Royal Statistical Society. Series A* **126**(2), 255.

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. J. (1969), ‘The combination of forecasts’, *The Journal of the Operational Research Society* **20**(4), 451–468.

Baumeister, C. and Kilian, L. (2015), ‘Forecasting the real price of oil in a changing world: A forecast combination approach’, *Journal of Business & Economic Statistics* **33**(3), 338–351.

Billah, B., Hyndman, R. J. and Koehler, A. B. (2005), ‘Empirical information criteria for time series forecasting model selection’, *Journal of Statistical Computation and Simulation* **75**(10), 831–840.

Blanc, S. M. and Setzer, T. (2016), ‘When to choose the simple average in forecast combination’, *Journal of Business Research* **69**(10), 3951–3962.

Brown, G., Wyatt, J., Harris, R. and Yao, X. (2005), ‘Diversity creation methods: a survey and categorisation’, *Information Fusion* **6**(1), 5–20.

Budescu, D. V. and Chen, E. (2015), ‘Identifying expertise to extract the wisdom of crowds’, *Management Science* .

Bunn, D. W. (1975), ‘A bayesian approach to the linear combination of forecasts’, *Journal of the Operational Research Society* .

Bunn, D. W. (1985), ‘Statistical efficiency in the linear combination of forecasts’, *International Journal of Forecasting* **1**(2), 151–163.

Burnham, K. P. and Anderson, D. R. (2002), ‘Model selection and multi-model inference: A practical information-theoretic approach (2nd ed.)’, *Berlin, New York: Springer* .

Chan, F. and Pauwels, L. L. (2018), ‘Some theoretical results on forecast combinations’, *International Journal of Forecasting* **34**(1), 64–74.

Chan, Y. L., Stock, J. H. and Watson, M. W. (1999), 'A dynamic factor model framework for forecast combination', *Spanish Economic Review* **1**(2), 91–121.

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.

Clark, T. E. and McCracken, M. W. (2010), 'Averaging forecasts from VARs with uncertain instabilities', *Journal of Applied Econometrics* .

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

Clemen, R. T. and Winkler, R. L. (1986), 'Combining economic forecasts', *Journal of Business & Economic Statistics* **4**(1), 39–46.

Clements, M. P. and Harvey, D. I. (2011), 'Combining probability forecasts', *International Journal of Forecasting* **27**(2), 208–223.

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.

Conflitti, C., De Mol, C. and Giannone, D. (2015), 'Optimal combination of survey forecasts', *International Journal of Forecasting* **31**(4), 1096–1103.

Crane, D. B. and Crotty, J. R. (1967), 'A two-stage forecasting model: Exponential smoothing and multiple regression', *Management Science* **13**(8), B–501.

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* .

Deutsch, M., Granger, C. W. J. and Teräsvirta, T. (1994), 'The combination of forecasts using changing weights', *International Journal of Forecasting* **10**(1), 47–57.

Diebold, F. X. (1988), 'Serial correlation and the combination of forecasts', *Journal of Business & Economic Statistics* **6**(1), 105–111.

Diebold, F. X. and Pauly, P. (1987), 'Structural change and the combination of forecasts', *Journal of Forecasting* **6**(1), 21–40.

Diebold, F. X. and Pauly, P. (1990), ‘The use of prior information in forecast combination’, *International Journal of Forecasting* **6**(4), 503–508.

Donaldson, R. G. and Kamstra, M. (1996), ‘Forecast combining with neural networks’, *Journal of Forecasting* .

Donate, J. P., Cortez, P., Sanchez, G. G. and De Miguel, A. S. (2013), ‘Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble’, *Neurocomputing* .

Edward Coulson, N. and Robins, R. P. (1993), ‘Forecast combination in a dynamic setting’, *Journal of Forecasting* **12**(1), 63–67.

Elliott, G. (2011), ‘Averaging and the optimal combination of forecasts’, *Manuscript, Department of Economics, UCSD* .

Elliott, G. and Timmermann, A. (2004), ‘Optimal forecast combinations under general loss functions and forecast error distributions’, *Journal of Econometrics* **122**(1), 47–79.

