Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

Combining prediction intervals in the M4 competition

Yael Grushka-Cockayne [a,b,*], Victor Richmond R. Jose [c]

[a] Harvard Business School, Harvard University, United States
[b] Darden School of Business, University of Virginia, United States
[c] McDonough School of Business, Georgetown University, United States

ARTICLE INFO

Keywords:
Hit rates
Interval combination methods
Forecasting
Calibration
Mean scaled interval score
Interior trimming
Overconfidence

ABSTRACT

The 2018 M4 Forecasting Competition was the first M Competition to elicit prediction intervals in addition to point estimates. We take a closer look at the twenty valid interval submissions by examining the calibration and accuracy of the prediction intervals and evaluating their performances over different time horizons. Overall, the submissions fail to estimate the uncertainty properly. Importantly, we investigate the benefits of interval combination using six recently-proposed heuristics that can be applied prior to learning about the realizations of the quantities. Our results suggest that interval aggregation offers improvements in terms of both calibration and accuracy. While averaging interval endpoints maintains its practical appeal as being simple to implement and performs quite well when data sets are large, the median and the interior trimmed average are found to be robust aggregators for the prediction interval submissions across all 100,000 time series.

1. Introduction

Decision makers can benefit from interval forecasts in many business settings, as they often provide more insights than point estimates. For instance, information beyond the average demand can improve the profitability when determining how many units of a perishable product to stock. In cases of project planning, worst-case duration scenarios might be important for establishing contingencies. According to Chatfield (1993), interval forecasts are useful for (1) assessing the future uncertainty, (2) enabling strategic planning for a range of possible outcomes, and (3) comparing forecasts from different methods thoroughly. Thus, the forecasting community has gradually recognized the importance of obtaining accurate interval forecasts (Gaba, Tsetlin, & Winkler, 2017; Hong et al., 2016; Yaniv, 1997). For this reason, it was exciting to see that, unlike previous M forecasting competitions, beginning with the initial M Competition in 1982, the 2018 M4 Competition collected, evaluated, and ranked prediction intervals (PIs) as well as point forecasts.

The M4 Forecasting Competition received a considerable amount of attention. A total of 50 valid point-estimate submissions were received for the 100,000 time series provided. However, only 20 submissions were received that specified PIs, even though 25% of the prize was to be awarded to the most accurate intervals. Overall, the submissions demonstrated that standard forecasting methods fail to estimate the uncertainty properly (Makridakis, Spiliotis, & Assimakopoulos, 2018). Only two of the 20 submissions of PIs were made solely by machine learning methods. We take a closer look at the PIs submitted by the various models, examining their calibrations and accuracies. Guided by the hypothesis set by the competition organizers, that “The 95% PIs will underestimate reality considerably, and this underestimation will increase as the forecasting horizon lengthens” (Makridakis et al., 2018, p. 805), we investigate the tendencies of PIs over different time horizons.

In an attempt to confirm Makridakis et al.’s (2018) Hypothesis 5, which speculates that “a combination of

[*] Corresponding author at: Harvard Business School, Harvard University, United States.
E-mail address: ygc@hbs.edu (Y. Grushka-Cockayne).

https://doi.org/10.1016/j.ijforecast.2019.04.015

Table 1
Average HR and MSIS results for the benchmark models and the individual submissions.

