HARVARD | BUSINESS | SCHOOL

Averaging Probability Forecasts: Back to the Future

Robert L. Winkler
Kenneth C. Lichtendahl Jr.
Yael Grushka-Cockayne
Victor Richmond R. Jose

Working Paper 19-039

Averaging Probability Forecasts: Back to the Future

Robert L. Winkler
Duke University

Yael Grushka-Cockayne
Harvard Business School

Kenneth C. Lichtendahl Jr.
University of Virginia

Victor Richmond R. Jose
Georgetown University

Working Paper 19-039

Averaging Probability Forecasts: Back to the Future

Robert L. Winkler

The Fuqua School of Business, Duke University, Durham, NC 27708, rwinkler@duke.edu

Yael Grushka-Cockayne

Harvard Business School, Harvard University, MA 02163, ygrushkacockayne@hbs.edu and

Darden School of Business, University of Virginia, Charlottesville, VA 22903, grushkay@darden.virginia.edu

Kenneth C. Lichtendahl Jr.

Darden School of Business, University of Virginia, Charlottesville, VA 22903, lichtendahlc@darden.virginia.edu

Victor Richmond R. Jose

McDonough School of Business, Georgetown University, Washington, DC 20057, vrj2@georgetown.edu

Abstract: The use and aggregation of probability forecasts in practice is on the rise. In this position piece, we explore some recent, and not so recent, developments concerning the use of probability forecasts in decision making. Despite these advances, challenges still exist. We expand on some important challenges such as miscalibration, dependence among forecasters, and selecting an appropriate evaluation measure, while connecting the processes of aggregating and evaluating forecasts to decision making. Through three important applications from the domains of meteorology, economics, and political science, we illustrate state-of-the-art usage of probability forecasts: how they are aggregated, evaluated, and communicated to stakeholders. We expect to see greater use and aggregation of probability forecasts, especially given developments in statistical modeling, machine learning, and expert forecasting; the popularity of forecasting competitions; and the increased reporting of probabilities in the media. Our vision is that increased exposure to and improved visualizations of probability forecasts will enhance the public's understanding of probabilities and how they can contribute to better decisions.

Key words: probability forecasts, forecast combination, forecast evaluation, decision analysis

Date: September 5, 2018

1. Introduction

Multiple opinions or estimates are available in a wide variety of situations. For example, we get second (or more) opinions when dealing with serious medical problems. We even do this for less serious decisions, such as when looking at multiple reviews of products on amazon.com or hotels and restaurants on tripadvisor.com. The motivation is that each additional opinion can provide more information, just as additional data points provide more information in a statistical study. Also, there is safety in numbers in the sense that considering multiple opinions can reduce the risk of a bad decision.

The same motivation extends to forecasts. When trying to forecast the path of a hurricane, for instance, weather forecasters consult forecasts from multiple meteorological models, considering the forecast path for the storm from each model. Often, these forecasters will create an average of the different forecast paths to provide a summary measure. Individuals, firms, and government agencies are increasingly comfortable relying on multiple opinions when forming estimates for key variables in important decisions. IARPA, the United States Intelligence Advanced Research Projects Activity, for instance, has invested heavily in multi-year research aimed at improving the government's use of crowds of forecasters for developing more accurate geopolitical forecasts (IARPA 2010, 2016).

The academic literature on the benefits of aggregating multiple opinions is vast, dating back at least to Galton (1907), who combined estimates of the weight of an ox at a county fair. The primary focus has been on the collection and aggregation of point forecasts. An important early paper was Bates and Granger (1969), who propose a method for determining the weights in a weighted average. Clemen (1989) and Armstrong (2001) provide reviews of the literature on averaging point forecasts. The public's fascination with the topic is evident through the success of popular press books, such as Surowiecki (2005) on the "wisdom of crowds." The literature has grown exponentially, supporting the eruption in uses of all things crowds, e.g., "crowdsourcing" (Howe 2006) or "crowdfunding" (Belleflamme et al. 2014).