Fildes, R. and Petropoulos, F. (2015), ‘Simple versus complex selection rules for forecasting many time series’, *Journal of Business Research* **68**(8), 1692–1701.

Fischer, I. and Harvey, N. (1999), ‘Combining forecasts: What information do judges need to outperform the simple average?’, *International Journal of Forecasting* **15**(3), 227–246.

Freitas, P. S. A. and Rodrigues, A. J. L. (2006), ‘Model combination in neural-based forecasting’, *European Journal of Operational Research* **173**(3), 801–814.

Gaba, A., Tsetlin, I. and Winkler, R. L. (2017), ‘Combining interval forecasts’, *Decision Analysis* **14**(1), 1–20.

Gastinger, J., Nicolas, S., Stepić, D., Schmidt, M. and Schülke, A. (2021), ‘A study on ensemble learning for time series forecasting and the need for meta-learning’.

Genest, C. and McConway, K. J. (1990), ‘Allocating the weights in the linear opinion pool’, *Journal of Forecasting* **9**(1), 53–73.

Genest, C. and Zidek, J. V. (1986), ‘Combining probability distributions: A critique and an annotated bibliography’, *Statistical Science* **1**(1), 114–135.

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.

Geweke, J. and Amisano, G. (2011), ‘Optimal prediction pools’, *Journal of Econometrics* **164**(1), 130–141.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007), ‘Probabilistic forecasts, calibration and sharpness’, *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **69**(2), 243–268.

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

Granger, C. W. J. (1989), ‘Invited review combining forecasts—twenty years later’, *Journal of Forecasting* **8**(3), 167–173.

Granger, C. W. J. and Jeon, Y. (2004), ‘Thick modeling’, *Economic Modelling* **21**(2), 323–343.

Granger, C. W. J. and Ramanathan, R. (1984), ‘Improved methods of combining forecasts’, *Journal of Forecasting* **3**(2), 197–204.

Grushka-Cockayne, Y. and Jose, V. R. R. (2020), ‘Combining prediction intervals in the M4 competition’, *International Journal of Forecasting* **36**(1), 178–185.

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

Gunter, S. I. (1992), ‘Nonnegativity restricted least squares combinations’, *International Journal of Forecasting* **8**(1), 45–59.

Harrald, P. G. and Kamstra, M. (1997), ‘Evolving artificial neural networks to combine financial forecasts’, *IEEE Transactions on Evolutionary Computation* **1**(1), 40–52.

Hsiao, C. and Wan, S. K. (2014), ‘Is there an optimal forecast combination?’, *Journal of Econometrics* **178**, 294–309.

Hyndman, R. J., Kang, Y., Montero-Manso, P., Talagala, T. S., 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*

Inoue, A., Jin, L. and Rossi, B. (2017), ‘Rolling window selection for out-of-sample forecasting with time-varying parameters’, *Journal of Econometrics* **196**(1), 55–67.

Jose, V. R. R., Grushka-Cockayne, Y. and Lichtendahl, K. C. (2014), ‘Trimmed opinion pools and the crowd’s calibration problem’, *Management Science* **60**(2), 463–475.

Jose, V. R. R. and Winkler, R. L. (2008), ‘Simple robust averages of forecasts: Some empirical results’, *International Journal of Forecasting* **24**(1), 163–169.

Kang, Y., Cao, W., Petropoulos, F. and Li, F. (2020), ‘Forecast with forecasts: Diversity matters’.

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., Hyndman, R. J. and Smith-Miles, K. (2017), ‘Visualising forecasting algorithm performance using time series instance spaces’, *International Journal of Forecasting* **33**(2), 345–358.

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* .

Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* ‘Ijcai’, Vol. 14, researchgate.net, pp. 1137–1145.

Kolassa, S. (2011), ‘Combining exponential smoothing forecasts using akaike weights’, *International Journal of Forecasting* **27**(2), 238–251.