	Hit rate (HR)					Mean scaled interval score (MSIS)				
	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Others (5k)	Average (100k)	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Others (5k)	Average (100k)
Benchmarks										
ARIMA	0.724	0.866	0.911	0.933	0.885	45.01	11.06	8.75	29.46	18.68
ETS	0.839	0.932	0.934	0.952	0.925	34.90	9.45	8.30	28.02	15.68
Naïve	0.716	0.866	0.927	0.932	0.895	56.55	14.07	12.30	35.31	24.05
Submission										
118 (Hybrid)	0.947	0.954	0.945	0.949	0.948	23.90	8.55	7.20	24.45	12.23
245 (Combination)	0.936	0.966	0.966	0.978	0.960	27.48	9.38	8.66	32.12	14.33
238 (Combination)	0.913	0.921	0.896	0.912	0.907	30.20	9.85	9.49	26.32	15.18
069 (Combination)	0.786	0.894	0.922	0.945	0.885	35.84	9.42	8.03	26.71	15.69
036 (Combination)	0.779	0.886	0.915	0.937	0.878	35.94	9.89	8.23	27.78	15.98
132 (Combination)	0.783	0.901	0.929	0.946	0.889	37.71	9.95	8.23	29.86	16.50
082 (Statistical)	0.785	0.876	0.894	0.902	0.865	39.79	11.24	9.82	37.19	18.43
251 (Statistical)	0.739	0.864	0.900	0.939	0.856	45.03	11.28	9.39	52.60	20.20
218 (Statistical)	0.736	0.903	0.902	0.900	0.864	53.47	11.30	11.16	32.68	22.00
024 (Statistical)	0.657	0.848	0.893	0.904	0.829	54.00	13.54	10.34	34.68	22.37
030 (Statistical)	0.655	0.85	0.894	0.922	0.830	58.60	12.43	9.71	31.03	22.67
239 (Combination)	0.818	0.911	0.936	0.897	0.901	52.32	15.06	11.22	33.68	22.72
211 (Mach. learning)	0.943	0.932	0.931	0.929	0.934	39.94	17.28	20.16	39.45	24.98
244 (Mach. learning)	0.640	0.854	0.892	0.974	0.829	72.83	14.30	11.35	28.24	27.04
227 (Combination)	0.823	0.915	0.913	0.939	0.894	38.78	9.53	17.47	170.89	28.14
252 (Statistical)	0.721	0.868	0.915	0.483	0.837	47.44	11.25	9.72	1317.56	84.16
255 (Statistical)	0.719	0.854	0.918	0.485	0.835	47.75	12.19	9.63	1317.47	84.40
009 (Statistical)	0.720	0.878	0.920	0.485	0.842	47.52	11.00	15.03	1317.21	86.64
253 (Statistical)	0.619	0.816	0.905	0.484	0.797	70.85	14.26	10.01	1317.28	90.39
256 (Statistical)	0.718	0.867	0.919	0.484	0.839	47.85	11.58	18.04	1520.67	98.48

statistical and/or [machine learning] methods will produce more accurate results than the best of the individual methods combined", we study the benefits of forecast combination using several interval aggregation heuristics. Inspired by research on probability forecast aggregation (Clemen, 2008; Cooke, 1991; Genest & Zidek, 1986; Hora, 2004; Jose, Grushka-Cockayne, & Lichtendahl Jr, 2013; Lichtendahl Jr, G.rushka-Cockayne, & Winkler, 2013; Stone, 1961), we consider six interval aggregation heuristics (Gaba et al., 2017; Park & Budescu, 2015) that can be calculated a priori, i.e., before learning about the realizations. The heuristics also do not rely on any of the methods' past performances. The six heuristics considered are (1) the simple average, (2) the median, (3) an envelope approach, (4) Gaba et al.'s (2017) probability averaging of endpoints and a simple average of the midpoints, (5) an exterior trimming method, and (6) an interior trimming method. While these interval aggregation methods are not new to the literature, we believe that there is value in exploring the application of recently proposed methods to a new and high-profile dataset such as the M4 PIs. The aim of our analysis is to investigate how intervals should be aggregated in a time series context, when considering both statistical and machine learning methods, in a forecasting competition setting.

The results suggest that time series PI aggregation heuristics offer an easy-to-implement way of improving both the calibration and the accuracy. Similarly to human experts (Du & Budescu, 2007; Lichtenstein, Fischhoff, & Phillips, 1982; Moore & Healy, 2008; Soll & Klayman, 2004), most individual methods generated intervals that were too narrow. In contrast, aggregations of the individual methods tend to be too wide (Hora, 2004; Lichtendahl Jr et al., 2013). We report average hit rates (HR), which are the percentages of realizations that fall within the central prediction interval (Grushka-Cockayne, Jose, & Lichtendahl Jr, 2017), as well as the absolute coverage deviation (ACD), which was used by the competition organizers and is the absolute difference between the hit rate and the target coverage (95%). Exploring hit rates enables us to distinguish between forecasts that are over- and underconfident. Finally, while averaging the interval endpoints maintains its practical appeal as being simple to implement and performing quite well on large data sets, the median and the interior trimmed average turn out to be the most robust aggregators across all 100,000 time series. This might be due to the high degree of correlation amongst the different methods' submissions and the individual methods' overconfidence, as we describe next.

2. The prediction interval submissions

The M4 Competition elicited forecasts for 100,000 time series, from various domains and at different time horizons. A total of 1,277,717 upper and lower interval estimates were submitted. Makridakis et al. (2018, Table 2) report the results of the two evaluation measures: the mean scaled interval score (MSIS) and the ACD. The MSIS, a proper scoring rule, was the sole measure used in the competition for ranking interval submissions, and rewards both sharpness and calibration. For definitions, see M4 Team (2018).