The focus in this paper is on averaging probability forecasts. "Averaging" will often refer to a simple average. With some abuse of terminology, however, we will use "averaging" to represent any method for combining probability forecasts, just as "average income" is often used to represent not just a simple average, but a median, mode, or other summary measure of location for a set of data on incomes. When it is important to do so, we will be more specific about the exact nature of the aggregation technique. We will use "averaging," "aggregating," and "combining" interchangeably to represent any method for combining probability forecasts. A probability forecast might refer to a single probability for the occurrence of a binary event, a complete probability mass or density function (pmf or pdf), or a cumulative distribution function (cdf). Here too, we will be more specific about the form when it is important to distinguish between various types of forecasts.

In this paper, we will highlight the importance of working with and aggregating multiple probability forecasts, and emphasize some key challenges that remain. The paper is intended as a position piece, not a review paper. Thus, we will provide appropriate references as needed but will not offer a comprehensive review of past work. Also, we will discuss various techniques for combining and evaluating probability forecasts, but not provide comprehensive lists of such techniques. The intent is to offer insights on important issues related to working with probability forecasts, particularly on their aggregation and evaluation. The forecasts being combined can come from various sources, including models, data, and human experts. For example, an average forecast for an election might combine model-based forecasts based on previous voting trends, forecasts from polling data, and subjective forecasts from experts.

Because probability forecasts provide a measure of uncertainty, they are much more informative and more useful for decision making under uncertainty than point forecasts. Much of the early work on averaging forecasts involved point forecasts, and the wisdom-of-crowds phenomenon is generally thought of as a characteristic of averaging point forecasts. Work on averaging point forecasts has informed how to think about averaging probability forecasts, and we will refer to results from averaging point forecasts at times to illustrate certain ideas. Averaging probability forecasts, however, adds an extra layer of complexity.

Increasing interest in probability forecasts was stimulated by Savage (1954) and the growth of Bayesian methods and decision theory/decision analysis, which are inherently probabilistic. Today, forecasts in the form of complete probability distributions are used by highly visible players such as Nate Silver and his FiveThirtyEight blog (fivethirtyeight.com) and by crowd prediction platforms such as Google-owned Kaggle (www.kaggle.com). The existence of ample data and the increased sophistication of forecasting and prediction techniques made possible by advances in computing have resulted in cheaper and quicker ways for firms to generate such probability forecasts. While probabilities, as compared to point estimates, are more complex to elicit, evaluate, and aggregate, and are harder for a layperson to understand, they do contain richer information about potential futures. Such information can be key for protection from poor decision making.

One of the most common ways to aggregate probability forecasts is the linear opinion pool, introduced by Stone (1961) and attributed by some to Laplace. It is a weighted average of the forecasts, which is a simple average if the weights are equal. Much has already been written and surveyed on aggregation mechanisms for probability forecasts. For reviews, see Genest and Zidek (1986), Cooke (1991), Clemen and Winkler (1999), and O'Hagan et al. (2006).

In Section 2, we will consider themes related to aggregation of probability forecasts. In Section 3, we will consider methods designed to evaluate probability forecasts. We will demonstrate the usefulness of working with and evaluating aggregate probability forecasts with three important applications in Section 4. In Section 5 we will aim, insofar as possible, to offer prescriptive advice about what we believe decision makers should or should not do when it comes to making the best possible use of all that probability forecasts have to offer, and we will provide some views on the future of probability forecasting and the aggregation of probability forecasts. Our intention is to offer inspiration for researchers in the field, as well as some prescriptive guidelines to practitioners working with probabilities.

2. Aggregation of Probability Forecasts

In this section we will consider some important issues that can affect the benefits of aggregating probabilities and influence the choice of methods for generating the probabilities. Some of the same issues arise when aggregating point forecasts, but may be more complex and less understood when we are dealing with probability forecasts. Greater familiarity with these issues and how they impact forecast quality can lead to improvements in probability forecasts.

