International Journal of Forecasting 36 (2020) 142–149

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

Why do some combinations perform better than others?

Kenneth C. Lichtendahl Jr. [a,b,*], Robert L. Winkler [c]

[a] Darden School of Business, University of Virginia, Charlottesville, VA 22903, USA
[b] Google, LLC, Mountain View, CA 94043, USA
[c] Fuqua School of Business, Duke University, Durham, NC 27708, USA

A B S T R A C T

The evidence from the literature on forecast combination shows that combinations generally perform well. We discuss here how the accuracy and diversity of the methods being combined and the robustness of the combination rule can influence performance, and illustrate this by showing that a simple, robust combination of a subset of the nine methods used in the M4 competition's best combination performs almost as well as that forecast, and is easier to implement. We screened out methods with low accuracy or highly correlated errors and combined the remaining methods using a trimmed mean. We also investigated the accuracy risk (the risk of a bad forecast), proposing two new accuracy measures for this purpose. Our trimmed mean and the trimmed mean of all nine methods both had lower accuracy risk than either the best combination in the M4 competition or the simple mean of the nine methods.

1. Introduction

Combinations of forecasting methods have generally performed quite well in the M competitions, which should not be too surprising. The intuition is simply that adding a method can provide some new information, just as we talk about seeking a second opinion for a health problem or looking at multiple reviews of a product on Amazon. The information being combined can come from various sources, including methods and human experts. In the case of the M competition, we are combining forecasts from multiple forecasting methods.

Based on extensive past work, it is generally accepted that using combinations of forecasting methods instead of a single forecasting method can lead to improvements in the forecast accuracy. However, this does not mean that all combinations perform well. There can be many reasons why some combinations perform better than others. Two important issues with respect to the forecasting methods being combined are accuracy and diversity. Including methods with extremely poor accuracy can hurt the performance of the combination. In addition, some combinations may perform poorly because the forecasting methods being combined are not sufficiently diverse. Including a poorer method that has forecasting errors which are highly correlated with those of a better method is redundant and can degrade the performance of the combination. On the other hand, a poorer method that has forecasting errors which are negatively correlated with those of a better method can improve a combination. We are seeking good forecasters and good forecasting methods, and these forecasters and methods have common training or commonalities, so it is natural to have positive dependence among their forecasting errors. However, when the performance of the individual forecasts varies and the dependence among their errors gets too high, including weaker forecasts in the combination can cause the performance of the combination to deteriorate. Using methods that involve fundamentally different approaches (e.g., statistical methods and ML methods, or statistical methods and judgmental forecasts) might help to reduce the dependence.

Another important issue is accuracy risk. Sophisticated forecasting methods and combination methods can provide very accurate forecasts, but they could also have a higher risk of bad forecasts, in which sense they are more risky than simple approaches. Simple combination rules

* Corresponding author at: Darden School of Business, University of Virginia, Charlottesville, VA 22903, USA.
E-mail addresses: lichtendahlc@darden.virginia.edu, lichtendahlc@google.com (K.C. Lichtendahl), rwinkler@duke.edu (R.L. Winkler).

https://doi.org/10.1016/j.ijforecast.2019.03.027
0169-2070/

have been shown to do quite well in terms of accuracy, often doing better than more complex methods, and the simple rules are quite robust in the sense of being less likely to result in bad forecasts (i.e., they have a lower accuracy risk). The simple rule that is used most commonly is the simple average of the forecasts, but other rules such as trimmed means of the forecasts (with the median as a special case) might do even better because they are more robust in the sense of being less sensitive to extreme forecasts. They also are much easier and less computationally burdensome, and can be implemented quickly.

Section 2 will elaborate briefly on the issues mentioned above. Then, the remainder of the paper will focus on the M4 competition's top combination method, which was submitted by Montero-Manso et al. Their combination method was a weighted average of nine methods: eight common statistical time series methods and one ML method. Guided by existing theory, we propose a simple combination of a subset of their nine methods that performs nearly as well as their combination. Considering the accuracy and diversity of the nine methods, we choose a subset of them and then combine the methods in the subset using a trimmed mean. Our principal goal here is to provide an extended example that illustrates the principles of accuracy, diversity, and accuracy risk. At the same time, this example demonstrates that a simple combination method can perform reasonably well. Section 3 presents results on the accuracy of these methods. Then, Section 4 shifts to results on the accuracy risk, proposing two new accuracy measures that make scores more comparable across time series and forecast horizons. Section 5 provides a brief conclusion.

