Decision Analysis

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Combining Interval Forecasts

Anil Gaba, Ilia Tsetlin, Robert L. Winkler

To cite this article:
Anil Gaba, Ilia Tsetlin, Robert L. Winkler (2017) Combining Interval Forecasts. Decision Analysis 14(1):1-20. https://doi.org/10.1287/deca.2016.0340

http://pubsonline.informs.org/journal/deca/

DECISION ANALYSIS
Vol. 14, No. 1, March 2017, pp. 1–20
ISSN 1545-8490 (print), ISSN 1545-8504 (online)

Combining Interval Forecasts

Anil Gaba,[a] Ilia Tsetlin,[a] Robert L. Winkler[b]

[a] INSEAD, Singapore 138676; [b] Fuqua School of Business, Duke University, Durham, North Carolina 27708
Contact: anil.gaba@insead.edu (AG); ilia.tsetlin@insead.edu (IT); rwinkler@duke.edu (RLW)

Received: July 5, 2016
Revised: September 28, 2016
Accepted: October 29, 2016
Published Online in Articles in Advance: February 21, 2017

https://doi.org/10.1287/deca.2016.0340

Abstract. When combining forecasts, a simple average of the forecasts performs well, often better than more sophisticated methods. In a prescriptive spirit, we consider some other parsimonious, easy-to-use heuristics for combining interval forecasts and compare their performance with the benchmark provided by the simple average, using simulations from a model we develop and data sets with forecasts made by professionals in their domain of expertise. We find that the empirical results closely match the results from our model, thus providing some validation for the theoretical model. The relative performance of the heuristics is influenced by the degree of overconfidence in and dependence among the individual forecasts, and different heuristics come out on top under different circumstances. The results provide some good, easy-to-use alternatives to the simple average with an indication of the conditions under which each might be preferable, enabling us to conclude with some prescriptive advice.

Keywords: interval forecasts • combining forecasts • heuristics • overconfidence • dependent forecasts

1. Introduction

Decisions are usually made in the face of uncertainty, and decision makers need some assessment of that uncertainty. Ideally, the assessments would involve a well-structured, facilitated elicitation process yielding full probability distributions of uncertain quantities, but in practice they are often given quickly without assistance and expressed as interval forecasts, which are less complicated than full distributions and are easily understood by both forecasters and decision makers. "Intervals are often communicated in the course of real life forecasting and decision making situations" (Yaniv 1997, p. 238): "For instance, project managers are encouraged to predict both the most likely effort of a new project (in work hours) and the 90% prediction interval" (Teigen and Jørgensen 2005, pp. 455–456; crediting Moder et al. 1995).

The forecast reported by the analyst could simply be subjective. Alternatively, the analyst might generate a model-based forecast, or obtain a forecast from a website. Moreover, the decision maker might obtain interval forecasts for an uncertain quantity from more than one analyst to obtain more information. How should such forecasts be evaluated and aggregated? Our concern is not with how the forecasts are obtained (e.g., how they are elicited in the case of subjective forecasts), but with evaluating and combining them once we have the forecasts.

There is an extensive literature on combining probability distributions; for reviews, see Genest and Zidek (1986), Cooke (1991), Clemen and Winkler (2007), and Ranjan and Gneiting (2010). Clemen (1989) and Armstrong (2001) review the broader area of combining forecasts later popularized by Surowiecki (2004), who coined the phrase "the wisdom of crowds." A key message from the extensive past work is that simple averaging is an easy and robust way to combine forecasts and accrue significant benefits in terms of accuracy. With simple averaging of the endpoints as a benchmark, we identify some other parsimonious, easy-to-use heuristics that perform well when combining interval forecasts.

One issue we consider is overconfidence, the tendency of distributions to be too tight and interval forecasts to be miscalibrated because they are too narrow. For example, 90% forecast intervals tend to capture fewer than the expected 90% of the realizations, often only 40%–70%. Since early papers by Alpert and Raiffa (1969) and Lichtenstein et al. (1982), overconfidence in subjective forecasts has been studied extensively. Moreover, Grushka-Cockayne et al. (2016, p. 2) state, "Overfitting and overconfidence are present even when we

consider models instead of human experts." Overfitting can happen with models ranging from linear regression to data mining and machine-learning algorithms.

A second issue is dependence among the forecast errors, which greatly reduces gains in performance from increasing the number of forecasts being combined (Clemen and Winkler 1985). This dependence is often high, with pairwise correlations of forecast errors creating redundancy. Experts have similar training and read the same journals; analysts use common data and similar models.

We consider the simple average and five other parsimonious, easy-to-use heuristics for combining interval forecasts. Key aspects and contributions of our work include:

- One of our heuristics (*PM*, defined in Section 2) is new and others have not been studied or used in practice extensively.
- Multiple evaluation measures are used, with a proper scoring rule as the primary measure, unlike many studies that have focused exclusively or primarily on calibration.
- We develop a new theoretical model of the forecast aggregation process that takes into account overconfidence and correlation.
- Simulations from the model enable us to predict which heuristics might perform better under different conditions in practice.
- The performance of the heuristics on two recent data sets, both involving real-time forecasts from experts, is compared with results from the simulations.
- The simulations and the analysis of the data sets enable us to offer some prescriptive advice.

The heuristics and evaluation measures are defined in Section 2. In Section 3 the theoretical model is described and simulation results are discussed. In Section 4 the relative performance of the heuristics on the data sets is explored and found to closely match the results from the model-based simulations, providing some validation for the model. The paper is summarized and prescriptive advice is discussed in Section 5.

2. Heuristics and Evaluation Measures

There are many ways that interval forecasts can be combined. We consider some heuristics that are very easy to apply and seem reasonable based on the past research on combining of probability distributions. Those heuristics are presented in Section 2.1, followed in Section 2.2 by measures that we use to summarize and evaluate the individual and combined interval forecasts.

2.1. Heuristics for Combining Interval Forecasts

Suppose we have $100(1-\alpha)\%$ forecast intervals $[L_i, U_i], i = 1, \ldots, k$, provided by k forecasters for a random variable $\tilde{x}$. Let $[L_H, U_H]$ denote the $100(1-\alpha)\%$ combined forecast interval for $\tilde{x}$ obtained from the k individual intervals with heuristic H.

Average (*Av*). $L_{Av} = (1/k)\sum_{i=1}^{k} L_i$, $U_{Av} = (1/k)\sum_{i=1}^{k} U_i$. This heuristic takes a simple average of the endpoints of the intervals. In combining point forecasts or probability forecasts, as in summarizing data, simple averages are often used because of their simplicity, good performance, and robustness. If we assume that the intervals are symmetric in probability ($F_i(L_i) = 1 - F_i(U_i) = \alpha/2$, where F_i is forecaster i's cumulative distribution function (cdf) for $\tilde{x}$), then *Av* corresponds to averaging quantiles, which has been shown to perform well (Lichtendahl et al. 2013). With the scoring rule $S(L, U, x)$ defined in Section 2.2, *Av* will always score at least as well as the average of scores for the individual intervals being combined, because $S(L, U, x)$ is concave in L and U.

Median (*Md*). $L_{Md} = Median\{L_1, \ldots, L_k\}$, $U_{Md} = Median\{U_1, \ldots, U_k\}$. The median is another measure commonly used in summarizing data, and it is less sensitive to extreme values than the mean. Hora et al. (2013) study the combination of probability distributions via the median cdf and show that it has many desirable properties.

Envelop (*En*). $L_{En} = \min\{L_1, \ldots, L_k\}$, $U_{En} = \max\{U_1, \ldots, U_k\}$. This heuristic is consistent with the idea that each interval represents a window on the world that is only a partial view and that the aggregate view should envelop all of these views so that no forecasts, however extreme, are discarded or discounted. Unless the k individual intervals are identical, enveloping will yield a wider interval than the other heuristics. Thus, it can overcome overconfidence that might be exhibited by the individual forecasts.

Probability averaging of endpoints and simple averaging of midpoints (*PM*). *PM* combines aspects of probability averaging (*PA*) and *Av*. First, L_{PA} and U_{PA} satisfy $(1/k)\sum_{i=1}^{k} F_i(L_{PA}) = 1-(1/k)\sum_{i=1}^{k} F_i(U_{PA}) = \alpha/2$, where we assume the individual interval forecasts are based on normal cdfs. Thus, L_{PA} and U_{PA} are the values at which the averages of the cumulative probabilities $F_1,\ldots,F_k$ are $\alpha/2$ and $1-\alpha/2$. Then we shift the *PA* interval so it is centered at the midpoint of the *Av* interval. The width and midpoint of the *PM* interval are $W_{PM} = U_{PA} - L_{PA}$ and $M_{PM} = (L_{Av} + U_{Av})/2$. *PA* gives *PM* a wider interval in a more tempered manner than *En*, and the shifting gives *PM* the good location of the *Av* interval.

Exterior trimming (*TE*). L_{TE} is the simple average of the lower endpoints after the lowest d are deleted, or trimmed. Similarly, U_{TE} is the simple average of the upper endpoints after the highest d have been deleted. In our analyses, which are for $k = 1,\ldots,20$, we use $d = 0$ for $k \le 3$, $d = 1$ for $4 \le k \le 7$, $d = 2$ for $8 \le k \le 11$, $d = 3$ for $12 \le k \le 15$, and $d = 4$ for $k \ge 16$. Exterior trimming yields intervals that are narrower and less sensitive to extreme endpoints than those from *Av*, with $L_{TE} \ge L_{Av}$ and $U_{TE} \le U_{Av}$. Except for the smaller values of k, roughly 18%–25% of the values are trimmed on each side. We tried different levels of trimming before settling on these choices, but potential levels of trimming are limited given the values of k.

Interior trimming (*TI*). *TI* is like *TE* but the highest d lower endpoints and the lowest d upper endpoints are trimmed. Thus, $L_{TI} \le L_{Av}$ and $U_{Ti} \ge U_{Av}$. Like *PM* but contrasting with *TE*, *TI* increases the width of the combined intervals, which can be helpful with overconfident forecasts. Exterior and interior trimming have received some attention in combining forecasts (Yaniv 1997, Jose et al. 2014). Note that in our analyses, the kinks for *TE* and *TI* in the figures presented later to show evaluation measures as a function of k are due to the discontinuous nature of d, which increases with k as a step function.

We have tried to use simple heuristics involving few or no assumptions. In this sense, *Av*, *Md*, and *En* are best, requiring absolutely no assumptions. Heuristic *PM* uses a normal distribution assumption when considering the tail probabilities, but other distributions could easily be used (e.g., gamma or lognormal distributions to reflect skewness); and *TE* and *TI* necessitate the choice of how many forecasts to trim. None of our heuristics require knowledge of any past performance data from the forecasters.