2.1. Miscalibration of the probability forecasts

In practice, probability forecasts are often poorly calibrated. When the forecasts are subjective, this poor calibration tends to be characterized by probability distributions that are too tight (e.g., realizations tend to be in the tails of the distribution more often than the distributions suggest they should be). This is typically attributed to overconfidence on the part of the forecasters. It can also occur when the forecasts are model-generated forecasts, in which case the attribution is to overfitting (Grushka-Cockayne et al. 2017a). In either case, the net result is that the forecasts are understating the uncertainty present in the forecasting situation. This in turn can cause decision makers using the forecasts to think that there is less risk associated with a decision than is really the case.

In principle, probability forecasts can be recalibrated to correct for miscalibration (Turner et al. 2014). However, it can be difficult to estimate the degree of miscalibration, which can vary considerably among forecasters and over time, and therefore to recalibrate properly. Complicating matters further is the result that averaging perfectly calibrated forecasts can lead to probability distributions that are underconfident, or not tight enough (Hora 2004, Ranjan and Gneiting 2010). More generally, the averaging may reduce any overconfidence, possibly to the point of yielding underconfident forecasts. Aggregation methods other than the simple average can behave differently (Lichtendahl et al. 2013b, Gaba et al. 2017), and miscalibration can also be affected by the issues discussed in the following subsections. These issues are all challenging when we aggregate.

2.2. Dependence among forecast errors

It is common to see dependence among forecasters, as indicated by positive correlations among forecast errors. We generally solicit forecasts from individuals who are highly knowledgeable in the field of interest. However, such experts are likely to have similar training, see similar data, and use similar forecasting methods, all of which are likely to create dependence in their forecasting errors.

This sort of dependence creates redundancy in the forecasts, which can greatly limit any increases in accuracy due to aggregation (Clemen and Winkler 1985). When the correlations are very high, as they often are, some commonly used aggregation methods yielding weighted averages of forecasts can have highly unstable and questionable weights, including negative weights or weights greater than one. Winkler and Clemen (1992) illustrate the impact of this phenomenon when combining point forecasts.

What can be done when aggregation methods using weighted averages provide unrealistic weights? Even though "better" experts would seem to deserve higher weights, identifying such experts can be difficult, and can be counterproductive in terms of improving the accuracy of the aggregated forecast. For example, if two "better" forecasters are highly dependent, including both of them will just include the same information twice. A combination of one of them with a less accurate forecaster who is not highly correlated with them can provide a better aggregated forecast. Alternatively, in terms of modeling, we can constrain the weights to be between zero and one or simply avoid weights entirely, using a simple average.

When combining point estimates, it has been shown that in order to reduce such dependence, we should aim for diversity among forecasters to the extent possible. This means including forecasters who differ in their forecasting style and methods, their relevant experiences, the data sets to which they have access, etc. Trading off some individual accuracy for reductions in dependence can be desirable, as noted by Lamberson and Page (2012). The challenge here is to find forecasters who have relevant expertise but also different viewpoints and approaches. Note that the desire for diversity in the forecasters is similar to the desire for diversification in investing. The motivation for diverse and independent opinions is key in the development of modern machine learning techniques. For instance, the random forest approach (Breiman 2001) generates multiple individual forecasts (trees), each based on a random subsample of the data and a subset of selected regressors, by design trading off individual accuracy for reduced dependence.

Larrick and Soll (2006) coin the term "bracketing" to describe the type of diversity that leads to an improved aggregate point forecast. Grushka-Cockayne et al (2017b) extend the notion of bracketing to probability forecasts and demonstrate how the recommended aggregation mechanism is impacted by the existence of bracketing among the forecasters' quantiles.

2.3. Instability in the forecasting process

A difficulty in trying to understand the forecasting process is instability that makes it a moving target. A prime source of this instability involves forecast characteristics. For subjective forecasts, learning over time can lead to changes in a forecaster's approach to forecasting and to characteristics such as accuracy, calibration, overconfidence, correlations with other forecasters, etc. These characteristics can also change as conditions change. For example, a stock market forecaster who produces good forecasts in a rising market might not do so in a declining market. For model-based forecasts, new modeling techniques and greater computer power can change the nature of a forecaster's modeling. As a result of this instability, uncertainty about forecast characteristics, which may be quite high when no previous evidence is available, might not be reduced too much even after data from previous forecasts are collected.