2. Accuracy, diversity, risk, and robustness

Of obvious importance in forecasting is accuracy. Including in a combination a method that is not very accurate on its own is not likely to improve the combined forecast. Thus, when there are a number of methods available, it makes sense to focus on the more accurate methods and exclude those that perform poorly. One is unlikely to want to get a second opinion about possible surgery from a surgeon who has a poor track record with the particular type of surgery in question. In the wisdom-of-crowds literature, Budescu and Chen (2015) and Mannes, Soll, and Larrick (2014) provide evidence that a "select-crowd strategy" of the five most accurate forecasters in the crowd works well for the future. In the forecasting literature, selecting a subset of forecasters is called pooling, and has been shown to improve the accuracy of a combination (Kourentzes, Barrow, & Petropoulos, 2019). More generally, there are decreasing returns to adding additional forecasts to a combination, even with accurate forecasts. Gaba, Tsetlin, and Winkler (2017) and Hora (2004) suggest using between five and 10 forecasts.

Also of importance is diversity. An increase in diversity among forecasting methods can improve the accuracy of their combination. Diversity among methods is usually measured in terms of correlations among their forecasting errors, with lower correlations indicating greater diversity. High correlations indicate that the forecasts are usually on the same side of the realization, whereas the gains in accuracy from combining are greatest when the forecasts are on opposite sides of the realization, so that they bracket the realization, in which case their average will often be closer to the truth than any individual forecasts (Larrick & Soll, 2006). Such bracketing is more likely to occur with low (or better yet, negative) correlations. Searching for methods with biases of different signs (as measured by the mean percentage error) would appear to be an alternative way of finding methods that bracket the truth. However, if the biased methods' errors are highly correlated, the methods will not tend to be on opposite sides of the truth and their combination may not lead to gains in accuracy.

Naturally, we like to rely on forecasters with expertise in forecasting methods, or, in the case of subjective forecasts, forecasters with substantive expertise. In either case, successful forecasters are likely to have had similar training and to be familiar with similar data and similar forecasting methods, leading to high levels of dependence among their forecasting errors. Low or negative correlations are the exception rather than the rule. When two methods or forecasters have high levels of dependence, including both in a combination creates redundancy. Thus, it makes sense to exclude one of them from the combination. If the redundancy is extreme, as it sometimes is, it can create multicollinearity problems with some combination methods, particularly those using weighted averages estimated by regression techniques (Bates & Granger, 1969; Winkler & Clemen, 1992).

Measures of accuracy such as *sMAPE* and *MASE* represent the average accuracy; however, the variance of the accuracy across series is also relevant. This variance provides an indication of the accuracy risk and is related to how likely a very poor forecast is for a given series. Using data from the first M-competition, Makridakis and Winkler (1983) found that the benefits of combining may be manifested more in terms of a reduced risk of poor forecasts than in terms of an improved accuracy. In choosing forecasting or combining methods for important forecasts, we should think about tradeoffs between the variance of accuracy and the average accuracy. This is analogous to balancing risks and returns in investing.

A final concern that involves the accuracy risk is especially relevant when we are dealing with time series forecasting, as is the case in the M competitions. This relates to the fact that the characteristics of a time series can change over time. However, the pre-training, validation, and testing sets cannot overlap in time; each set ends before the next begins. This sequencing means that a pattern – such as a trend or a seasonal component – that a method detects in the pre-training set may not hold up well in the validation and testing sets. Moreover, the further that one is forecasting into the future, the more likely it is that a pattern will evolve. When we have such an instability in the pattern, a method needs to be robust to pattern evolution.

Simple, robust combining methods have been used frequently in practice. They perform very well on average, and also in terms of reducing the risk because of their

robustness (Winkler, Grushka-Cockayne, Lichtendahl Jr., & Jose, 2019). The most commonly-used such method is a simple average of the forecasts being combined. Recently, more attention has been given to other options, such as trimmed means (Grushka-Cockayne, Jose, & Lichtendahl Jr., 2017; Jose, Grushka-Cockayne, & Lichtendahl Jr., 2014), of which the median is a special case. More sophisticated methods, or even slightly less simple methods such as weighted means, can perform better than simple methods, but they can also produce forecasts that perform poorly (as has been the case in some of the M competitions), perhaps because the pattern in the time series is unstable and/or there is overfitting when training the method.