What about weighted averages, which are used quite often? The idea of giving "better forecasters" higher weight is intuitively appealing, but it is not as simple as the heuristics considered here because it requires estimation of weights, whether subjectively, via models, or via data on past performance. Rowe (1992, p. 161) notes that "A considerable number of studies have examined the relative worth of various weighting schemes, and have generally found there to be little advantage (if any) in using differential over equal weighting."

An approach involving performance-weighted averages, Cooke's (1991, pp. 187–198) classical model, has many adherents. Cooke and Goossens (2008) compare it favorably with *Av*, although it is hard to evaluate the comparison because Cooke's measure of performance is very different from ours, involving hypothesis testing and giving more emphasis to calibration. In any event, it involves a detailed process using performance on seed variables, so it doesn't fit the goals of our study.

Our analyses involve randomly selected groups of forecasters. Recent studies (Mannes et al. 2014, Budescu and Chen 2015) found that groups chosen from larger sets of forecasters based on past performance outperformed randomly selected groups, with equal weighting as well as with performance-based weighting. This approach is promising and deserves further study. It requires data on past forecasts, ideally from a larger set of forecasters than the "sweet spot" of 5 to 10, so like Cooke's model, it goes beyond the scope of our study.

2.2. Measures for Summarizing and Evaluating Interval Forecasts

Before the value of $\tilde{x}$ is known, an interval forecast for $\tilde{x}$ can be summarized by statistics such as its midpoint and width. Summary measures of combined interval forecasts from the heuristics can be compared and related to the set of interval forecasts being combined.

Of course, the key question of interest for a forecast is how well it performs in view of the realized value x of $\tilde{x}$. For measuring this kind of performance, a value for just one or a few forecasts is not very helpful. Instead, we compute average values of the measures over a series of forecasts.

Average *S*-score ($\bar{S}$). An interval forecast can be evaluated by the S-score, a strictly proper quantile scoring rule (Jose and Winkler 2009), which is our primary measure of overall performance:

$$S(L,U,x) = -(\alpha/2)(U-L) - (L-x)^+ - (x-U)^+,$$

where $t^+ = \max\{t, 0\}$ and a higher (less negative) score is better. In our simulations and empirical studies, we work with 90% interval forecasts, which correspond to $\alpha = 0.10$.

The $-(\alpha/2)(U-L)$ term represents a penalty associated with the width of the interval, and the last two terms represent penalties imposed if x falls below or above the interval. The tradeoff is that a wider interval is given a higher penalty for width but avoids or reduces the penalty for not "capturing" the realized value in the interval. For a series of interval forecasts for different quantities, forecasters, or heuristics, $\bar{S}$ is a good measure of the accuracy of the interval forecasts, taking into account both sharpness and calibration. Good interval forecasts are sharp in the sense of having narrow intervals and well-calibrated in the sense of having a "capture rate" close to the $100(1-\alpha)\%$ indicated by the forecast.

Relative frequency (*RF*). Over a series of interval forecasts, RF is the relative frequency of times the interval captures the realized value. It measures how well calibrated the forecasts are. With 90% interval forecasts, the forecasts are perfectly calibrated if $RF = 0.90$, or 90%. We operationalize overconfidence in terms of intervals that are too narrow to capture the expected percentage of realized values; thus, $RF < 0.90$ indicates overconfidence, and $RF > 0.90$ indicates underconfidence.

Average interval width ($\bar{W}$). The average interval width $\bar{W}$ over a series of interval forecasts indicates the degree of uncertainty about a variable that is implied by the intervals and can help explain any overconfidence or underconfidence revealed by RF.

Mean absolute error (*MAE*). The absolute error of the midpoint of an interval is $AE = |M - x|$. The mean absolute error (MAE), a common measure of accuracy for point forecasts, is the average AE over a series of forecasts. It indicates how "well-located" the intervals are.

In addition to these primary summary measures, we consider some other measures as needed. For example, the average pairwise correlations of forecast errors among the forecasts being combined can impact $RF, \bar{W}, MAE$, and $\bar{S}$. Higher correlations are detrimental to the performance of the combined forecasts (Clemen and Winkler 1985) because of redundancy of information.

3. Relative Performance of the Heuristics Under a Multinormal Model

We develop a new multivariate normal model of the forecasting process, considering the possibility of overconfident and correlated intervals. A symmetric model that treats the forecasters as exchangeable is presented in this section, and we use simulation to explore the relative performance of our heuristics under different conditions. To investigate the robustness of the performance of the heuristics, a more general model with heterogeneous forecasters is presented in Appendix A, and a lognormal model with skewed distributions in Appendix B.

Let the random variable of interest to a decision maker be $\tilde{x}$, e.g., a future observation of the price of oil, gross domestic product (GDP) growth, or inflation. The decision maker consults k forecasters, each of whom provides a $100(1-\alpha)\%$ interval for $\tilde{x}$. We model the data-generating process for the k intervals as follows. Suppose that forecaster $i, i = 1, \ldots, k$, has a uniform diffuse uniform prior distribution for $\tilde{x}$ and observes a signal y_i with mean x and variance σ_i^2 from a normal process. The signal could be, for instance, a summary statistic such as a sample mean from a random sample. After seeing the signal, forecaster i's distribution for $\tilde{x}$ should be normal with mean y_i and standard deviation σ_i given the uniform prior.

Our model assumes that after seeing y_i, forecaster i reports a $100(1-\alpha)\%$ interval for $\tilde{x}$ that is of the form $y_i \pm z_{\alpha/2}(1-\gamma_i)\sigma_i, \gamma_i < 1$, where γ_i is an overconfidence parameter and $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. If $\gamma_i > 0$, forecaster i's interval is overconfident, and a higher γ_i corresponds to more overconfidence. Overconfidence implies that a forecaster's distribution for $\tilde{x}$ is tighter than it should be given the signal.

To investigate the aggregation of the intervals from the k forecasters, we assume that the decision maker's prior distribution for $\tilde{x}$ is diffuse, which implies that any information the decision maker has regarding $\tilde{x}$ is

overwhelmed by the information from the forecasters. (If it is not, that can be accounted for by treating the decision maker as a $(k+1)$st forecaster.) We assume a normal model for the distribution of $(y_1, \ldots, y_k)$ given x with mean vector $x\mathbf{e}$, where $\mathbf{e} = (1, \ldots, 1)$, and positive-definite covariance matrix $\mathbf{\Sigma}$ with variances σ_i^2 and correlations $\rho_{ij}, i, j = 1, \ldots, k, i \neq j$. Because the distribution is conditional on x, the correlations are pairwise correlations of forecast errors, not of the forecasts themselves.

Our symmetric normal model has common parameters $\sigma_i = \sigma$, $\gamma_i = \gamma$, and $\rho_{ij} \geq 0$ for $i, j = 1, \ldots, k$, $i \neq j$, implying that the forecasters can be viewed as exchangeable. With a diffuse prior on $\tilde{x}$ and known σ and ρ, the posterior distribution for $\tilde{x}$ given the midpoints $y_1, \ldots, y_k$ of the k intervals is normal with mean $\bar{y} = (y_1 + \cdots + y_k)/k$ and variance $\sigma^2(\rho + (1-\rho)/k) > 0$.

With this simple model, we can derive some analytical results for *Av*, but it is more complicated to do so for the other heuristics. Hence, we explore comparisons of the heuristics with simulation. We let $\sigma = 1$ and $x = 0$ without loss of generality and simulate observations from the model for $k = 1, \ldots, 20$; $\gamma = 0, 0.25, 0.5, 0.75$; and $\rho = 0, 0.25, 0.75$. In each instance, we simulate 10,000 groups of size k, generate individual 90% intervals ($\alpha = 0.10$) according to the model, and combine the intervals using our heuristics. Statistical significance is not our main concern, but with 10,000 simulations, all standard errors in this paper are small enough so that visible differences between curves are significant.

We characterize the overall performance of a heuristic by its average score $\bar{S}$, which is influenced by the interplay of different characteristics of an interval. Before looking at $\bar{S}$, we build some intuition regarding the combined intervals by exploring $\bar{W}$, *MAE*, and *RF*.

The width of an interval is related to S through the term $-(\alpha/2)(U - L) = -(\alpha/2)W$. Also relevant are the penalties $(L - x)^+$ and $(x - U)^+$ incurred when x is not captured by the interval. The wider the interval, the farther x has to be from the midpoint before one of these penalties is activated, and the smaller the penalty is for a given x when it is activated. Thus, W affects both the penalty for width and the penalty for x being outside the interval, with the former penalty increasing and the latter penalty potentially decreasing as W increases.

Av yields a combined interval forecast $\bar{y} \pm z_{\alpha/2} \cdot (1-\gamma)\sigma$. In our model, $W_{Av} = W_{Md} = 2z_{\alpha/2}(1-\gamma)\sigma$ for all k because W is the same for all individual intervals. Of the heuristics, *En* has the largest W. By definition, $W_{PM} = W_{PA}$, and from Lichtendahl et al. (2013, p. 1600), $W_{PA} \geq W_{Av}$ for the model. Thus, $W_{En} \geq W_{PM} \geq W_{Av} = W_{Md}$ for all σ and γ. Also, $W_{TI} \geq W_{Av} \geq W_{TE}$ by the way the trimming is done.

Figure 1 shows the average widths $\bar{W}$ of the combined intervals from the heuristics as a function of k for the base case of $\rho = \gamma = 0$ (independent forecasters, no overconfidence), along with four contrasting cases of low/high correlation ($\rho = 0.25, 0.75$) and low/high overconfidence ($\gamma = 0.25, 0.75$). The $\bar{W}$s follow the rank orders given previously, do not change with k for *Av* and *Md*, are increasing in k for *En* and *PM*, and are steady but "bumpy" due to the kinks for *TI* and *TE*. The average widths decrease for all heuristics as overconfidence increases with correlation fixed, and decrease for *En*, *PM*, and *TI* as correlation increases with overconfidence fixed. If the individual intervals are overconfident, then a heuristic leading to wider intervals might perform better than *Av* and *Md* by correcting for overconfidence. However, this correction might be overdone and lead to underconfidence.