Table 1 presents the performances of the twenty PI submissions with respect to the average hit rates and MSIS values of the various methods. Hit rates are presented rather than ACD in order to highlight when the coverage is too narrow or too wide, which is not noted directly by either the ACD or the MSIS. The hit rate is

Table 2
Average MSIS for benchmark models, average performance of the individual submissions, and the interval combinations.

	Yearly	Quarterly	Monthly	Others	Average
No. of obs.	23k	24k	48k	5k	100k
Horizon length	6	8	18	13–48	6–48
	Mean scaled interval score (MSIS)				
Benchmarks					
ARIMA	45.01	11.06	8.75	29.46	18.68
ETS	34.90	9.45	8.30	28.02	15.68
Naïve	56.55	14.07	12.30	35.31	24.05
Average performance of individual submissions	45.36	11.66	11.15	370.91	37.13
Combinations					
Average	27.99	8.82	9.22	277.76	26.87
Median	33.52	9.04	7.72	26.39	14.90
Envelope	48.37	19.25	24.50	87.16	31.86
PM	27.48	9.51	11.72	75.97	18.02
IT(0.2)	25.71	8.91	8.58	55.78	14.96
IT(0.4)	25.35	9.32	9.25	41.22	14.57
ET(0.2)	37.50	9.48	9.14	376.89	34.13
ET(0.4)	47.12	10.64	10.15	532.37	44.88

defined as the percentage of realizations that fall within the interval. A well-calibrated set of 95% prediction interval forecasts will see 95% of the realizations fall within the intervals. We also explore the idea of the end points being well-calibrated, i.e., the percentages of realizations that fall beyond those end points being consistent with the quantiles that they represent (the 2.5% and 97.5% quantiles). Identifying which forecasts are well-calibrated and which under- or overconfident will be useful in Section 3 when determining which combination methods to use.

Ten of the submissions were made by statistical methods, two were from machine learning methods, and the remaining eight were either combination or hybrid methods. The overall competition winner, Submission 118 (Smyl, Ranganathan, & Pasqua, 2019), was well-calibrated ($HR = 0.948$ and ACD of 0.002). The second place, Submission 245 (Montero-Manso, Athanasopoulos, Hyndman, & Talagala, 2020), was close but slightly underconfident, as is indicated by its hit rate of 0.960, which is 0.01 higher than the desired target of 0.95. This submission had the second-largest average interval width on average, and most methods were wider than this submission only about 23% of the time. However, having wide intervals on average does not guarantee a hit rate above the target: the submission with the widest average interval width (Submission 211) had a hit rate of 93.4%, which is still slightly below the target level of 95%.

The results show that the two machine learning methods are outperformed by all three benchmarks: the Naïve, ARIMA, and ETS methods. Across all twenty submissions, only one submission was consistently underconfident. Averaging the scores and hit rates for all submissions (excluding the benchmarks), we find that the overall average MSIS was 37.13 and the average hit rate for the prediction intervals was 87%, which is 8% lower than the target level (see also the rows labelled "Average performance of individual submissions" in Tables 2 and 3). The variation in the hit rates is low for the quarterly and monthly series, with a higher variation in the yearly series and the highest in the "Other" series, the smallest set in the group (5000 time series).

