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?

Several 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, 1998) and the Bayesian Information Criterion (BIC, Schwarz et al., 1978), are commonly used 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 et al., 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 is 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 can be substantially improved by combining multiple forecasts. Timmermann (2006) summarized the benefits of forecast combinations into the fact that individual forecasts are 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 are beneficial due to diversification gains. More recently, Atiya (2020) illustrated graphically why forecast combinations are 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 en-

semble 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; 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 by the importance of parameter estimation error—that is, simple combination schemes do not require estimating parameters such as combination weights based on forecast errors, thus avoiding parameter estimation error that often exists in the weighted combination.

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 1001 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 dominate more refined forecast combinations. Makridakis and Winkler (1983) concluded empirically that the accuracy of combined forecasts is improved and the variability associated with the choice of methods is reduced, as the number of individual methods included in a simple average increases. Palm and Zellner (1992) concisely summarized the advantages of adopting a simple average into three points: (i) combination weights are equal and do not have to be estimated, (ii) a simple average significantly reduces 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 is taken into account. Furthermore, Timmermann (2006) pointed out that the good average performance of the simple average depends 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 affected by 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 are appealing 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 deal 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 have been largely overlooked in 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 et al., 2020) show that simple combinations can 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 sets 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 et al., 2020; Montero-Manso et al., 2020; Kang et al., 2020; Wang et al., 2021).

2.2 Combination weighting schemes

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 in an N-vector of individual forecasts, $\hat{\mathbf{y}}_{t+h,t} = (\hat{y}_{t+h,t,1}, \hat{y}_{t+h,t,2}, \ldots, \hat{y}_{t+h,t,N})'$, to a combined point 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, non-linear, and time-varying combinations. Below we discuss in detail the use of various combination weight schemes to determine combination weights.

2.2.1 Linear combinations

Typically, the combined forecast is commonly constructed as a linear combination of the individual forecasts. To this end a combined forecast of the linear form can be written as

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

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

1. 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 variance of errors is 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

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

where $\Sigma_{t+h,t}$ 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 $\Sigma_{t+h,t}$ 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 find 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 y_{t+h} and $\hat{\boldsymbol{y}}_{t+h,t}$.

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 depend 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 are 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.

2. Performance-based weights

Estimation errors of the optimal weights given by Equation (2) tend to be particularly large due to difficulties in properly estimating the covariance matrix $\Sigma_{t+h,t}$. In consequence Bates and Granger (1969) and Newbold and Granger (1974) suggested treating the covariance matrix as a diagonal matrix, leading to the combination weights with the expression

$$w_{t+h,t,i} = \frac{\left(\sum_{\tau=t-\nu}^{t} e_{\tau,\tau-h,i}^2\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=t-\nu}^{t} e_{\tau,\tau-h,j}^2\right)^{-1}}, \tag{3}$$

where $e_{t+h,t,i}$ is the h-period forecast error at time t from ith individual model. It is clear that the combination weights ignore correlations across individual forecasts and are inversely proportional to the corresponding MSE performance. Winkler and Makridakis (1983) provided extensive empirical results using the suggested procedures for estimating combination weights. Their results reconfirmed Newbold and Granger's (1974) argument that correlations can be poorly estimated in practice and should be ignored in calculating combination weights.

Stock and Watson (1998): Forecast pooling procedure

Stock and Watson (2004): Discounted MSFE forecasts

?: MSE-weighted (application) Kolassa (2011): Combining exponential smoothing forecasts using AIC, AICc, and BIC

Genre et al. (2013): performance-based weighting

Nowotarski et al. (2014): IRMSE averaging

Baumeister and Kilian (2015): Although the accuracy of the forecast combinations remains remarkably robust across alternative specifications, our results indicate that inverse MSPE weights

based on recursive or rolling windows of past data generate less accurate forecasts than constant equal weights.

Pawlikowski and Chorowska (2020): utilizes several commonly used statistical models, which are weighted according to their performance on historical data.

Rank-based combinations

───

3. Regression weights

Least squares regression of Granger and Ramanathan (1984).

Granger and Ramanathan (1984) showed that the method of optimal weights (Bates and Granger, 1969; Newbold and Granger, 1974) is equivalent to a least squares regression in which the constant is suppressed and the weights are constrained to sum to one.

Genre et al. (2013)

? Optimal combination of point forecasts and density forecasts (OLS)

4. Bayesian weights

Bayesian techniques by Bunn (1975) and Clemen and Winkler (1993)

2.2.2 Nonlinear combinations

2.2.3 Time-varying combinations

2.2.4 Combining by learning

- Stacking: introduce stacking as a combination/aggregation method rather than a generalization of many ensemble methods in ML.

- Regression-based combinations: a simple stacking.

- Meta-learning that only takes individual forecasts as input.

- Stacking extension: use other information in the meta-learner, such as feature and diversity (FFORMA).

2.3 Forecast Combination Puzzle

(why don't weights work?)

(De Menezes et al., 2000; Genre et al., 2013; Post et al., 2019; Chan and Pauwels, 2018; Lichtendahl and Winkler, 2020; Kourentzes et al., 2019)

3 Probabilistic forecast combinations

- Combining quantiles and prediction intervals.

- Combining distributions as in fable.

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

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

Akaike, H. (1998), Information theory and an extension of the maximum likelihood principle, *in* 'Selected papers of hirotugu akaike', Springer, pp. 199–213.

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

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

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.

Bunn, D. W. (1975), 'A bayesian approach to the linear combination of forecasts', *Operational Research Quarterly* **26**(2), 325.

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

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.

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.

Clemen, R. T. and Winkler, R. L. (1993), 'Aggregating point estimates: A flexible modeling approach', *Management Science* **39**(4), 501–515.

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* .

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.

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.

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

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

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

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., 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.

Kohavi, R. et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'IJCAI', Vol. 14, Montreal, Canada, 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.

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.

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), 'The M4 competition: 100,000 time series and 61 forecasting methods', *International Journal of Forecasting* **36**(1), 54–74.

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

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.

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.

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.

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

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. 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.

Post, T., Karabatı, S. and Arvanitis, S. (2019), 'Robust optimization of forecast combinations', *International Journal of Forecasting* **35**(3), 910–926.

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.

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

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

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?

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.

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.

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* .

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

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.