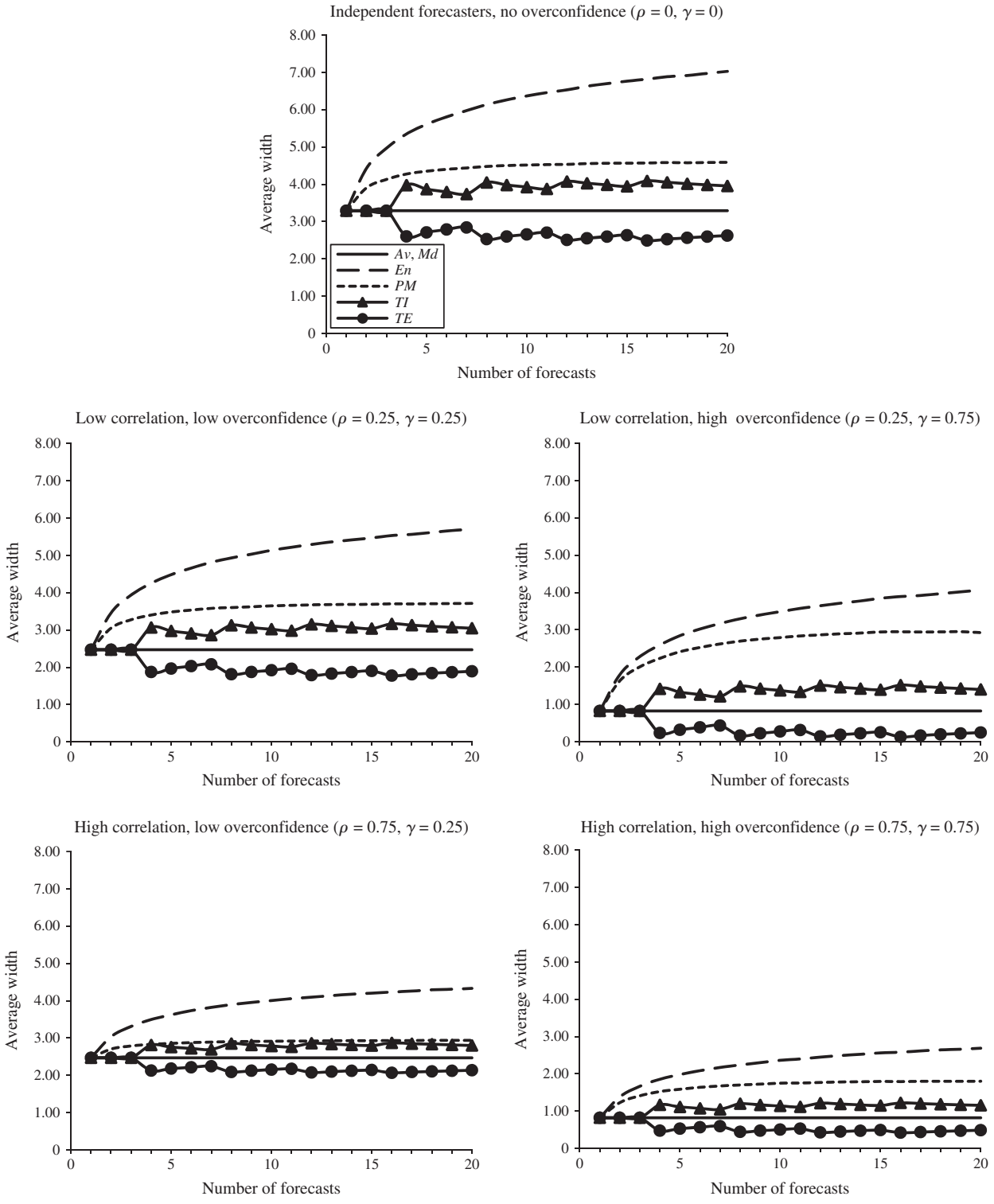
The *MAE* also relates to S. From $AE = |0.5(U + L) - x|$ and $W = U - L, S$ can be expressed as $S(AE, W) = -(\alpha/2)W - (AE - 0.5W)^+$. The penalty associated with the second term is more likely to kick in when AE is larger, and a larger AE makes it more likely that interval will not capture x.

For the model, the expected *MAE* of the combined interval for *Av* is $\sigma\sqrt{(2/\pi)(\rho + (1-\rho)/k)}$, which does not depend on γ. It decreases in k because the interval is better located as k increases, which should lead to a higher S Further, it increases in ρ, implying that S should decrease, consistent with the notion that dependence leads to loss of information due to redundancy.

Because the sample mean is a good estimator of the population mean, *Av* should have well-located midpoints and low *MAE*s. By definition, *Av* and *PM* have the same midpoints and *MAE*s. We might expect *Md*, *TI*, and *TE* to have *MAE*s reasonably close to those of *Av*. Heuristic *En*, being based only on two extreme values, should have the highest *MAE*.

Figure 2 shows the *MAE*s of midpoints of the combined intervals as a function of k for the same cases

Figure 1. Average Width ($\bar{W}$) for Combined Intervals as a Function of k for the Symmetric Normal Model



shown in Figure 1. When $\rho = \gamma = 0$, *MAE* decreases in k for all heuristics, rapidly at first before starting to level off. As expected, *En* has the highest *MAE*, and *Md* is next. The other heuristics are very close to each other. The pattern is similar with correlation and overconfidence, with the *MAE*s leveling off sooner and at a higher value with higher ρ for fixed γ but not being affected by higher γ for a given ρ.

The relative frequency affects S because the higher *RF* is, the less often the penalty for x being outside the interval comes into play. The wider the intervals, the higher *RF* will be, all other things equal, so a higher *RF* is related to the good and bad implications of a larger W for S. For the 90% intervals, the optimal balance between these good and bad implications occurs when $RF = 0.90$.

The expected relative frequency of the combined interval $\bar{y} \pm z_{\alpha/2}(1-\gamma)\sigma$ for *Av* is $1 - 2\Phi(-z_{\alpha/2}((1-\gamma)/\sqrt{\rho + (1-\rho)/k}))$, where Φ is the cdf of the standard normal distribution. This expected *RF* decreases with γ due to greater overconfidence, increases with k due to a lower *MAE*, and decreases with ρ due to a higher *MAE*. For $\gamma > (<) 1 - \sqrt{\rho + (1-\rho)/k}$, the combined interval for *Av* is expected to be over(under)confident.

Figure 3 shows the *RF*s of the combined intervals as a function of k for the same cases shown in Figure 1. When $\rho = \gamma = 0$, *RF* increases in k for all heuristics, starting at 0.90 and approaching 1 rapidly, thereby becoming underconfident. As γ increases, *RF* starts at a lower level and then increases more rapidly for *En* and *PM* than for *Av* and *Md*, resulting in the heuristics exhibiting larger differences, especially with $\gamma = 0.75$. Heuristic *En* has the highest *RF*, followed by *PM* and then *TI*, *Av*, *Md*, and *TE*. For example, when $\rho = \gamma = 0.75$, the *RF*s start at 0.32 for $k = 1$ and increase to 0.86 for *En*, 0.69 for *PM*, 0.49 for *TI*, and only to 0.36 for *Av* and *Md*, while declining to 0.22 for *TE*.

These observations on $\bar{W}$, *MAE*, and *RF* are insightful for understanding the overall performance of the heuristics on the score S. The expected S for *Av* is

$$E(S) = -\alpha(1-\gamma)z_{\alpha/2}\sigma\Phi\left(-z_{\alpha/2}\frac{1-\gamma}{\sqrt{\rho+(1-p)/k}}\right) - 2\sigma\left(\sqrt{\rho+(1-p)/k}\right)\phi\left(-z_{\alpha/2}\frac{1-\gamma}{\sqrt{\rho+(1-p)/k}}\right),$$

where ϕ is the pdf of the standard normal distribution. The expected score $E(S)$ is increasing with k and decreasing with ρ,reinforcing the intuition that having more forecasters implies more information and positive correlation reduces the overall information content. Moreover, $E(S)$ increases (decreases) with γ when $\gamma > (<) 1 - \sqrt{\rho + (1-\rho)/k}$. Given k and ρ, perfect calibration is achieved and $E(S)$ is maximized at $\gamma = 1 - \sqrt{\rho + (1-\rho)/k}$.

Figure 4 shows the average scores $\bar{S}$ for the combined intervals as a function of k for the same cases shown in Figure 1. For independent forecasters who are perfectly calibrated, *TE* does best, with *Av* and *Md* next, then *TI*, which increases at first and then returns to about its original level; *PM* is even lower and decreases slightly with k after a very slight increase; and *En* is worst, with $\bar{S}_{En}$ rapidly decreasing in k.

The relative performance of the heuristics in terms of $\bar{S}$ is similar for low correlation and overconfidence, with $\bar{S}_{TE}$ leading the pack. However, as either correlation or overconfidence increases, the relative performance of the heuristics begins to change. It is only in the extreme case of high correlation and high overconfidence that *En* has the highest $\bar{S}$ beyond very small values of k, with $\bar{S}_{PM}$ next and the others trailing. In other cases, rapid increases in $\bar{W}_{En}$ lead to rapid increases in RF_{En} that reduce overconfidence, but this quickly leads to underconfidence and declining $\bar{S}_{En}$ with higher k.

With low correlation and high overconfidence, $\bar{S}_{PM}$ does well for low k but is surpassed by $\bar{S}_{TI}$ for $k > 7$. Heuristics *TI* and *PM* also do quite well if high correlation is paired with low overconfidence, and *Av* is not far behind.

In sum, when the correlation between experts is low and the degree of overconfidence in individual intervals is low in the model, *TE* performs the best, followed closely by *Av* and *Md*, then *TI* and *PM*, with *En* performing the worst. In contrast, when either correlation or overconfidence is high and the other remains low, but not both, *TI* and *PM* do best, followed by *Av* and *Md*. In the extreme case of both high overconfidence and high correlation, *En* performs the best, followed by *PM*, *TI*, *Av*, *Md*, and *TE*, in that order.

To conclude, we note that the patterns of performance of different heuristics for the symmetric model are quite robust with respect to heterogeneity (Appendix A) and asymmetry of the distribution (Appendix B). In our empirical studies, we will observe similar patterns.

Figure 2. Mean Absolute Error (*MAE*) for Combined Intervals as a Function of k for the Symmetric Normal Model

