ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecast combinations: An over 50-year review

Xiaoqian Wang [a], Rob J. Hyndman [b], Feng Li [c], Yanfei Kang [a,*]

[a] School of Economics and Management, Beihang University, Beijing 100191, China
[b] Department of Econometrics & Business Statistics, Monash University, Clayton VIC 3800, Australia
[c] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China

ARTICLE INFO

Keywords:
Combination forecast
Cross learning
Forecast combination puzzle
Forecast ensembles
Model averaging
Open-source software
Pooling
Probabilistic forecasts
Quantile forecasts

ABSTRACT

Forecast combinations have flourished remarkably in the forecasting community and, in recent years, have become part of mainstream forecasting research and activities. Combining multiple forecasts produced for a target time series is now widely used to improve accuracy through the integration of information gleaned from different sources, thereby avoiding the need to identify a single "best" forecast. Combination schemes have evolved from simple combination methods without estimation to sophisticated techniques involving time-varying weights, nonlinear combinations, correlations among components, and cross-learning. They include combining point forecasts and combining probabilistic forecasts. This paper provides an up-to-date review of the extensive literature on forecast combinations and a reference to available open-source software implementations. We discuss the potential and limitations of various methods and highlight how these ideas have developed over time. Some crucial issues concerning the utility of forecast combinations are also surveyed. Finally, we conclude with current research gaps and potential insights for future research.

1. Introduction

The idea of combining multiple individual forecasts dates back to Francis Galton, who in 1906 visited an ox-weight-judging competition and observed that the average of 787 estimates of an ox's weight was remarkably close to the ox's actual weight; see Surowiecki (2005) for more details. About sixty years later, the work of Bates and Granger (1969) popularized the idea and spawned a rich literature on forecast combinations. More than fifty years have passed since Bates and Granger's (1969) seminal work, and it is now well established that those forecast combinations are beneficial. They offer substantially improved forecasts on average relative to constituent models; see Clemen (1989) and Timmermann (2006) for extensive literature reviews.

This paper aims to present an up-to-date modern review of the literature on forecast combinations over the past five decades. We cover various forecast combination methods for both point and probabilistic forecasts, contrasting them and highlighting how different related ideas have developed in parallel.

Combining multiple forecasts derived from numerous forecasting methods is often better than identifying a single "best forecast". These are usually called "combination forecasts" or "ensemble forecasts" in different domains. Observed time series data are unlikely to be generated by a simple process specified with a specific functional form because of the possibility of time-varying trends, seasonality changes, structural breaks, and the complexity of real data generating processes (Clements & Hendry, 1998). Thus, selecting a single "best model" to approximate the unknown underlying data generating process may be misleading and is subject to at least three sources of uncertainty: data uncertainty, parameter uncertainty, and model uncertainty (Kourentzes et al., 2019; Petropoulos et al., 2018a). Given these challenges, it is

* Corresponding author.
E-mail addresses: xiaoqianwang@buaa.edu.cn (X. Wang), rob.hyndman@monash.edu (R.J. Hyndman), feng.li@cufe.edu.cn (F. Li), yanfeikang@buaa.edu.cn (Y. Kang).

https://doi.org/10.1016/j.ijforecast.2022.11.005
0169-2070/

Please cite this article as: X. Wang, R.J. Hyndman, F. Li et al., Forecast combinations: An over 50-year review. International Journal of Forecasting (2022), https://doi.org/10.1016/j.ijforecast.2022.11.005.

often better to combine multiple forecasts to incorporate multiple drivers of the data generating process and mitigate uncertainties regarding model form and parameter specification.

Potential explanations for the strong performance of forecast combinations are manifold. First, the combination is likely to improve forecasting performance when multiple forecasts to be combined incorporate partial (but incompletely overlapping) information. Second, structural breaks are a common motivation for combining forecasts from different models (Timmermann, 2006). In the presence of structural breaks and other instabilities, combining forecasts from models with varying degrees of misspecification and adaptability can mitigate the problem, helping to explain the empirical success of forecast combinations. See, e.g., Rossi (2013, 2021) for an extensive discussion on forecast combinations in the presence of instabilities. One can consider the competing forecasts as intercept correction relative to a baseline forecast. This provides potential gains in forecast accuracy if there are either structural breaks or deterministic misspecifications (Hendry & Clements, 2004). Finally, Hendry and Clements (2004) noted that forecast combination could be an application of Stein–James shrinkage estimation (Judge & Bock, 1978). Specifically, if the unknown future value is considered as a "meta-parameter" of which all the individual forecasts are estimates, then averaging has the potential to provide an improved estimate.

In light of their superiority, forecast combinations have appeared in a wide range of applications such as retail (Ma & Fildes, 2021), energy (Xie & Hong, 2016), economics (Aastveit et al., 2019), and epidemiology (Ray et al., 2022). Among all published forecasting papers included in the Web of Science, the proportion of papers concerning forecast combinations has been trending upward over the past 50 years, reaching 13.80% in 2021, as shown in Fig. 1. Consequently, reviewing the extant literature on this topic is timely and necessary.

The gains from forecast combinations rely on not only the quality of the individual forecasts to be combined but the estimation of the combination weights assigned to each forecast (Cang & Yu, 2014; Timmermann, 2006). Numerous studies have been devoted to discussing critical issues concerning the constitution of the model pool and the selection of the optimal model subset, including but not limited to the accuracy, diversity, and robustness of individual models (Batchelor & Dua, 1995; Kang et al., 2021; Lichtendahl & Winkler, 2020; Mannes et al., 2014; Thomson et al., 2019). On the other hand, combination schemes vary across studies. They have evolved from simple combination methods that avoid weight estimation (e.g., Clemen & Winkler, 1986; Genre et al., 2013; Grushka-Cockayne, Jose & Lichtendahl, 2017; Palm & Zellner, 1992; Petropoulos & Svetunkov, 2020) to sophisticated methods that tailor weights for different individual models (e.g., Bates & Granger, 1969; Kang et al., 2021; Kolassa, 2011; Li et al., 2020; Montero-Manso et al., 2020; Newbold & Granger, 1974; Wang, Kang, Petropoulos & Li, 2022). Accordingly, forecast combinations can be linear or nonlinear, static or time-varying, series-specific or cross-learning, and ignore or cover correlations among individual forecasts. Despite the diverse set of forecast combination schemes, forecasters still have little guidance on solving the "forecast combination puzzle" (Chan & Pauwels, 2018; Claeskens et al., 2016; Smith & Wallis, 2009; Stock & Watson, 2004) — simple averaging often empirically dominates sophisticated weighting schemes that should (asymptotically) be superior.

Initial work on forecast combinations after the seminal work of Bates and Granger (1969) focused on dealing with point forecasts (see, e.g., Clemen, 1989; Timmermann, 2006). In recent years considerable attention has moved towards the use of probabilistic forecasts (e.g., Gneiting & Ranjan, 2013; Hall & Mitchell, 2007; Kapetanios et al., 2015; Martin et al., 2021) as they enable a rich assessment of forecast uncertainties. When working with probabilistic forecasts, issues such as diversity among individual forecasts can be more complex and less understood than combining point forecasts (Ranjan & Gneiting, 2010). Additional problems such as calibration and sharpness need to be considered when assessing or selecting a combination scheme (Gneiting et al., 2007). Probabilistic forecasts can be elicited in different forms (i.e., density forecasts, quantiles, prediction intervals, etc.), and the resulting combinations may have different properties such as calibration, sharpness, and shape; see Lichtendahl et al. (2013) for further analytical details.

We should clarify that we take the individual forecasts to be combined as given and do not discuss how the forecasts are generated. We focus on combinations of multiple forecasts derived from *separate* and *non-interfering* models for a given time series. Nevertheless, the literature involves at least two other types of combinations that are not covered in the present review. The first is the case of generating multiple series from the single (target) series, forecasting each of the generated series independently, and then combining the outcomes. Such data manipulation extracts more information from the target time series, which, in turn, can be used to enhance the forecasting performance. Petropoulos and Spiliotis (2021) referred to this category of forecast combinations generally as "wisdom of the data" and provided an overview of approaches in this category. In this particular context, the combination methods reviewed in this paper can function as tools to aggregate (or combine) the forecasts computed from different perspectives of the same data. The second type of forecast combination we do not cover is forecast reconciliation for hierarchical time series, which has developed over the past ten years since the pioneering work of Hyndman et al. (2011). Forecast reconciliation involves reconciling forecasts across the hierarchy to ensure that the forecasts sum appropriately across the hierarchy levels, and hence is a type of forecast combination.

We note that forecast combination and model averaging are sometimes used without distinction in the literature. The two terms overlap, but their focuses are different. "Model averaging" is a general term allowing for model uncertainty, particularly in parameter estimation, which can lead to better estimates and more reliable forecasts and prediction intervals than model selection (selecting a single best model). Several approaches to model averaging have been developed in statistics, econometrics, and machine learning. Two main strands can be

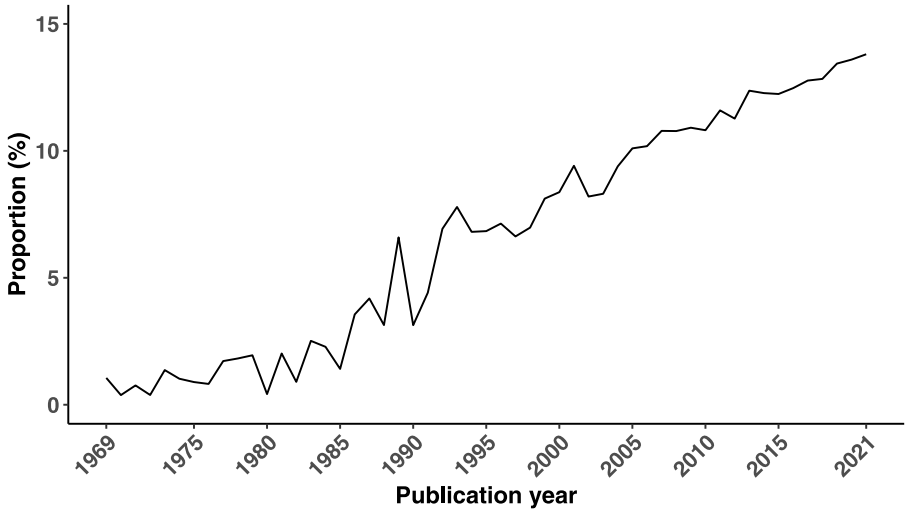

Fig. 1. The proportion of papers that concern forecast combinations among all published forecasting papers included in the Web of Science databases during the publication year range 1969–2021. Specifically, we use the search query `TS = (forecast*)` to find all forecasting papers. To find papers concerning forecast combinations, we use `TS = ((forecast* NEAR/5 combin*) OR (forecast* NEAR/5 ensemble*) OR (forecast* NEAR/5 averag*) OR (forecast* NEAR/5 aggregat*) OR (forecast* NEAR/5 pool*) OR (forecast* AND ((model* NEAR/5 combin*) OR (model* NEAR/5 ensemble*) OR (model* NEAR/5 averag*) OR (model* NEAR/5 aggregat*) OR (model* NEAR/5 pool*))))`.

identified: frequentist approaches (e.g., Fletcher, 2018) and Bayesian approaches (e.g. Steel, 2020). "Forecast combination" is a more focused terminology describing the combination of forecasts to generate a better forecast; the component forecasts could be outcomes from model averaging, individual models, or expert forecasts, e.g.. As with model averaging, weights can be used to combine the component forecasts. Unlike model averaging, however, forecast combination also has some underlying assumptions to ensure that the forecast combinations are unbiased or optimal.

This paper aims to contribute a broad perspective and historical overview of the main developments in forecast combinations. The paper is organized into two main sections on point forecast combinations (Section 2) and probabilistic forecast combinations (Section 3). Section 4 concludes the paper and identifies possible future developments.

2. Point forecast combinations

2.1. Simple point forecast combinations

A considerable literature has accumulated over the years regarding how individual forecasts are combined, with the unanimous conclusion that simple combination schemes are hard to beat (Clemen, 1989; Fischer & Harvey, 1999; Kang, 1986; Lichtendahl & Winkler, 2020; Stock & Watson, 2004). Equally weighted averages, which ignore past information regarding the precision of individual forecasts and correlations between forecast errors, work reasonably well compared to more sophisticated combination schemes.

The vast majority of studies on combining multiple forecasts have dealt with point forecasting, even though point forecasts (without associated measures of uncertainty) provide insufficient information for decision-making. The simple arithmetic average of forecasts based on equal weights stands out as the most popular and surprisingly robust combination rule (see Bunn, 1985; Clemen & Winkler, 1986; Genre et al., 2013; Stock & Watson, 2003), and can be effortlessly implemented.

An early example of an equally weighted combination is from the M-competition, the first forecasting competition run by Spyros Makridakis and Michèle Hibon, involving 1001 time series; see Makridakis et al. (1982) and Hyndman (2020) for more details of the competition. Makridakis et al. (1982) reported that the simple average outperformed the individual forecasting models. Clemen (1989) provided an extensive bibliographical review of the early work on the combination of forecasts and addressed the issue that the arithmetic means often dominate more refined forecast combinations. Makridakis and Winkler (1983) concluded that a larger number of individual methods included in the simple average scheme would help improve the accuracy of combined forecasts and reduce the variability associated with the selection of methods. Palm and Zellner (1992) concisely summarized the advantages of adopting simple averaging into three aspects: (i) combination weights are equal and do not have to be estimated; (ii) simple averaging significantly reduces variance and bias by averaging out individual bias in many cases; and (iii) simple averaging should be considered when the uncertainty of weight estimation is taken into account. Additionally, Timmermann (2006) pointed out that the outstanding average performance of simple averaging depends strongly on model instability and the ratio of forecast error variances associated with different forecasting models.

More attention has been given to other strategies, including using the median and mode, as well as trimmed and winsorized means (e.g., Chan et al., 1999; Genre et al., 2013; Grushka-Cockayne, Jose & Lichtendahl, 2017; Jose et al., 2014; Stock & Watson, 2004), due to their

robustness in the sense of being less sensitive to extreme forecasts than a simple average (Lichtendahl & Winkler, 2020). For example, the early work of Galton (1907b) observed that the "middlemost" of 787 estimates of an ox's weight is within nine pounds of the ox's actual weight, and thus advocated for the median forecast as the "vox populi" (Galton, 1907a). However, there is little consensus in the literature on whether the mean or the median of individual forecasts performs better in 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), Galton (1907b) 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 appeal when there is high variability among the individual forecasts due to their simplicity and robust performance. Kourentzes, Barrow and Crone (2014) empirically compared 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 the relevant literature.

Compared to various complicated combination approaches and machine learning algorithms, simple combinations seem outdated and uncompetitive in the big data era. However, the results from the recent M4 competition (Makridakis et al., 2020a) showed that simple combinations continue to achieve relatively good forecasting performance and are still competitive. Specifically, a simple equal-weight combination ranked third for yearly time series (Shaub, 2019) and a median combination of four simple forecasting models achieved sixth place for point forecasting (Petropoulos & 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 challenging benchmark, with few combination schemes outperforming it. Moreover, simple combinations have a lower computational burden and can be implemented more efficiently than alternatives. Therefore, simple combination rules have consistently been the choice of many researchers and practitioners, providing a challenging benchmark to measure the effectiveness of the newly proposed weighted forecast combination algorithms (e.g., Kang, Hyndman & Li, 2020; Makridakis & Hibon, 2000; Makridakis et al., 2020a; Montero-Manso et al., 2020; Stock & Watson, 2004; Wang, Kang, Petropoulos & Li, 2022).

Despite the ease of implementing simple combination schemes, their success still depends largely on the choice of the forecasts to be combined. Intuitively, we prefer that the component forecasts fall on opposite sides of the truth (the realization) so that the forecast errors tend to cancel each other out (Bates & Granger, 1969; Larrick & Soll, 2006). However, this rarely occurs in practice, as the component forecasts are usually trained based on overlapping information sets and use similar forecasting methods. If all component forecasts are established similarly based on the same, or highly overlapping sets of information, forecast combinations are unlikely to improve forecast accuracy. Mannes et al. (2014) and Lichtendahl and Winkler (2020) emphasized two critical issues concerning the performance of simple combination rules: one for the level of accuracy (or expertise) of the forecasts in the pool and another for diversity among individual forecasts. Involving forecasts with low accuracy in the pool can decrease the combination performance. Additionally, a high degree of diversity among component models facilitates the achievement of the best possible forecast accuracy from simple combinations (Thomson et al., 2019). In conclusion, simple, easy-to-use combination rules can provide excellent and robust forecasting performance, especially when properly considering the accuracy and diversity of the individual forecasts to be combined.

2.2. Linear combinations

Despite the simplicity and performance of simple combination rules, it makes sense to assign greater weight to the most accurate forecast methods. But how to choose those weights? The problem of point forecast combinations can be defined as seeking a one-dimensional aggregator that integrates an N-dimensional vector of h-step-ahead forecasts involving the information up to time T, $\hat{\boldsymbol{y}}_{T+h|T} = \left(\hat{y}_{T+h|T,1}, \hat{y}_{T+h|T,2}, \ldots, \hat{y}_{T+h|T,N}\right)'$, into a single combined h-step-ahead forecast $\tilde{y}_{T+h|T} = C\left(\hat{\boldsymbol{y}}_{T+h|T}; \boldsymbol{w}_{T+h|T}\right)$, where N is the number of forecasts to be combined and $\boldsymbol{w}_{T+h|T}$ is an N-dimensional vector of combining weights. The class of combination methods represented by the mapping, C, comprises linear and nonlinear combinations, as well as series-specific and cross-learning combinations. Additionally, the combination weights can be static or time-varying along the forecasting horizon. Below we discuss various approaches for determining combination weights associated with individual forecasts.

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

$$\tilde{y}_{T+h|T} = \boldsymbol{w}'_{T+h|T}\hat{\boldsymbol{y}}_{T+h|T},$$

where $\boldsymbol{w}_{T+h|T} = \left(w_{T+h|T,1}, \ldots, w_{T+h|T,N}\right)'$ is an N-dimensional 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 the so-called "optimal" weights by minimizing the variance of the combined forecast error and discussed only combinations of pairs of forecasts. Newbold and Granger (1974) then extended the method to combinations of more than two forecasts. Specifically, if the individual forecasts are unbiased and their error variances are consistent over time, then 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

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

where $\boldsymbol{\Sigma}_{T+h|T}$ is the $N \times N$ covariance matrix of the h-step forecast errors and $\mathbf{1}$ is an N-dimensional unit vector. This is implemented, e.g., in the R package **ForecastComb** (Weiss et al., 2018). In practice, the elements of the covariance matrix $\boldsymbol{\Sigma}_{T+h|T}$ are usually unknown and need to be estimated.

It follows that if $\boldsymbol{w}_{T+h|T}$ is determined by Eq. (1), 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 explored by Timmermann (2006) to illustrate the diversification gains offered by forecast combinations by simply considering combinations of pairs of forecasts. Under mean squared error (MSE) loss, Timmermann (2006) characterized the general solution of the optimal linear combination weights by assuming a joint Gaussian distribution of the outcome y_{T+h} and available forecasts $\hat{\boldsymbol{y}}_{T+h|T}$.

The loss assumed in Bates and Granger (1969) and Newbold and Granger (1974) is quadratic and symmetric. Elliott and Timmermann (2004) examined forecast combinations under more general loss functions accounting for asymmetries and skewed forecast error distributions. They demonstrated that the optimal combination weights strongly depend on the degree of asymmetry in the loss function and skewness in the underlying forecast error distributions. 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. In addition, the properties of optimal forecasts were generalized to consider asymmetric loss and nonlinear data generating processes.