Kourentzes, N., Barrow, D. K. and Crone, S. F. (2014), ‘Neural network ensemble operators for time series forecasting’, *Expert Systems with Applications* **41**(9), 4235–4244.

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**(February 2018), 226–235.

Krasnopolsky, V. M. and Lin, Y. (2012), ‘A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental US’, *Advances in Meteorology* **2012**.

Kullback, S. and Leibler, R. A. (1951), ‘On information and sufficiency’, *Annals of Mathematical Statistics* .

Larrick, R. P. and Soll, J. B. (2006), 'Intuitions about combining opinions: Misappreciation of the averaging principle', *Management Science* .

Lemke, C. and Gabrys, B. (2010), 'Meta-learning for time series forecasting and forecast combination', *Neurocomputing* **73**(10), 2006–2016.

Leutbecher, M. and Palmer, T. N. (2008), 'Ensemble forecasting', *Journal of Computational Physics* **227**(7), 3515–3539.

Lewis, J. M. (2005), 'Roots of ensemble forecasting', *Monthly Weather Review* **133**(7), 1865–1885.

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

Lichtendahl, Jr, K. C., Grushka-Cockayne, Y. and Winkler, R. L. (2013), 'Is it better to average probabilities or quantiles?', *Management Science* **59**(7), 1594–1611.

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

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982), 'The accuracy of extrapolation (time series) methods: Results of a forecasting competition', *Journal of Forecasting* **1**(2), 111–153.

Makridakis, S. and Hibon, M. (2000), 'The M3-Competition: results, conclusions and implications', *International Journal of Forecasting* **16**(4), 451–476.

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

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2020*b*), 'The M5 accuracy competition: Results, findings and conclusions', *International Journal of Forecasting* .

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I. and Winkler, R. (2020), 'The M5 Uncertainty competition: Results, findings and conclusions', *International Journal of Forecasting* pp. 1–24.

Makridakis, S. and Winkler, R. L. (1983), 'Averages of forecasts: Some empirical results'.

Mannes, A. E., Soll, J. B. and Larrick, R. P. (2014), 'The wisdom of select crowds', *Journal of Personality and Social Psychology* **107**(2), 276–299.

McNees, S. K. (1992), 'The uses and abuses of 'consensus' forecasts', *Journal of Forecasting* **11**(8), 703–710.

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.

Moon, J., Jung, S., Rew, J., Rho, S. and Hwang, E. (2020), 'Combination of short-term load forecasting models based on a stacking ensemble approach', *Energy and Buildings* **216**, 109921.

Newbold, P. and Granger, C. W. J. (1974), 'Experience with forecasting univariate time series and the combination of forecasts', *Journal of the Royal Statistical Society. Series A* **137**(2), 131.

Newbold, P. and Harvey, D. I. (2002), 'Forecast combination and encompassing', *A companion to economic forecasting* .

Nowotarski, J., Raviv, E., Trück, S. and Weron, R. (2014), 'An empirical comparison of alternative schemes for combining electricity spot price forecasts', *Energy Economics* **46**, 395–412.

Öller, L.-E. (1978), 'A method for pooling forecasts', *Journal of the Operational Research Society* **29**(1), 55–63.

Palm, F. C. and Zellner, A. (1992), 'To combine or not to combine? issues of combining forecasts', *Journal of Forecasting* **11**(8), 687–701.

Park, S. and Budescu, D. V. (2015), 'Aggregating multiple probability intervals to improve calibration', *Judgment and Decision Making* **10**(2), 130.

Patton, A. J. and Timmermann, A. (2007), 'Properties of optimal forecasts under asymmetric loss and nonlinearity', *Journal of Econometrics* **140**(2), 884–918.

Pawlikowski, M. and Chorowska, A. (2020), 'Weighted ensemble of statistical models', *International Journal of Forecasting* **36**(1), 93–97.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Bergmeir, C., Bessa, R. J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., Baets, S. D., Dokumentov, A., Fiszeder, P., Franses,

P. H., Gilliland, M., Gönül, M. S., 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., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. 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. (2020), 'Forecasting: theory and practice'.