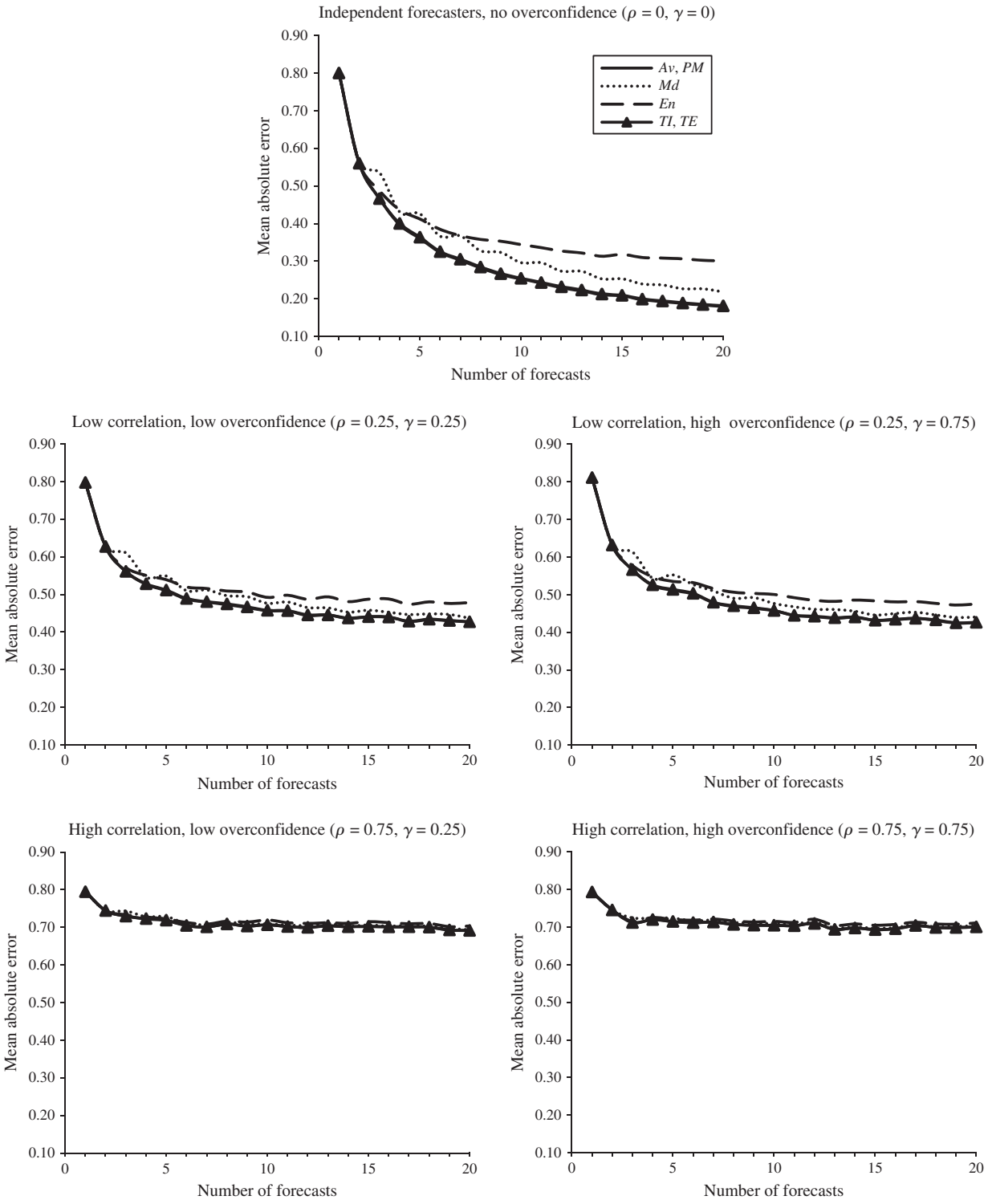


Figure 3. Relative Frequency (*RF*) for Combined Intervals as a Function of k for the Symmetric Normal Model

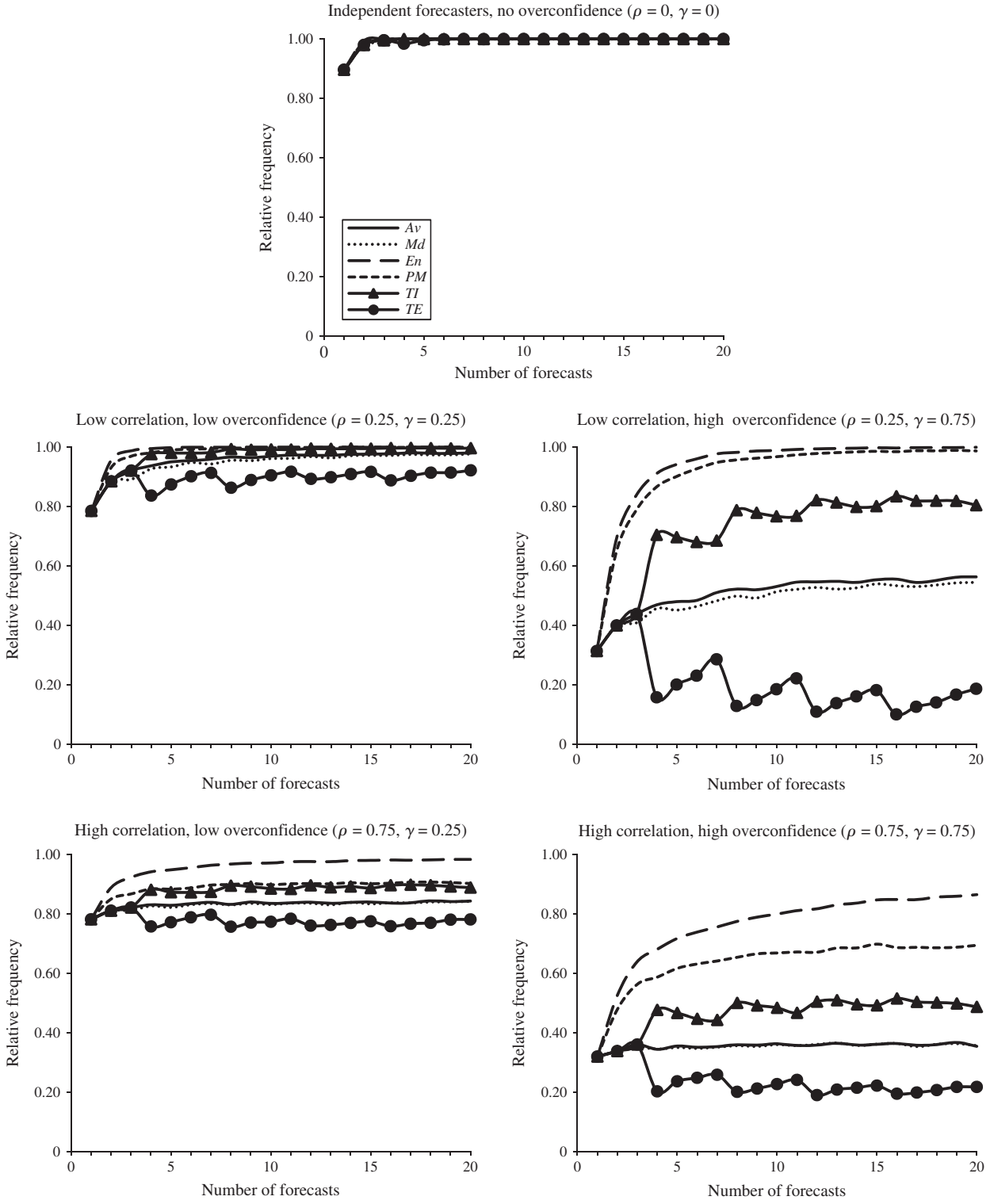
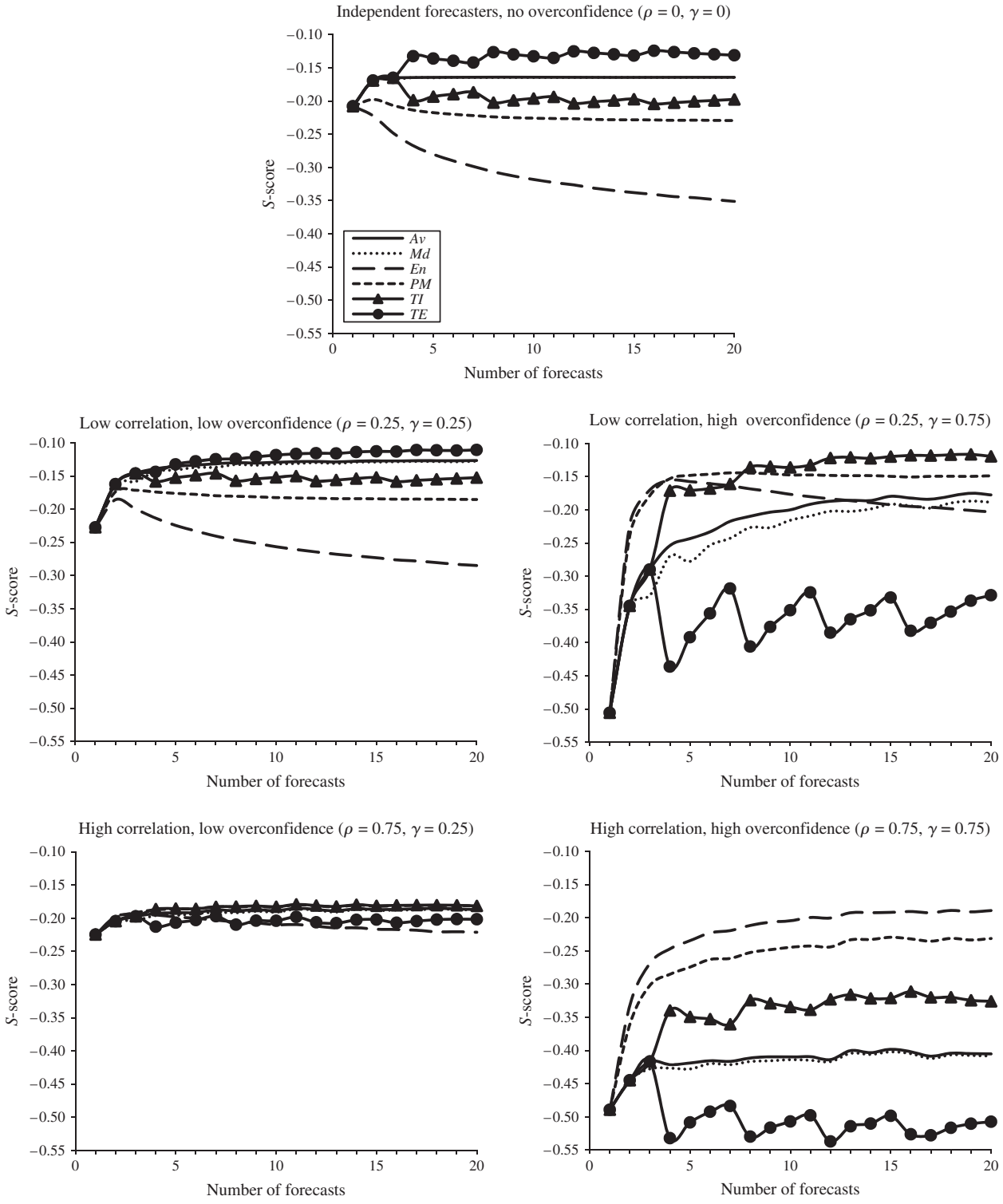


Figure 4. Average S-score ($\bar{S}$) for Combined Intervals as a Function of k for the Symmetric Normal Model



4. Empirical Studies

We analyzed data from two data sets of forecasts. The first data set is from a study we designed and conducted with analysts at a brokerage and investment group. The analysts made interval forecasts involving stock exchange indices and oil and gold prices. The second data set is based on forecasters' probability forecasts of GDP growth and inflation for the Survey of Professional Forecasters (SPF). In both data sets, the forecasters followed the variables of interest on a regular basis. Details concerning these data sets and summary statistics related to the forecasts are presented in Section 4.1, and our analysis of the performance of the heuristics is presented and discussed in Section 4.2.