3. Results: accuracy

This section describes how to form a combination based on the principles of expertise, diversity, and robustness. In a nutshell, our combination method is a lightly trimmed mean of forecasts from those of Montero-Manso et al.'s nine methods that survive two screens: one screen for individual accuracy (or expertise) and another for diversity. We describe our process in more detail below. We begin with some definitions of measures that were used in the M4 competition for evaluating the forecasting accuracy.

The *OWA* measure – the measure used in the competition to determine the overall point forecasting winner – is based on the symmetric mean absolute percentage error (*sMAPE*) and the mean absolute scaled error (*MASE*). For forecasts of time series i at various forecast horizons, the *sMAPE* is given by

$$sMAPE\left(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i\right) = \frac{1}{H_i}\sum_{h=1}^{H_i}\frac{2\left|y_{i,h}-\hat{y}_{i,h}\right|}{\left|y_{i,h}\right|+\left|\hat{y}_{i,h}\right|},$$

where $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,H_i})$ is the vector of time series i's actual values in the testing set and $\hat{\boldsymbol{y}}_i = (\hat{y}_{i,1}, \ldots, \hat{y}_{i,H_i})$ is the vector of forecasts from one to H_i steps ahead. The *MASE* is given by

$$MASE\left(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_i\right) = \frac{1}{H_i}\sum_{h=1}^{H_i}\frac{2\left|y_{i,h}-\hat{y}_{i,h}\right|}{\frac{1}{n_i-m_i}\sum_{j=m_i+1}^{n_i}\left|x_{i,j}-x_{i,j-m_i}\right|},$$

where m_i is time series i's frequency (e.g., 12 for monthly data) and $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,n_i})$ is the training set. Then, the *OWA* measure is given by

$$OWA = \frac{1}{2}\frac{\sum_{i=1}^{T} sMAPE(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)}{\sum_{i=1}^{T} sMAPE(\boldsymbol{y}_i, \hat{\boldsymbol{z}}_i)} + \frac{1}{2}\frac{\sum_{i=1}^{T} MASE(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)}{\sum_{i=1}^{T} MASE(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{z}}_i)},$$

where $\hat{\boldsymbol{z}}_i$ is the vector of Naïve2's forecasts for time series i from one to H_i steps ahead, $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,n_i})$ is time series i's actual values in the training set, and T is the number of time series in the competition (or series type).

We eliminate relatively inaccurate methods by calculating each method's accuracy on Montero-Manso et al.'s validation set using these measures. We do this by using a 5% trimmed mean of the *sMAPE* and *MASE* values over the time series and calculating a trimmed version of the overall weighted average (*OWA*). Before applying this trimmed mean, we remove time series for which the denominator of *MASE* in the validation set is zero. In the M4 competition, there are four series with no changes in the smaller training set (i.e., the full training set minus Montero-Manso et al.'s validation set): yearly series nos. 12,146 and 21,168, quarterly series no. 5619, and daily series no. 2085. Having no changes in the smaller training set means that the denominator of *MASE* is zero. Later in the paper, we propose an alternative measure that is designed to address this issue.

Table 1
Trimmed *OWA* in the validation set.

	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Weekly (359)	Daily (4227)	Hourly (414)	Average (100k)
ARIMA	0.705	0.928	0.952	0.914	1.025	0.614	0.855
ETS	0.729	0.900	0.941	0.857	1.005	0.831	0.853
NNETAR	0.902	1.159	1.117	1.143	1.139	0.608	1.050
TBATS	0.710	0.915	0.923	0.868	1.009	0.637	0.841
STLM	2.007	1.475	1.256	8.091	9.294	0.719	1.643
RWw/D	0.723	1.002	1.174	0.981	1.015	3.388	0.970
Theta	0.835	0.923	0.979	0.891	1.005	1.009	0.914
Naïve	1.000	1.065	1.095	1.000	1.000	3.424	1.059
SNaive	1.000	1.169	1.205	1.000	1.000	0.671	1.110
Naive2	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Next, we calculate the *OWA* by time series frequency (yearly, monthly, quarterly, weekly, daily, and hourly). We remove methods that have *OWA* scores above 1; that is, methods that do not perform at least as well as the benchmark. In measuring accuracy on the validation set, we use a trimmed mean in order to remove the effect of outliers.