Regression-based weights

The seminal work by Granger and Ramanathan (1984) provided an important impetus for approximating the "optimal" weights under a linear regression framework. They recommended that the combination weights be estimated by ordinary least squares (OLS) in regression models with the vector of past observations as the response variable and the matrix of past individual forecasts as the predictor variables. Three alternative approaches imposing various possible restrictions were considered

$$y_{T+h} = \boldsymbol{w}'_{T+h|T}\hat{\boldsymbol{y}}_{T+h|T} + \varepsilon_{T+h}, \quad s.t. \quad \boldsymbol{w}'\mathbf{1} = 1, \tag{2}$$

$$y_{T+h} = \boldsymbol{w}'_{T+h|T}\hat{\boldsymbol{y}}_{T+h|T} + \varepsilon_{T+h}, \tag{3}$$

$$y_{T+h} = w_{T+h|T,0} + \boldsymbol{w}'_{T+h|T}\hat{\boldsymbol{y}}_{T+h|T} + \varepsilon_{T+h}. \tag{4}$$

The R package **ForecastComb** (Weiss et al., 2018) provides the corresponding implementations. The constrained OLS estimation of the regression in Eq. (2), 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). Granger and Ramanathan (1984) further suggested that the unrestricted OLS regression in Eq. (4), which allows for a constant term and does not require the weights to sum to one, is superior to the popular "optimal" method regardless of whether the constituent forecasts are biased. However, De Menezes et al. (2000) argue that when using the unrestricted regression, one needs to consider the stationarity of the series being forecast, the possible presence of serial correlation in forecast errors (see also Coulson & Robins, 1993; Diebold, 1988), and the issue of multicollinearity.

Generalizations of the combination regressions have been considered in a large body of literature. Diebold (1988) exploited serial correlated errors in the least squares framework by characterizing the combined forecast errors as 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 are constrained to sum to unity, and the restricted OLS regression where the weights are constrained to be nonnegative. The results revealed that constraining weights to be nonnegative is at least as robust and accurate as the simple average and yields superior results compared to other combinations based on a regression framework. Conflitti et al. (2015) addressed the problem of determining the combination weights by imposing both restrictions (that the weights should be nonnegative and sum to one), which turns out to be a special case of a LASSO regression. Coulson and Robins (1993) found that allowing a lagged dependent variable in forecast combination regressions can improve performance. Instead of using the quadratic loss function, Nowotarski et al. (2014) applied the absolute loss function in the unrestricted regression, also implemented in the **ForecastComb** package for R, to yield the least absolute deviation regression, which is more robust to outliers than OLS combinations.

Forecast combinations using changing weights have also been developed to solve structural changes in constituent forecasts. For instance, Diebold and Pauly (1987) explored rolling weighted least squares and time-varying parameter techniques in the basic regression framework, including both deterministic and stochastic time-varying parameters. Specifically, the combination weights are either described as deterministic nonlinear (polynomial) functions of time or allowed to involve random variation. They showed, via numerical examples based on various types of structural change in the constituent forecasts, that time-varying weights substantially help improve forecasting ability in instabilities. Deutsch et al. (1994) allowed the combination weights to evolve immediately or smoothly using switching regression models and smooth transition regression models. Terui and van Dijk (2002) generalized the regression method by incorporating time-varying coefficients assumed to follow a random walk process. The generalized model can be interpreted as a state space model and then estimated using Kalman filter updating. Following the spirit of Terui and van Dijk (2002), Raftery et al. (2010) achieved an accelerated inference process by using forgetting factors in the recursive Kalman filter updating.

Researchers have also worked on including many forecasts in a regression framework to take advantage of many models. However, Chan et al. (1999) examined a wide range of combination methods and showed that OLS combinations have very poor performance when N (the number of forecasts to be combined) is very large. Factor methods are a common way of condensing information when modeling and forecasting. They have also been used explicitly in forecast combination settings and are especially attractive when the number of forecasts to be combined is very large ($N > T$); see, e.g., Chan et al. (1999) for a dynamic factor model framework for forecast combinations. The common factors in approximate dynamic factor models can be estimated by principal components (Stock & Watson, 1999). Principal components regression (PCR) is typically motivated as an ad hoc tool for the solution of multicollinearity. Chan et al. (1999) and Stock and Watson (2004) explicitly applied PCR to forecast combinations, resulting in a two-step procedure. The first step extracts the principal components, while the second produces the final forecasts utilizing OLS regression. The superiority of PCR over OLS combinations was also supported by Rapach and Strauss (2008) and Poncela et al. (2011). In turn, these methods relate to the question of whether one should forecast with variables (competing point forecasts in our paper's context), factors (extracted from the N competing forecasts), or both; see, e.g., Castle et al. (2013) for a detailed discussion.

In large N cases, given the estimation problems that arise when $N > T$, researchers also frequently relate forecast combinations to shrinkage-type approaches (whether frequentist or Bayesian) that facilitate estimation of the forecast combination regression even when $N > T$, e.g., see Stock and Watson (2004). Diebold and Shin (2019) considered methods for selection and shrinkage in regression-based forecast combinations to address the estimation problem. They shed light on how machine learning can optimally combine a large set of forecasts by introducing a LASSO-based procedure consisting of two steps. The first step involves setting some combination weights to zero using LASSO, and the second step shrinks the combination weights of the survivors toward equal weights. Additionally, Aiolfi and Timmermann (2006) argued in favor of clustering the individual forecasts using the k-means clustering algorithm based on their historical performance. For each cluster, a pooled (average) forecast is computed, which precedes the estimation of combination weights for the constructed clusters.

Performance-based weights

Estimation errors in the "optimal" and regression-based weights tend to be particularly large due to difficulties in adequately estimating the covariance matrix $\boldsymbol{\Sigma}_{T+h|T}$, especially in situations with many forecasts to combine. Instead, Bates and Granger (1969) suggested weighing the constituent forecasts in inverse proportion to their historical performance, ignoring mutual dependence. In follow-up studies, Newbold and Granger (1974) and Winkler and Makridakis (1983) generalized this idea by considering more time series, more individual forecasts, and multiple forecast horizons. Their extensive results demonstrated that combinations ignoring correlations are more successful than those attempting to take account of correlations, and consequently reconfirmed Bates and Granger's (1969) argument that correlations can be poorly estimated in practice and should be ignored when calculating combination weights.

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

$$w^{\text{bg1}}_{T+h|T,i} = \frac{\left(\sum_{t=T-\nu+1}^{T} e^2_{t|t-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{t=T-\nu+1}^{T} e^2_{t|t-h,j}\right)^{-1}}; \tag{5}$$

$$\boldsymbol{w}^{\text{bg2}}_{T+h|T} = \frac{\hat{\boldsymbol{\Sigma}}^{-1}_{T+h|T}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}_{T+h|T}\mathbf{1}},$$

$$\text{where} \quad (\hat{\boldsymbol{\Sigma}}_{T+h|T})_{i,j} = \nu^{-1}\sum_{t=T-\nu+1}^{T} e_{t|t-h,i}e_{t|t-h,j}; \tag{6}$$

$$w^{\text{bg3}}_{T+h|T,i} = \alpha\hat{w}_{T+h-1|T-1,i} + (1-\alpha) \times \frac{\left(\sum_{t=T-\nu+1}^{T} e^2_{t|t-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{t=T-\nu+1}^{T} e^2_{t|t-h,j}\right)^{-1}}, \quad \text{where} \quad 0 < \alpha < 1; \tag{7}$$

$$w^{\text{bg4}}_{T+h|T,i} = \frac{\left(\sum_{t=1}^{T} \gamma^t e^2_{t|t-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{t=1}^{T} \gamma^t e^2_{t|t-h,j}\right)^{-1}}, \quad \text{where} \quad \gamma \geq 1; \tag{8}$$

$$\boldsymbol{w}^{\text{bg5}}_{T+h|T} = \frac{\hat{\boldsymbol{\Sigma}}^{-1}_{T+h|T}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}_{T+h|T}\mathbf{1}},$$

$$\text{where} \quad (\hat{\boldsymbol{\Sigma}}_{T+h|T})_{i,j} = \frac{\sum_{t=1}^{T} \gamma^t e_{t|t-h,i}e_{t|t-h,j}}{\sum_{t=1}^{T} \gamma^t} \quad \text{and} \quad \gamma \geq 1. \tag{9}$$

These weighting schemes differ in the factors and 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 via the usual sample estimator (which 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 individual forecasts over rolling windows of the most recent ν observations. In contrast, others base the weights on exponential discounting with higher values of γ giving larger weights to more recent observations. Consequently, these weighting schemes are well adapted to allow a non-stationary relationship between the individual forecasting procedures over time (Newbold & Granger, 1974). However, they tend to increase the variance of the parameter estimates and work quite poorly if the data generating process is truly covariance stationary (Timmermann, 2006).

A broader set of combination weights based on the relative performance of individual forecasting techniques

has been developed and examined in a series of studies. For example, Stock and Watson (1998) generalized the rolling window scheme in Eq. (5) in the sense that the weights on the individual forecasts are 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 forecasts 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), have also been considered to lead to performance-based combination weights (e.g., Nowotarski et al., 2014; Pawlikowski & Chorowska, 2020). A weighting scheme with the weights depending inversely on the exponentially discounted errors was proposed by Stock and Watson (2004) as an upgraded version of the scheme in Eq. (8). It was used in several subsequent studies (e.g., Clark & 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., Clark & McCracken, 2010; Genre et al., 2013; Stock & Watson, 1998). It is natural to adopt rolling windows in estimating the weights to deal with structural changes, but the window length should not be too short without the estimates of the weights becoming too noisy (Baumeister & Kilian, 2015).

Compared to directly constructing the weights using historical forecast errors, a new form of combinations that is more robust and less sensitive to outliers was introduced based on the "ranking" of individual forecasts. Again this kind of combination ignores correlations among forecast errors. The class's simplest and most commonly used method uses 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. The R package **ForecastComb** (Weiss et al., 2018) provides tools for rank-based combinations. Another weighting scheme which attaches a weight proportional to $\exp(\beta(N + 1 - i))$ to the ith ordered constituent forecast was adopted in Yao and Islam (2008) and Donate et al. (2013) to combine forecasts obtained from artificial neural networks (ANNs), where β is a scaling factor. However, as mentioned by Andrawis et al. (2011), this class of combination methods limits the weights to only a discrete set of possible values.

Criteria-based weights

Information criteria, such as Akaike's information criterion (AIC, Akaike, 1974), the corrected Akaike information criterion (AICc, Sugiura, 1978), and the Bayesian information criterion (BIC, Schwarz, 1978), are often used for model selection in forecasting. However, choosing a single model out of the candidate model pool may be misleading because of the information loss from the alternative models. An alternative approach proposed by Burnham and Anderson (2002) is to combine multiple models based on information criteria to mitigate the risk of selecting a single model. It is also worth mentioning that the R packages **MuMIn** (Bartoń, 2022) and **mmSAR** (Guilhaumon, 2019) have been developed to perform model selection and multimodel averaging based on the use of information-theoretic approaches introduced by Burnham and Anderson (2002).

One such common approach is using Akaike weights. Specifically, because AIC estimates the Kullback–Leibler distance (Kullback & Leibler, 1951) between a model and the true data generating process, differences in the AIC can be used to weight different models, providing a measure of the evidence for supporting a given model relative to other constituent models. Given N individual models, the Akaike weight of model i can be derived as:

$$w^{\text{aic}}_{T+h|T,i} = \frac{\exp(-0.5\Delta\text{AIC}_i)}{\sum_{k=1}^{N}\exp(-0.5\Delta\text{AIC}_k)},$$

where $$\Delta\text{AIC}_i = \text{AIC}_i - \min_{k\in\{1,2,\ldots,N\}} \text{AIC}(k).$$

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

The outstanding performance of weighted combinations based on information criteria has been supported in several studies. For instance, Kolassa (2011) used weights derived from AIC, AICc, and BIC to combine exponential smoothing forecasts and obtained superior accuracy over selecting a model using the same information criteria. A similar strategy was adopted by Petropoulos et al. (2018a) to separately explore the benefits of bootstrap aggregation (bagging) for time series forecasting. Additionally, an empirical study by Petropoulos, Kourentzes et al. (2018) showed that a weighted combination based on AIC improves the performance of the statistical benchmark they used.

Bayesian weights

Some effort has been directed towards Bayesian approaches to updating forecast combination weights in the face of new information gleaned 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 is a major challenge in practice, regardless of whether correlations among forecast errors are ignored or not. With this in mind, Bunn (1975) suggested the idea of Bayesian combinations based on the probability of each forecasting model performing the best on any given occasion. By considering the beta and Dirichlet distributions as the conjugate priors for the binomial and multinomial processes, the suggested non-parametric method performs well when there is relatively little past data by attaching prior subjective probabilities to individual forecasts (Bunn, 1985; De Menezes et al., 2000). Öller (1978) presented another approach to using 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 strand 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 & Harvey, 2004). Assuming that the vector of forecast errors is 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 T observations are used to estimate $\boldsymbol{\Sigma}$, the combination weights derived from the posterior distribution for $\boldsymbol{\Sigma}$ are

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

where $\boldsymbol{\Sigma}^* = \left(\nu_0\boldsymbol{\Sigma}_0^{-1} + T\hat{\boldsymbol{\Sigma}}^{-1}\right)/(\nu_0 + T)$ is the precision matrix and $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix. Compared to estimating $\boldsymbol{\Sigma}$ using the standard sample covariance estimator or treating it as a diagonal matrix, the proposed approach provides a more stable estimation and allows for correlations between methods. 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

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

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are T-dimensional vectors of historical data and residuals, respectively, and $\hat{\boldsymbol{Y}}$ is the $T \times N$ matrix of one-step constituent forecasts. This approach results in estimated combination weights that 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 the subjective and data-based estimation of the combination weights. Because Bayesian approaches have been mostly employed to construct combinations of probability forecasts, we will elaborate on other newly developed methods of determining combination weights from a foundational Bayesian perspective in Section 3.7.

2.3. Nonlinear combinations

Linear combination approaches implicitly assume a linear dependence between constituent forecasts and the variable of interest (Donaldson & Kamstra, 1996; Freitas & Rodrigues, 2006), and may not result in the best forecast (Shi et al., 1999). This is especially true if the individual forecasts come from nonlinear models or if the relationship between combination members and the best forecast is characterized by nonlinear systems (Babikir & Mwambi, 2016). In such cases, it is natural to relax the linearity assumption and consider nonlinear combination schemes of higher complexity; these have received minimal research attention.

As Timmermann (2006) identified, two types of nonlinearities can be incorporated in forecast combinations. One involves nonlinear functions of the individual forecasts but with the unknown parameters of the combination weights given in the linear form. The other allows a more general combination with nonlinearities directly considered in the combination parameters. Neural networks are often employed to estimate nonlinear mapping because they offer the potential to learn the underlying nonlinear relationship between the future outcome and individual forecasts. The design of a neural network model is time-consuming and sometimes leads to overfitting and poor forecasting performance as more parameters need to be estimated.

Donaldson and Kamstra (1996) used ANNs to obtain the combined forecasts $\tilde{y}_{T+h|T}$ by the following form

$$\tilde{y}_{T+h|T} = \beta_0 + \sum_{j=1}^{k}\beta_j\hat{y}_{T+h|T,j} + \sum_{i=1}^{p}\delta_i g(\boldsymbol{z}_{T+h|T}\boldsymbol{\gamma}_i), \tag{10}$$

$$g(\boldsymbol{z}_{T+h|T}\boldsymbol{\gamma}_i) = \left(1 + \exp\left\{-\left(\gamma_{0,i} + \sum_{j=1}^{N}\gamma_{1,j}z_{T+h|T,j}\right)\right\}\right)^{-1}, \tag{11}$$

where $z_{T+h|T,j} = (\hat{y}_{T+h|T,j} - \bar{y})/\hat{\sigma}$, $\bar{y}$ and $\hat{\sigma}$ denote the in-sample mean and in-sample standard deviation respectively, $k \in \{0, N\}$, and $p \in \{0, 1, 2, 3\}$. This approach permits special cases of both purely linear combinations ($k = N$ and $p = 0$) and nonlinear combinations ($k = 0$ and $p \neq 0$). Building on this, Harrald and Kamstra (1997) proposed to evolve ANNs and demonstrated their utility, but only using a single time series. 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 nonlinear mapping, resulting in variants of Eq. (10). The empirical results of nonlinear combinations from these studies generally dominate those from traditional linear combination strategies, such as simple average, OLS weights, and performance-based weights. However, the empirical evidence is based on fewer than ten time series, possibly hand-picked to lead to this result. Additionally, these nonlinear combination methods suffer from other drawbacks, including the neglect of correlations among forecast errors, the instability of parameter estimation, and the multicollinearity caused by the overlap in the information sets used to produce individual forecasts. Thus, the performance of nonlinear combinations relative to linear combinations needs further investigation.

Some researchers have sought to construct nonlinear combinations by including an additional nonlinear term to cope with the case where individual forecast errors are correlated. The combination mechanism can be generalized to the following form

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

where v_{ij} is some nonlinear combination of forecasts i and j. This way, the general framework for linear combinations is extended to deal with nonlinearities.

For example, Freitas and Rodrigues (2006) defined v_{ij} as the product of individual forecasts from different models, $v_{ij} = \hat{y}_{T+h|T,i}\hat{y}_{T+h|T,j}$. In contrast, Adhikari and Agrawal (2012) took into account the linear correlations among the forecast pairs by including the term, $v_{ij} = (\hat{y}_{T+h|T,i} - \bar{y}_i)(\hat{y}_{T+h|T,j} - \bar{y}_j)/(\sigma_i\sigma_j)^2$, where $\bar{y}_i$ and σ_i are the mean and standard deviation of the ith model, respectively. Moreover, Adhikari (2015) defined the nonlinear term using $v_{ij} = \left(\hat{z}_i - m_{ij}\hat{z}_j\right)\left(\hat{z}_j - m_{ji}\hat{z}_i\right)$, where $\hat{z}_i$ denotes the standardized ith individual forecast using the mean $\bar{y}_i$ and standard deviation σ_i, and the term m_{ij} denotes the degree of mutual dependency between the ith and jth individual forecasts.

Nonlinearly combining forecasts requires further research. In particular, the forecasting performance of the various proposed nonlinear combination schemes should be investigated appropriately with an extensive, diverse collection of time series datasets and appropriate statistical inference. There is also a need to develop nonlinear combination approaches that take account of correlations across forecast errors and the multicollinearity of forecasts.

2.4. Combining by learning

Stacked generalization (stacking, Wolpert, 1992) provides a strategy to combine the available forecasting models adaptively. Stacking is frequently employed on a wide variety of classification tasks (Zhou, 2012); in the time series forecast context, it uses the concept of meta-learning to boost forecasting accuracy beyond that achieved by any of the individual models. Stacking is a general framework that comprises at least two levels. The first level involves training the individual forecasting models using the original data. In contrast, the second and subsequent levels utilize an additional "meta-model", using the prior level forecasts as inputs to form a set of forecasts. Thus, the stacking approach to forecast combinations weights individual forecasts adaptively using meta-learning processes.