4.1. Data Description and Summary Statistics

Data from Analysts. Fifty-nine analysts at CLSA (https://www.clsa.com), a major international brokerage and investment group, participated in a one-time online study. The median age of the participants, all with university or advanced degrees, was 36 years, and the median years of service with CLSA (not including experience at other financial houses) was five years. The analysts were based in New York, Tokyo, and Hong Kong. They were asked to provide 90% interval forecasts one, two, and three months ahead for some of the financial quantities that were of most interest to and were continuously tracked by them: the price of oil in US$/barrel from the Brent EUCRBRDT Index (Oil1 to Oil3), the price of gold in US$/oz from the Bloomberg GOLDS COMDTY Index (Gold1 to Gold3), the Dow Jones Industrial Average Index (DJ1 to DJ3), the Nikkei NKY Index (NK1 to NK3), and the Hang Seng HIS Index (HS1 to HS3).

SPF Data. We also analyzed the forecasts of annual GDP growth (1992–2009) and annual inflation (1992–2010) reported by forecasters surveyed in the first through fourth quarters (Q1–Q4) of each year as part of the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters. Panelists reported probabilities that percentage changes in U.S. real GDP and U.S. inflation for the full year would fall in each of 10 predetermined bins. The forecasters did not all participate in the survey every year. Our data involved 25–50 forecasters per year for both GDP and inflation forecasts. The data are part of the data set downloaded from http://www.phil.frb.org/econ/spf and analyzed by Lichtendahl et al. (2013). We used data starting from 1992 to have the same 10 bins over the period of our analysis. Since the panelists were not asked directly about interval forecasts, we approximated the SPF panelists' continuous predictive distributions by fitting piecewise-linear cdfs to their reported discrete distributions, as in Lichtendahl et al. (2013), taking the 0.05 and 0.95 quantiles of the fitted cdfs to construct 90% prediction intervals.

Summary Statistics. Summary statistics involving the analysts' forecasts are presented in Table 1 for the three lead times (e.g., DJ1 to DJ3) for all five quantities of interest. Values of $\bar{S}$ cannot be compared for different quantities because they depend on the scaling of $\tilde{x}$, but $\bar{S}$ for forecasts of the same variable with different lead times are comparable. As anticipated, $\bar{S}$ tends to be better (less negative) for shorter lead times, with the only violations occurring in going from HS2 to HS3 and Gold2 to Gold3. The *MAE* is also better for shorter lead times except for the same violations as $\bar{S}$.

The average interval widths in Table 1 decrease with shorter lead time for all quantities, which is not surprising. The overall *RF* is 0.37, indicating high overconfidence, and *RF* varies by variable and lead time,

Table 1. Summary Statistics for the 59 Analysts' Assessed 90% Intervals

Quantity	Average S-score	Average width	Relative frequency	MAE of midpoints	Misses $< L$, misses $> H$
DJ1	−127.6	1,155	0.59	350	0, 24
DJ2	−498.6	1,540	0.34	938	38, 1
DJ3	−757.2	1,848	0.24	1,296	44, 1
HS1	−292.5	2,148	0.59	709	22, 2
HS2	−1,282.2	2,726	0.14	2,230	49, 2
HS3	−919.2	3,275	0.39	1,821	32, 4
NK1	−107.1	1,296	0.86	250	7, 1
NK2	−991.4	1,739	0.10	1,622	52, 1
NK3	−1,464.3	2,116	0.14	2,272	51, 0
Gold1	−42.4	126	0.19	84	1, 47
Gold2	−56.9	175	0.24	111	1, 44
Gold3	−36.6	206	0.68	91	4, 15
Oil1	−1.4	12	0.58	4	0, 25
Oil2	−5.0	17	0.25	10	44, 0
Oil3	−7.5	21	0.29	14	42, 0
Average			0.37		

Notes. DJ: Dow Jones Industrial Average Index; HS: Hang Seng Index; and NK: Nikkei Index.

from 0.10 for NK2 to 0.86 for NK1. For each quantity except Gold, the shortest lead time has the least overconfident intervals, suggesting that the increases in interval widths observed for longer lead times are not sufficient.

If we confine attention to "misses" (cases where the intervals do not capture the realized value), x is predominantly below or predominantly above the intervals, as seen in the last column of Table 1. For example, HS2 and Gold1 both had very low values of *RF*, 0.14 for HS2 and 0.19 for Gold1. The 59 intervals for HS2 had 51 misses, with x below 49 of those intervals and above 2, whereas for Gold1's 48 misses, x was below 1 interval and above 47. This suggests high correlations among the forecast errors from the different analysts, verified by the high average pairwise correlation of forecast errors, 0.81.

Tables 2 and 3 give the same set of statistics for the SPF forecasts of GDP and inflation, respectively, with year-by-year results shown only for Q1 to save space because the year-by-year results for Q2–Q4 are very similar in nature. Averages across the years are shown for Q1–Q4 to look at lead time effects. Note that unlike the analysts' forecasts, with Oil1 indicating the shortest lead time and Oil3 the longest for forecasts of oil prices, the SPF forecasts are all for the same full-year values of GDP and inflation, with the longest lead times corresponding to Q1 forecasts and the shortest to Q4 forecasts.

For inflation, $\bar{S}$ and *MAE* improve for shorter lead times. For GDP, the same thing is true for *MAE*, but $\bar{S}$ is slightly worse in going from Q2 to Q3 and Q3 to Q4. Similar to the analysts, $\bar{W}$ decreases with shorter lead times for both variables.

The three sets of forecasts differ in terms of *RF*. The GDP forecasts exhibit moderate overconfidence (average *RF* near 0.70 for Q1–Q3, declining to 0.59 for Q4), much less severe than for the analysts (average *RF* of 0.37). For inflation, the forecasts are not overconfident, hovering around 0.90 for all four lead times. The misses for forecasts of a given quantity tend to cluster

Table 2. Summary Statistics for Percentage Change in Annual U.S. Real GDP: Calculated 90% Intervals from SPF Forecasters' Binned Probability Assessments Year by Year for Q1 and Overall Averages for Q1–Q4

Year	Number of forecasters	Average S-score	Average width	Relative frequency	MAE of midpoints	Misses $< L$, misses $> H$
1992	36	−0.54	3.25	0.39	1.61	0, 22
1993	31	−0.15	2.83	0.97	0.39	0, 1
1994	27	−0.19	2.93	0.81	0.85	0, 5
1995	25	−0.20	2.72	0.92	0.57	2, 0
1996	35	−0.82	3.03	0.23	2.08	0, 27
1997	33	−0.77	2.91	0.18	2.03	0, 27
1998	29	−0.74	2.62	0.14	1.83	0, 25
1999	30	−0.68	3.43	0.27	2.19	0, 22
2000	33	−0.17	3.03	0.91	0.73	0, 3
2001	30	−0.23	3.47	0.80	0.81	6, 0
2002	30	−0.22	3.19	0.90	0.76	1, 2
2003	33	−0.18	3.49	0.97	0.531	0, 1
2004	27	−0.24	3.85	0.85	0.86	4, 0
2005	32	−0.19	2.84	0.91	0.48	3, 0
2006	49	−0.18	2.99	0.92	0.58	4, 0
2007	46	−0.17	2.88	0.85	0.70	7, 0
2008	43	−0.62	3.15	0.37	1.75	27, 0
2009	39	−0.55	3.52	0.74	1.94	10, 0
Q1 Average	33.8	−0.38	3.12	0.67	1.15	
Q2 Average	36.1	−0.26	2.86	0.75	0.90	
Q3 Average	34.9	−0.26	2.43	0.70	0.83	
Q4 Average	34.8	−0.30	1.79	0.59	0.77	
Average	34.9	−0.30	2.55	0.68	0.91	

Table 3. Summary Statistics for Percentage Change in Annual U.S. Inflation: Calculated 90% Intervals from SPF Forecasters' Binned Probability Assessments Year by Year for Q1 and Overall Averages for Q1–Q4.

Year for Q1	Number of forecasters	Average S-score	Average width	Relative frequency	MAE of midpoints	Misses $< L$, misses $> H$
1992	36	−0.15	2.77	0.94	0.68	2, 0
1993	30	−0.19	2.58	0.87	0.82	4, 0
1994	26	−0.13	2.52	1.00	0.58	0, 0
1995	26	−0.24	2.38	0.81	0.83	5, 0
1996	33	−0.14	2.62	0.91	0.57	3, 0
1997	33	−0.17	2.56	0.82	0.71	6, 0
1998	29	−0.28	2.27	0.72	0.96	8, 0
1999	30	−0.14	2.85	1.00	0.43	0, 0
2000	33	−0.14	2.53	0.94	0.42	0, 2
2001	30	−0.14	2.62	0.97	0.46	0, 1
2002	30	−0.15	2.70	0.97	0.50	1, 0
2003	32	−0.15	2.82	0.94	0.45	0, 2
2004	26	−0.22	2.96	0.81	1.07	0, 5
2005	32	−0.30	2.50	0.59	0.96	0, 13
2006	50	−0.16	2.81	0.92	0.59	0, 4
2007	46	−0.14	2.58	0.85	0.58	0, 7
2008	45	−0.19	2.95	0.91	0.61	3, 1
2009	39	−0.17	3.07	0.92	0.55	1, 2
2010	38	−0.15	2.81	0.97	0.56	1, 0
Q1 Average	33.9	−0.18	2.68	0.89	0.65	
Q2 Average	36.5	−0.16	2.55	0.91	0.58	
Q3 Average	34.1	−0.14	2.27	0.92	0.46	
Q4 Average	35.1	−0.12	1.81	0.90	0.39	
Average	34.9	−0.15	2.33	0.91	0.52	

on one side of the interval. The average pairwise correlations of the forecast errors are 0.79 for GDP, similar to that for the analysts, and a more moderate 0.52 for inflation.

Some differences between the characteristics of the analysts' forecasts and the SPF forecasts might be expected. The analysts had little or no experience at making formal interval forecasts, and there was no chance to learn from experience. Many of the SPF forecasters, on the other hand, had previous experience at making probability forecasts through the SPF program itself and perhaps otherwise. Also, the analysts provided 90% intervals directly and the SPF forecasters provided probabilities for fixed bins, from which their 90% intervals were constructed. Different elicitation techniques can influence the individual forecasts (e.g., Budescu and Du 2007, Abbas et al. 2008). As noted earlier, our focus is not on elicitation but on combining once intervals have been determined. An important similarity is that both the analysts and SPF forecasters are familiar with the quantities they were forecasting and often make forecasts of those quantities, whether formal or not, as part of their normal occupations.