We eliminate methods that lack diversity by looking at the error correlations for each pair of methods that survive the individual accuracy screen. For each pair of methods that has an error correlation coefficient of over 0.95, we remove the less accurate method. The lightly trimmed mean that we apply in the end for combining these methods is as light as possible, removing only the lowest and highest surviving forecasts.

Table 1 presents the trimmed *OWA* scores for Montero-Manso et al.'s nine methods in their validation set for each series type. For example, consider the yearly series, for which all but the STLM method do at least as well as the Naive2 benchmark and are candidates for inclusion in our combination method.

Table 2 shows the pairwise error correlations in the methods' percentage errors for the yearly time series at their longest horizon (six steps ahead) at the end of the validation set. We consider percentage errors in order to reduce the effect of the time series' different scales. Of the eight methods that survived the individual accuracy screen for the yearly series, we next checked each pair that had an error correlation coefficient greater than 0.95 and removed the less accurate method. Our choice of 0.95 is merely a starting point. We chose a high correlation for illustrative purposes, so as to avoid removing too many more models. Although we did not tune this parameter (or any of the other parameters in this example), one certainly could in practice.

In Table 2, several pairs of methods have error correlations above 0.95. Of the four methods with highly

Table 2
Correlation coefficients between percentage errors at six steps ahead for yearly series.

	ARIMA	ETS	NNETAR	TBATS	RWw/D	Theta	Naïve	SNaive
ARIMA	1.000	0.631	0.596	0.719	0.808	0.818	0.810	0.810
ETS	0.631	1.000	0.488	0.693	0.814	0.829	0.847	0.847
NNETAR	0.596	0.488	1.000	0.543	0.654	0.631	0.641	0.641
TBATS	0.719	0.693	0.543	1.000	0.824	0.827	0.795	0.795
RWw/D	0.808	0.814	0.654	0.824	1.000	0.967	0.949	0.949
Theta	0.818	0.829	0.631	0.827	0.967	1.000	0.969	0.969
Naïve	0.810	0.847	0.641	0.795	0.949	0.969	1.000	1.000
SNaive	0.810	0.847	0.641	0.795	0.949	0.969	1.000	1.000

correlated pairs (i.e., the last four methods), only the random walk with drift (RWw/D) survives. Thus, our combination for yearly time series is a 20% trimmed mean of forecasts from the ARIMA, ETS, NNETAR, TBATS, and RWw/D methods.

We follow the same steps for the other five series types: we first remove methods that do not have enough expertise, then remove those that do not add enough diversity, and eventually take a robust average of the surviving methods as our combination. The surviving methods for quarterly and monthly time series are ARIMA, ETS, TBATS, and Theta, so we take 25% trimmed means of their forecasts as our combinations for these two series types. The surviving methods for weekly time series are ARIMA, ETS, TBATS, RWw/D, Theta, and Naïve. The surviving method for daily time series is the Naïve method. The surviving methods for hourly time series are ARIMA, NNETAR, TBATS, STLM, and SNaïve.

Table 3 presents the results for Montero-Manso et al.'s combination, our combination, and the combinations from the fifth- and sixth-ranked teams in the M4 competition. As can be seen, our combination (hereafter, the trimmed mean of screened methods) would have ranked sixth in the competition overall, with a 5.6% improvement over the Comb benchmark, compared to Montero-Manso et al.'s 6.2% improvement. All of the code required to reproduce the results in this paper is available at the link in Appendix A.

One of the most attractive aspects of our combination is that it is easier to implement than Montero-Manso et al.'s combination. Their combination involves the use of a machine learning approach for selecting the weights to place on their nine methods. They begin by extracting 42 features from each time series in the pre-training set, then score the methods' forecasts on the validation set and identify the best method. They use the methods' errors in the validation set to define a customized objective function for a boosted trees method (via xgboost in R). This boosted trees method is fitted as a classifier using a customized objective involving the *OWA* scores. In the end, their method issues probability forecasts as predictions, and these probabilities become the weights in their combination.