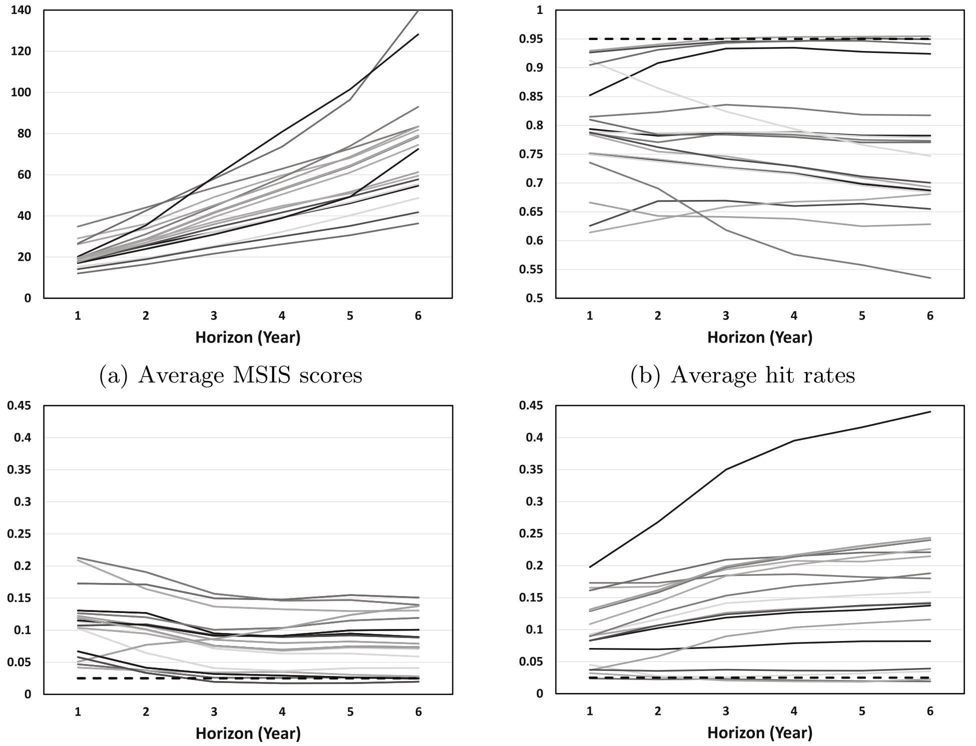
In all of these time series, there is a consistent sense of overconfidence in the PI submissions on average. This seems to be consistent with the findings of previous competitions (Makridakis, Hibon, Lusk, & Belhadjali, 1987), and was one of the predictions made by the competition organizers (Makridakis et al., 2018). This may also be partially a result of the benchmarks themselves being overconfident, thus providing some form of anchor for the submissions. However, we note that our ability to draw generalizable conclusions might be limited, since only twenty submissions of PIs were made.

Interestingly, several of the submissions included PIs with negative widths, suggesting that these methods did not include procedures for testing the feasibility or correctness of their entries, as the lower end of their interval submission was higher than the upper end. In addition, nearly half of the methods had at least one instance with an interval width of zero, i.e., the upper end of the interval was equal to the lower, suggesting some form of extreme overconfidence.

Fig. 1 plots the MSIS, HR, and percentage of the realizations below (above) the 2.5% (97.5%) quantiles, over years 1 to 6, for the 23,000 yearly time series. When considering the performances of the submissions over the forecasting time horizon, Fig. 1(a) demonstrates that all submissions' MSIS values increase over the horizon, which points to a reduced accuracy when forecasting farther into the future. Fig. 1(b) confirms the hypothesis that "The 95% PIs will underestimate reality considerably, and this underestimation will increase as the forecasting horizon lengthens" (Makridakis et al., 2018, p. 805). Many of the submissions have declining HRs over the time horizon, implying that their hit rates are moving away from 95% (the dashed line towards the top of the figure) and their calibrations are worsening. However, there are a few submissions that show improvements in calibration from year to year.

Table 3
Average hit rates (HR) and ACD for benchmark models, the average performance of the individual submissions, and the interval combinations.

	Yearly		Quarterly		Monthly		Others		Average	
No. of obs.	23k		24k		48k		5k		100k	
Horizon length	6		8		18		13–48		6–48	
	HR	ACD	HR	ACD	HR	ACD	HR	ACD	HR	ACD
Benchmarks										
ARIMA	0.724	0.226	0.866	0.084	0.911	0.039	0.933	0.017	0.885	0.065
ETS	0.839	0.111	0.932	0.018	0.934	0.016	0.952	0.002	0.925	0.025
Naïve	0.716	0.234	0.866	0.084	0.927	0.023	0.932	0.018	0.895	0.055
Average performance of individual submissions	0.772	0.178	0.888	0.062	0.915	0.035	0.820	0.130	0.871	0.079
Combinations										
Average	0.896	0.054	0.953	0.003	0.968	0.018	0.617	0.333	0.930	0.020
Median	0.818	0.132	0.921	0.029	0.945	0.005	0.943	0.007	0.910	0.040
Envelope	0.993	0.043	0.996	0.046	0.997	0.047	0.997	0.047	0.996	0.046
PM	0.959	0.009	0.978	0.028	0.985	0.035	0.966	0.016	0.976	0.026
IT(0.2)	0.929	0.021	0.967	0.017	0.979	0.029	0.843	0.107	0.958	0.008
IT(0.4)	0.951	0.001	0.978	0.028	0.985	0.035	0.979	0.029	0.975	0.025
ET(0.2)	0.778	0.172	0.900	0.050	0.930	0.020	0.561	0.389	0.870	0.080
ET(0.4)	0.694	0.256	0.858	0.092	0.899	0.051	0.518	0.432	0.823	0.127



(a) Average MSIS scores

(b) Average hit rates

(c) Percentage of realizations below the 2.5% quantile

(d) Percentage of realizations above the 97.5% quantile

Fig. 1. Average MSIS values, hit rates, and realizations in the tails for all 20 submissions over years 1–6. The dashed lines represent perfect calibration.