There are many ways to implement the stacking strategy. Its primary implementation is to combine individual models in a series-by-series fashion. Individual forecasting models in the method pool are trained using only data from the single series they are going to forecast. In contrast, their forecast outputs are subsequently fed to a meta-model tailored to calculate the combined forecasts for the target series. This means that n meta-models are required for n separate time series data. Unsurprisingly, regression-based weight combinations discussed in Section 2.2 (e.g., Granger & Ramanathan, 1984; Gunter, 1992) fall into this category and can be viewed as the most simple, common learning algorithm used in stacking. Instead of applying multiple linear regressions, Moon et al. (2020) suggested a 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, ANN, wavelet neural network (WNN), and support vector regression (SVR) can be conducted in a series-by-series fashion to achieve the same goal (e.g., Conflitti et al., 2015; Donaldson & Kamstra, 1996; Ribeiro et al., 2019; Ribeiro & dos Santos Coelho, 2020). One could use an 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" (Hyndman & Athanasopoulos, 2021), is also recommended in the training procedures for individual models and meta-models to help with the parameter estimation. Nevertheless, stacking approaches implemented in a series-by-series fashion still suffer from some limitations, such as requiring a long computation time and time series and inefficiently using the training data.

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 time series rather than employing only a single series. Thus, various patterns can be captured along different series. The M4 competition organized by Spyros Makridakis et al. (2020a), comprising 100,000 time series, recognized the benefits of cross-learning in the sense that the top three performing methods of the competition utilize 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 techniques for extracting information from large, diverse time series datasets are adopted (Kang, Spiliotis et al., 2020; Semenoglou et al., 2020). Zhao and Feng (2020) trained a neural network model across the M4 competition dataset to learn how to combine individual 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 for the neural network model.

An increasing stream of studies has shown that time series features characterizing each series in a dataset provide valuable information for forecast combinations in a cross-learning fashion, leading to an extension of stacking. Numerous software packages have been developed for time series feature extraction, including the R packages **feasts** (O'Hara-Wild et al., 2021) and **tsfeatures** (Hyndman et al., 2019), the Python packages **Kats** (team, 2021), **tsfresh** (Christ et al., 2018) and **TSFEL** (Barandas et al., 2020), the Matlab package **hctsa** (Fulcher & Jones, 2017), and the C-coded package **catch22** (Lubba et al., 2019). These sets of time series features were empirically evaluated by Henderson and Fulcher (2021).

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, Makridakis 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 combinations. More recently, Montero-Manso et al. (2020) introduced a Feature-based FORecast Model Averaging (FFORMA) approach, available in the R package **M4metalearning** (Montero-Manso, 2019). This approach employs 42 statistical features (implemented using the R package **tsfeatures**) to estimate

the optimal weights for combining nine different traditional models trained per series based on an XGBoost model. The FFORMA method reported the second-best forecasting accuracy in the M4 competition. Additionally, Ma and Fildes (2021) highlighted the potential of convolutional neural networks as a meta-model to link the learned features with a set of combination weights. Li et al. (2020) extracted time series features automatically with the idea of time series imaging, and then these features were used for forecast combinations. Gastinger et al. (2021) demonstrated the value of a collection of combination methods on a large and diverse amount of time series from the M3 (Makridakis & Hibon, 2000), M4, M5 (Makridakis et al., 2022) datasets and FRED datasets[1]. Because it is unclear which combination strategy should be selected, they introduced a meta-learning step to choose a promising subset of combination methods for a newly given dataset based on extracted features.

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 combinations (Atiya, 2020; Batchelor & Dua, 1995; Lichtendahl & Winkler, 2020; Thomson et al., 2019). An increase in diversity among forecasting models has the potential to improve the accuracy of their combination. In this respect, features measuring 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 of the performance of individual models with the extracted features and to improve forecasting performance. Kang et al. (2021) 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 limitations. The first limitation is the requirement for a large, diverse time series dataset to enable meaningful training outcomes. This issue can be addressed by simulating series based on some assumed data generating processes (Talagala et al., 2018) (implemented using the R package **forecast**, Hyndman et al., 2021), or by generating time series with diverse and controllable characteristics (Kang, Hyndman & Li, 2020) (implemented in the R package **gratis**, Kang, Li et al., 2020). Moreover, given the considerable literature on feature identification and feature engineering (e.g., Kang et al., 2017; Lemke & Gabrys, 2010; Li et al., 2020; Montero-Manso et al., 2020; Wang et al., 2009), the feature-based forecast combination methods naturally raise some issues that are yet to receive much research attention. These include how to design an appropriate feature pool to achieve the best out of such methods and which is the best loss function for the meta-model.

It is also worth mentioning that many neural network models rely on a model combination strategy, namely "ensembling" (see, e.g., Caruana et al., 2004, a popular work in the machine learning context), that is applied internally to improve the overall forecasting performance. Due to the weak learning process in deep learning models, the overall forecasting results heavily depend on the combination of each forecasting result. They diversify the individual forecast via (1) varying the training data, (2) varying the model pool, and (3) varying the evaluation metric. For example, the N-BEATS model (Oreshkin et al., 2019) utilized different strategies to diversify the forecasting results. For each forecasting horizon, individual models are trained with six window lengths. It also used three metrics sMAPE, MASE, and MAPE, to validate each model. In the end, various models are used to make the median ensemble for results on the test set. One may refer to Ganaie et al. (2022) for a general view of deep learning ensembles.

2.5. Which forecasts should be combined?

Including forecast methods with poor accuracy degrades the performance of the forecast combination. One prefers to exclude component forecasts that perform poorly and to combine only the top performers. In judgmental forecasting, Mannes et al. (2014) highlighted the importance of the crowd's mean level of accuracy (expertise). They argued that the mean level of knowledge sets a floor for the performance of combining. The gains in accuracy from selecting top-performing forecasts 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 time series, which indicates the accuracy risk, exerts a significant influence on the performance of the combined forecasts. They suggested balancing the trade-offs between the average accuracy and the variance of accuracy when choosing component models from a set of available models.

Another critical issue is diversity. Diversity among the individual forecasts is often recognized as one of the elements required for accurate forecast combination (Batchelor & 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 that an increase in diversity among the individual forecasts is responsible for the error reduction achieved in combined forecasts. Diversity among individual forecasts is frequently measured in terms of correlations among their forecast errors, with lower correlations indicating a higher degree of diversity. The distance of top-performing clusters introduced by Lemke and Gabrys (2010), where a 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.

[1] The FRED (Federal Reserve Economic Data) dataset is openly available at https://fred.stlouisfed.org.

In an analysis of a winner-take-all forecasting competition, Lichtendahl Jr et al. (2013) found that the optimal strategy for reporting forecasts is to exaggerate the forecasters' private information and down-weight any common information. This exaggeration results in gains in the accuracy of the simple average by amplifying the diversity of the individual forecasts. The gains were confirmed by Grushka-Cockayne, Jose and Lichtendahl (2017), who looked more closely at the impact of private-signal exaggeration on forecast combinations, which translates into averaging forecasts that are overfitted and overconfident.

Ideally, we would choose independent forecasts to amplify the diversity of the component forecasts when forming a combination. However, the available individual forecasts are often produced based on similar training, similar models, and overlapping information sets, leading to highly positively correlated forecast errors. Including forecasts that have highly correlated forecast errors in a combination creates redundancy and may result in unstable weights, especially in the class of regression-based combinations (see Section 2.2). In this respect, using different types of forecasting models (e.g., statistical, machine learning, and judgmental) or different sources of information (e.g., exogenous variables), can help improve diversity (Atiya, 2020). The results of the M4 competition reconfirmed the benefits of combinations of statistical and machine learning models (Makridakis et al., 2020a).

It is often suggested that a subset of individual forecasts be combined, rather than the full set of forecasts, as there are decreasing returns to adding additional forecasts (Armstrong, 2001; Diebold & Shin, 2019; Geweke & Amisano, 2011; Hibon & Evgeniou, 2005; Lichtendahl & Winkler, 2020; Zhou et al., 2002). Simply put, *many could be better than all*. In this regard, given a method pool with many forecasting models available, one can consider an additional step ahead of combining: subset selection. Instead of using all available forecasts in a combination, the step aims to eliminate some forecasts from the combination and select only a subset of the available forecasts.

The most common technique of subset selection is to include only the most accurate methods in the combination, discarding the worst-performing individual forecasts (e.g., Granger & Jeon, 2004). Mannes et al. (2014) investigated the gains in accuracy from this *select-crowd* strategy. Kourentzes et al. (2019) proposed a heuristic where we exclude component forecasts that show a sharp drop in performance by using the outlier detection methods in boxplots. Their empirical results over four diverse datasets showed that this subset selection approach outperforms selecting a single forecast or combining all available forecasts. Nonetheless, the approach may suffer from a lack of diversity when formulating appropriate pools.

Early studies considering diversity used forecast encompassing tests for combining forecasts (e.g., Costantini & Pappalardo, 2010; Kışınbay, 2010). The forecast encompassing literature ties in very closely with forecast combinations. Several forecast encompassing tests have been developed to test whether one forecast (or a set of forecasts) encompasses all information contained in another forecast (or another set of forecasts); see, e.g., Chong and Hendry (1986) and Harvey et al. (1998). A classical argument suggests that when fixed weights are used (as in an average), only non-encompassed individual models are worth combining (Diebold, 1989). However, Hendry and Clements (2004) provided a counter-example in processes subject to location shifts where previously encompassed models may later dominate, while the earlier dominant model may fail badly.

The diversity of an available forecast pool has occasionally been explicitly considered for subset selection. Cang and Yu (2014) proposed an optimal subset selection algorithm for forecast combinations based on mutual information, which takes account of diversity among different forecasts. More recently, Lichtendahl and Winkler (2020) developed a subset selection approach comprising two screens: one screen for removing individual models that perform worse than the Naïve2 benchmark and another for excluding pairs of models with highly correlated forecast errors. This way, accuracy and diversity issues are addressed when forming a combination.

Subset selection techniques take advantage of allowing many forecasts to be considered when combining, reducing weight estimation errors and improving computational efficiency. However, subset selection has received scant attention in the context of forecast combinations, and it is mainly focused on trimming based on the principles of expertise. Therefore, automatic selection techniques considering expertise and diversity merit further attention and development.

One approach is to note that subset selection is equivalent to assigning zero weights to some individual forecasts, which could be determined either statistically or judgmentally. Diebold and Shin (2019) focused on weights that solve a penalized estimation problem. Specifically, they proposed a two-step LASSO-based procedure that selects a subset of forecasts to combine in the first step and shrinks the weights of the selected candidates toward equality. An alternative idea can be using a pre-set threshold to select individual models with weights greater than the threshold to join the subsequent combination; see, e.g., Wang, Kang, Petropoulos and Li (2022), Zhou et al. (2002). Of course, there is no guarantee that the zero weight over the training period will also be zero over the forecast horizon. Hence, time-varying subset selection is certainly one solution to this problem and can be achieved by applying a pre-set threshold to forecast combinations with time-varying weights (Li et al., 2022).

2.6. Forecast combination puzzle

Despite the explosion of various 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 vast literature, including the early studies by Stock and Watson (1998, 2003, 2004), the series of Makridakis competitions (Makridakis et al., 1982;

Makridakis & Hibon, 2000; Makridakis et al., 2020a), and also the more recent articles by Blanc and Setzer (2016, 2020), etc. Clemen (1989) surveyed the early combination studies and raised a variety of issues that remain to be addressed, one of which is "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 many time series from diverse sources and found that the winning combination methods differ for the different data sources. At the same time, the simple average strategies show, on average, more gains in improving accuracy than other, more 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.

Most explanations of why simple averaging might dominate complex combinations in practice have centered on the errors that arise when estimating the combination weights. e.g., Timmermann (2006) noted that the success of simple combinations is due to the increased parameter estimation error with weighted combinations — simple combination schemes do not require estimating combination parameters, such as weights based on forecast errors. Smith and Wallis (2009) demonstrated that the simple average is expected to overshadow the weighted average in a situation where the weights are theoretically equivalent. The simulations and an empirical study showed the estimation cost of weighted averages when the optimal weights are close to equality, thus providing an empirical explanation of the puzzle. Later, Claeskens et al. (2016) provided a theoretical basis for these empirical results. Taking the estimation of "optimal" weights (see Section 2.2) into account, Claeskens et al. (2016) considered random weights rather than fixed weights during the optimality derivation. They showed that, in this case, the forecast combination might 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 combinations. The proposed framework verified the estimation error explanation of the "forecast combination puzzle" and, more crucially, provided additional insights into the puzzle. Specifically, the mean squared forecast error (MSFE) can be considered 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. Blanc and Setzer (2020) explained why, in practice, equal weights are often a good choice using the tradeoff between bias (reflecting the error resulting from underfitting training data when choosing equal weights) and variance (quantifying the error resulting from the uncertainty when estimating other weights).

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 where the covariance matrix of the forecast errors has equal variances on the diagonal, and all off-diagonal covariances are 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 are often too small to balance estimation error, providing a supplementary explanation of the puzzle for the explanation of large estimation error.

Rather than focusing on the impact of combination weight estimation, Zischke et al. (2022) explored the implications of sampling variability in forecast combinations. They demonstrated that, asymptotically, the sampling variability in the performance of the combination forecast is driven entirely by the variability arising from the estimation of the constituent models, and combination weight estimation imparts no bias or variability to the performance of forecast combinations, which lies in opposition to the finding of Claeskens et al. (2016). These findings imply that, when the combination weights are theoretically equivalent, there will be little performance difference between a sophisticated forecast combination and an equally weighted combination, providing new insights into the "forecast combination puzzle".

Examining and explaining the "forecast combination puzzle" can give decision-makers the following 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 an insufficiently small sample size may be unable to provide robust weight estimates. Thus, if one has access to 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 a means to cope with structural changes, as suggested in Diebold and Pauly (1987) and Deutsch et al. (1994).
- If one has access to many component forecasts, the PCR and the clustering strategy (for details, see Section 2.2) might be useful to diminish estimation errors and solve the multicollinearity problem by reducing the number of parameters needed to be estimated.

- Involving time series features (see Section 2.4) and diverse individual forecasts (see Section 2.5) in the process of weight estimation can enlarge the gains of forecast combinations, providing a possible way to untangle the "forecast combination puzzle".

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

3. Probabilistic forecast combinations

3.1. Probabilistic forecasts

In recent years, probabilistic forecasts have received increasing attention. For example, the recent Makridakis competitions, the M4 and the M5 Uncertainty (Makridakis et al., 2020b) competitions, encouraged participants to provide probabilistic forecasts of different types as well as point forecasts. Probabilistic forecasts are appealing for enabling optimal decision-making with a better understanding of uncertainties and the resulting risks. A brief survey of extensive applications of probabilistic forecasting was offered by Gneiting and Katzfuss (2014).

Probabilistic forecasts can be reported in various forms, including density forecasts, distribution forecasts, quantiles, and prediction intervals, and how to combine them can vary. For example, although a quantile forecast is the inverse of the corresponding forecast represented by the cumulative distribution function, the combined quantile forecast and the combined probability forecast may not be equivalent. Examples of averaging quantiles and probabilities with equal weights are provided by Lichtendahl et al. (2013).

Interval forecasts form a crucial special case and are often constructed using quantile forecasts where the endpoints are specific quantiles of a forecast distribution. For example, the lower and upper endpoints of a central $(1-\alpha)\times 100\%$ prediction interval can be defined via the quantiles at levels $\alpha/2$ and $1-\alpha/2$.

As with point forecasts, combining multiple probabilistic forecasts allows for diverse information sets and different types of forecasting models, as well as the mitigation of potential misspecifications derived from a single model. Empirical studies suggest that the relative performance of different models often varies over time due to structural instabilities in the unknown data generating process (e.g., Billio et al., 2013). Thus, there has been a growing interest in combining multiple probabilistic forecasts to produce combined forecasts that integrate information from separate sources.

3.2. Scoring rules

Decision makers mainly focus on accuracy when combining point forecasts, while other measures such as calibration and sharpness need to be considered when working with combinations of probabilistic forecasts (Gneiting et al., 2007; Gneiting & Raftery, 2007; Lahiri et al., 2015). **Calibration** concerns the statistical consistency between the probabilistic forecasts and the corresponding realizations, thus serving as a joint property of forecasts and observations. In practice, a probability integral transform (PIT) histogram is commonly employed informally as a diagnostic tool to assess the calibration of probability forecasts regardless of whether they are continuous (Dawid, 1984; Diebold et al., 1998) or discrete (Gneiting & Ranjan, 2013). A uniform histogram indicates a probabilistically calibrated forecast. **Sharpness** refers to the concentration of probabilistic forecasts and thus serves as a property of the forecasts only; the sharper a forecast is, the better it is. Sharpness is easily comprehended when considering prediction intervals: the sharper the forecasts, the narrower the intervals. In the case of probability forecasts, sharpness can be assessed in terms of the width of central prediction intervals. For more thorough definitions and diagnostic tools of calibration and sharpness, we refer to Gneiting and Katzfuss (2014).

According to Gneiting et al. (2007), the intent of probabilistic forecasting is to *maximize the sharpness of the forecast distributions subject to calibration* based on the available information set. In this light, scoring rules that reward calibration and sharpness are appealing in providing summary measures for the quality of probabilistic forecasts, with a higher score indicating a better forecast. For a probabilistic forecast F, a scoring rule is proper if it satisfies the condition that the expected score for an observation drawn from distribution G is maximized when $F = G$. It is strictly proper if the maximum is unique. Gneiting and Raftery (2007) provides an excellent review and discussion on a diverse collection of proper scoring rules for probabilistic forecasts.

The schemes for combining multiple probabilistic forecasts have evolved from a simple distribution mixture to more sophisticated combinations accounting for correlations between distributions. The type of strategy one might choose to use depends largely on the computational burden and the overall performance of the combined forecasts regarding accuracy, calibration, and sharpness.

3.3. Linear pooling

Probability forecasts strive to predict the probability distribution of quantities or events of interest. In line with the notations in previous sections, here we consider N individual forecasts specified as cumulative probability distributions of a random variable Y at time $T+h$, denoted $F_i(y_{T+h}|I_T)$, $i = 1,\ldots,N$, using the information available up to time T, I_T. One popular approach is to directly take a mixture distribution of these N individual probability forecasts with estimated weights, neglecting correlations between these individual components. This approach is commonly referred to as the "linear opinion pool" in the literature on combining experts' subjective probability distributions, dating back at least to Stone (1961). The linear pool of probability forecasts is defined as the finite

mixture

$$\tilde{F}(y_{T+h}|I_T) = \sum_{i=1}^{N} w_{T+h|T,i} F_i(y_{T+h}|I_T), \quad (12)$$

where $w_{T+h|T,i}$ is the weight assigned to the ith probability forecast. These weights are often set to be non-negative and sum to one to guarantee that the pooled forecast preserves properties of both non-negativity and integrating to one. The pooled probability forecast satisfies numerous properties such as the *unanimity* property (if all individual forecasters agree on a probability, then the pooled forecast agrees also); see Clemen and Winkler (1999) for more details.

Linear pooling of probability forecasts allows us to accommodate skewness and kurtosis (fat tails), and also multi-modality, even under normal distributions of individual forecasts; see Wallis (2005) and Hall and Mitchell (2007) for further discussion on this point.

Define μ_i and σ_i^2 as the mean and variance of the ith component forecast distribution and drop the time and horizon subscripts for simplicity. Then the linear combined probability forecast has the mean and variance,

$$\tilde{\mu} = \sum_{i=1}^{N} w_i \mu_i, \quad (13)$$

and

$$\tilde{\sigma}^2 = \sum_{i=1}^{N} w_i \sigma_i^2 + \sum_{i=1}^{N} w_i (\mu_i - \tilde{\mu})^2 . \quad (14)$$

Note that the mean of the combined distribution is equivalent to the linear combination of the individual means. Thus, the associated combination point forecast is consistent with the linear combination point forecast.