Not a lot can be done to remedy instabilities like these. The challenge, then, is to try to take account of them when aggregating forecasts. In building a Bayesian forecasting model, for instance, this suggests the use of a prior that suitably reflects the uncertainties, which may be difficult to assess. Machine learning algorithms, and the data scientists who use them, focus on avoiding overfitting their models to the data at hand by testing their models' accuracy on out of sample predictions.

2.4. How many forecasts should be combined?

The question of how many forecasts to aggregate is like the age-old question of how large a sample to take. In statistical sampling, where independence is generally assumed, there are decreasing improvements in accuracy as the sample size is increased. When positive dependence among forecast errors is present, the improvements in accuracy decrease more rapidly as the degree of dependence increases. For example, with the model of exchangeable forecasters in Clemen and Winkler (1985), the accuracy in combining k forecasts with pairwise error correlations of ρ is equivalent in the limit as $k \to \infty$ to accuracy when combining $1/\rho$ independent forecasts with the same individual accuracy. With $\rho = 0.5$, not an unusually high correlation, combining any number of forecasts will always be equivalent to less than combining 2 independent forecasts. Unless ρ is small, little is gained by averaging more experts.

Of course, these results are based on an idealized model. Empirical studies of aggregating actual probability forecasts (e.g., Hora 2004, Budescu and Chen 2015, Gaba et al. 2017) suggest that k between 5 and 10 might be a good choice. Most potential gains in accuracy are typically attained by $k = 5$ and smaller gains are achieved in the 6-10 range, after which any gains tend to be quite small. Some might be surprised that small samples of forecasts like this are good choice. However, Figure 2 shows that even with moderate levels of dependence, gains from additional forecasts can be quite limited. When obtaining forecasts is costly and time-consuming, the challenge is to find the number of forecasts providing an appropriate tradeoff between accuracy of an aggregated forecast and the costs of obtaining the individual forecasts.

2.5. Robustness and the role of simple rules

There are many ways to aggregate probability forecasts. At one extreme are basic rules using summary measures from data analysis: the mean of the forecasts, the median, a trimmed mean, etc. The most common method in practice is just the mean, a simple average of the forecasts. It can be generalized to a weighted average if there is reason to give some forecasts greater emphasis. At the other extreme are complex methods using statistical modeling, stacking, machine learning, and other sophisticated techniques to aggregate the forecasts.

One might think that the more sophisticated methods would produce better forecasts, and they often can, but they face some nontrivial challenges. As we move from simple models to more sophisticated models, careful modeling is required and more parameters need to be chosen or estimated, with relevant past data not always available. These things do not come without costs in terms of money and time. They also lead to the possibility of overfitting, especially given potential instabilities in the process that cause the situation being forecasted to behave differently than past data would imply.

The more sophisticated rules, then, can produce superior forecasts, but because of the uncertainties and instability of the forecasting process, they can sometimes produce forecasts that perform poorly. In that sense, they have both an upside and a downside and are thus more risky.

Simpler rules such as the simple average of the forecasts are worthy of consideration. They are very easy to understand and implement and are very robust, usually performing quite well. For combining point forecasts, an example of a simple and powerful rule is the trimmed mean (Jose and Winkler 2008). Robust averages like the trimmed mean have been shown to work when averaging probabilities as well (Jose et al. 2014, Grushka-Cockayne et al. 2017a).

Simple rules are often touted as desirable because they perform very well on average, but they also perform well in terms of risk reduction because of their robustness. They won't necessarily match the best forecasts but will generally come close while reducing the risk of bad forecasts. Even moving from a simple average to a weighted average of forecasts can lead to more volatile forecasts, as noted above. The challenge is to find more complex aggregation procedures that produce increased accuracy without the increased risk of bad forecasts.