The lack of calibration seems more pronounced and deteriorates more over time in the upper tail (Fig. 1(d)) than in the lower tail (Fig. 1(c)), which is consistent with the findings of Makridakis et al. (1987). On average, the percentage of realizations that fall below the 2.5% quantile is 8.8%, while the percentage that fall above the 97.5% quantile is 14.0%. The trends in the quarterly and monthly time series follow similar patterns to those presented in Fig. 1, but the overall degree of miscalibration is lower.

3. Interval combinations

Earlier M-Competitions have established that combining forecasts can improve the forecast accuracy

(Makridakis & Hibon, 2000). However, intervals differ from point forecasts in the sense that there is an associated probability related to the reported values. Thus, a decision maker needs to consider issues such as calibration and overconfidence when selecting the best combination method.

While the question of how best to aggregate forecasts in the form of intervals has not been explored in previous M Competitions, there has been extensive academic work on the combination of probabilistic forecasts (Clemen & Winkler, 2007; Genest & Zidek, 1986), upon which we will draw for this section's analysis. The literature on combining quantiles is also relevant, since the endpoints of a prediction interval can be interpreted as the specific quantiles of a predictive distribution (Gaba et al., 2017; Grushka-Cockayne et al., 2017; Lichtendahl Jr et al., 2013).

3.1. Simple combination heuristics

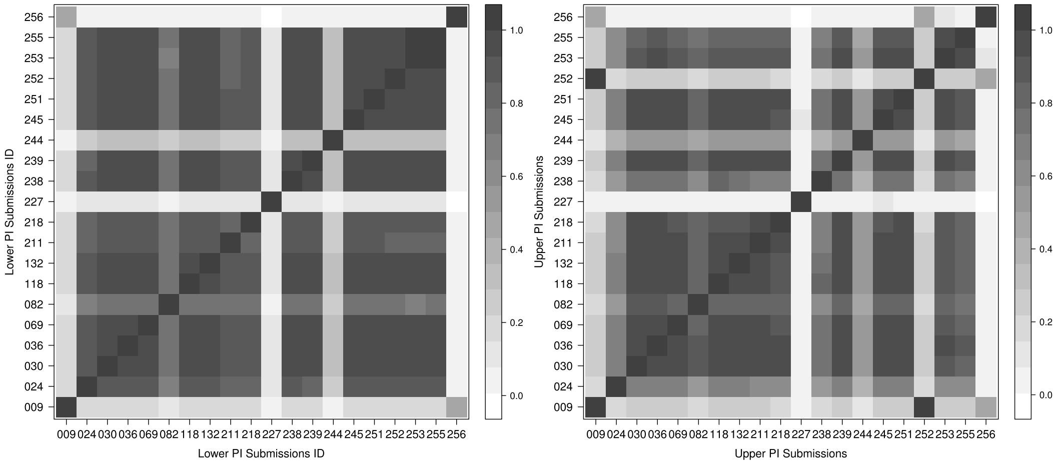
There are numerous ways of combining interval forecasts. Consider two sets of forecasts: the lower end points of the n reported prediction intervals $\mathbf{L} = \{L_1, \ldots, L_n\}$ and their corresponding upper end points $\mathbf{U} = \{U_1, \ldots, U_n\}$. We utilize the following heuristics, as per Gaba et al. (2017) and Park and Budescu (2015):