4. Accuracy risk

In many contexts, a planner will be concerned about the risk of relying on a poor forecast. In forecasting competitions, it makes sense to rank competitors based on average forecasting errors, but in practice, the variability in those forecasting errors will sometimes matter just as much, if not more. We examine the variability in forecasting errors from the M4 competition by proposing two new accuracy measures. Both of the new measures are designed so that the errors between pairs of series are more comparable than is the case with *sMAPE* or *MASE*.

We call our first new measure the series-level *OWA* (*sOWA*). This measure will be familiar because it is related closely to the *OWA* measure used in the M4 competition. However, it suffers from the same issue that we raised earlier concerning a time series with no changes in its training set. Thus, we propose a second new measure, called the relative average absolute error (RAAE), in order to address both this issue and an issue regarding comparability among different forecast horizons. The next two subsections introduce these two new measures, after which we examine the risks in these measures of various combination methods.

4.1. Series-level OWA

We propose the series-level *OWA* (*sOWA*), which we define as

$$sOWA_i = \frac{1}{2}\frac{sMAPE(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)}{sMAPE(\boldsymbol{y}_i, \hat{\boldsymbol{z}}_i)} + \frac{1}{2}\frac{MASE(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)}{MASE(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{z}}_i)}.$$

This score is for a single time series, so we label it with a subscript i for the ith time series. Using this score, one could have settled the competition using the mean series-level overall weighted average (*MsOWA*):

$$MsOWA = \frac{1}{T}\sum_{i=1}^{T} sOWA_i.$$

With $sOWA_i$, we normalize the *sMAPE* and *MASE* within a series, before we take a mean over all series in the competition. This normalization makes scores more comparable between the series types. For instance, we can see from the competition's leaderboard that yearly and quarterly series are not comparable on *sMAPE*. The yearly series have much larger *sMAPE* values than the quarterly series do (13.528 vs. 9.733 for the top-ranked combination).

sOWA has at least two drawbacks. The first is that, when the time series in the training set does not change, the denominator in *MASE* is zero, and therefore *MASE* is infinite. The second is that forecasts at longer time horizons, which typically have greater absolute errors, will inherently receive more emphasis in both the *sMAPE* and *MASE* within a series. We address these two concerns by proposing the following measure.

4.2. Relative average absolute error

We start our development of this measure by normalizing each step-ahead forecast's absolute error by the average of its absolute error and the benchmark's absolute error at that same horizon. For a single time series that is split into a training set $\boldsymbol{x}_i$ and a testing set $\boldsymbol{y}_i$, the

Table 3
OWA in the testing set.

	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Others (5k)	Average (100k)	Rank
Montero-Manso et al. combination	0.799	0.847	0.858	0.915	0.838	2
Fiorucci & Louzada combination	0.802	0.855	0.868	0.935	0.843	5
Trimmed mean of screened methods	0.812	0.854	0.861	0.913	0.844	
Petropoulos & Svetunkov combination	0.806	0.853	0.876	0.906	0.848	6

relative average absolute error (*RAAE*) of some method's h-step-ahead forecast is defined as

$$RAAE_{i,h} = \begin{cases} 1 & \text{for } y_{i,h} = \hat{y}_{i,h} = \hat{z}_{i,h} \\ \frac{2|y_{i,h}-\hat{y}_{i,h}|}{|y_{i,h}-\hat{y}_{i,h}|+|y_{i,h}-\hat{z}_{i,h}|} & \text{otherwise,} \end{cases}$$

where $\hat{y}_{i,h}$ is the method's h-step-ahead forecast of $y_{i,h}$ and $\hat{z}_{i,h}$ is the benchmark's h-step-ahead forecast of $y_{i,h}$. The benchmark here could be the Naïve2 method from the M4 competition or some other naïve forecasting method. Note that the measure is equal to one when the method's h-step-ahead forecast is as accurate as the benchmark's h-step-ahead forecast. Also, the measure is bounded below by zero and above by two. Importantly, this measure compares each step-ahead's forecast to the benchmark at that same forecast horizon, both within the testing set. We contrast this "apples to apples" comparison with the way in which the *MASE* compares, for example, a method's 6-step-ahead forecast in the testing set relative to the average of a benchmark's 1-step-ahead forecasts in the training set. *MASE*'s comparison is an "apples to oranges" comparison in two ways: it is based on both different forecast horizons (for $h > 1$) and different datasets.