4.2. Performance of the Heuristics

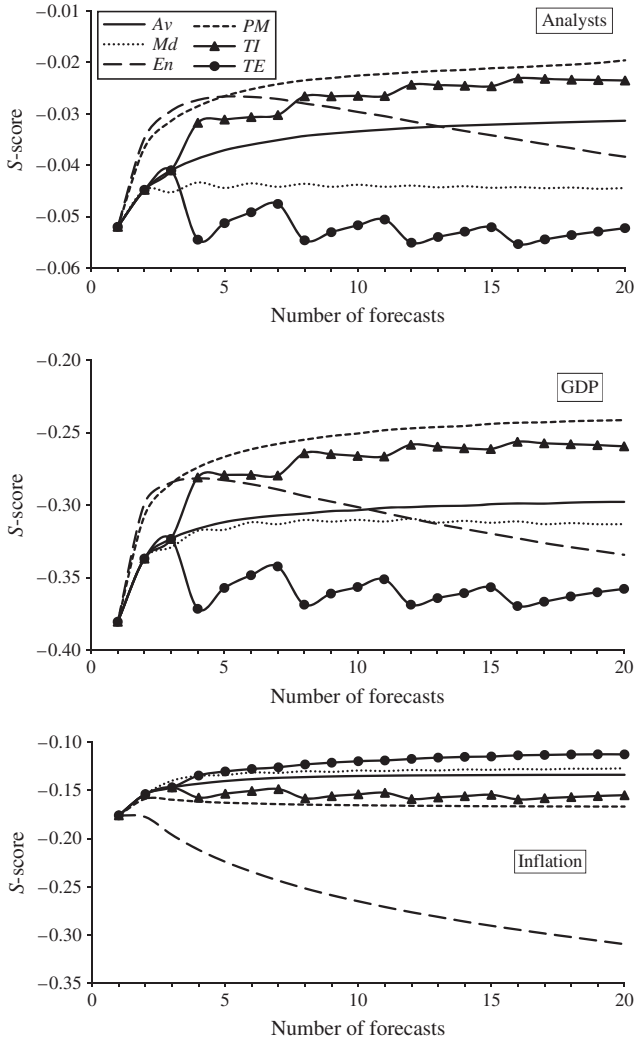
In this section, we discuss the performance of the different heuristics for combining intervals from the analysts and SPF forecasters. For each variable, we form random subgroups of $k = 2, \ldots, 20$ forecasters, with 10,000 simulations for each k, and create combined intervals using the six heuristics. We then compute the four performance measures for the combined intervals: $\bar{S}$, $\bar{W}$, MAE, and RF.

For reasons of space and brevity, we present results aggregated across the five forecasting variables and three time horizons for the analysts and across the different years with Q1 only for GDP and inflation; similar graphs for the Q2–Q4 SPF forecasts look virtually the same. Before combining intervals for the

analysts, we rescale their forecasts in terms of return, where return = (forecast/actual value at the time of forecast) −1, to place all the forecast intervals on the same scale. The realized values are similarly rescaled to yield realized returns. No such rescaling is necessary for GDP and inflation, as those variables were the same from year to year.

Figure 5 shows $\bar{S}$ for the combined intervals as a function of k for the analysts, GDP, and inflation. For the analysts, *En* has the highest $\bar{S}$ for $k < 5$ and then decreases. Once *En* fades, *PM* and *TI* increase at a decreasing rate and are the top two, in that order. The lower *Av* follows a similar pattern, whereas *Md* levels off at an even lower

Figure 5. Average *S*-Score ($\bar{S}$) for Combined Intervals as a Function of k for Analysts, GDP, and Inflation



value and *TE* performs worst. The pattern is similar for GDP, with *PM* performing best, followed closely by *TI*. For inflation, *TE* performs the best, followed closely by *Md* and *Av*. These three heuristics all level off more quickly than the best heuristics do for the analysts and GDP. Heuristics *TI* and *PM* barely increase at all, and *En* decreases from the start.

A comparison of Figure 5 with Figures 4, A.1, and B.1 reveals some striking similarities. In comparing these figures, we can compare the shapes of the curves and the relative order of the heuristics. Note, however, that the values of $\bar{S}$ are not comparable because the scale is different for each variable. To save space, graphs for $\bar{W}$, *MAE*, and *RF* for the analysts, GDP, and inflation are not presented here, but they are quite similar to Figures 1–3. These measures influence $\bar{S}$, as discussed in Section 3.

The graphs for the analysts and GDP in Figure 5 both closely resemble the graphs for low correlation and high overconfidence in Figures 4 and A.1 and are similar to the case of high correlation and overconfidence in Figure A.1. This is consistent with the overconfidence and the high correlation exhibited by the analysts and GDP. The overall shapes of the curves for the six heuristics are very similar. The only slight differences are that for the analysts and GDP, *PM* stays a bit above *TI* instead of being passed by *TI*, and *Md* for the analysts is quite a bit below *Av* and remains below *En*.

The graph for inflation, on the other hand, is most similar to the graph for independence and no overconfidence in Figure 4 and the graphs for low correlation and low overconfidence in Figures 4 and A.1. This makes sense given that the inflation forecasts exhibit very little if any overconfidence as well as a lower correlation than the analysts and GDP. Finally, all three graphs in Figure 5 show some resemblance to graphs in Figure B.1, but to a lesser degree.

The analyses from the empirical studies provide some validation for the multinormal model developed in Section 3 and reinforce many of the implications of the model for the performance of the heuristics under different conditions. There is a close correspondence between the results presented in Sections 3 and 4. Next we briefly summarize our conclusions with respect to the different heuristics.

Heuristic *Av* can improve the placement of the combined interval by improving the accuracy of the midpoint. However, moderate or high overconfidence in

the individual intervals is not corrected for, even with high k. On the other hand, if there is no overconfidence in the individual intervals, as in the case of inflation, *Av* performs much better, as the combined interval gets better placed as k increases.

Heuristic *Md* also increases the accuracy of the midpoint with larger k, but at the same time it creates marginally smaller interval widths than *Av*, thus performing even worse in accounting for overconfidence. This is seen in Figure 5 for the analysts and GDP, where $\bar{S}$ for *Md* is almost flat over k while most of the other heuristics improve with k. It does quite well for inflation, where overconfidence is not an issue and the marginally narrower intervals make it slightly better than *Av*.

The extreme heuristic *En* can do well for very small values of k when there is substantial overconfidence. For example, with the analysts and GDP, the average relative frequency of the individual intervals (for $k = 1$) is well below 0.90 for the 90% intervals. As k increases, RF_{En} rather quickly goes to 0.90 and beyond due to the rapidly increasing width of the combined interval, resulting in very underconfident forecasts, and MAE_{En} increases with k. Hence, $\bar{S}_{En}$ does well for very small k due to the needed correction for overconfidence but thereafter begins to decline quickly due to overcorrection. For inflation, on the other hand, the individual intervals are on average already well calibrated. Thus, *En* creates excessive width in the combined interval even for small values of k, which leads to $\bar{S}_{En}$ decreasing with k from the start and performing by far the worst among all heuristics.

Heuristic *PM* accounts for overconfidence by increasing the width of the intervals in a less extreme manner than *En* through probability averaging, which yields combined forecasts that are less overconfident or more underconfident than the individual forecasts (Hora 2004, Ranjan and Gneiting 2010, Lichtendahl et al. 2013). Heuristic *PM* also improves the location of the intervals by borrowing *Av*'s midpoint. This approach results in $\bar{S}_{PM}$ performing the best for the analysts and GDP. For inflation, where the individual forecasts are not overconfident, *PM* is not at the top but still does reasonably well.

Heuristic *TE* has the narrowest intervals due to trimming off the most extreme endpoints, which in turn makes it harder to overcome overconfidence but helps by preventing combined forecasts from getting underconfident quickly if the individual interval forecasts are not very overconfident. The trimming yields a more accurate midpoint, which provides better-located intervals. Heuristic *TE* performs the best for inflation, but the worst for the analysts and GDP. It is closest to *Md* and *Av*, doing better than them when there is no overconfidence but worse with overconfidence.

Heuristic *TI* trims in the opposite manner to *TE*, trimming endpoints closest to the midpoint. This results in combined intervals that are wider than *Av* and *Md* instead of narrower. The wider intervals make it similar to *PM*, performing well for situations with overconfidence like the analysts and GDP but not as well for situations with no overconfidence like inflation.

The ranking of the heuristics for the empirical studies on the basis of correlation and overconfidence is very consistent with the conclusions given at the end of Section 3 for the symmetric normal model. The only minor exception is that even for the analysts, whose forecasts show the highest correlation and the highest overconfidence of the three empirical studies, $\bar{S}_{En}$ is only marginally higher than $\bar{S}_{PM}$ for $k \leq 4$ before dropping behind $\bar{S}_{PM}$ and then $\bar{S}_{TI}$ and continuing to decline.

Park and Budescu (2015) borrowed four heuristics (*Av*, *Md*, *En*, and *PA*) from an earlier version of this paper (which did not include *TE* and *TI*), added others, and reanalyzed data from Glaser et al. (2013) and Soll and Klayman (2004). They focused on calibration (using hit rates, which correspond to our *RF*), reaching very similar conclusions to our Figure 3 and to *RF* results from our empirical studies.

What are the implications of these results for the choice of k, the number of forecasters to consult when obtaining forecasts? Looking at Figure 5, we see that for the analysts and GDP, most of the gains in $\bar{S}$ for *PM* and *TI*, the best-performing heuristics in those cases, are attained by $k = 5$, with much smaller gains from $k = 5$ to $k = 10$ and beyond. The results are similar for inflation, where *TE* and *Md* perform best, with *Av* close, and also for the simulations from the models in section 3 and Appendices A and B. This suggests that regardless of the degrees of correlation and overconfidence, any k from 5 to 10 might be a good choice. Others studying

the aggregation of forecasts have come to similar conclusions about k (Armstrong 2001, Hora 2004, Budescu and Chen 2015, Mannes et al. 2014).

5. Summary and Discussion

Extensive work on combining forecasts shows that simple combining methods are more robust, easier to use, and often perform better than more complex methods. Moreover, decision makers often may be less inclined to use more complex methods because they tend to involve detailed modeling and the collection of data to estimate model parameters. We consider some parsimonious, easy-to-use heuristics for combining interval forecasts, with the intent of seeing if other simple combining approaches can compete successfully with the commonly used simple average. Our orientation is prescriptive, and our concern is with combining the forecasts once they are obtained, not with how they are obtained.

Our model of the forecast aggregation process provides a theoretical underpinning and enables us to confirm that overconfidence and correlation are drivers of what makes some heuristics outperform others. The results from the empirical studies are very consistent with those from the model-based simulations, thus validating the model, confirming the role of overconfidence and correlation, and indicating heuristics that are viable alternatives to the simple average. In terms of the number of forecasts to combine, most gains are attained by five and there is little reason to go beyond 10.

The results show that the simple average is still a good choice unless overconfidence is high. They also enable us to identify other heuristics that also can be good choices. The greatest shifts in the differences in performance among the heuristics and in the rank ordering of the heuristics occur when overconfidence is high, so information about the degree of overconfidence can be helpful.