1. Simple average. $\bar{L} = (1/n)\sum_{i=1}^{n} L_i$ and $\bar{U} = (1/n)\sum_{i=1}^{n} U_i$. A simple way of combining intervals is to average the elements of $\mathbf{L}$ and the elements of $\mathbf{U}$. Averaging endpoints is akin to averaging quantiles, which has been shown to perform well (Lichtendahl Jr et al., 2013).
2. Median (Md). $L_{Md} = \text{Median}(\mathbf{L})$ and $U_{Md} = \text{Median}(\mathbf{U})$. The median approach provides a simple way of addressing possible outliers that can easily skew the average (Hora, Fransen, Hawkins, & Susel, 2013). For the M4 Competition, the fact that only 20 forecasts were submitted means that a single outlier can affect the estimates significantly.
3. Envelope (En). $L_{En} = \text{Min}(\mathbf{L})$ and $U_{En} = \text{Max}(\mathbf{U})$. The envelope approach is designed to provide an easy way of addressing the issue of overconfidence by providing the widest possible interval among any linear combination of $\mathbf{L}$ and $\mathbf{U}$.
4. Interior trimming (IT). $L_{IT(\beta)} = [1/(n-k)]\sum_{i=1}^{n-k} L_{[i]}$ and $U_{IT(\beta)} = [1/(n-k)]\sum_{i=k+1}^{n} U_{[i]}$, where β is the percentage of forecasts trimmed, $k = \lfloor \beta n \rfloor$, and $L_{[i]}$ (or $U_{[i]}$) is the ith order statistic of $\mathbf{L}$ (or $\mathbf{U}$). Interior trimming asymmetrically eliminates points on one side of $\mathbf{L}$ and $\mathbf{U}$ to generate an aggregate interval that is wider than the simple average of endpoints. The main motivation for this is to reduce overconfidence in the interval $(\bar{L}, \bar{U})$ (Jose et al., 2013; Yaniv, 1997). Since we have only 20 submissions, we do not use levels that are too refined. We examine $\beta = 0.2$, which is what was used by Gaba et al. (2017) for $n = 20$. We also show a more extreme case with $\beta = 0.4$ for comparison. In the limit, $\beta \to 1$ will yield $L_{IT(\beta)} \to L_{En}$ and $U_{IT(\beta)} \to U_{En}$.
5. Exterior trimming (ET). $L_{ET(\beta)} = [1/(n-k)]\sum_{i=k+1}^{n} L_{[i]}$ and $U_{ET(\beta)} = [1/(n-k)]\sum_{i=1}^{n-k} U_{[i]}$, where β is the percentage of forecasts trimmed, $k = \lfloor \beta n \rfloor$, and $L_{[i]}$ (or $U_{[i]}$) is the ith order statistic of $\mathbf{L}$ (or $\mathbf{U}$). Similar to interior trimming, exterior trimming uses the notion of asymmetric trimming. However, the difference is that the combined interval estimate will be narrower than the simple average of endpoints because the points taken out in the estimation are on the outside, i.e., the lower of lower interval submissions and the higher of upper interval submissions. ET addresses the possibility of underconfidence in the simple average of forecasts. We illustrate this heuristic using $\beta = 0.20$ and 0.40, similarly to the IT approach. As $\beta \to 1$, one should check that $L_{ET(\beta)} \leq U_{ET(\beta)}$, to ensure that the resulting combination is a valid interval forecast.
6. Probability averaging of endpoints and simple averaging of midpoints (PM). Gaba et al. (2017) provide this heuristic by assuming that the endpoints of the prediction intervals come from a normal distribution. Under normality, the midpoint represents a reasonable estimate of the center of any symmetric prediction interval. We implement this heuristic by computing the center (C) of the aggregated interval as the average midpoint of the reported intervals, i.e., $C = \sum_{i=1}^{n}(L_i + U_i)/2n = (\bar{L} + \bar{U})/2$. Then, the endpoints L_{PM} and U_{PM} are chosen such that $(1/n)\sum_{i=1}^{n} F_i(L_{PM}) = \alpha/2$ and $(1/n)\sum_{i=1}^{n} F_i(U_{PM}) = 1 - \alpha/2$, where F_i is a normal cdf whose $\alpha/2$-quantile is L_i and whose $(1 - \alpha/2)$-quantile is U_i.

The methods considered are limited to those that can be applied when no information is available regarding the historical performances of the individual forecasts. Thus, these methods are not dynamic and do not involve learning. When available, past performance information enables the use of unequal weighting combinations; however, we note that simple combination methods have been shown to be robust in many settings (Clemen, 1989).

The past literature suggests that these heuristics can be useful in different scenarios for dealing with under-/overconfidence and correlations among the individual submissions. In the context of averaging probability distributions, the simple average probability will be too wide when individual submissions are well-calibrated (Hora, 2004), and exterior trimming methods (Jose et al., 2013) and median-based approaches (Hora et al., 2013) have been shown to be useful in such cases. When the individual submissions exhibit high levels of overconfidence, the envelope and interior trimming methods are useful. Though there is a tight connection between averaging probability distributions and averaging quantiles/intervals (Lichtendahl Jr et al., 2013), the extension of these results to quantile combination has only been explored empirically. The advantages of the various methods in the context of averaging quantiles have been explored empirically by Gaba et al. (2017), who recommend the simple average, with PM, Md, and IT as "worthy competitors".