2.6. Summary

The issues described in this section pose important challenges present when aggregating probability forecasts. Moreover, these issues interact with each other. For example, the presence of instability in the underlying process can increase the already difficult tasks of trying to estimate the degrees of miscalibration and dependence associated with a given set of forecasts, and adding more forecasters can complicate things further. The good news is that just being aware of these issues can be helpful, and more is being learned about them and how to deal with them.

3. Evaluation of Probability Forecasts

Too often forecasts are made but soon forgotten and never evaluated after the actual outcomes are observed. This is true for all forecasts but is especially so for probability forecasts. Sometimes this is intentional. We might remember forecasts that turned out to look extremely bad or extremely good, and the source responsible for such forecasts might brag proudly about a good forecast and try to avoid mentioning a forecast that turns out to be bad. That's how soothsayers and fortune tellers survive.

Most of the time the lack of record-keeping and evaluation of forecasts is not due to any self-serving motive. Some might believe that once the event of interest has occurred, there is no need to conduct a formal evaluation or keep records. However, keeping track of forecasts and evaluating them after we learn about the corresponding outcomes is important for two reasons. First, it provides a record of how good the forecasts were and makes it possible to track forecast performance over time. Second, it can encourage forecasters to improve future forecasts and, with appropriate evaluation measures, can help them learn how they might do so.

Common evaluation measures for point forecasts, such as mean square error (MSE), are well known and easy to understand. There is less familiarity with how probability forecasts can be evaluated, in part because evaluating probability forecasts has not been very common and in part because the evaluation measures are a little more complex than those for point forecasts. Different measures are needed for different types of forecasts (e.g., probabilities for single events versus entire probability distributions). In this section we will discuss some issues related to the evaluation of probability forecasts.

3.1. Selecting appropriate evaluation measures

The primary measure of "goodness" of a probability forecast used in practice is a strictly proper scoring rule, which yields a score for each forecast. For example, with a forecast of the probability of rain, a scoring rule is strictly proper if the forecaster's ex ante expected score is maximized only when her reported probability equals her "true probability." An early strictly proper scoring rule developed by a meteorologist to discourage weather forecasters from "'hedging' or 'playing the system'" (Brier 1950, p. 1) is the Brier score. It is a special case of the commonly used quadratic scoring rule. Another early rule is the logarithmic rule (Good 1952), which has connections with Shannon entropy. For some reviews of the scoring rule literature, see Winkler (1996), O'Hagan et al. (2006), and Gneiting and Raftery (2007).

One thing influencing the choice of an evaluation measure is the nature of the reported probability forecast. The quadratic and logarithmic scores for probabilities of a single event such as the occurrence of rain have extensions to probabilities of multiple events and to discrete and continuous distributions for a random variables. For random variables, straightforward scoring rules designed for forecasts of the pmf or pdf are supplemented by rules designed for the cdf, such as the continuous ranked probability score (CRPS) based on the quadratic score (Matheson and Winkler 1976). Rules based on the cdf take into account the ordering inherent in the variable of interest.

Not all scoring rules used in practice are strictly proper. For instance, a linear scoring rule with a score equal to the reported probability or density for the actual outcome (e.g., rain or no rain) sounds appealing, but it incentivizes the reporting of probabilities of zero or one. A rule developed in weather forecasting to evaluate a forecast relative to a benchmark, or baseline, forecast (often climatology, which is the climatological relative frequency) is the skill score, which is the percentage improvement of the Brier score for the forecast relative to the Brier score for climatology. A percentage improvement like this seems intuitively appealing, but it is not strictly proper. If the Brier score is transformed linearly to another quadratic score with different scaling, the resulting quadratic score is strictly proper. A skill score based on that quadratic score is not strictly proper, however.