For the competition, the mean relative average absolute error (*MRAAE*) is given by

$$MRAAE = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{H_i}\sum_{h=1}^{H_i} RAAE_{i,h}.$$

4.3. Risk in the series-level OWA

Next we look at the risk in the series-level *OWA*. To do so, we plot the empirical cumulative distribution function (cdf) for each of four combination methods: Montero-Manso et al.'s combination, the trimmed mean of screened methods (screened for expertise and diversity as described above), a 1/9-trimmed mean of all nine methods, and the simple mean of all nine methods. Fig. 1 depicts these four combinations' empirical cdfs, and Table 4 presents both the *OWA* measure used in the competition and the mean and standard deviation of *sOWA* for these four combinations.

As we can see, both trimmed-mean combinations' cdfs cross Montero-Manso et al.'s combination's cdf once from below. Also, the trimmed-mean combinations and Montero-Manso et al.'s combination have roughly the same *MsOWA*s (see Table 4). Thus, both trimmed-mean combinations are (approximately) *less dangerous* than Montero-Manso et al.'s combination. In other words, they are less spread out for the same mean.

Formally, the *less dangerous* ordering of two random variables with different cdfs comes from the literature

Table 4
OWA, mean (M) of *sOWA*, and standard deviation (*SD*) of *sOWA* for four combinations.

	OWA	*MsOWA*	*SDsOWA*
Montero-Manso et al. combination	0.838	0.933	0.571
Trimmed mean of screened methods	0.844	0.925	0.516
Trimmed mean of all nine methods	0.876	0.920	0.339
Simple mean of all nine methods	0.952	1.020	0.754

on stochastic orders. One random variable is said to be *less dangerous* than another if the first random variable's cdf crosses the second random variable's cdf once from below and both random variables have the same expected values (Müller & Stoyan, 2002, p. 23). The "less dangerous" ordering here implies that any risk-averse decision maker will prefer either trimmed-mean combination to Montero-Manso et al.'s combination, no matter how averse to risk in (negative) accuracy the decision maker may be (Müller & Stoyan, 2002, p. 267). While the "less dangerous" ordering is less well-known than the mean-preserving spread and second-order stochastic dominance, these types of stochastic orders are often used in economic settings to trade-off risk and expected return (Rothschild & Stiglitz, 1970).

We include the combination based on trimming alone (without screening) in Fig. 1 and Table 4 because it is even more robust than our combination based on screening/trimming. However, while it is more robust, it does not do as well in terms of the competition's *OWA*, obtaining a score of 0.876 (which would have tied for 12th). In terms of *sOWA*, though, Table 4 shows that Montero-Manso et al.'s combination scores worse than both trimmed means on *MsOWA*, as well as in terms of risk (measured by *SDsOWA*). We include the simple mean in Fig. 1 and Table 4 to demonstrate that some level of trimming is helpful.

4.4. Risk in relative average absolute error

Similar to the risk analysis for *sOWA*, the results for *RAAE* are given in Figs. 2 and 3 and Table 5. Figs. 2 and 3 show the histograms of *RAAE* for Montero-Manso et al.'s combination and the trimmed mean of screened methods, respectively. These histograms are the decumulative views of the cdfs presented in Fig. 1, and may be a more familiar way of comparing empirical risks. In the case of *RAAE*, the trimmed mean of screened methods is (approximately) *less dangerous* than Montero-Manso et al.'s combination.

Table 5 shows that the means of *RAAE* are similar for Montero-Manso et al.'s combination and the trimmed mean of screened methods, but that the trimmed mean