As was seen in Section 2, the individual methods' PIs in the M4 Competition tended to be overconfident



(a) Pairwise correlations among lower intervals

(b) Pairwise correlations among upper intervals

Fig. 2. Pairwise correlations between reported interval endpoints.

(too narrow). We attempt to hypothesize which aggregation heuristic will work best by examining the diversity amongst the PIs, using correlation as a proxy for diversity. We assess this by considering the pairwise (Pearson's) correlations amongst the lower interval points and the upper interval points separately. Fig. 2 presents the pairwise correlations among all submissions for the lower (panel (a)) and upper (panel (b)) intervals. The x and y axes of Fig. 2 list the submission IDs. It is apparent from Fig. 2(a) that the lower end points of Submissions 9, 227, 244, and 256 have low correlations with the others' lower end points, given the light shaded columns/rows. The figure suggests that the correlations among the models' PIs are high overall (a darker color represents a higher correlation): the majority of the correlations are above 0.5, with the average pairwise correlations among the lower and upper intervals being equal to 0.65 and 0.59, respectively. Such moderate to high positive correlations hint that the diversity in this dataset is low, suggesting that the En and IT methods are likely to be effective in improving the calibration.

3.2. Performance of simple combinations

Table 2 summarizes the performances of the different combination heuristics. The table also includes the performances of the three benchmark models and the average performance (or average score) of the twenty individual submissions, which we denote by "Average performance of individual submissions". In the literature (e.g. Larrick & Soll, 2006), the performances of forecasting combination methods are often compared to that of a random forecaster, which is captured by the "average forecaster", who scores the same as the average of the individual submissions.

The results demonstrate that the simple average beats the "average forecaster" in terms of MSIS. Some combination heuristics perform slightly better than the simple average. Overall, the Md, IT, and PM heuristics perform better (with MSIS scores ranging from 14.9 to 18.02) than the simple average (with an MSIS score of 26.87), while ET and En perform worse (with MSIS scores ranging from 31.86 to 44.88). This pattern generally holds for the different periodic subseries, with the smallest subset of "others" (consisting of weekly, daily, and hourly time series) showing a stark difference, especially for methods that are not robust to outliers.

The main explanation for this behavior is that ET tends to generate interval estimates that are narrower than the simple average, while IT, En, and PM on average make the interval estimates wider than the simple average. In this case, the median does something similar to ET, but to a lesser extent, making the estimates slightly sharper (i.e., narrower intervals) but still with a reasonably good calibration. MSIS as a score reflects the tradeoff between the width of the interval and the likelihood of being in the interval. On average, a comparison of Tables 1 to 3 shows that the individual submissions tend to be too narrow and that heuristics which improve the calibration of the combined forecasts should perform better. Of the heuristics that widen the intervals relative to the simple average, we note that En and PM tend to make the intervals too wide compared to IT, as can be seen from their very high hit rates in Table 3.

In terms of calibration, we now examine the two metrics discussed earlier in more detail: hit rates and ACD. Perhaps not surprisingly, the combination methods that yielded the best MSIS scores also produced better hit rates and ACDs. This suggests that the tendency of scoring rules to "reward sharpness subject to calibration" is in play (Gneiting & Raftery, 2007). Calibration as a sole measure for performance may be problematic, because it can be gamed. For example, one can achieve a perfect calibration almost surely by making the intervals extremely wide (e.g., setting the estimated interval to be the theoretical range for the time series) for 95% of the time series and extremely narrow (say, an interval width of zero) for 5% of the time series. Of the aggregation heuristics, we see that modest interior trimming appears to be a good compromise, achieving a good hit rate and an ACD that is comparable to those of the top two performing methods, but also an MSIS that is generally good and consistent for the various subsets of series with different time horizons.

We note that the top two methods (Submissions 118 and 245) beat all of the simple combination heuristics. Interestingly, these two winning methods are themselves ensembles, or combinations, combining the benchmarks or base models to reach a final prediction. The determination ex ante of which submission will perform best is difficult and unrealistic in a single period setting such as a one-shot competition. Thus, using a simple heuristic such as the median (Md) is robust. While it is not necessarily guaranteed to beat all individual submissions, it will provide improvements in MSIS and hit rates over randomly selecting a single submission ex ante (represented by the "average forecaster"). In addition, if we are able to predict that on average the estimates will be overconfident, as is the case in many competitions (Moore et al., 2016), a simple heuristic such as interior trimming with moderate asymmetric trimming is likely to help.