However, the variance of the combined distribution is larger than the linear combination of the individual variances when the individual means differ. Consequently, seeking diverse forecasts may harm the probabilistic forecast while helping the point forecast; see Ranjan and Gneiting (2010) for a theoretical illustration and simulation study. Simply put, as the diversity among individual probability forecasts increases, the mixed forecast will lose sharpness and may become under-confident because of the spread driven by the disagreement between the individual probability forecasts (Hora, 2004; Ranjan & Gneiting, 2010; Wallis, 2005).

Even in the ideal case where individual forecasts are well calibrated, the resulting linear pooling combination may be poorly calibrated. Theoretical aspects of this finding and properties of linear pools of probability forecasts have been further studied in Hora (2004), Ranjan and Gneiting (2010), and Lichtendahl et al. (2013).

On the other hand, Hora (2004) demonstrated, both from theoretical and empirical aspects, that linear pooling may provide better-calibrated forecasts than the individual distributions when individual forecasts tend to be overconfident. This finding helps to account for the success of linear pooling in varied applications. Jose et al. (2014) highlighted that if the experts are overconfident but have low diversity, the linear pool may remain overconfident. Lichtendahl et al. (2013) identified three factors that manipulate the calibration of the probability forecast derived from linear pooling: (i) the number of constituent forecasts, (ii) the degree to which the constituent forecasts are overconfident, and (iii) the degree of the constituents' disagreement on the location (e.g., mean) of the distribution.

In principle, probability forecasts can be recalibrated before or after the pooling to correct for miscalibration (Turner et al., 2014). However, it is challenging to appraise the degree of miscalibration (which may vary considerably among different forecasts and over time) and, therefore, to recalibrate accordingly. Some effort has been directed toward developing alternative combination methods to address the calibration issue. For example, Jose et al. (2014) suggested the "trimmed opinion pool", which trims away some individual forecasts from the "linear opinion pool" before mixing the component forecasts. Specifically, exterior trimming that trims away forecasts with low or high means values serves as a way to address under-confidence by decreasing the variance. Conversely, interior trimming that trims away forecasts with moderate means is suggested to mitigate overconfidence by increasing the variance. At a more foundational level, the improvement in forecasting performance offered by trimming was confirmed by Grushka-Cockayne, Jose and Lichtendahl (2017).

Some researchers prefer nonlinear alternatives, including a generalized linear pool, the spread-adjusted linear pool, and the beta-transformed linear pool, to deliver better calibrated combined probability forecasts; these are discussed in Section 3.5. Instead of mixing probability forecasts mentioned above, Lichtendahl et al. (2013) recommended averaging quantile forecasts (see Section 3.9) based on the supportive results both theoretically and empirically.

The key practical issue determining the success (or failure) of linear pooling is how the weights for the individual probability forecasts in the finite mixture should be estimated. As with point forecast combinations, equal weights are worthy of consideration, while determining optimal weights is particularly challenging in the case of having access to probability forecasts with limited historical data.

Linear pooling with equal weights is easy to understand and implement, commonly yielding robust and stable outcomes. For reviews, see, e.g., Wallis (2005) and O'Hagan et al. (2006). A leading example is the US survey of professional forecasters (SPF), which publishes mixed probability forecasts (in the form of histograms) for inflation and GDP growth using equal weights. As the experience of combining point forecasts has taught us, the equally weighted approach often turns out to be hard to beat. An important reason is that it avoids parameter estimation error that usually exists in weighted approaches; see Section 2.6 for more details and illustrations.

Motivated by the "optimal" weights obtained in point forecast combinations by minimizing the MSE loss (see Section 2.2), Hall and Mitchell (2007) proposed obtaining the set of weights by minimizing the Kullback–Leibler information criterion (KLIC) distance between the combined probability forecast density $\tilde{f}(y_{\tau+h}|I_\tau)$ and the true (but

unknown) probability density $f(y_{\tau+h})$, $\tau = 1, \ldots, T$. The KLIC distance is defined as

$$\begin{aligned}\text{KLIC} &= \int f(y_{\tau+h}) \log \left\{ \frac{f(y_{\tau+h})}{\tilde{f}(y_{\tau+h}|I_\tau)} \right\} \mathrm{d}y_{\tau+h} \\ &= E\left[\log f(y_{\tau+h}) - \log \tilde{f}(y_{\tau+h}|I_\tau)\right].\end{aligned}$$

Under the asymptotic assumption that the number of observations T grows to infinity, the problem of minimizing the KLIC distance reduces to the maximization of the average logarithmic score of the combined probability forecast. Therefore, the optimal weight vector $\boldsymbol{w}_{T+h|T}$ is given by

$$\boldsymbol{w}_{T+h|T} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{1}{T-h} \sum_{t=1}^{T-h} \log \tilde{f}(y_{t+h}|I_t), \tag{15}$$

where $\boldsymbol{w}_{T+h|T} = \left(w_{T+h|T,1}, \ldots, w_{T+h|T,N}\right)'$. Using the logarithmic scoring rule eliminates the need to estimate the unknown true probability distribution and simplifies the weight estimation for the component forecasts. This was followed by Pauwels and Vasnev (2016) documenting the properties of the optimal weights in Eq. (15), centering on the asymptotic assumption used by Hall and Mitchell (2007). Their simulations and empirical results indicated that the combination with optimal weights is inferior for small T, while it is valid in minimizing the KLIC distance when T is sufficiently large. Therefore, a sufficient training sample is recommended when solving the optimization problem.

Following in the footsteps of Hall and Mitchell (2007), many extensions and refinements of the combination strategy have been suggested. Conflitti et al. (2015) devised a simple iterative algorithm to compute the optimal weights in Eq. (15). The algorithm scales well with the dimension N and enables the combination of many individual probability forecasts. Geweke and Amisano (2011) provided a Bayesian perspective on an optimal linear pool and a theoretical justification for using optimal weights. Li et al. (2022) conducted time-varying weights based on time-varying features from historical information, where the weights in the forecast combination were estimated via Bayesian logarithmic predictive scores. Jore et al. (2010) put forward an exponential weighting scheme based on the recursive weights constructed using the relative past performance of each probability forecast in terms of the logarithmic score. In contrast to the optimal opinion pool based on the weights in Eq. (15), in this case, the logarithmic score of the combined probability forecast is not necessarily maximized. The logarithmic scoring rule is appealing as it intuitively assigns a higher weight to a component forecast that better fits the realized value. On the other hand, forecast combinations with weights optimized by minimizing the continuously ranked probability score (CRPS, Gneiting & Raftery, 2007), which is a strictly proper scoring rule for distribution forecasts, have been considered in some research, see, e.g., Raftery et al. (2005), Thorey et al. (2017), and Thorey et al. (2018).

Furthermore, some special treatments were given to accommodate probability forecast combinations in applications such as energy forecasting, retail forecasting, and economic forecasting. For instance, Opschoor et al. (2017) extended the idea of optimal combinations but estimated optimal weights by either maximizing the censored likelihood scoring rule (Diks et al., 2011) or minimizing a weighted version of the CRPS, allowing forecasters to limit themselves to a specific region of the target distribution. For example, we are more likely to be interested in avoiding out-of-stocks when working with retail forecasting. The tail of the distribution is also the main feature of interest when measuring downside risk in equity markets. Additionally, Zischke et al. (2022) showed that when forecasting during times of high volatility, forecast combinations produced by optimizing according to the censored likelihood scoring rule always lead to a better out-of-sample performance than "optimal" forecast combinations with weights optimized using the logarithmic score. This supports using a scoring rule that prioritizes accurate forecasts in a specific region. Diebold et al. (2022) instead constructed regularized mixtures of density forecasts using a variety of objectives and regularization penalties. The optimal regularization tends to spread probability mass from the center into both tails of the distribution, correcting for overconfidence and adjusting kurtosis. Besides, Pauwels et al. (2020) proposed an approach to computing the optimal weights by maximizing the average logarithmic score subject to additional higher moments restrictions. Through constrained optimization, the combined probability forecast can preserve specific characteristics of the distribution, such as fat tails or asymmetry. Martin et al. (2021) looked at model misspecification and showed via simulation and empirical results that score-specific optimization of linear pooling weights does not consistently improve forecasting accuracy.

3.4. Bayesian model averaging

Bayesian model averaging (BMA) provides an alternative means of mixing individual probability forecasts with respect to their posterior model probabilities. BMA offers a conceptually elegant and logically coherent solution to the issue of accounting for model uncertainty (see, e.g., Draper, 1995; Garratt et al., 2003; Leamer, 1978; Raftery et al., 1997). Under this approach, the posterior probability forecast is computed by mixing a set of individual probability forecasts distributions, $F_i(y_{T+h}|I_T) = F(y_{T+h}|I_T, M_i)$, from model M_i, and can be given as

$$\tilde{F}(y_{T+h}|I_T) = \sum_{i=1}^{N} P(M_i|I_T) F(y_{T+h}|I_T, M_i), \tag{16}$$

where $P(M_i|I_T)$ is the posterior probability of model M_i. The decision-makers update the prior probability of model M_i being the true model, $P(M_i)$, via Bayes' Theorem to compute the posterior probability

$$P(M_i|I_T) = \frac{P(M_i)P(I_T|M_i)}{\sum_{i=1}^{N} P(M_i)P(I_T|M_i)}, \tag{17}$$

where

$$P(I_T|M_i) = \int_{\boldsymbol{\theta}_i} P\left(\boldsymbol{\theta}_i|M_i\right) P\left(I_T|M_i, \boldsymbol{\theta}_i\right) d\boldsymbol{\theta}_i \tag{18}$$

is the marginal likelihood of model M_i, $P(\boldsymbol{\theta}_i|M_i)$ is the prior on the unknown parameters $\boldsymbol{\theta}_i$ conditional on model M_i, and $P(I_T|M_i, \boldsymbol{\theta}_i)$ is the likelihood function of model M_i. See, e.g., Koop (2003) for textbook illustrations of BMA.

BMA in Eq. (16) can be viewed as a form of linear pooling of individual probability forecasts (12), weighted by their posterior model probabilities given in Eq. (17). Note that the weights characterized by posterior probabilities do not account for correlations among individual probability forecasts. The approach provides a general way to deal with model uncertainty and does not necessarily require using conjugate families of distributions. The BMA procedure is consistent in the sense that the posterior probability in Eq. (17) indicates the probability that model M_i is the best under the KLIC measure distance and shows how well the model fits the observations (Fernández-Villaverde & Rubio-Ramırez, 2004; Raftery et al., 2005; Wright, 2008).

While theoretically attractive, BMA suffers three major challenges when implemented in practice. One is how to correctly specify the model space of interest to avoid model incompleteness. It is often impractical to cover the complete set of models when the number of possible models is large, or their structures are complex. This difficulty can mostly be resolved via selecting a subset of models supported by the data or through stochastic search algorithms over the model space; see Hoeting et al. (1999) and Koop and Potter (2003) for more details on model search strategies. A second well-known challenge relates to eliciting two types of priors (on parameters and models) for many models of interest (Aastveit et al., 2019; Moral-Benito, 2015). Another practical concern lies in the computation of the integrals in Eq. (18). The integrals required for deriving the marginal likelihood may be analytically intractable in many cases, except for the generalized linear regression models with conjugate priors. The Laplace method, as well as the Markov chain Monte Carlo (MCMC) methods, are therefore frequently used to provide an excellent approximation to $P(I_T|M_i)$; see, e.g., Hoeting et al. (1999) and Bassetti et al. (2020) for discussions of these approximations.

One drawback of the BMA approach is the implicit assumption that the true model is included in the model space to be considered (Wright, 2008). Under this assumption, when the sample size tends to infinity, the posterior probabilities converge to zero, except for one, which converges to unity. Thus, BMA reduces to model selection for a large sample size, with the best model (which is the true model if that exists, but is still well-defined if none of the models are true) receiving a weight very close to one; see Geweke and Amisano (2010) for an empirical demonstration. In this regard, the combined forecast derived from BMA may be misspecified when the model space is incomplete (i.e., all models under consideration are incorrect), raising the issue of model incompleteness. Recently, Yao et al. (2018) took the idea of stacking from the literature on point forecast combinations (see Section 2.4) and generalized it to the combinations of forecast distributions in the Bayesian setting, which can essentially be regarded as a minor tweak on BMA. However, as the critique given at the end of Yao et al. (2018) says, averaging distribution functions may be inferior to averaging quantiles (see Section 3.9). This is especially true when the combined problem is more like an information aggregation problem than a BMA problem. In this case, BMA (or minor tweaks) does not seem like the proper framework since we are almost always in a world without a true model.

In contrast, as defined in Eq. (15), optimal weights do not suffer from the issue of model incompleteness because the weights need not converge to zero or unity regardless of whether the component models are correct or not. This allows for a convex combination (rather than a selection) of the individual probability forecast distributions. In a binary-event context, Lichtendahl Jr et al. (2022) introduced a new class of Bayesian combinations in which stacking is used to form the approach by aggregating the probabilities provided by the experts or models. But it should not be confused with BMA and the approach developed by Yao et al. (2018) since it does not have to assume, as BMA does, that one of the models being combined is the true model. Its setting is information aggregation rather than model selection. Lichtendahl Jr et al. (2022) showed that extremizing (i.e. shifting the average probability closer to its nearest extreme, see, e.g., Satopää et al., 2016) is not always appropriate when combining binary-event forecasts.

The other drawback of the BMA approach may be related to the fixed probabilities assigned to component models, as documented in Aastveit et al. (2019). The uncertainty of the weights is ignored in this case, leading to unstable combined forecasts in a forecasting environment characterized by significant instability and structural changes in the forecast performance of the individual models. Thus, it is plausible to let the pooling weights evolve. Raftery et al. (2010) developed a model combination strategy for doing dynamic model averaging (DMA), allowing the forecasting model and the coefficients in each model to evolve over time. By considering multiple models, the goal of DMA is to calculate the probabilities that the process is governed by model M_i for $i = 1, \ldots, N$ at time $T + 1$, given the information available up to time T, and average forecasts across individual models using these probabilities. When the forecasting model and model parameters do not change, DMA reduces to a recursive implementation of standard BMA. The strategy advocated by Raftery et al. (2010) can also be used for dynamic model selection (DMS), where a single model with the highest probability is selected and used to forecast. Note that these calculated probabilities will vary over time; thus, different forecasting models hold at each point in time. Such specifications are of particular interest in economics, see, e.g., Koop and Korobilis (2012) and Del Negro et al. (2016) for notable macroeconomic applications. One contribution of Raftery et al. (2010) is that a forgetting factor is used to develop a computationally efficient recursive algorithm that allows for fast calculation of the required probabilities when model uncertainty and the number of models considered are large.

3.5. Nonlinear pooling

Despite their simplicity and popularity, the classical linear pooling methods have several shortcomings, such as the calibration problem discussed previously. A linear pooling of probability forecasts increases the variance of the forecasts and may result in a suboptimal solution, lacking both calibration and sharpness. Several nonlinear alternatives to linear pooling methods have been developed to address these shortcomings for recalibration purposes.

Motivated by the seminal work of Dawid et al. (1995), Gneiting and Ranjan (2013) developed the generalized linear pool (GLP) to incorporate a parametric family of combination formulas. Let $F_i(y_{T+h}|I_T)$ denote the CDF (cumulative distribution function) of the probability forecast ($i = 1, \ldots, N$), and $\tilde{F}(y_{T+h}|I_T)$ denote the CDF of the combined forecast. The generalized pooling scheme takes the following form

$$\tilde{F}(y_{T+h}|I_T) = g^{-1}\left(\sum_{i=1}^{N} w_{T+h|T,i}\, g\big(F_i(y_{T+h}|I_T)\big)\right),$$

where $w_{T+h|T,1}, \ldots, w_{T+h|T,N}$ are nonnegative weights that sum to one, and g denotes a continuous and strictly monotonic function with the inverse g^{-1}. The linear, harmonic and logarithmic (geometric) pools become special cases of the GLP for $g(x) = x$, $g(x) = 1/x$, and $g(x) = \log(x)$, respectively. Gneiting and Ranjan (2013) highlighted that the generalized pooling strategy might fail to be sufficiently flexibly dispersive for calibration.

As a result, they also proposed the spread-adjusted linear pool (SLP) to allow one to address the calibration problem. Define F_i^0 and corresponding density f_i^0 via $F_i(y_{T+h}|I_T) = F_i^0(y_{T+h} - \eta_i|I_T)$ and $f_i(y_{T+h}|I_T) = f_i^0(y_{T+h} - \eta_i|I_T)$, where η_i is the unique median of $F_i(y_{T+h}|I_T)$. Then the SLP has the combined CDF and the corresponding density,

$$\tilde{F}(y_{T+h}|I_T) = \sum_{i=1}^{N} w_{T+h|T,i} F_i^0\left(\frac{y_{T+h} - \eta_i}{c}\middle| I_T\right) \quad \text{and}$$

$$\tilde{f}(y_{T+h}|I_T) = \frac{1}{c}\sum_{i=1}^{N} w_{T+h|T,i} f_i^0\left(\frac{y_{T+h} - \eta_i}{c}\middle| I_T\right),$$

respectively, where $w_{T+h|T,1}, \ldots, w_{T+h|T,N}$ are nonnegative weights with $\sum_{i=1}^{N} w_{T+h|T,i} = 1$, and c is a strictly positive spread adjustment parameter. The traditional linear pool is a special case for $c = 1$. A value of $c < 1$ is suggested for neutrally confident or underconfident component forecasts, while a value of $c \geq 1$ is suggested for overconfident components. Moreover, one can introduce spread adjustment parameters varying with the components in case the degrees of miscalibration of the components differ substantially.

The cumulative beta distribution is widely employed for recalibration because of the flexibility of its shape (see, e.g., Graham, 1996). Ranjan and Gneiting (2010) introduced a beta-transformed linear pool (BLP) that merges the traditional linear pool with a beta transform to achieve calibration. The BLP takes the form

$$\tilde{F}(y_{T+h}|I_T) = B_{\alpha,\beta}\left(\sum_{i=1}^{N} w_{T+h|T,i} F_i(y_{T+h}|I_T)\right),$$

where $w_{T+h|T,1}, \ldots, w_{T+h|T,N}$ are nonnegative weights that sum to one, and $B_{\alpha,\beta}$ is the CDF of the beta distribution with the shape parameters $\alpha > 0$ and $\beta > 0$. Full generality of the BLP enables an asymmetric beta-transformation based on the linear pooling of probability forecasts. In its most simplistic case, the BLP approach nests the traditional linear pool under the restriction $\alpha = \beta = 1$. The beta-transformation tunes up a linear pooled probability forecast if it is larger than 0.5 and tunes it down otherwise when imposing the constraint $\alpha = \beta \geq 1$. The approach can combine probability forecasts from both calibrated and uncalibrated sources. The estimates of the beta-transformation, along with the mixture weights for linear pooling, can be obtained by maximum likelihood, as suggested by Ranjan and Gneiting (2010). Recent work by Lahiri et al. (2015) demonstrated the superiority of the BLP approach, based on identifying the most valuable individual forecasts by a Welch-type test, over the equally weighted approach for calibration and sharpness.