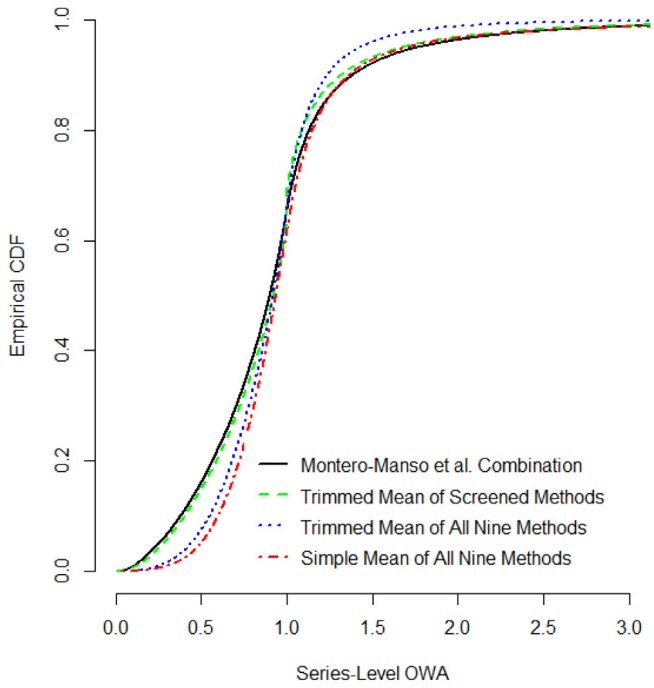

Fig. 1. Empirical CDFs of series-level ***OWA*** for four combinations.

Table 5
OWA, mean (*M*) of *RAAE*, and standard deviation (*SD*) of *RAAE* for four combinations.

	OWA	*MRAAE*	*SDRAAE*
Montero-Manso et al. combination	0.838	0.913	0.256
Trimmed mean of screened methods	0.844	0.914	0.242
Trimmed mean of all nine methods	0.876	0.938	0.185
Simple mean of all nine methods	0.952	0.973	0.197

of screened methods has a lower standard deviation. The other two combinations have larger *MRAAE*s but lower standard deviations than either Montero-Manso et al.'s combination or the trimmed mean of screened methods.

5. Conclusions

We discuss some important issues that a forecaster should consider when forming a combination. Two important concerns regarding the forecasting methods to be combined are their accuracy and their diversity. Adding an extra method does not always improve the accuracy when combining forecasts. There are many factors that can influence this, relating to both the characteristics of the forecasting methods being combined and the choice of a method for combining their forecasts. Regarding the combination methods, their robustness is an important consideration. Simple combinations such as the simple mean of the forecasts are often touted for their robustness. All of these factors have received considerable attention in the combining literature. A final issue that has not received as much attention in the forecasting literature is the accuracy risk, which relates to the variability of the accuracy of a forecasting method or combination across series. This provides information about the risk of a very poor forecast for an individual series.

In an attempt to illustrate the above points, we considered the M4 competition's top combination method, a weighted average of eight statistical time series methods and one ML method, submitted by Montero-Manso et al. Guided by existing theory involving the above issues, we showed that a simple combination of a subset of their nine methods performs nearly as well as their combination. We first used a simple screen on accuracy to remove any methods that were not as accurate as the Naive2 benchmark, then used a screen on diversity to eliminate pairs of methods with highly dependent forecast errors. Next, we used a lightly trimmed mean to combine forecasts from the methods that survived the screen. We followed this procedure separately for each series type (yearly, quarterly, etc.). The resulting set of combined forecasts – the trimmed mean of screened methods – would have finished 6th in the M4 competition, with an *OWA* of 0.844 compared with 0.838 for the Montero-Manso et al. combination.

We investigated the accuracy risk in more detail by proposing the use of a series-level *OWA* (*sOWA*), which involved computing the *OWA* for each series and then averaging over all series. We also explored a second accuracy measure, which we call the relative average absolute error (*RAAE*). This measure is designed to overcome two issues with *MASE*, which is a component of the *sOWA*. We then looked at the empirical distributions of the *sOWA* and *RAAE* for the Montero-Manso et al. combination and the trimmed mean of screened methods. We found that the trimmed mean of screened methods performed about the same as Montero-Manso et al.'s combination on the

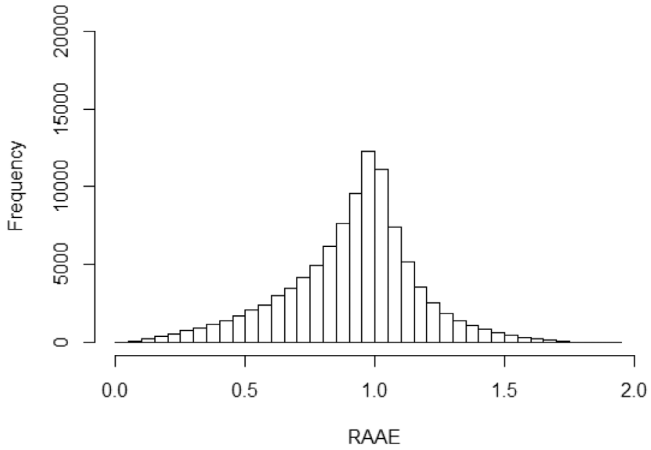

Fig. 2. Histogram of *RAAE* for Montero-Manso et al.'s combination.