To summarize, we see that the median and interior trimming heuristics performed well in this competition. Surprisingly, the two approaches tackle calibration in quite different ways, with interior trimming making the intervals wider while the median makes the intervals narrower relative to the simple average. This highlights the challenge of balancing the tradeoff between sharpness (i.e., width of the interval) and calibration when maximizing the average MSIS. Though interior trimming provides flexibility by having a parameter that can be adjusted, the median provides a simpler and more robust way to gain accuracy.

4. Reflections

The introduction of prediction intervals to the M competitions is a welcome development in the advancement of probabilistic forecasting in both research and practice. Given the number of time series being forecast (100,000), over various time horizons and domains, the M4 PI dataset provides a rich new data set of over 1.2 million forecasts. Half of the methods used by the participants were statistical, only two were pure machine learning approaches and the rest were combination methods.

Many of the models submitted performed poorly (except, notably, for the top two or three submissions), in that they often performed worse than the benchmarks provided by the competition organizers. A possible explanation for this poor performance might be a mix of overconfidence and overfitting. The two best models each incorporate some combination step in the generation of their final reported interval estimate, with the best model incorporating some machine learning elements to weight these combinations. However, it should be noted that we may not have a sufficiently large sample to enable us to draw generalizable conclusions about the individual methods, since only twenty submissions of PIs were made.

The competition organizers hypothesized that the prediction intervals would be overconfident and would become increasingly so over longer horizons. We see that this indeed is the case, on average, though a few submissions did experience some improvements in their hit rates over time. Though ACD is useful for ranking submissions in terms of their calibrations compared to hit rates, the use of hit rates may provide a clearer picture for our understanding of overconfidence. Similarly, we see that MSIS scores deteriorate over the time horizon.

Combinations provide a good way of improving forecasts in terms of both accuracy and calibration. The average MSIS and calibration scores for the simple averages generally work well when the data sets are large. This suggests that the power of the simple average combination heuristic still holds in this domain. When data sets are small, the median may be a better alternative, but if we do have some indication of overconfidence and low diversity among the individual submissions, we recommend a modest amount of interior trimming.

Similarly to the earlier M Competitions, this new data set of prediction interval estimates will certainly form a data set that researchers will use for benchmarking future advances in probabilistic forecasting and aggregation. We look forward to the novel ways that these prediction interval submissions to the M4 Competition will be used in the future by researchers and practicing forecasters.

References

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, *11*(2), 121–135.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety*, *93*(5), 760–765.

Clemen, R. T., & Winkler, R. L. (2007). Aggregating probability distributions. In W. Edwards, R. F. Miles, & D. von Winterfeldt (Eds.), *Advances in decision analysis: from foundations to applications* (pp. 154–176). Cambridge University Press.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

Du, N., & Budescu, D. V. (2007). Does past volatility affect investors' price forecasts and confidence judgements?. *International Journal of Forecasting*, *23*(3), 497–511.

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

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

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

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

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913.

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

Hora, S. C., Fransen, B. R., Hawkins, N., & Susel, I. (2013). Median aggregation of distribution functions. *Decision Analysis*, *10*(4), 279–291.

Jose, V. R. R., Grushka-Cockayne, Y., & Lichtendahl Jr, K. C. (2013). Trimmed opinion pools and the crowd's calibration problem. *Management Science*, *60*(2), 463–475.

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

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

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 306–334). Cambridge University Press.

M4 Team, (2018). M4 competitor's guide: Prizes and rules, URL https://www.m4.unic.ac.cy/wp-content/uploads/2018/03/M4-Competitors-Guide.pdf.

Makridakis, S., & Hibon, M. (2000). The m3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Makridakis, S., Hibon, M., Lusk, E., & Belhadjali, M. (1987). Confidence intervals: An empirical investigation of the series in the M-competition. *International Journal of Forecasting*, *3*(3–4), 489–508.

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

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.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., et al. (2016). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552–3565.

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

Smyl, S., Ranganathan, J., & Pasqua, A. (2019). M4 forecasting competition: Introducing a new hybrid ES-RNN model. *International Journal of Forecasting, this issue*.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(2), 299.

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

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, *69*(3), 237–249.

Yael Grushka-Cockayne is a visiting Associate Professor of Business Administration at Harvard Business School and an associate Professor of Business Administration, Darden School of Business. Her research and teaching activities focus on decision analysis, forecasting and estimation, project management and data science.

Victor Richmond R. Jose is an associate professor and the William and Karen Sonneborn Term Chair in the operations and information management area of the Robert Emmett McDonough School of Business at Georgetown University. His main research interests lie in decision analysis and the use of statistical methods in management science, operations research, and risk analysis. His recent works have been in the areas of data science and machine learning.