To achieve improved calibration properties, Bassetti et al. (2018) proposed a Bayesian nonparametric approach based on Gibbs and slice sampling to realize the calibration and combination of probability forecasts by introducing an additional beta mixture in the BLP method. The resulting predictive CDF is

$$\tilde{F}(y_{T+h}|I_T) = \sum_{k=1}^{K} \omega_k B_{\alpha_k,\beta_k}\left(\sum_{i=1}^{N} w_{T+h|T,ki} F_i(y_{T+h}|I_T)\right),$$

where $\omega_1, \ldots, \omega_K$ denote beta mixture weights. The proposed approach enables one to treat the parameter K as bounded or unbounded, and it reduces to the BLP for $K = 1$. The Bayesian inference approach achieved a compromise between parsimony and flexibility and produced well-calibrated and accurate forecasts in their simulations and the empirical examples, outperforming the linear pool substantially.

The essence of these nonlinear pooling methods is to perform various transformations, which may be nonlinear, to either the component forecasts or the linearly pooled forecasts to restore calibration and sharpness. Kapetanios et al. (2015) generalized the literature by incorporating the dependence of the mixture weights on the variable one is trying to forecast, allowing the weights to introduce the nonlinearities and thus leading to outcome-dependent density pooling. The forecast performance of nonlinear pooling approaches largely depends on diverse factors, including the features of the target data, mixture component models, and training periods, and thereby deserves further research. This is in agreement with Baran and Lerch (2018), who investigated the performance of state-of-the-art forecast combination methods through case studies and found no substantial differences in forecast performance between the simple linear pool and the theoretically superior but cumbersome nonlinear pooling approaches.

3.6. Meteorological ensembles

The term "combination" and "ensemble" are often used interchangeably in the literature on forecast combinations. However, "ensemble" was initially developed in the meteorological literature from combinations of probabilistic forecasts that we have discussed.

Instead of combining multiple probabilistic forecasts available for forecasters, as with the approaches reviewed in preceding sections, an ensemble weather forecast is constructed from a set of point forecasts of the same weather quantity of interest, based on perturbed initial atmospheric states (e.g., Gneiting & Raftery, 2005; Maqsood et al., 2004) and/or different model formulations (e.g., Buizza et al., 2005, 1999). In this light, two major sources of forecast uncertainty, initial condition uncertainty resulting from the chaotic nature of the atmosphere, and model uncertainty arising from imperfect numerical models, are addressed (Baran, 2014; Lorenz, 1963; Weigel et al., 2008). This enables a measure of uncertainty to be attached to an ensemble weather forecast, making it more valuable than a single "deterministic" forecast, providing an inherently probabilistic assessment.

A meteorological ensemble forecast is a probabilistic forecast in the sense described here, assuming that there is no inherent uncertainty other than that contained in the initial conditions and the model formulation. In contrast, most statistical forecasting methods include an important additional source of uncertainty due to random noise innovations but do not usually include uncertainty due to initial conditions. The distinction is important enough, and the literature sufficiently distinct, that we have chosen to discuss meteorological ensemble forecasts in this separate section.

It has been demonstrated that the raw meteorological ensemble forecasts typically present systematic errors regarding bias (Atger, 2003; Mass, 2003) and dispersion (Buizza et al., 2005; Sloughter et al., 2010), with a tendency for the truth frequently falling outside of the range of the ensemble. Various statistical postprocessing methods have been introduced to improve the forecast quality and correct these errors by estimating representable relationships between the response variable of interest and predictors. Most postprocessing methods can be categorized into two groups: parametric approaches with distribution-based assumptions, such as ensemble model output statistics (EMOS, Gneiting et al., 2005) models and BMA (Raftery et al., 2005), and nonparametric approaches with distribution-free assumptions, such as analog-based methods (e.g., Delle Monache et al., 2013) and quantile regression forests (Taillardat et al., 2019). See Vannitsem et al. (2021) for a recent review of statistical postprocessing methods and their potential and challenges.

Recently, the community of weather forecasting is starting to explore the potentials of machine learning techniques, especially in the context of ensemble forecasting, in the sense of including arbitrary predictors and accounting for nonlinear dynamics of the Earth system that are not captured by existing numerical models (Dueben et al., 2021). One use of machine learning techniques is to complement ensemble NWP (numerical weather prediction, see, e.g., Bauer et al., 2015; Benjamin et al., 2018, for a summary of its revolution) forecasts using an additive postprocessing step for correction of ensemble bias and spread (Grönquist et al., 2021; Rasp & Lerch, 2018; Scher & Messori, 2018). Machine learning techniques, such as neural networks, have also been used as data-driven forecast tools, an alternative to NWP models based on the physical laws governing the atmosphere, to generate base forecasts. These techniques improve computational efficiency in creating ensemble forecasts with much larger ensemble sizes (Dueben & Bauer, 2018; Rasp & Thuerey, 2021; Scher, 2018; Scher & Messori, 2021).

3.7. Combinations constructed via Bayes' theorem

Pooling approaches, elaborated in Sections 3.3–3.5, pool/mix multiple probability forecasts with equal weights, weights evaluated using various scoring rules, or posterior probabilities sequentially updated via Bayes' Theorem. They inherently neglect correlations among the component probability forecasts. Nevertheless, forecasts derived from different sources are likely to share the same data, overlapping information, similar forecasting models, and common training processes. Thus, some sort of dependence among individual probability forecasts is extremely likely, and such dependence can severely impact the aggregated distributions. This section reviews an alternative class of combination techniques in which dependence among component probability forecasts can be incorporated. The major issue that makes this class of combinations difficult lies in how to model the dependence among individual distributions to achieve a good performance.

The extensive literature on probability forecast combinations considering correlations among individual distributions has primarily been driven from a foundational Bayesian perspective. The literature has originated in agent opinion analysis theory, free from the time series context, dating back to at least the pioneering work of Winkler (1968). We remark that in pooling techniques, the contribution of each probability forecast to the final aggregated probability forecast is measured explicitly via weights. In contrast, it is not specified by a specific form in the Bayesian combination techniques discussed in this section.

Early work in the Bayesian vein focused on a Bayesian paradigm developed by Morris (1974, 1977) in which a decision-maker views available probability forecasts from various sources simply as data and updates their prior distribution using Bayes' Theorem. At time T, the decision-maker aims to forecast y_{T+h} and receives current h-step-ahead probability forecasts $\mathcal{H}_{T+h} = \{f_1(y_{T+h}|I_T), \ldots, f_N(y_{T+h}|I_T)\}$ from the set of models. The posterior probability forecast of y_{T+h} is then

$$\tilde{f}\left(y_{T+h}|I_T, \mathcal{H}_{T+h}\right) \propto p\left(y_{T+h}|I_T\right) f_N\left(\mathcal{H}_{T+h}|y_{T+h}, I_T\right), \qquad (19)$$

where $p\left(y_{T+h}|I_T\right)$ denotes the decision-maker's prior probability for y_{T+h} given the available information I_T, and $f_N\left(\mathcal{H}_{T+h}|y_{T+h}, I_T\right)$ denotes the joint likelihood function derived from the individual distributions.

The problem of eliciting the posterior probability forecast in Eq. (19) is therefore broken down into the problem of specifying the prior distribution and assessing the form of the joint distribution, or likelihood, derived from the component probability forecasts. A flat (possibly improper) prior is often considered in the literature because: (i) it is reasonable to assume that everything the decision-maker knows is integrated into the individual distributions; and (ii) if not, the extra knowledge from the decision-maker can be incorporated in the likelihood as an additional individual distribution; see, e.g., Winkler (1968), Clemen and Winkler (1993), Clemen and Winkler (1985), and Jouini and Clemen (1996). Thus, the application of Bayes' Theorem presents the most taxing difficulties in delicately specifying the likelihood function, which requires consideration of the bias and precision of the individual distributions as well as their dependence (Hall & Mitchell, 2007).

One line of research has considered specifying the likelihood as a joint distribution of forecast errors and supported using the correlation between individuals' forecast errors to represent the dependence among individual distributions. Emphasis has been placed on making the likelihood computation tractable by adopting certain distributional assumptions. For example, Winkler (1981) assessed the likelihood as a multivariate normal distribution. Restricting the focus to individual models with unbiased forecasts, he derived tractable expressions for the posterior probability forecast, a normal distribution with mean $\tilde{\mu} = \mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}$ and variance $\tilde{\sigma}^2 = 1/\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}$, where $\mathbf{1}$ is an N-dimensional unit vector, $\boldsymbol{\mu}$ is an N-dimensional vector of individuals' mean, and $\boldsymbol{\Sigma}$ is a known covariance matrix of forecast errors. The mean of the posterior probability forecast is essentially a linear combination of the individuals' means with weights $\mathbf{1}'\boldsymbol{\Sigma}^{-1}/\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}$ identical to the "optimal" weights proposed by Bates and Granger (1969) and the weights derived from the constrained regression proposed by Granger and Ramanathan (1984) (see Section 2.2). The latter two approaches do not require normality. The estimation of $\boldsymbol{\Sigma}$, therefore, becomes crucial when $\boldsymbol{\Sigma}$ is unknown. Winkler (1981) suggested estimating $\boldsymbol{\Sigma}$ from data and using an inverted Wishart distribution as a prior for $\boldsymbol{\Sigma}$. The procedure is computationally intensive when the number of individual distributions to combine increases; see Hall and Mitchell (2007) for more discussion of the covariance matrix estimation. Following Winkler (1981), Palm and Zellner (1992) extended the approach to allow for biased individual forecasts, providing a complete solution to the forecast combination problem that takes into account the joint distribution of forecast errors from the individual models.

Jouini and Clemen (1996) took a different perspective and looked at the likelihood function derived from a copula-based joint distribution, in which dependence among individual distributions is encoded into the copula. The procedure is appealing in the sense of being able to deal with individual forecasts with arbitrary distributions. A recent study by Wilson (2017) gave an expert judgment study to assess the practical significance of the individual's dependency by comparing the Bayesian combination methods, developed by Winkler (1981) and Jouini and Clemen (1996) and common pooling methods.

3.8. Combinations constructed via integration

A fully specified Bayesian model is difficult to conceptualize, especially when biases and miscalibration of individual distributions (and, critically, dependencies among them) are time-varying. In this light, the probability forecast combination method of McAlinn and West (2019) may be helpful. They adapted and extended the basic Bayesian predictive synthesis (BPS) framework developed in agent opinion analysis (see, e.g., Genest & Schervish, 1985; West, 1992; West & Crosse, 1992) to sequential forecasting in time series. In the dynamic extension of the BPS model, the posterior probability forecast takes the form

$$\tilde{f}\left(y_{T+h}|I_T, \mathcal{H}_{1:T+h}\right) = \int_{\boldsymbol{x}_{T+h}} \alpha\left(y_{T+h}|\boldsymbol{x}_{T+h}\right) \prod_{i=1:N} f_i\left(x_{T+h,i}|I_T\right) d\boldsymbol{x}_{T+h},$$

where, to use our earlier notation, $\mathcal{H}_{1:T+h}$ denotes the full set of individual probability forecasts available for the decision-maker up to forecast origin T, $\boldsymbol{x}_{T+h} = x_{T+h,1:N}$ is an N-dimensional vector of latent variables at time $T+h$, and $\alpha\left(y_{T+h}|\boldsymbol{x}_{T+h}\right)$ is a conditional distribution for y_{T+h} given $\boldsymbol{x}_{T+h}$ defining the synthesis function.

Instead of constructing Bayesian combinations by multiplying a likelihood by a prior, the dynamic BPS method follows a subclass of Bayesian updating rules, i.e., updating by integration, in the form of latent factor models. Information about biases, miscalibration, and dependencies among the individual distributions can then be incorporated directly through the specification of the synthesis function. Specifically, McAlinn and West (2019) developed a time-varying (non-convex/nonlinear) combination of probability forecasts by defining a normally distributed synthesis function

$$\alpha\left(y_{T+h}|\boldsymbol{x}_{T+h}\right) = N\left(y_{T+h}|\boldsymbol{A}_{T+h}'\boldsymbol{\theta}_{T+h}, v_{T+h}\right)$$

with $\boldsymbol{A}_{T+h} = \left(1, \boldsymbol{x}_{T+h}'\right)'$ and $\boldsymbol{\theta}_{T+h} = \left(\theta_{T+h,0}, \theta_{T+h,1}, \ldots, \theta_{T+h,N}\right)'$. A dynamic linear model is built to model the time evolution of these parameter processes, which is defined as

$$y_{T+h} = \boldsymbol{A}_{T+h}'\boldsymbol{\theta}_{T+h} + \nu_{T+h}, \quad \nu_{T+h} \sim N\left(0, v_{T+h}\right),$$
$$\boldsymbol{\theta}_{T+h} = \boldsymbol{\theta}_{T+h-1} + \boldsymbol{\omega}_{T+h}, \quad \boldsymbol{\omega}_{T+h} \sim N\left(\mathbf{0}, v_{T+h}\boldsymbol{W}_{T+h}\right),$$

where $\boldsymbol{\theta}_{T+h}$ evolves in time according to a normal random walk with innovations variance matrix $v_{T+h}\boldsymbol{W}_{T+h}$, and v_{T+h}, identifies the residual variance in forecasting y_{T+h} given past information and the set of agent forecast distributions. It is suggested that BPS models be customized specifically to the forecast horizon, as a forecasting model may provide different forecast performances at different forecast horizons. MCMC methods are required for this posterior inference, and dependencies among agents are involved in sequentially updated estimates of the BPS parameters. Their results of forecasting a quarterly series of inflation rates showed that

the proposed dynamic BPS model significantly outperforms benchmark methods, such as the pooling and BMA combination techniques described in Sections 3.3 and 3.4. This also held true in a multivariate extension studied by McAlinn et al. (2020). McAlinn and West (2019) also showed that the dynamic BPS framework encompasses many existing combination methods, including linear pooling and BMA methods, by specifying different forms of the BPS synthesis function.

3.9. Quantile forecast combinations

Probabilistic forecasts can also be elicited in the form of quantiles, which are the inverse of the corresponding probability forecasts characterized by the CDFs. Quantile combinations involve averaging the individuals' quantile functions rather than their inverses as in linear pooling (see Section 3.3). In other words, quantile combinations entail horizontally averaging the individuals' CDFs while linear pooling entails vertically averaging (Lichtendahl et al., 2013).

The standard combination strategy for quantiles is to *allocate identical weights over all quantile levels for each model*. For $i = 1, \ldots, N$, let $F_i(y_{T+h}|I_T)$ denote the individual CDF with corresponding probability density function given by $f_i(y_{T+h}|I_T)$ and let $Q_{T+h|T,i}(\tau) = F^{-1}_{T+h|T,i}(\tau)$ denote the corresponding quantile function. Quantile averaging is then given by

$$\tilde{Q}_{T+h|T}(\tau) = \sum_{i=1}^{N} w_{T+h|T,i} Q_{T+h|T,i}(\tau), \quad 0 < \tau \leq 1, \tag{20}$$

where the weight $w_{T+h|T,i} \geq 0$ such that $\sum_{i=1}^{N} w_{T+h|T,i} = 1$. This combination strategy is also referred to as Vincentization (Vincent, 1912).

Interestingly, unlike linear pooling in Eq. (12), if individual distributions belong to the same location-scale family (such as normal, Logistic, Cauchy, etc.), then quantile averaging yields a combined distribution from the same family, with parameters given by weighted averages of the individuals' parameters (Ratcliff, 1979; Thomas & Ross, 1980). Consequently, quantile averaging of normal distributions is always uni-modal (and normal), while linear pooling, in general, may be multi-modal. Moreover, quantile averaging and linear pooling share the same mean, while quantile averaging tends to be sharper and more confident due to the additional spread driven by the disagreement on the mean in linear pooling.

Is it better to average quantiles (as in quantile averaging) or average probabilities (as in linear pooling)? By restricting themselves to simple averages, Lichtendahl et al. (2013) and Busetti (2017) theoretically and empirically compared the properties of these two combination strategies and suggested that quantile averaging seems overall a preferable and viable approach. Lichtendahl et al. (2013) attributed this, in part, to the fact that the average probability forecast is, in general, underconfident while the average quantile forecast is always sharper. Even when individual forecasts agree on the location and the average probability forecast is overconfident, the more overconfident average quantile forecast still offers the possibility of forecast improvements. This is due to its shape properties, specifically a higher density in the shoulders and a lower density in the tails. Busetti (2017) reconfirmed the argument and demonstrated that quantile averaging performs better than linear pooling and logarithmic pooling when combining individual forecast distributions with large biases. Incorporating quantile and probability averaging taken together may be useful to provide additional insight. Accordingly, three simple methods were suggested by Lichtendahl et al. (2013) to blend these two different combination strategies.

Rather than assuming that the entire individual quantile functions are available, one is often provided with a collection of quantiles corresponding to an equidistant dense grid of probabilities $\mathcal{T} \subseteq [0, 1]$, leading to a loss of information compared with consideration of the whole distribution. For instance, the 0.05, 0.25, 0.50, 0.75, and 0.95 quantiles are often elicited in practice, and another popular choice contains quantiles on all percentiles, i.e., $\mathcal{T} = (0.01, 0.02, \ldots, 0.99)$. Quantile combination, in this case, turns out to be a special case of the general aggregation rule in Eq. (20). For each quantile level, equally weighing the quantiles across all individual models is simple and quite robust, yielding improved forecast skill relative to the individual models and competitive performance relative to a variety of more sophisticated combination strategies (e.g., Busetti, 2017; Ray et al., 2022; Smyl & Hua, 2019). In cases where sufficient past data is available, some effort has been directed toward determining combination weights via a cross-validation framework to reflect the past out-of-sample performance of the different models and improve the utility of combinations. Scoring rules for quantile forecast evaluation can be harnessed for weight construction (Gneiting & Raftery, 2007; Grushka-Cockayne, Lichtendahl et al., 2017; Trapero et al., 2019).

One line of research has looked at *tailoring the individual weights for different quantile levels*, i.e., a separate weight is allocated for each model and quantile level by replacing $w_{T+h|T,i}$ with $w_{T+h|T,i}(\tau)$ in Eq. (20). For example, individual quantiles can be weighted by the reciprocal of the value of the pinball loss function (also referred to as the quantile loss) (Browell et al., 2020; Wang et al., 2019; Zhang et al., 2020). This flexible strategy enables the combination to accommodate the fact that individual forecasting models may have varying performances at different quantile levels. However, the number of weights to be learned scales with the number of quantile levels considered, which makes it challenging to achieve forecast improvements. Computationally intensive techniques, such as grid search and linear programming (LP), are applied for weight estimation, which is hardly scalable to large datasets. Moreover, the datasets involved in their empirical studies are not large enough to demonstrate the potential benefits of estimating such a large number of weights.

As discussed previously in the literature on point forecast combinations, the error in estimating optimal weights often exacerbates out-of-sample combined forecasts. The issue is even more problematic when it comes to

quantile combinations because it is a much more challenging task to estimate combination weights for a collection of quantiles, especially in the tails of forecast distributions, than merely for point forecasts defined using distributions. For example, Ray et al. (2022) failed to empirically demonstrate the utility of weighting different quantiles separately by optimizing the weighted interval score (WIS, Bracher et al., 2021) of the corresponding combined result. On the other hand, a promising line of research by Fakoor et al. (2021), in the context of quantile regression rather than time series forecasting, produced superior estimates using an aggregation strategy of greater flexibility than previously introduced. Separate weights depend on the features of individual models and (pairs of) quantile levels. They used a scalable stochastic gradient descent (SGD) algorithm with a monotone operator to solve the weight optimization problem and prevent the need for non-crossing constraints. More recently, Berrisch and Ziel (2021) introduced a new weighting method that allows individual forecasters to perform differently over time and within the distribution. This was the first to consider methods that aggregate across quantiles for optimal CRPS-based combinations using pointwise evaluation across the pinball loss. They also demonstrated the optimal convergence properties and the potential for performance improvements by considering pointwise optimization of the weight functions.