Petropoulos, F., Hyndman, R. J. and Bergmeir, C. (2018), 'Exploring the sources of uncertainty: Why does bagging for time series forecasting work?', *European Journal of Operational Research* **268**(2), 545–554.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K. and Siemsen, E. (2018), 'Judgmental selection of forecasting models', *Journal of Operations Management* **60**, 34–46.

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

Poler, R. and Mula, J. (2011), 'Forecasting model selection through out-of-sample rolling horizon weighted errors', *Expert Systems with Applications* **38**(12), 14778–14785.

Poncela, P., Rodríguez, J., Sánchez-Mangas, R. and Senra, E. (2011), 'Forecast combination through dimension reduction techniques', *International Journal of Forecasting* **27**(2), 224–237.

Qi, M. and Zhang, G. P. (2001), 'An investigation of model selection criteria for neural network time series forecasting', *European Journal of Operational Research* **132**(3), 666–680.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997), 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association* **92**(437), 179–191.

Ranjan, R. and Gneiting, T. (2010), 'Combining probability forecasts', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(1), 71–91.

Rapach, D. E. and Strauss, J. K. (2008), 'Forecasting US employment growth using forecast combining methods', *Journal of Forecasting* **27**(1), 75–93.

Ribeiro, G. T., Mariani, V. C. and Coelho, L. d. S. (2019), 'Enhanced ensemble structures using wavelet neural networks applied to short-term load forecasting', *Engineering Applications of Artificial Intelligence* **82**, 272–281.

Ribeiro, M. H. D. M. and dos Santos Coelho, L. (2020), 'Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series', *Applied Soft Computing* **86**, 105837.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**(2), 461–464.

Semenoglou, A.-A., Spiliotis, E., Makridakis, S. and Assimakopoulos, V. (2020), 'Investigating the accuracy of cross-learning time series forecasting methods', *International Journal of Forecasting* .

Shaub, D. (2019), 'Fast and accurate yearly time series forecasting with forecast combinations', *International Journal of Forecasting* .

Shi, S. and Liu, B. (1993), Nonlinear combination of forecasts with neural networks, *in* 'Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)', Vol. 1, ieeexplore.ieee.org, pp. 959–962 vol.1.

Shi, S. M., Da Xu, L. and Liu, B. (1999), 'Improving the accuracy of nonlinear combined forecasting using neural networks', *Expert Systems with Applications* **16**(1), 49–54.

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.

Stock, J. H. and Watson, M. W. (1998), A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, Technical Report w6607, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2003), 'How did leading indicator forecasts perform during the 2001 recession?', *FRB Richmond Economic Quarterly* **89**(3), 71–90.

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

Sugiura, N. (1978), ‘Further analysts of the data by akaike’s information criterion and the finite corrections: Further analysts of the data by akaike’s’, *Communications in Statistics-Theory and Methods* .

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

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

Timmermann, A. (2006), Chapter 4 forecast combinations, *in* G. Elliott, C. W. J. Granger and A. Timmermann, eds, ‘Handbook of Economic Forecasting’, Vol. 1, Elsevier, pp. 135–196.

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

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), 2581–2594.

Winkler, R. L. and Makridakis, S. (1983), ‘The combination of forecasts’, *Journal of the Royal Statistical Society. Series A* **146**(2), 150.

Wolpert, D. H. (1992), ‘Stacked generalization’, *Elsevier Oceanography Series* **87545**(505), 1–57.

Yang, Y. (2005), ‘Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation’, *Biometrika* **92**(4), 937–950.

Yao, X. and Islam, M. M. (2008), ‘Evolving artificial neural network ensembles’, *IEEE Computational Intelligence Magazine* **3**(1), 31–42.

Zhao, S. and Feng, Y. (2020), ‘For2For: Learning to forecast from forecasts’.

Zhou, Z.-H. (2012), ‘Ensemble methods: foundations and algorithms’.

Zhou, Z.-H., Wu, J. and Tang, W. (2002), ‘Ensembling neural networks: Many could be better than all’, *Artificial Intelligence* **137**(1), 239–263.