We focus on combining forecasts once they have been obtained, not on elicitation, but overconfidence can vary depending on the forecaster, the quantity, and the way the forecasts are assessed (e.g., Klayman et al. 1999, Soll and Klayman 2004, Teigen and Jørgensen 2005, Budescu and Du 2007, Jain et al. 2013). Previous work indicates that assessing the endpoints of the intervals separately instead of asking directly for an interval reduces overconfidence. Haran et al. (2010) suggest that a procedure like the way the SPF forecasts are assessed into bins can lead to lower overconfidence. These results could explain in part why the SPF forecasters exhibit lower overconfidence than the analysts, and the choice of an elicitation procedure might be able to reduce overconfidence somewhat.

High correlations of forecast errors, as in the case of the analysts and GDP, are not unusual. Although high correlations do not have as great an influence on the performance of the heuristics as high overconfidence, they do have some influence. The primary impact of this dependence is to reduce the potential improvement in performance from additional forecasters. A decision maker can attempt to reduce dependence by choosing a diverse group of forecasters who have different backgrounds and training, follow different theories, and use a variety of modeling approaches and data sets.

We close with a summary of some prescriptive advice. If the decision maker does not have a good idea of the degrees of overconfidence and correlation in a given situation, a heuristic that is more robust and never performs extremely poorly might be preferred to reduce the risk of a bad forecast. From our results, that would suggest avoiding *En* and *TE*, both of which perform well under some conditions but do the worst by far in other cases. The other four heuristics are more robust. Our benchmark, *Av*, is a good choice, but *PM*, *Md*, and *TI* are worthy competitors. Heuristics *PM* and *TI* are not quite as simple as *Av* and *Md*, but if that is not a problem then they deserve serious consideration. Of course, since these heuristics are all very easy to implement, we could consider a second-order combining, using all of these robust heuristics and taking a simple average of their forecasts.

If some information is available about the degrees of overconfidence and correlation, then more targeted prescriptive advice can be offered. With both overconfidence and correlation low, the best bets are *TE*, *Md*, and *Av*. For low overconfidence and high correlation, *PM*, *TI*, *Av*, and *Md* are all very close at the top, and for high overconfidence and low correlation, *PM* and *TI* are best. Finally, *PM* and *En* perform best when both overconfidence and correlation are high.

Acknowledgments

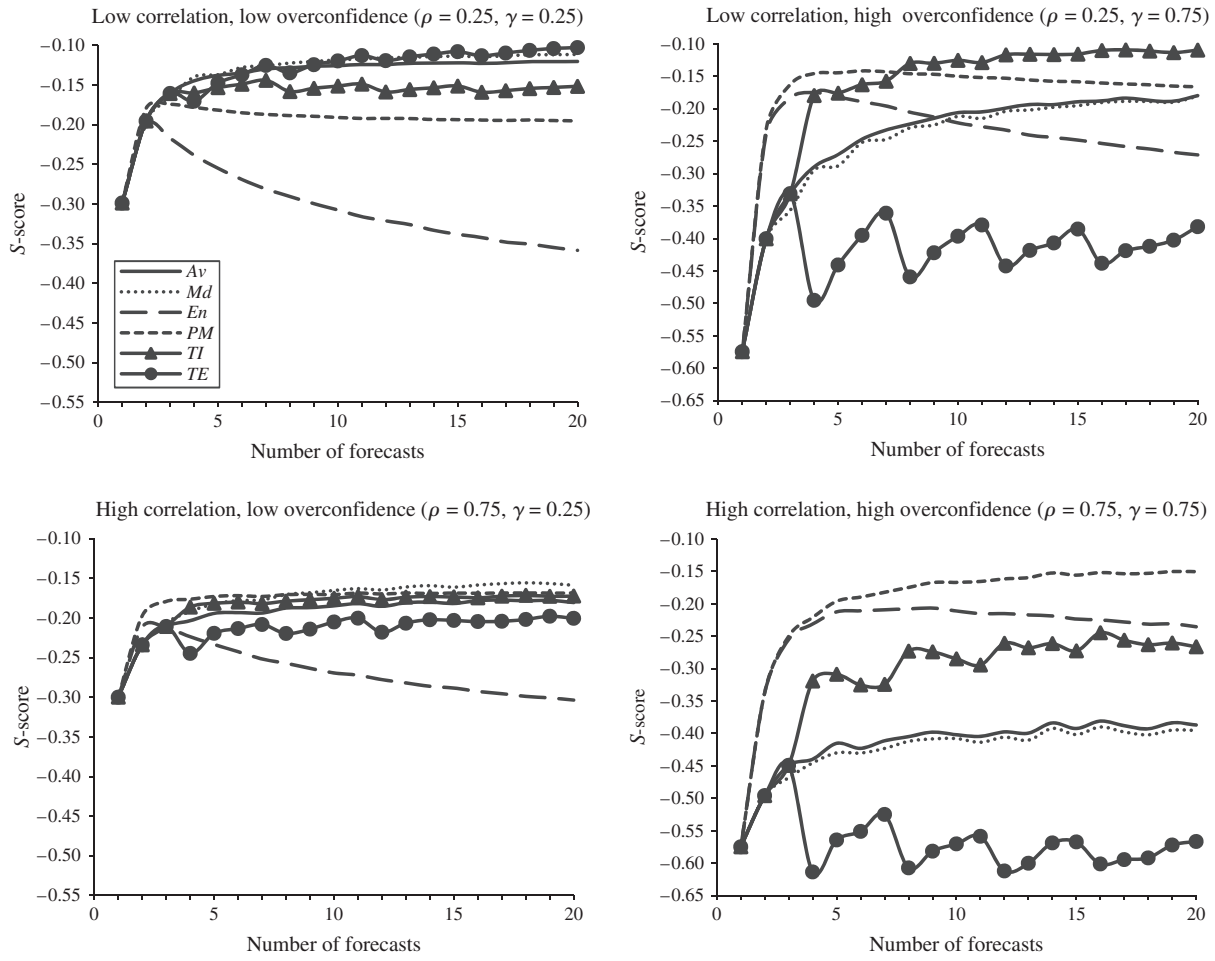
We are grateful to CLSA for arranging to have their analysts make the forecasts used in our study, to Casey Lichtendahl and Yael Grushka-Cockayne for sharing their "cleaned" data set of SPF forecasts, and to an associate editor and three referees for helpful suggestions. We also thank Zhi Chen for his excellent assistance in the data analysis. The Centre for Decision Making and Risk Analysis (CDMRA) at INSEAD provided financial support for this project.

Appendix A. Multivariate Normal Model with Heterogeneity on σ, γ, and ρ

The symmetric model in Section 3 assumes that all forecasters share the same values of γ, σ, and ρ. This assumption is convenient for a baseline analysis, but in practice it is likely that the members of a group of forecasters will have different values of these parameters. Here, we introduce heterogeneity among forecasters to investigate the robustness of the results from the symmetric model.

Using the simulation approach from Section 3, we first consider different levels of overconfidence. For each group of k forecasters, $\gamma_1, \ldots, \gamma_k$ are drawn independently from a beta distribution, $f(\gamma) \propto \gamma^{\alpha-1}(1-\gamma)^{\beta-1}$, with $\alpha, \beta > 0$, $E(\gamma) = \alpha/(\alpha+\beta)$, and $V(\gamma) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. For a variety of choices of (α, β), we simulate 10,000 groups of size $k, k = 1, \ldots, 20$, for each (α, β). Holding $E(\gamma)$ constant, we vary $\alpha+\beta$ from 32 to 0.4 to provide different levels of heterogeneity for γ, with lower $\alpha+\beta$ implying higher $V(\gamma)$ (greater heterogeneity). This is done for $E(\gamma) = 0.25$, 0.50, and 0.75, $\rho = 0.25$, 0.50, and 0.75, and $\sigma = 1$. In all these cases, the relative performance of the heuristics and general shapes of the curves are quite similar to the equivalent cases in Section 3 with γ for the homogeneous model equal to $E(\gamma)$ for the heterogeneous model.

Figure A.1. Average S-Score ($\bar{S}$) for Combined Intervals as a Function of k for the Normal Model with High Heterogeneity on γ and σ



Next, we introduce heterogeneity with respect to σ in the same manner while considering the same values of γ and ρ. For each group, $\sigma_1, \ldots, \sigma_k$ are drawn independently from a gamma distribution with $E(\sigma) = 1$ fixed and $V(\sigma)$ varied from 0.02 to 0.50. As with γ, heterogeneity with respect to σ has little impact on the performance of the heuristics, which remains similar to the equivalent cases in Section 3.

We also explore heterogeneity in pairwise correlations. For different values of (γ, σ), pairwise correlations for the k forecasters in each group are independently drawn from a beta distribution. If a draw results in a covariance matrix that is not positive definite, this draw is discarded. The proportion of covariance matrices that are positive definite is small if the level of heterogeneity is high and k is not too small. As a result, for low heterogeneity on ρ we can only say that the relative performance of the heuristics remained similar to the equivalent cases in Section 3 for $k \le 20$, and the same is true for moderate heterogeneity for $k \le 11$.

Finally, we consider heterogeneity on both γ and σ using the distributions for γ and σ given above with the greatest heterogeneity ($\alpha + \beta = 0.4$ for γ and $V(\sigma) = 0.5$ for σ). Figure A.1 shows the average scores $\bar{S}$ for the combined intervals as a function of k for the same four combinations of low/high correlation and low/high overconfidence shown in Figure 4. Note that the case of independent forecasters and no overconfidence is not included in Figure A.1; heterogeneity on γ is not possible with $E(\gamma) = 0$ because γ cannot be negative.

There is a high degree of similarity between Figures 4 and A.1. The values of $\bar{S}$ for a particular heuristic are sometimes slightly higher or lower in Figure A.1, but the shapes of the curves are quite similar. In the biggest switch at the top, *PM* jumps ahead of *En* when $\rho = \gamma = 0.75$. Also, *Md* edges slightly ahead of *TI*, *PM*, and *AV* when $\rho = 0.75$ and $\gamma = 0.25$, but here these heuristics are too close to say that one is clearly better than the others.

The overall performance of the heuristics in the simulations is therefore very robust with respect to heterogeneity on γ, σ, and ρ. Even with the highest levels of heterogeneity on both γ and σ simultaneously, the results and implications for the relative performance of the heuristics under different conditions are very similar to the comparable cases for the homogeneous model. This robustness to heterogeneity is reassuring because it is a more realistic assumption than homogeneity.

Appendix B. Lognormal Model

In Section 3, the distribution of the signal $\tilde{y}_i$ given x is normal, which results in individual intervals being symmetric around y_i. That may often be a reasonable assumption, but there are many situations where the underlying distributions would be skewed. Examples are situations when the support of $\tilde{x}$ is bounded at one end, usually the lower end, as in distributions of family incomes in a city. Now we consider a lognormal model to investigate the robustness of the results from Section 3 to skewed distributions.