An independent line of research has looked at model-free combination heuristics, which frequently serve as benchmarks to measure the effectiveness of newly developed combination strategies. From a statistical point of view, these heuristics involve pooling together Q quantiles derived from N individual forecast distributions that are assumed to have the same values of characteristics (e.g., the quantile function at different quantile levels). The stacked larger pool is used to draw more precise estimates of those characteristics without training. Wang et al. (2019) formally introduced naïve sorting and median-based sorting methods, in which a total of $N \times Q$ quantiles are stacked and sorted by ascending order to pick the first and median values respectively in consecutive blocks of N forecasts. Naturally, averaging blocks of N forecasts is also another option.

While aggregating quantiles tightly connects with aggregating probability distributions, there has been little theoretical work in this area compared to combinations of probability forecasts which have seen considerable theoretical advances. One exception is Lichtendahl et al. (2013) which looked at the statistical properties of the simple average of probability forecasts and the ability to benefit from averaging. The choice of the combination weights has only been explored empirically, mainly in the context of energy forecasting (e.g., Browell et al., 2020; Wang et al., 2019) and epidemiological forecasting (e.g., Ray et al., 2022). Some of these proposals appear practical and beneficial, while others appear less useful. Further research is required to explore their utility.

Quantile crossing is a well-known problem caused by the lack of monotonicity in quantile estimates, which may arise when different combination weights are utilized for different quantile levels. Model-free heuristics, such as the median, are also included in such cases. Quantile crossing can be avoided by, e.g., (i) integrating the aggregation problems for individual models into one optimization problem subject to more non-crossing constraints (e.g., Bondell et al., 2010; Fakoor et al., 2021), and (ii) conducting naïve rearrangement after all the combined quantiles are obtained (e.g., Berrisch & Ziel, 2021; Chernozhukov et al., 2010). The rearrangement operation, though simple, is frequently recommended in practice since it will never deteriorate the forecasting performance in terms of the pinball loss (Chernozhukov et al., 2010).

Interval forecasts form a crucial special case of quantile forecasts, which makes the preceding combination approaches for quantile forecasts naturally apply to interval forecasts as well. When forming combinations of interval forecasts, attention should be paid to the fact that the combined interval forecasts are not guaranteed to provide target coverage rates (Grushka-Cockayne & Jose, 2020; Timmermann, 2006; Wallis, 2005). As a result, when evaluating the combined interval forecasts, proper scoring approaches that consider both width and coverage are appealing and can serve as objective functions to determine combination weights; see, e.g., Gneiting and Raftery (2007) and Jose and Winkler (2009).

For interval forecasts, six heuristics have been outlined: (1) simple average, (2) median, (3) envelope, (4) interior trimming, (5) exterior trimming, and (6) probability averaging of endpoints (Gaba et al., 2017; Park & Budescu, 2015). These six heuristics are virtually free of computational costs and have subsequently been promoted by recent research due to their robustness and benefits in different scenarios for addressing underconfidence/overconfidence; e.g., Smyl and Hua (2019), Petropoulos and Svetunkov (2020), and Grushka-Cockayne and Jose (2020). They can easily be extended to address the combinations of quantiles by aggregating individual quantiles in several ways for each quantile level.

Determining combination weights for interval forecasts is easier to implement than quantile forecasts since one only has to consider two quantiles. For example, by assuming the intervals to be symmetric around the point forecast, Montero-Manso et al. (2020) used the combined point forecast produced by a feature-based meta-learner as the center of the combined interval and generated the radius as a linear combination of the individual radii to minimize the MSIS (mean scaled interval score, Gneiting & Raftery, 2007) of the interval. The approach achieved the second position in the M4 competition with 100,000 time series involved. Subsequent work by Wang, Kang, Petropoulos and Li (2022) introduced a feature-based weight determination approach to directly combine lower and upper bounds of individual interval forecasts, leading to significant performance improvements compared to individual forecasts and the simple average.

4. Conclusions and a look to the future

Forecasting plays an indispensable role in decision-making, where success depends heavily on the accuracy of

the available forecasts. Even with a slight increase in accuracy, remarkable gains may be achieved in activities such as management planning and strategy setting (Makridakis, 1996; Syntetos et al., 2009). In this regard, forecast combinations provide an easy path to improving forecast accuracy by integrating the available information used in individual forecasts.

In this review, our goal has been to show how forecast combinations have evolved over time, identify the potential and limitations of various methods, and highlight the areas needing further research. Forecast combinations can be model-free or model-fitting, linear or nonlinear, static or time-varying, series-specific or cross-learning, and frequentist or Bayesian. The toolbox of combination methods has grown in size and sophistication, each with its merits. Which combination method to choose depends on several factors, such as the form of forecasts (point forecasts, probabilistic forecasts, quantiles, etc.), the quality and size of the model pool, the information available, and the specific forecasting problems. There is no clear consensus on which forecast combination method can be expected to perform best in one particular setting. Based on this review, we summarise some of the current research gaps and potential insights for future research in the following paragraphs.

Continuing to examine simple averaging. Over fifty years after Bates and Granger's (1969) pioneering work on forecast combinations, it is impressive that, in empirical studies, simple averaging still repeatedly dominates sophisticated weighted combinations, which are theoretically preferred, posing a challenging benchmark to beat. Although it is well known that the "forecast combination puzzle" stems from the unstable estimates of combination weights, researchers still lack comprehensive quantitative decision guidance on when to choose a simple averaging strategy over more complex strategies. One exception is Blanc and Setzer (2016), who merely looked at the combination of two individual forecasts and proposed decision rules to decide when to choose simple averaging over the "optimal" weights introduced by Bates and Granger (1969). In addition, the examination of simple averaging in the context of probabilistic forecast combinations deserves further attention and development, both theoretical and empirical.

Keeping combinations sophisticatedly simple. Forecasting models and forecast combination methods have grown swiftly in size and sophistication. Nevertheless, empirical results are ambiguous, and there is no coherent evidence that complexity systematically improves forecast accuracy; see, e.g., Green and Armstrong (2015) for a review comparing simple and complex forecasting methods. Following Zellner (2001), we suggest the blooming of sophisticatedly simple combinations to balance the trade-off between the benefits of tailoring weights for different individual models and the instability of learned weights in sophisticated weighting schemes. Additionally, it is strongly recommended that a detailed analysis is required to explore in depth how and why various sophisticated combination strategies work and, thus, provide more insights into which combination method to choose in a particular situation; Petropoulos et al. (2018a) provided a good example of this kind of work.

Obtaining statistical inference for the combination forecasts. The "forecast combination puzzle" revolves primarily around choosing fixed simple weights or random "optimal" ones. A related aspect to the puzzle, but somehow different from it, is that the randomness of the combination weights (and, in particular, the correlation with the forecasts) makes it difficult to perform statistical inference for the weighted combined forecast. A standard error is nontrivial to obtain in most cases, let alone the sampling distribution. Getting the combined forecast is one aspect; what to do with it in a statistical sense is another aspect. Therefore, future studies on the randomness of combination weights and the statistical inference for the combined forecasts would be of interest.

Selecting forecasts to be combined. The pool of individual forecasts lays the foundation for the utility of forecast combinations. These forecasts may come from statistical or machine learning models based on observed data, or be elicited from experts. Empirical evidence suggests that the future lies in the combination of statistical and machine learning generated forecasts and the incorporation of human judgment (Makridakis et al., 2020a; Petropoulos et al., 2022; Petropoulos, Kourentzes et al., 2018). Given a large number of available forecasts, selecting a subset of combinations becomes particularly pivotal to improving forecast skills and reducing computational costs. Various crucial issues in selecting forecasts have to be addressed, such as accuracy, robustness, diversity, and calibration when considering probabilistic forecasts (Lichtendahl & Winkler, 2020; Wang, Kang & Li, 2022). However, most of the existing algorithms perform ad hoc selections and lack statistical rationale (Kourentzes et al., 2019). Therefore, further attention should be paid to developing empirical guidelines and quantitative metrics to help forecasters select forecasts before combinations. Since a zero weight in the past does not indicate a zero weight in the future, time-varying subset selection for forecast combination would also be an interesting research direction.

Advancing the theory of nonlinear combinations. While the benefits of linear combinations of multiple forecasts are well appreciated in the forecasting literature, less attention has been paid to nonlinear combination schemes for modeling nonlinear dependencies among individual forecasts. This is possibly due to the lack of theoretical foundations and poor records of success; see Timmermann (2006) for a brief review of the related literature. Nonlinearities are currently addressed using neural networks or an additional nonlinear term in the combination equation. However, the limited evidence on the benefits of involving nonlinear combinations is mainly derived from only a few time series and is not entirely unconvincing. Consequently, we expect more theoretical and empirical work in this area soon.

Focusing more on probabilistic forecast combinations. In probabilistic forecast combinations, linear pooling and quantile averaging suggest two different ways of thinking

— linear pooling entails vertically averaging the individuals' CDFs, while quantile aggregation entails horizontally averaging. Accordingly, their combined forecasts hold other properties and benefit differently from the combination. For example, the shape-preserving property of quantile averaging may be appealing in certain settings (Lichtendahl et al., 2013). Over the past decade, linear pooling has attracted considerable attention, theoretically and empirically achieving appreciable advancements. Quantile averaging, however, has not received much attention, especially in the theoretical realm. Furthermore, when tailoring combination weights for different quantile levels, the instability of the estimated weights is especially problematic since many parameters have to be estimated. This issue will likely harm the calibration and sharpness of the out-of-sample combined forecasts, making quantile averaging challenging. Taken together, we expect combinations of quantiles to be an important area of research in the future.

Discussing if, how, and when it is helpful to interpret combination weights. In probability forecast combinations, some combination approaches have the property that poorly performing forecasts will almost always be rejected in favor of the best one as the sample size tends to infinity. For example, BMA reduces to model selection for a large sample size, with the best model receiving a weight very close to one. See Section 3.4 for more detailed discussions. However, it is sometimes found that individually "bad" forecasts may still be helpful in combinations (e.g., Geweke & Amisano, 2011). In this case, one does not want to zero-weight these bad forecasts (in the limit, as the sample size goes to infinity). This relates to the question of if, how, and when it is helpful to interpret combination weights, another future research direction worth exploring.

Taking account of correlations among individual forecasts. Some sort of correlations among individual forecasts are expected as they are likely to share the same data, overlapping information, similar forecasting models, and a common training process. Such correlations can be critical and seriously impact the utility of forecast combinations (De Menezes et al., 2000). An extensive body of literature on point forecast combinations has attempted to account for correlations in terms of weight estimation, even though these correlations can be poorly estimated. Despite the existence of such correlations, the literature on probabilistic forecast combinations has paid scant attention to addressing them; they are discussed primarily from a Bayesian perspective (e.g., McAlinn & West, 2019; Winkler, 1981). Therefore, another interesting path for further research would be considering correlations among individual forecasts in weighting schemes for probabilistic forecast combinations.

Cross-learning and feature engineering. Instead of combinations in a series-by-series fashion, numerous studies have confirmed the beneficial usage of information from multiple series to study common patterns among series, thereby facilitating the determination of combination weights and exploiting the benefits of cross-learning. The evidence of the potential of cross-learning has largely come from competitions (e.g., Makridakis et al., 2020a, 2022) and empirical studies (e.g., Ma & Fildes, 2021). Moreover, access to feature engineering can lead to improved forecasting performance, providing valuable information for forecast combinations in a cross-learning fashion (Kang et al., 2021; Montero-Manso et al., 2020). In this regard, we believe that further research needs to be done on feature engineering for time series data to unlock the potential of cross-learning.

Encouraging researchers to contribute open-source software and datasets. In this paper, we list some open-source packages linking to the developed approaches for forecast combinations (e.g., **fable**, **ForecastComb**, and **forecastHybrid** packages for R), time series features (e.g., **feasts** and **tsfeatures** packages for R and **tsfresh** and **Kats** packages for Python), and time series generations (e.g., **forecast** and **gratis** packages for R). We emphasize that open-source research software is a pathway to impact. Recent decades have witnessed a dramatically accelerating pace of advancements in computing. Consequently, it is time to promote the idea of researchers producing open-source software that provides evidence and support behind all the statements. Publicly releasing new software benefits researchers and end users. It reduces research costs, allows for quick implementation, helps people modify the existing software, and adapts it to other research ideas. We also encourage researchers to contribute open-source datasets because of the benefits of investigating and comparing the performance of newly developed methods; see, e.g., Godahewa et al. (2021a, 2021b) for a time series forecasting archive containing 20 publicly available time series datasets from different domains.

In this paper, we take the multiple forecasts to be combined essentially as given and limit ourselves to combinations of forecasts derived from separate models for a given series. These separate models can be identified with different model forms and/or the same model form with different parameters. However, we highlight that there are other types of forecast combinations in the forecasting literature. For example, one approach involves constructing replicas of the original time series through various manipulations of local curvatures, frequency transformation, or bootstrapping. Subsequently, multiple forecasts are produced to form the final combined forecasts, leading to a wide variety of approaches such as the theta method (Assimakopoulos & Nikolopoulos, 2000), temporal aggregation (e.g., Kourentzes & Petropoulos, 2016; Kourentzes, Petropoulos & Trapero, 2014; Kourentzes et al., 2017), bagging (e.g., Bergmeir et al., 2016; Petropoulos et al., 2018a), and structural combination (e.g., Rendon-Sanchez & De Menezes, 2019). Besides, one can construct a combination by averaging model parameters of multiple subseries to achieve ultra-long time series forecasting (e.g., Wang et al., 2022a). Another approach involves forming a hierarchical structure using multiple time series that are structurally connected based on geographic or logical reasons and reconciling multiple forecasts across the hierarchy, leading to various hierarchical aggregation methods (e.g., Ben Taieb et al., 2021; Hollyman et al., 2021; Hyndman et al., 2011; Wickramasuriya et al., 2019).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Adrian Raftery, Casey Lichtendahl, Yael Grushka-Cockayne, Fotios Petropoulos, and other experts in this area for providing helpful feedback on an earlier version of this paper. We thank the editors and two anonymous reviewers for their valuable comments and suggestions that improved the paper.

Feng Li's research was supported by the National Social Science Foundation of China (22BTJ028).

References

Aastveit, K. A., Mitchell, J., Ravazzolo, F., & van Dijk, H. K. (2019). The evolution of forecast density combinations in economics. http://dx.doi.org/10.1093/acrefore/9780190625979.013.381.

Adhikari, R. (2015). A mutual association based nonlinear ensemble mechanism for time series forecasting. *Applied Intelligence*, *43*(2), 233–250. http://dx.doi.org/10.1007/s10489-014-0641-y.

Adhikari, R., & Agrawal, R. K. (2012). A novel weighted ensemble technique for time series forecasting. In *Advances in knowledge discovery and data mining* (pp. 38–49). Berlin, Heidelberg: Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-30217-6_4.

Agnew, C. E. (1985). Bayesian consensus forecasts of macroeconomic variables. *Journal of Forecasting*, *4*(4), 363–376. http://dx.doi.org/10.1002/for.3980040405.

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*(1), 31–53. http://dx.doi.org/10.1016/j.jeconom.2005.07.015.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705.

Aksu, C., & 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. http://dx.doi.org/10.1016/0169-2070(92)90005-T.

Andrawis, R. R., Atiya, A. F., & 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. http://dx.doi.org/10.1016/j.ijforecast.2010.05.019.

Armstrong, J. S. (2001). Combining forecasts. In *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Boston, MA: Springer, http://dx.doi.org/10.1007/978-0-306-47630-3_19.

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*(4), 521–530. http://dx.doi.org/10.1016/S0169-2070(00)00066-2.

Atger, F. (2003). Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review*, *131*(8), 1509–1523. http://dx.doi.org/10.1175//1520-0493(2003)131<1509:SAIVOT>2.0.CO;2.

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, *36*(1), 197–200. http://dx.doi.org/10.1016/j.ijforecast.2019.03.010.

Babikir, A., & 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*, *51*(4), 1541–1556. http://dx.doi.org/10.1007/s00181-015-1049-1.

Baran, S. (2014). Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics & Data Analysis*, *75*, 227–238. http://dx.doi.org/10.1016/j.csda.2014.02.013.

Baran, S., & Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, *34*(3), 477–496. http://dx.doi.org/10.1016/j.ijforecast.2018.01.005.

Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., Liu, H., Schultz, T., & Gamboa, H. (2020). *SoftwareX, TSFEL: Time Series Feature Extraction Library, vol. 11* (pp. 2352–7110). Elsevier, http://dx.doi.org/10.1016/j.softx.2020.100456, 100456.

Bartoń, K. (2022). MuMIn: Multi-model inference.

Bassetti, F., Casarin, R., & Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association*, *113*(522), 675–685. http://dx.doi.org/10.1080/01621459.2016.1273117.

Bassetti, F., Casarin, R., & Ravazzolo, F. (2020). Density forecasting. In P. Fuleky (Ed.), *Macroeconomic forecasting in the era of big data: theory and practice* (pp. 465–494). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-31150-6_15.

Batchelor, R., & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, *41*(1), 68–75. http://dx.doi.org/10.1287/mnsc.41.1.68.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468. http://dx.doi.org/10.1057/jors.1969.103.

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. http://dx.doi.org/10.1038/nature14956.

Baumeister, C., & 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. http://dx.doi.org/10.1080/07350015.2014.949342.

Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2021). Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*, *116*(533), 27–43. http://dx.doi.org/10.1080/01621459.2020.1736081.

Benjamin, S. G., Brown, J. M., Brunet, G., Lynch, P., Saito, K., & Schlatter, T. W. (2018). 100 Years of progress in forecasting and NWP applications. *Meteorological Monographs*, *59*, 13.1–13.67. http://dx.doi.org/10.1175/AMSMONOGRAPHS-D-18-0020.1.

Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2016). Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *International Journal of Forecasting*, *32*(2), 303–312. http://dx.doi.org/10.1016/j.ijforecast.2015.07.002.

Berrisch, J., & Ziel, F. (2021). CRPS learning. *Journal of Econometrics*, http://dx.doi.org/10.1016/j.jeconom.2021.11.008.

Billio, M., Casarin, R., Ravazzolo, F., & van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, *177*(2), 213–232. http://dx.doi.org/10.1016/j.jeconom.2013.04.009.

Blanc, S. M., & Setzer, T. (2016). When to choose the simple average in forecast combination. *Journal of Business Research*, *69*(10), 3951–3962. http://dx.doi.org/10.1016/j.jbusres.2016.05.013.

Blanc, S. M., & Setzer, T. (2020). Bias–Variance Trade-Off and shrinkage of weights in forecast combination. *Management Science*, *66*(12), 5720–5737. http://dx.doi.org/10.1287/mnsc.2019.3476.

Bondell, H. D., Reich, B. J., & Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika*, *97*(4), 825–838. http://dx.doi.org/10.1093/biomet/asq048.

Bracher, J., Ray, E. L., Gneiting, T., & Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, *17*(2), Article e1008618. http://dx.doi.org/10.1371/journal.pcbi.1008618.

Browell, J., Gilbert, C., Tawn, R., & May, L. (2020). Quantile combination for the EEM20 wind power forecasting competition. In *2020 17th International conference on the european energy market* EEM, (pp. 1–6). ieeexplore.ieee.org, http://dx.doi.org/10.1109/EEM49802.2020.9221942.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information Fusion*, *6*(1), 5–20. http://dx.doi.org/10.1016/j.inffus.2004.04.004.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280. http://dx.doi.org/10.1287/mnsc.2014.1909.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., & Wei, M. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, *133*(5), 1076–1097. http://dx.doi.org/10.1175/MWR2905.1.