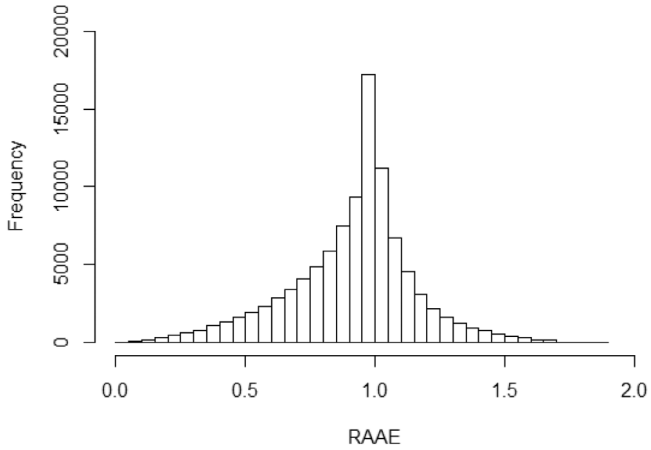

Fig. 3. Histogram of *RAAE* for the trimmed mean of screened methods.

mean *sOWA* and *RAAE*, and had smaller standard deviations of *sOWA* and *RAAE*. In terms of *sOWA* and *RAAE*, our trimmed mean of screened methods has a lower accuracy risk and is more robust.

It is important to point out that our illustrative method is quite ad hoc. Our combination method was designed primarily for descriptive purposes, in an attempt to answer the question posed in the title of this paper, and not for the sake of predictive accuracy. One could improve upon our method's accuracy by testing different cutoffs when screening for accuracy and diversity, and different choices of a trimmed mean or other simple combination method. One could also test the trimmed means of all nine methods with different degrees of trimming. We did not take the time to investigate other options. We leave for future work explorations of the fine tuning of methods along the lines that we have proposed in this discussion paper.

Undoubtedly, the proposed method is a greedy algorithm—a heuristic designed to find a good subset of models to combine. Because the method is a step-wise procedure, any model discarded early on because its accuracy is poor may be useful later because of the diversity that it brings to some subset. Similarly, discarding the worse of two models with highly positively correlated errors may reduce the diversity and degrade the combination's performance. Nonetheless, the point here was to show that simple, easy-to-use combination methods are robust and can perform quite well, and that considering issues such as accuracy, diversity, and accuracy risk can be helpful when thinking about combination methods.

Finally, we restricted our attention in this paper to point forecasts, due to time and space constraints, but many of the points that we have raised apply to prediction intervals and other probability forecasts as well. We are encouraged by the increased use of probability forecasts and the combination of these forecasts in important decision-making situations (e.g., the combination of the forecast paths from multiple meteorological methods for hurricane path prediction) and the increased communication of such probabilities to the general public in the media. Probability forecasts are much more informative than point forecasts because they provide important information about uncertainty. The inclusion of prediction intervals in the M4-competition is also very encouraging, and we hope to see this expanded in future competitions.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2019.03.027.

References

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, *20*(4), 451–468.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

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

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

Hora, S. C. (2004). Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science*, *50*(5), 597–604.

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

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

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*(9), 987–996.

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.

Müller, A., & Stoyan, D. (2002). *Comparison methods for stochastic models and risks*. Chichester, England: John Wiley & Sons, Ltd..

Rothschild, M., & Stiglitz, J. E. (1970). Increasing risk: I. A definition. *Journal of Economic Theory*, *2*(3), 225–243.

Winkler, R. L., & Clemen, R. T. (1992). Sensitivity of weights in combining forecasts. *Operations Research*, *40*(3), 609–614.

Winkler, R. L., Grushka-Cockayne, Y., Lichtendahl Jr., K. C., & Jose, V. R. R. (2019). Probability forecasts and their combination: A research perspective. In *Decision Analysis*, (in press).