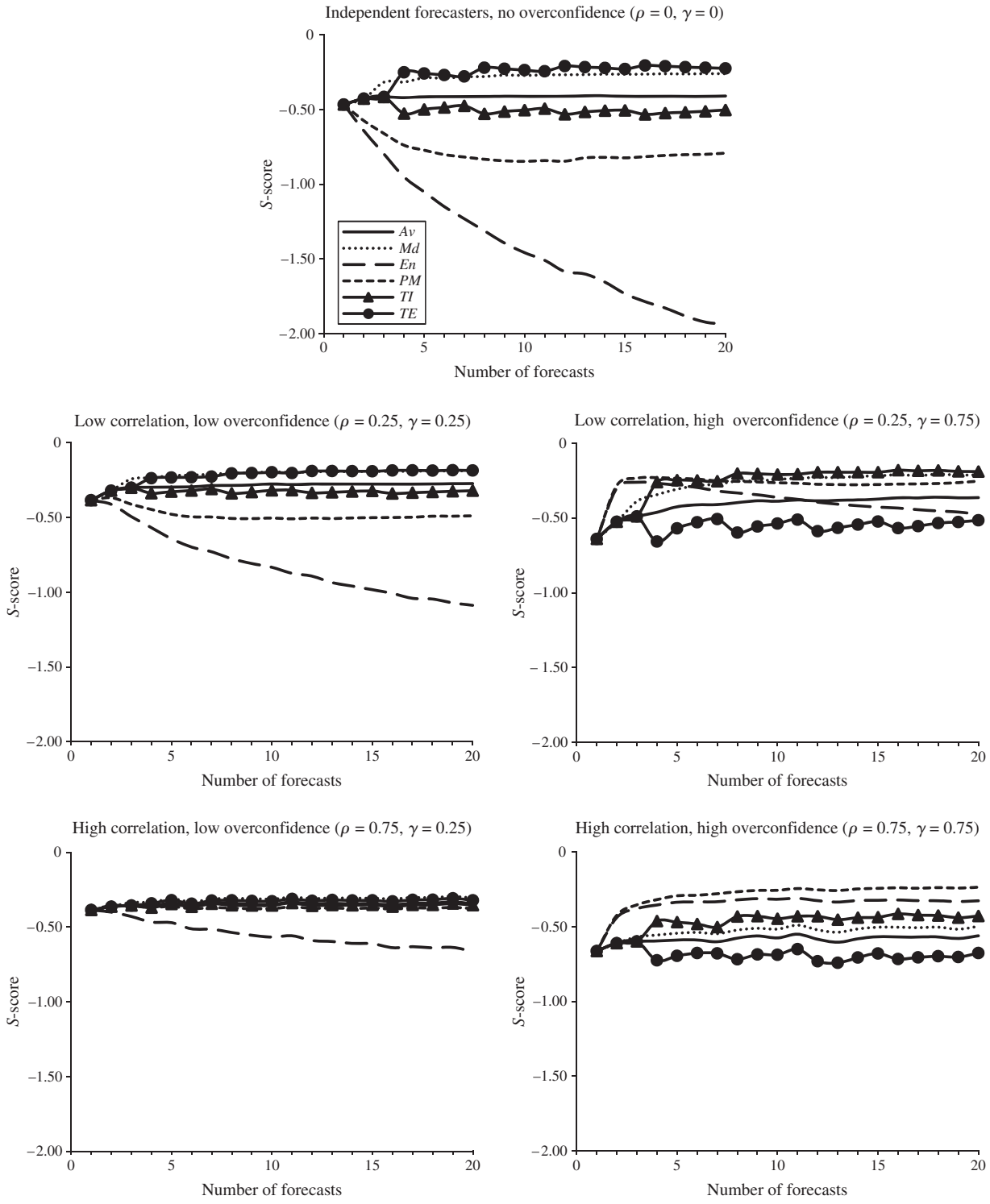
Our lognormal model modifies the symmetric model from Section 3 by assuming that the signal $\tilde{y}_i$ given x comes from a normal process with mean $\ln x$ and variance σ_i^2. With a diffuse prior on $\ln \tilde{x}$, forecaster i's distribution for $\ln \tilde{x}$ after seeing the signal is normal and for $\tilde{x}$ is lognormal, and the reported interval for $\tilde{x}$ is of the form $\exp(y_i \pm z_{\alpha/2}(1 - \gamma_i)\sigma_i)$. We assume common parameters γ, σ, and ρ, and for the purposes of our simulation we consider the same values of k ,γ, σ, and ρ as in Section 3 and set $x = 1$ (so that $\ln x = 0$) without loss of generality.

Figure B.1 shows the average scores $\bar{S}$ for the combined intervals as a function of k for the five combinations of correlation and overconfidence in Figure 4. In terms of the shapes of the curves and the relative performance of the heuristics, Figure B.1 is very similar to Figure 4 with one main exception. Whereas $\bar{S}_{Av}$ is equal to or slightly better than $\bar{S}_{Md}$ in Figure 4, $\bar{S}_{Md}$ is the better of the two in Figure B.1, often considerably so. In the lognormal simulations, $MAE_{Md} < MAE_{Av}$ and $\bar{W}_{Md} < \bar{W}_{Av}$ due to *Md*'s upper bounds being less influenced by extreme values in the right-hand tail of the lognormal distribution for $\tilde{x}$.

Moreover, except for the case of high correlation and high overconfidence, *Md* is often at or near the top in terms of $\bar{S}$. It competes with *TE* (which is also less influenced by the extreme values in the right-hand tail) with no or low overconfidence, sometimes slightly above and slightly below. With high overconfidence and low correlation, it is beaten consistently just by *TI*. Only with high correlation and high overconfidence is it not near the top. This performance represents considerable improvement over the results for *Md* in the symmetric normal model.

In addition to the better performance of *Md*, *TE* also shows some improvement. In the base case of independence with no overconfidence and the cases with low overconfidence, *Md* and *TE* perform best, followed by *Av* and *TI*, which are close to the leaders only when high correlation accompanies low overconfidence. With low correlation and high overconfidence, *PM* and *TI* are best for $k \le 7$, with *Md* moving past *PM* and staying close to *TI* for larger groups. Finally, *PM* has the best $\bar{S}$ when correlation and overconfidence are both high, and *En*, which was the leader in this case for the symmetric normal model of Section 3, is next best. The wider intervals from *En* and *PM* enable them to do a better job of overcoming the high overconfidence and correlation.

Figure B.1. Average S-Score ($\bar{S}$) for Combined Intervals as a Function of k for the Lognormal Model



References

Abbas AE, Budescu DV, Yu H-T, Haggerty R (2008) A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Anal.* 5(4):190–202.

Alpert M, Raiffa H (1969) A progress report on the training of probability assessors. Working paper, Harvard Business School, Cambridge, MA.

Armstrong JS (2001) Combining forecasts. Armstrong JS, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Kluwer Academic, Norwell, MA), 417–439.

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

Budescu DV, Du N (2007) Coherence and consistency of investors' probability judgments. *Management Sci.* 53(11):1731–1744.

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* 5(4):559–583.

Clemen RT, Winkler RL (1985) Limits for the precision and value of information from dependent sources. *Oper. Res.* 33(2): 427–442.

Clemen RT, Winkler RL (2007) Aggregating probability distributions. Edwards W, Miles RF, von Winterfeldt D, eds. *Advances in Decision Analysis* (Cambridge University Press, Cambridge, UK), 154–176.

Cooke RM (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford University Press, Oxford, UK).

Cooke RM, Goossens LLHJ (2008) TU Delft expert judgment data base. *Reliability Engrg. System Safety* 93(5):657–674.

Genest C, Zidek JV (1986) Combining probability distributions: A critique and an annotated bibliography. *Statist. Sci.* 1(1): 114–135.

Glaser M, Langer T, Weber M (2013) True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *J. Behav. Decision Making* 26(5):405–417.

Grushka-Cockayne Y, Jose VRR, Lichtendahl KC Jr (2016) Ensembles of overfit and overconfident forecasts. *Management Sci.*, ePub ahead of print April 20, http://dx.doi.org/10.1287/mnsc.2015.2389.

Haran U, Moore DA, Morewedge CK (2010) A simple remedy for overprecision in judgment. *Judgment Decision Making* 5(7): 467–476.

Hora SC (2004) Probability judgments for continuous quantities: Linear combinations and calibration. *Management Sci.* 50(5): 597–604.

Hora SC, Fransen BR, Hawkins N, Susel I (2013) Median aggregation of distribution functions. *Decision Anal.* 10(4):279–291.

Jain K, Mukherjee K, Bearden JN, Gaba A (2013) Unpacking the future: A nudge toward wider subjective confidence intervals. *Management Sci.* 59(9):1970–1987.

Jose VRR, Winkler RL (2009) Evaluating quantile assessments. *Oper. Res.* 57(5):1287–1297.

Jose VRR, Grushka-Cockayne Y, Lichtendahl KC Jr (2014) Trimmed opinion pools and the crowd's calibration problem. *Management Sci.* 60(2):463–475.

Klayman J, Soll JB, González-Vallejo C, Barlas S (1999) Overconfidence: It depends on how, what, and whom you ask. *Organ. Behav. Human Decision Processes* 79(3):216–247.

Lichtendahl KC Jr, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles? *Management Sci.* 59(7): 1594–1611.

Lichtenstein S, Fischhoff B, Phillips LD (1982) Calibration of probabilities: The state of the art to 1980. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, UK), 305–334.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Soc. Psych.* 107(2):276–299.

Moder JJ, Phillips CR, Davis EW (1995) *Project Management with CPM, PERT and Precedence Diagramming* (Blitz Publishing Co., Middleton, WI).

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

Ranjan R, Gneiting T (2010) Combining probability forecasts. *J. Roy. Statist. Soc. B* 72(1):71–91.

Rowe G (1992) Perspectives on expertise in the aggregation of judgments. Wright G, Bolger F, eds. *Expertise and Decision Support* (Plenum, New York), 155–180.

Soll JB, Klayman J (2004) Overconfidence in interval estimates. *J. Experiment. Psych.: Learn., Memory, Cognition* 30(2):299–314.

Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (Doubleday, New York).

Teigen KH, Jørgensen N (2005) When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Appl. Cognitive Psych.* 19(4):455–475.

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

Anil Gaba is the Orpar Chaired Professor in Risk Management and academic director of the Centre for Decision Making and Risk Analysis at INSEAD. His research interests are in judgments under uncertainty, with current focus on combining judgments and on group versus individual judgments.

Ilia Tsetlin is professor of decision sciences at INSEAD. He is interested in prescriptive decision making emerging from normative analysis. His recent research focuses on generic properties of preferences (multiattribute utility, stochastic dominance) and on search, deadlines, and the role of uncertainty.

Robert L. Winkler is the James B. Duke Professor in the Fuqua School of Business and the Department of Statistical Science at Duke University. His research interests include decision analysis, Bayesian statistics, forecasting, and risk analysis. Recent work involves probability forecasting, the combination of probabilities, probability evaluation, stochastic dominance, and multiattribute utility.