Buizza, R., Milleer, M., & Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887–2908. http://dx.doi.org/10.1002/qj.49712556006.

Bunn, D. W. (1975). A Bayesian approach to the linear combination of forecasts. *Journal of the Operational Research Society*, *26*, 325–329. http://dx.doi.org/10.1057/jors.1975.67.

Bunn, D. W. (1985). Statistical efficiency in the linear combination of forecasts. *International Journal of Forecasting*, *1*(2), 151–163. http://dx.doi.org/10.1016/0169-2070(85)90020-2.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2nd ed.). Springer New York, NY, http://dx.doi.org/10.1007/b97636.

Busetti, F. (2017). Quantile aggregation of density forecasts. *Oxford Bulletin of Economics and Statistics*, *79*(4), 495–512. http://dx.doi.org/10.1111/obes.12163.

Cang, S., & Yu, H. (2014). A combination selection algorithm on forecasting. *European Journal of Operational Research*, *234*(1), 127–139. http://dx.doi.org/10.1016/j.ejor.2013.08.045.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on machine learning* (p. 18). Association for Computing Machinery, http://dx.doi.org/10.1145/1015330.1015432.

Castle, J. L., Clements, M. P., & Hendry, D. F. (2013). Forecasting by factors, by variables, by both or neither? *Journal of Econometrics*, *177*(2), 305–319. http://dx.doi.org/10.1016/j.jeconom.2013.04.015.

Chan, F., & Pauwels, L. L. (2018). Some theoretical results on forecast combinations. *International Journal of Forecasting*, *34*(1), 64–74. http://dx.doi.org/10.1016/j.ijforecast.2017.08.005.

Chan, Y. L., Stock, J. H., & Watson, M. W. (1999). A dynamic factor model framework for forecast combination. *Spanish Economic Review*, *1*(2), 91–121. http://dx.doi.org/10.1007/s101080050005.

Chernozhukov, V., Fernández-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, *78*(3), 1093–1125. http://dx.doi.org/10.3982/ecta7880.

Chong, Y. Y., & Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, *53*(4), 671–690. http://dx.doi.org/10.2307/2297611.

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, *307*, 72–77. http://dx.doi.org/10.1016/j.neucom.2018.03.067.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762. http://dx.doi.org/10.1016/j.ijforecast.2015.12.005.

Clark, T. E., & McCracken, M. W. (2010). Averaging forecasts from VARs with uncertain instabilities. *Journal of Applied Econometrics*, *25*(1), 5–29. http://dx.doi.org/10.1002/jae.1127.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. http://dx.doi.org/10.1016/0169-2070(89)90012-5.

Clemen, R. T., & Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. *Operations Research*, *33*(2), 427–442. http://dx.doi.org/10.1287/opre.33.2.427.

Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, *4*(1), 39–46. http://dx.doi.org/10.1080/07350015.1986.10509492.

Clemen, R. T., & Winkler, R. L. (1993). Aggregating point estimates: A flexible modeling approach. *Management Science*, *39*(4), 501–515. http://dx.doi.org/10.1287/mnsc.39.4.501.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*(2), 187–203. http://dx.doi.org/10.1023/A:1006917509560.

Clements, M., & Hendry, D. (1998). *Forecasting economic time series*. Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511599286.

Collopy, F., & 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. http://dx.doi.org/10.1287/mnsc.38.10.1394.

Conflitti, C., De Mol, C., & Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, *31*(4), 1096–1103. http://dx.doi.org/10.1016/j.ijforecast.2015.03.009.

Costantini, M., & Pappalardo, C. (2010). A hierarchical procedure for the combination of forecasts. *International Journal of Forecasting*, *26*(4), 725–743. http://dx.doi.org/10.1016/j.ijforecast.2009.09.006.

Coulson, N. E., & Robins, R. P. (1993). Forecast combination in a dynamic setting. *Journal of Forecasting*, *12*(1), 63–67. http://dx.doi.org/10.1002/for.3980120106.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, *147*(2), 278–290. http://dx.doi.org/10.2307/2981683.

Dawid, A. P., DeGroot, M. H., Mortera, J., Cooke, R., French, S., Genest, C., Schervish, M. J., Lindley, D. V., McConway, K. J., & Winkler, R. L. (1995). Coherent combination of experts' opinions. *Test*, *4*, 263–313. http://dx.doi.org/10.1007/BF02562628.

De Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*(1), 190–204. http://dx.doi.org/10.1016/S0377-2217(98)00380-4.

Del Negro, M., Hasegawa, R. B., & Schorfheide, F. (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics*, *192*(2), 391–405. http://dx.doi.org/10.1016/j.jeconom.2016.02.006.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, *141*(10), 3498–3516. http://dx.doi.org/10.1175/mwr-d-12-00281.1.

Deutsch, M., Granger, C. W. J., & Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting*, *10*(1), 47–57. http://dx.doi.org/10.1016/0169-2070(94)90049-3.

Diebold, F. X. (1988). Serial correlation and the combination of forecasts. *Journal of Business & Economic Statistics*, *6*(1), 105–111. http://dx.doi.org/10.1080/07350015.1988.10509642.

Diebold, F. X. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, *5*(4), 589–592. http://dx.doi.org/10.1016/0169-2070(89)90014-9.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, *39*(4), 863–883. http://dx.doi.org/10.2307/2527342.

Diebold, F. X., & Pauly, P. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, *6*(1), 21–40. http://dx.doi.org/10.1002/for.3980060103.

Diebold, F. X., & Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, *6*(4), 503–508. http://dx.doi.org/10.1016/0169-2070(90)90028-A.

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting*, *35*(4), 1679–1691. http://dx.doi.org/10.1016/j.ijforecast.2018.09.006.

Diebold, F. X., Shin, M., & Zhang, B. (2022). On the aggregation of probability assessments: Regularized mixtures of predictive densities for eurozone inflation and real interest rates. *Journal of Econometrics*, http://dx.doi.org/10.1016/j.jeconom.2022.06.008.

Diks, C., Panchenko, V., & van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, *163*(2), 215–230. http://dx.doi.org/10.1016/j.jeconom.2011.04.001.

Donaldson, R. G., & Kamstra, M. (1996). Forecast combining with neural networks. *Journal of Forecasting*, *15*(1), 49–61. http://dx.doi.org/10.1002/(SICI)1099-131X(199601)15:1<49::AID-FOR604>3.0.CO;2-2.

Donate, J. P., Cortez, P., Sanchez, G. G., & De Miguel, A. S. (2013). Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble. *Neurocomputing*, *109*, 27–32. http://dx.doi.org/10.1016/j.neucom.2012.02.053.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *57*(1), 45–70. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02015.x.

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*(10), 3999–4009. http://dx.doi.org/10.5194/gmd-11-3999-2018.

Dueben, P. D., Bauer, P., & Adams, S. (2021). Deep learning to improve weather predictions. In *Deep Learning for the Earth Sciences* (pp. 204–217). Wiley, http://dx.doi.org/10.1002/9781119646181.ch14,

Elliott, G. (2011). Averaging and the optimal combination of forecasts. University of California, San Diego.

Elliott, G., & Timmermann, A. (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics*, *122*(1), 47–79. http://dx.doi.org/10.1016/j.jeconom.2003.10.019.

Fakoor, R., Kim, T., Mueller, J., Smola, A., & Tibshirani, R. J. (2021). Flexible model aggregation for quantile regression. http://dx.doi.org/10.48550/ARXIV.2103.00083, arXiv:2103.00083.

Fernández-Villaverde, J., & Rubio-Ramırez, J. F. (2004). Comparing dynamic equilibrium models to data: a Bayesian approach. *Journal of Econometrics*, *123*(1), 153–187. http://dx.doi.org/10.1016/j.jeconom.2003.10.031.

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, *15*(3), 227–246. http://dx.doi.org/10.1016/S0169-2070(98)00073-9.

Fletcher, D. (2018). *Model Averaging*. Berlin: Springer, http://dx.doi.org/10.1007/978-3-662-58541-2.

Freitas, P. S. A., & Rodrigues, A. J. L. (2006). Model combination in neural-based forecasting. *European Journal of Operational Research*, *173*(3), 801–814. http://dx.doi.org/10.1016/j.ejor.2005.06.057.

Fulcher, B. D., & Jones, N. S. (2017). Hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell Systems*, *5*(5), 527–531.e3. http://dx.doi.org/10.1016/j.cels.2017.10.001.

Gaba, A., Tsetlin, I., & Winkler, R. L. (2017). Combining interval forecasts. *Decision Analysis*, *14*(1), 1–20. http://dx.doi.org/10.1287/deca.2016.0340.

Galton, F. (1907a). One vote, one value. *Nature*, *75*, 414. http://dx.doi.org/10.1038/075414a0.

Galton, F. (1907b). Vox populi. *Nature*, *75*, 450–451. http://dx.doi.org/10.1038/075450a0.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, Article 105151. http://dx.doi.org/10.1016/j.engappai.2022.105151.

Garratt, A., Lee, K., Pesaran, M. H., & Shin, Y. (2003). Forecast uncertainties in macroeconomic modeling. *Journal of the American Statistical Association*, *98*(464), 829–838. http://dx.doi.org/10.1198/016214503000000765.

Gastinger, J., Nicolas, S., Stepić, D., Schmidt, M., & Schülke, A. (2021). A study on ensemble learning for time series forecasting and the need for meta-learning. In *2021 International joint conference on neural networks* (pp. 1–8). http://dx.doi.org/10.1109/IJCNN52387.2021.9533378.

Genest, C., & Schervish, M. J. (1985). Modeling expert judgments for Bayesian updating. *The Annals of Statistics*, *13*(3), 1198–1212. http://dx.doi.org/10.1214/aos/1176349664.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121. http://dx.doi.org/10.1016/j.ijforecast.2012.06.004.

Geweke, J., & Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, *26*(2), 216–230. http://dx.doi.org/10.1016/j.ijforecast.2009.10.007.

Geweke, J., & Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, *164*(1), 130–141. http://dx.doi.org/10.1016/j.jeconom.2011.02.017.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *69*(2), 243–268. http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. http://dx.doi.org/10.1146/annurev-statistics-062713-085831.

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*(5746), 248–249. http://dx.doi.org/10.1126/science.1115255.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. http://dx.doi.org/10.1198/016214506000001437.

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118. http://dx.doi.org/10.1175/MWR2904.1.

Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, *7*, 1747–1782. http://dx.doi.org/10.1214/13-EJS823.

Godahewa, R., Bergmeir, C., Webb, G. I., Hyndman, R. J., & Montero-Manso, P. (2021a). Monash time series forecasting archive. In J. Vanschoren, & S. Yeung (Eds.), *1*, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Godahewa, R., Bergmeir, C., Webb, G. I., Hyndman, R. J., & Montero-Manso, P. (2021b). Monash time series forecasting repository. URL https://forecastingdata.org/.

Graham, J. R. (1996). Is a group of economists better than one? Than none? *Journal of Business*, *69*(2), 193–232.

Granger, C. W. J., & Jeon, Y. (2004). Thick modeling. *Economic Modelling*, *21*(2), 323–343. http://dx.doi.org/10.1016/S0264-9993(03)00017-8.

Granger, C. W. J., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, *3*(2), 197–204. http://dx.doi.org/10.1002/for.3980030207.

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, *68*(8), 1678–1685. http://dx.doi.org/10.1016/j.jbusres.2015.03.026.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society, Series A*, *379*(2194), Article 20200092. http://dx.doi.org/10.1098/rsta.2020.0092.

Grushka-Cockayne, Y., & Jose, V. R. R. (2020). Combining prediction intervals in the M4 competition. *International Journal of Forecasting*, *36*(1), 178–185. http://dx.doi.org/10.1016/j.ijforecast.2019.04.015.

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl, K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, *63*(4), 1110–1130. http://dx.doi.org/10.1287/mnsc.2015.2389.

Grushka-Cockayne, Y., Lichtendahl, K. C., Jose, V. R. R., & Winkler, R. L. (2017). Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Operations Research*, *65*(3), 712–728. http://dx.doi.org/10.1287/opre.2017.1588.

Guilhaumon, F. (2019). mmSAR: multimodel Species-Area Relationships. URL http://mmsar.r-forge.r-project.org/ R package version 1.0.

Gunter, S. I. (1992). Nonnegativity restricted least squares combinations. *International Journal of Forecasting*, *8*(1), 45–59. http://dx.doi.org/10.1016/0169-2070(92)90006-U.

Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, *23*(1), 1–13. http://dx.doi.org/10.1016/j.ijforecast.2006.08.001.

Harrald, P. G., & Kamstra, M. (1997). Evolving artificial neural networks to combine financial forecasts. *IEEE Transactions on Evolutionary Computation*, *1*(1), 40–52. http://dx.doi.org/10.1109/4235.585891.

Harvey, D. I., Leybourne, S. J., & Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics*, *16*(2), 254–259. http://dx.doi.org/10.1080/07350015.1998.10524759.

Henderson, T., & Fulcher, B. D. (2021). An empirical evaluation of time-series feature sets. In *2021 International conference on data mining workshops* (pp. 1032–1038). IEEE, http://dx.doi.org/10.1109/ICDMW53433.2021.00134.

Hendry, D. F., & Clements, M. P. (2004). Pooling of forecasts. *The Econometrics Journal*, *7*(1), 1–31.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting*, *21*(1), 15–24. http://dx.doi.org/10.1016/j.ijforecast.2004.05.002.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, *14*(4), 382–417. http://dx.doi.org/10.1214/ss/1009212519.

Hollyman, R., Petropoulos, F., & Tipping, M. E. (2021). Understanding forecast reconciliation. *European Journal of Operational Research*, *294*(1), 149–160. http://dx.doi.org/10.1016/j.ejor.2021.01.017.

Hora, S. C. (2004). Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science*, *50*(5), 597–604. http://dx.doi.org/10.1287/mnsc.1040.0205.

Hsiao, C., & Wan, S. K. (2014). Is there an optimal forecast combination? *Journal of Econometrics*, *178*, 294–309. http://dx.doi.org/10.1016/j.jeconom.2013.11.003.

Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36*(1), 7–14. http://dx.doi.org/10.1016/j.ijforecast.2019.03.015.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589. http://dx.doi.org/10.1016/j.csda.2011.03.006.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd). OTexts: Melbourne, Australia, URL https://otexts.com/fpp3/.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2021). forecast: Forecasting functions for time series and linear models. URL https://pkg.robjhyndman.com/forecast/ R package version 8.15.

Hyndman, R. J., Kang, Y., Montero-Manso, P., Talagala, T. S., Wang, E., Yang, Y., O'Hara-Wild, M., Ben Taieb, S., Hanqing, C., Lake, D. K., Laptev, N., & Moorman, J. R. (2019). tsfeatures: Time series feature extraction. URL https://CRAN.R-project.org/package=tsfeatures R package version 1.0.2.

Jore, A. S., Mitchell, J., & Vahey, S. P. (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Economics*, *25*(4), 621–634. http://dx.doi.org/10.1002/jae.1162.

Jose, V. R. R., Grushka-Cockayne, Y., & Lichtendahl, K. C. (2014). Trimmed opinion pools and the Crowd's calibration problem. *Management Science*, *60*(2), 463–475. http://dx.doi.org/10.1287/mnsc.2013.1781.

Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, *24*(1), 163–169. http://dx.doi.org/10.1016/j.ijforecast.2007.06.001.

Jose, V. R. R., & Winkler, R. L. (2009). Evaluating quantile assessments. *Operations Research*, *57*(5), 1287–1297. http://dx.doi.org/10.1287/opre.1080.0665.

Jouini, M. N., & Clemen, R. T. (1996). Copula models for aggregating expert opinions. *Operations Research*, *44*(3), 444–457. http://dx.doi.org/10.1287/opre.44.3.444.

Judge, G. G., & Bock, M. E. (1978). *The statistical implications of pre-test and stein-rule estimators in econometrics*. Amsterdam: North Holland.

Kang, H. (1986). Unstable weights in the combination of forecasts. *Management Science*, *32*(6), 683–695. http://dx.doi.org/10.1287/mnsc.32.6.683.

Kang, Y., Cao, W., Petropoulos, F., & Li, F. (2021). Forecast with forecasts: Diversity matters. *European Journal of Operational Research*, http://dx.doi.org/10.1016/j.ejor.2021.10.024.

Kang, Y., Hyndman, R. J., & Li, F. (2020). GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. *Statistical Analysis and Data Mining*, *13*(4), 354–376. http://dx.doi.org/10.1002/sam.11461.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*(2), 345–358. http://dx.doi.org/10.1016/j.ijforecast.2016.09.004.

Kang, Y., Li, F., Hyndman, R. J., O'Hara-Wild, M., & Zhao, B. (2020). gratis: GeneRAting TIme Series with diverse and controllable characteristics. URL https://CRAN.R-project.org/package=gratis R package version 0.2-1.

Kang, Y., Spiliotis, E., Petropoulos, F., Athiniotis, N., Li, F., & Assimakopoulos, V. (2020). Déjà vu: A data-centric forecasting approach through time series cross-similarity. *Journal of Business Research*, http://dx.doi.org/10.1016/j.jbusres.2020.10.051.

Kapetanios, G., Mitchell, J., Price, S., & Fawcett, N. (2015). Generalised density forecast combinations. *Journal of Econometrics*, *188*(1), 150–165. http://dx.doi.org/10.1016/j.jeconom.2015.02.047.

Kışınbay, T. (2010). The use of encompassing tests for forecast combinations. *Journal of Forecasting*, *29*(8), 715–727. http://dx.doi.org/10.1002/for.1170.

Kolassa, S. (2011). Combining exponential smoothing forecasts using Akaike weights. *International Journal of Forecasting*, *27*(2), 238–251. http://dx.doi.org/10.1016/j.ijforecast.2010.04.006.

Koop, G. (2003). *Bayesian Econometrics*. Wiley.

Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, *53*(3), 867–886. http://dx.doi.org/10.1111/j.1468-2354.2012.00704.x.

Koop, G., & Potter, S. (2003). Forecasting in large macroeconomic panels using Bayesian model averaging. [ISSN: 1556-5068] http://dx.doi.org/10.2139/ssrn.892860, FRB NY Staff Report No. 163.

Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, *41*(9), 4235–4244. http://dx.doi.org/10.1016/j.eswa.2013.12.011.

Kourentzes, N., Barrow, D., & Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, *209*(February 2018), 226–235. http://dx.doi.org/10.1016/j.ijpe.2018.05.019.

Kourentzes, N., & Petropoulos, F. (2016). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*, *181*, 145–153. http://dx.doi.org/10.1016/j.ijpe.2015.09.011.

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*(2), 291–302. http://dx.doi.org/10.1016/j.ijforecast.2013.09.006.

Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research*, *78*, 1–9. http://dx.doi.org/10.1016/j.jbusres.2017.04.016.

Krasnopolsky, V. M., & Lin, Y. (2012). A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental US. *Advances in Meteorology*, *2012*, http://dx.doi.org/10.1155/2012/649450.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1).

Lahiri, K., Peng, H., & Zhao, Y. (2015). Testing the value of probability forecasts for calibrated combining. *International Journal of Forecasting*, *31*(1), 113–129. http://dx.doi.org/10.1016/j.ijforecast.2014.03.005.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. http://dx.doi.org/10.1287/mnsc.1050.0459.

Leamer, E. E. (1978). *Specification searches: Ad Hoc inference with nonexperimental data*. Wiley.

Lemke, C., & Gabrys, B. (2010). Meta-learning for time series forecasting and forecast combination. *Neurocomputing*, *73*(10), 2006–2016. http://dx.doi.org/10.1016/j.neucom.2009.09.020.

Li, X., Kang, Y., & Li, F. (2020). Forecasting with time series imaging. *Expert Systems with Applications*, *160*(113680), Article 113680. http://dx.doi.org/10.1016/j.eswa.2020.113680.

Li, L., Kang, Y., & Li, F. (2022). Bayesian forecast combination using time-varying features. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2022.06.002.

Lichtendahl, K. C., Grushka-Cockayne, Y., & Winkler, R. L. (2013). Is it better to average probabilities or quantiles? *Management Science*, *59*(7), 1594–1611. http://dx.doi.org/10.1287/mnsc.1120.1667.

Lichtendahl, K. C., & Winkler, R. L. (2020). Why do some combinations perform better than others? *International Journal of Forecasting*, *36*(1), 142–149. http://dx.doi.org/10.1016/j.ijforecast.2019.03.027.

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., Jose, V. R., & Winkler, R. L. (2022). Extremizing and Antiextremizing in Bayesian Ensembles of Binary-Event Forecasts. *Operations Research*, http://dx.doi.org/10.1287/opre.2021.2176.

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., & Pfeifer, P. E. (2013). The wisdom of competitive crowds. *Operations Research*, *61*(6), 1383–1398. http://dx.doi.org/10.1287/opre.2013.1213.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, *20*(2), 130–141. http://dx.doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., & Jones, N. S. (2019). Catch22: Canonical time-series characteristics. *Data Mining and Knowledge Discovery*, *33*(6), 1821–1852. http://dx.doi.org/10.1007/s10618-019-00647-x.

Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, *288*(1), 111–128. http://dx.doi.org/10.1016/j.ejor.2020.05.038.

Makridakis, S. (1996). Forecasting: its role and value for planning and strategy. *International Journal of Forecasting*, *12*(4), 513–537. http://dx.doi.org/10.1016/S0169-2070(96)00677-2.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153. http://dx.doi.org/10.1002/for.3980010202.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476. http://dx.doi.org/10.1016/S0169-2070(00)00057-1.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020a). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. http://dx.doi.org/10.1016/j.ijforecast.2019.04.014.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2021.11.013.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., & Winkler, R. L. (2020b). The M5 Uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, 1–24. http://dx.doi.org/10.1016/j.ijforecast.2021.10.009.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*(9), 987–996. http://dx.doi.org/10.1287/mnsc.29.9.987.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. http://dx.doi.org/10.1037/a0036677.

Maqsood, I., Khan, M. R., & Abraham, A. (2004). An ensemble of neural networks for weather forecasting. *Neural Computing & Applications*, *13*(2), 112–122. http://dx.doi.org/10.1007/s00521-004-0413-4.

Martin, G. M., Loaiza-Maya, R., Maneesoonthorn, W., Frazier, D. T., & Ramírez-Hassan, A. (2021). Optimal probabilistic forecasts: When do they work? *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2021.05.008.

Mass, C. F. (2003). IFPS and the future of the national weather service. *Weather and Forecasting*, *18*(1), 75–79. http://dx.doi.org/10.1175/1520-0434(2003)018<0075:IATFOT>2.0.CO;2.

McAlinn, K., Aastveit, K. A., Nakajima, J., & West, M. (2020). Multivariate Bayesian predictive synthesis in macroeconomic forecasting. *Journal of the American Statistical Association*, *115*(531), 1092–1110. http://dx.doi.org/10.1080/01621459.2019.1660171.

McAlinn, K., & West, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, *210*(1), 155–169. http://dx.doi.org/10.1016/j.jeconom.2018.11.010.

McNees, S. K. (1992). The uses and abuses of 'consensus' forecasts. *Journal of Forecasting*, *11*(8), 703–710. http://dx.doi.org/10.1002/for.3980110807.

Montero-Manso, P. (2019). M4metalearning: Metalearning tools for time series forecasting.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. http://dx.doi.org/10.1016/j.ijforecast.2019.02.011.

Moon, J., Jung, S., Rew, J., Rho, S., & Hwang, E. (2020). Combination of short-term load forecasting models based on a stacking ensemble approach. *Energy and Buildings*, *216*, Article 109921. http://dx.doi.org/10.1016/j.enbuild.2020.109921.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys*, *29*(1), 46–75. http://dx.doi.org/10.1111/joes.12044.

Morris, P. A. (1974). Decision analysis expert use. *Management Science*, *20*(9), http://dx.doi.org/10.1287/mnsc.20.9.1233.

Morris, P. A. (1977). Combining expert judgments: A Bayesian approach. *Management Science*, *23*(7), 667–787. http://dx.doi.org/10.1287/mnsc.23.7.679.

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)*, *137*(2), 131–146. http://dx.doi.org/10.2307/2344546.

Newbold, P., & Harvey, D. I. (2004). Forecast combination and encompassing. In M. P. Clements, & D. F. Hendry (Eds.), *A companion to economic forecasting*. Blackwell Publishing, http://dx.doi.org/10.1002/9780470996430.ch12.

Nowotarski, J., Raviv, E., Trück, S., & Weron, R. (2014). An empirical comparison of alternative schemes for combining electricity spot price forecasts. *Energy Economics*, *46*, 395–412. http://dx.doi.org/10.1016/j.eneco.2014.07.014.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Richard Eiser, J., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons, http://dx.doi.org/10.1002/0470033312.

O'Hara-Wild, M., Hyndman, R. J., Wang, E., Cook, D., Talagala, T. S., & Chhay, L. (2021). feasts: Feature extraction and statistics for time series. URL https://CRAN.R-project.org/package=feasts R package version 0.2.2.

Öller, L.-E. (1978). A method for pooling forecasts. *Journal of the Operational Research Society*, *29*(1), 55–63. http://dx.doi.org/10.1057/jors.1978.8.

Opschoor, A., van Dijk, D., & van der Wel, M. (2017). Combining density forecasts using focused scoring rules. *Journal of Applied Economics*, *32*(7), 1298–1313. http://dx.doi.org/10.1002/jae.2575.

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. http://dx.doi.org/10.48550/ARXIV.1905.10437.

Palm, F. C., & Zellner, A. (1992). To combine or not to combine? issues of combining forecasts. *Journal of Forecasting*, *11*(8), 687–701. http://dx.doi.org/10.1002/for.3980110806.

Park, S., & Budescu, D. V. (2015). Aggregating multiple probability intervals to improve calibration. *Judgment and Decision Making*, *10*(2), 130–143.

Patton, A. J., & Timmermann, A. (2007). Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics*, *140*(2), 884–918. http://dx.doi.org/10.1016/j.jeconom.2006.07.018.

Pauwels, L. L., Radchenko, P., & Vasnev, A. L. (2020). *Higher moment constraints for predictive density combination*. (Working Paper 2020–45), CAMA, http://dx.doi.org/10.2139/ssrn.3593124.

Pauwels, L. L., & Vasnev, A. L. (2016). A note on the estimation of optimal weights for density forecast combinations. *International Journal of Forecasting*, *32*(2), 391–397. http://dx.doi.org/10.1016/j.ijforecast.2015.09.002.

Pawlikowski, M., & Chorowska, A. (2020). Weighted ensemble of statistical models. *International Journal of Forecasting*, *36*(1), 93–97. http://dx.doi.org/10.1016/j.ijforecast.2019.03.019.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., Baets, S. D., Dokumentov, A., ... Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. http://dx.doi.org/10.1016/j.ijforecast.2021.11.001.

Petropoulos, F., Hyndman, R. J., & 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. http://dx.doi.org/10.1016/j.ejor.2018.01.045.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management*, *60*, 34–46. http://dx.doi.org/10.1016/j.jom.2018.05.005.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). 'Horses for Courses' in demand forecasting. *European Journal of Operational Research*, *237*(1), 152–163. http://dx.doi.org/10.1016/j.ejor.2014.02.036.

Petropoulos, F., & Spiliotis, E. (2021). The wisdom of the data: Getting the most out of univariate time series forecasting. *Forecasting*, *3*(3), 478–497. http://dx.doi.org/10.3390/forecast3030029.

Petropoulos, F., & Svetunkov, I. (2020). A simple combination of univariate models. *International Journal of Forecasting*, *36*(1), 110–115. http://dx.doi.org/10.1016/j.ijforecast.2019.01.006.

Poncela, P., Rodríguez, J., Sánchez-Mangas, R., & Senra, E. (2011). Forecast combination through dimension reduction techniques. *International Journal of Forecasting*, *27*(2), 224–237. http://dx.doi.org/10.1016/j.ijforecast.2010.01.012.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. http://dx.doi.org/10.1175/mwr2906.1.

Raftery, A. E., Kárný, M., & Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, *52*(1), 52–66. http://dx.doi.org/10.1198/TECH.2009.08104.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191. http://dx.doi.org/10.1080/01621459.1997.10473615.

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *72*(1), 71–91. http://dx.doi.org/10.1111/j.1467-9868.2009.00726.x.

Rapach, D. E., & Strauss, J. K. (2008). Forecasting US employment growth using forecast combining methods. *Journal of Forecasting*, *27*(1), 75–93. http://dx.doi.org/10.1002/for.1051.

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. http://dx.doi.org/10.1175/MWR-D-18-0187.1.

Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, *13*(2), http://dx.doi.org/10.1029/2020ms002405.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446–461.

Ray, E. L., Brooks, L. C., Bien, J., Biggerstaff, M., Bosse, N. I., Bracher, J., Cramer, E. Y., Funk, S., Gerding, A., Johansson, M. A., Rumack, A., Wang, Y., Zorn, M., Tibshirani, R. J., & Reich, N. G. (2022). Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. http://dx.doi.org/10.48550/ARXIV.2201.12387.

Rendon-Sanchez, J. F., & De Menezes, L. M. (2019). Structural combination of seasonal exponential smoothing forecasts applied to load forecasting. *European Journal of Operational Research*, *275*(3), 916–924. http://dx.doi.org/10.1016/j.ejor.2018.12.013.

Ribeiro, G. T., Mariani, V. C., & 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. http://dx.doi.org/10.1016/j.engappai.2019.03.012.

Ribeiro, M. H. D. M., & 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*, Article 105837. http://dx.doi.org/10.1016/j.asoc.2019.105837.

Rossi, B. (2013). Chapter 21 - advances in forecasting under instability. In G. Elliott, & A. Timmermann (Eds.), *Handbook of Economic Forecasting*: *vol. 2*, *Handbook of Economic Forecasting* (pp. 1203–1324). Elsevier, http://dx.doi.org/10.1016/B978-0-444-62731-5.00021-X.

Rossi, B. (2021). Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature*, *59*(4), 1135–1190. http://dx.doi.org/10.1257/jel.20201479.

Satopää, V. A., Pemantle, R., & Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, *111*(516), 1623–1633. http://dx.doi.org/10.1080/01621459.2015.1100621.

Scher, S. (2018). Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*(22), 12,616–12,622. http://dx.doi.org/10.1029/2018gl080704.

Scher, S., & Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2830–2841. http://dx.doi.org/10.1002/qj.3410.

Scher, S., & Messori, G. (2021). Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, *13*(2), http://dx.doi.org/10.1029/2020ms002331.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. http://dx.doi.org/10.1214/aos/1176344136.

Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2020). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, *37*(3), 1072–1084. http://dx.doi.org/10.1016/j.ijforecast.2020.11.009.

Shaub, D. (2019). Fast and accurate yearly time series forecasting with forecast combinations. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2019.03.032.

Shi, S. M., Da Xu, L., & Liu, B. (1999). Improving the accuracy of nonlinear combined forecasting using neural networks. *Expert Systems with Applications*, *16*(1), 49–54. http://dx.doi.org/10.1016/S0957-4174(98)00030-X.

Sloughter, J. M., Gneiting, T., & Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, *105*(489), 25–35. http://dx.doi.org/10.1198/jasa.2009.ap08615.

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355. http://dx.doi.org/10.1111/j.1468-0084.2008.00541.x.

Smyl, S., & Hua, N. G. (2019). Machine learning methods for GEFCom2017 probabilistic load orecasting. *International Journal of Forecasting*, *35*(4), 1424–1431. http://dx.doi.org/10.1016/j.ijforecast.2019.02.002.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*(3), 644–719. http://dx.doi.org/10.1257/jel.20191385.

Stock, J. H., & Watson, M. W. (1998). *A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series*: *Working Paper 6607*, National Bureau of Economic Research, http://dx.doi.org/10.3386/w6607.

Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, *44*(2), 293–335. http://dx.doi.org/10.1016/S0304-3932(99)00027-6.

Stock, J. H., & 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., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430. http://dx.doi.org/10.1002/for.928.

Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*, *32*(4), 1339–1342.

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*, http://dx.doi.org/10.1080/03610927808827599.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Syntetos, A. A., Boylan, J. E., & Disney, S. M. (2009). Forecasting for inventory planning: a 50-year review. *Journal of the Operational Research Society*, *60*(1), S149–S160. http://dx.doi.org/10.1057/jors.2008.173.

Taillardat, M., Fougères, A.-L., Naveau, P., & Mestre, O. (2019). Forest-Based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, *34*(3), 617–634. http://dx.doi.org/10.1175/WAF-D-18-0149.1.

Talagala, T. S., Hyndman, R. J., & Athanasopoulos, G. (2018). Meta-learning how to forecast time series. *Monash Econometrics and Business Statistics Working Papers*, *6*, 18.

team, F. I. D. S. (2021). Kats. URL https://facebookresearch.github.io/Kats/ Python package version 0.1.0.

Terui, N., & van Dijk, H. K. (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, *18*(3), 421–438. http://dx.doi.org/10.1016/S0169-2070(01)00120-0.

Thomas, E. A. C., & Ross, B. H. (1980). On appropriate procedures for combining probability distributions within the same family. *Journal of Mathematical Psychology*, *21*(2), 136–152. http://dx.doi.org/10.1016/0022-2496(80)90003-6.

Thomson, M. E., Pollock, A. C., Önkal, D., & Gönül, M. S. (2019). Combining forecasts: Performance and coherence. *International Journal of Forecasting*, *35*(2), 474–484. http://dx.doi.org/10.1016/j.ijforecast.2018.10.006.

Thorey, J., Chaussin, C., & Mallet, V. (2018). Ensemble forecast of photovoltaic power with online CRPS learning. *International Journal of Forecasting*, *34*(4), 762–773. http://dx.doi.org/10.1016/j.ijforecast.2018.05.007.

Thorey, J., Mallet, V., & Baudin, P. (2017). Online learning with the continuous ranked probability score for ensemble forecasting. *Quarterly Journal of the Royal Meteorological Society*, *143*(702), 521–529. http://dx.doi.org/10.1002/qj.2940.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting, vol. 1* (pp. 135–196). Elsevier, http://dx.doi.org/10.1016/S1574-0706(05)01004-9.

Trapero, J. R., Cardós, M., & Kourentzes, N. (2019). Quantile forecast optimal combination to enhance safety stock estimation. *International Journal of Forecasting*, *35*(1), 239–250. http://dx.doi.org/10.1016/j.ijforecast.2018.05.009.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*(3), 261–289. http://dx.doi.org/10.1007/s10994-013-5401-4.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Bouallègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Plenković, I. O., Schmeits, M., ... Ylhaisi, J. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, *102*(3), E681–E699. http://dx.doi.org/10.1175/BAMS-D-19-0308.1.

Vincent, S. B. (1912). *The functions of the vibrissae in the behavior of the white rat*. Kessinger Publishing.

Wallis, K. F. (2005). Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics*, *67*(s1), 983–994. http://dx.doi.org/10.1111/j.1468-0084.2005.00148.x.

Wang, X., Kang, Y., Hyndman, R. J., & Li, F. (2022). Distributed ARIMA models for ultra-long time series. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2022.05.001.

Wang, X., Kang, Y., & Li, F. (2022). Another look at forecast trimming for combinations: robustness, accuracy and diversity. arXiv preprint arXiv:2208.00139.

Wang, X., Kang, Y., Petropoulos, F., & Li, F. (2022). The uncertainty estimation of feature-based forecast combinations. *Journal of the Operational Research Society*, *73*(5), 979–993. http://dx.doi.org/10.1080/01605682.2021.1880297.

Wang, X., Smith-Miles, K., & Hyndman, R. J. (2009). Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. *Neurocomputing*, *72*(10), 2581–2594. http://dx.doi.org/10.1016/j.neucom.2008.10.017.

Wang, Y., Zhang, N., Tan, Y., Hong, T., Kirschen, D. S., & Kang, C. (2019). Combining probabilistic load forecasts. *IEEE Transactions on Smart Grid*, *10*(4), 3664–3674. http://dx.doi.org/10.1109/TSG.2018.2833869.

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260. http://dx.doi.org/10.1002/qj.210.

Weiss, C. E., Roetzer, G. R., & Raviv, E. (2018). ForecastComb: Forecast combination methods. URL https://CRAN.R-project.org/package=ForecastComb R package version 1.3.1.

West, M. (1992). Modelling agent forecast distributions. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *54*(2), 553–567. http://dx.doi.org/10.1111/j.2517-6161.1992.tb01896.x.

West, M., & Crosse, J. (1992). Modelling probabilistic agent opinion. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *54*(1), 285–299. http://dx.doi.org/10.1111/j.2517-6161.1992.tb01882.x.

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*(526), 804–819. http://dx.doi.org/10.1080/01621459.2018.1448825.

Wilson, K. J. (2017). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*, *33*(1), 325–336. http://dx.doi.org/10.1016/j.ijforecast.2015.11.014.

Winkler, R. L. (1968). The consensus of subjective probability distributions. *Management Science*, *15*(2), B–61–B–75. http://dx.doi.org/10.1287/mnsc.15.2.B61.

Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science*, *27*(4), 479–488. http://dx.doi.org/10.1287/mnsc.27.4.479.

Winkler, R. L., & Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)*, *146*(2), 150–157. http://dx.doi.org/10.2307/2982011.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. http://dx.doi.org/10.1016/S0893-6080(05)80023-1.

Wright, J. H. (2008). Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics*, *146*(2), 329–341. http://dx.doi.org/10.1016/j.jeconom.2008.08.012.

Xie, J., & Hong, T. (2016). GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation. *International Journal of Forecasting*, *32*(3), 1012–1016. http://dx.doi.org/10.1016/j.ijforecast.2015.11.005.

Yao, X., & Islam, M. M. (2008). Evolving artificial neural network ensembles. *IEEE Computational Intelligence Magazine*, *3*(1), 31–42. http://dx.doi.org/10.1109/MCI.2007.913386.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007. http://dx.doi.org/10.1214/17-BA1091.

Zellner, A. (2001). Keep it sophisticatedly simple. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, inference and modelling* (pp. 242–262). Cambridge: Cambridge University Press.

Zhang, S., Wang, Y., Zhang, Y., Wang, D., & Zhang, N. (2020). Load probability density forecasting by transforming and combining quantile forecasts. *Applied Energy*, *277*, Article 115600. http://dx.doi.org/10.1016/j.apenergy.2020.115600.

Zhao, S., & Feng, Y. (2020). For2For: Learning to forecast from forecasts. http://dx.doi.org/10.48550/ARXIV.2001.04601, arXiv:2001.04601.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC Press.

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, *137*(1), 239–263. http://dx.doi.org/10.1016/S0004-3702(02)00190-X.

Zischke, R., Martin, G. M., Frazier, D. T., & Poskitt, D. (2022). The impact of sampling variability on estimated combinations of distributional forecasts. http://dx.doi.org/10.48550/ARXIV.2206.02376, arXiv:2206.02376.