One issue arising with the most common strictly proper scoring rules is that the resulting scores are not always comparable across forecasting situations. For all strictly proper scoring rules, the forecaster's expected score with "honest" forecasting as a function of the value of the forecast probability is a convex function. With a probability forecast of a single probability such as the probability of rain, this convex function is symmetric on [0,1], minimized when the probability is 0.5, and maximized at 0 and 1. Thus, a forecaster in a location with a baseline near 0.5 will tend to have lower scores than a forecaster in a location with a baseline near 0 or 1, so their scores are not really comparable. A family of strictly proper asymmetric scores based on the quadratic score shifts the expected score function with honest forecasting so that it is minimized at the baseline forecast and has different quadratic functions above and below that baseline (Winkler 1994). This makes the scores for forecasters at different locations more comparable, and the asymmetric rule can be based on any strictly proper rule, not just the quadratic rule.

A final issue in choosing a scoring rule is that it should fit not just the situation, but the way the probability forecast is reported. If a forecast is for a discrete random variable and the forecaster is asked to report probabilities for the possible values (a pmf), the rules discussed above are appropriate. If the forecaster is asked to report quantiles (a cdf), those rules will not provide the proper incentives despite the fact that once either the pmf or cdf is known, the other can be determined. In the first case, the scores are based on probabilities, which are on [0,1]; in the second case, the scores are based on quantiles, which depend on the scaling of the random variable. Strictly proper scoring rules for quantiles are developed in Jose and Winkler (2009).

Grushka-Cockayne et al. (2017b) encourage the use of quantile scoring rules. Focusing on evaluating the performance of the aggregate forecast, they suggest that the score of a crowd's combined quantile should be better than that of a randomly selected forecaster's quantile only when the forecasters' quantiles bracket the realization. If a score satisfies this condition, we say it is sensitive to bracketing.

3.2 Using multiple measures for evaluation

A strictly proper scoring rule is an overall measure of the accuracy of probability forecasts and is therefore the most important type of evaluation measure. Just as it is helpful to consider multiple forecasts for the same uncertain situation, it can be helpful to consider multiple scoring rules for a given situation. We do not combine the scores from different rules, but they provide slightly different ways of evaluating the forecasts. Thus, using multiple scoring rules when evaluating individual or aggregate probability forecasts can be helpful.

In addition to the overall evaluation provided by scoring rules, measures for certain forecast characteristics of interest such as calibration and sharpness are important in order to better understand different characteristics of individual forecasts and aggregate forecasts. Calibration involves whether the forecasts are consistent with the outcomes. Sharpness involves how variable the forecasts are, and is not connected with the outcomes. A goal to strive for in probability forecasting is to maximize the sharpness of the probabilities while maintaining good calibration (Gneiting and Raftery, 2007).

Strictly proper scoring rules can be related to measures of calibration and sharpness through decompositions of the rules into components, with a common decomposition expressing a scoring rule as a function of a calibration measure and a sharpness measure. For a quadratic scoring rule, the overall score equals the sum of three components: the score under perfect calibration and sharpness, a calibration measure (a penalty representing the degree of miscalibration), and a sharpness measure (a penalty representing the lack of sharpness). Both penalties are non-positive and are zero only for perfect forecasts, which are forecasts providing a degenerate distribution that puts probability one on the value that actually occurs.

For a probability forecast of an event occurring, calibration can be expressed graphically. A calibration diagram is a plot of the relative frequency of occurrence of the event as a function of the probability forecast. Perfect calibration is represented by the 45˚ line on the graph, and deviations from that line represent miscalibration. Of course, because of sampling error, we would not expect the plot to follow the 45˚ line exactly.

An important issue discussed in Section 2 is overconfidence, which occurs when probability forecasts are miscalibrated in the sense of being too extreme. On a calibration diagram, that corresponds to relative frequencies above (below) the 45˚ line for low (high) probabilities.

For probability forecasts of a continuous quantity, calibration can be expressed graphically with a probability integral transform (PIT) chart. A PIT chart is a histogram of historical cdfs evaluated at the realization. Perfect calibration is represented by a uniform histogram. A bathtub-shaped PIT chart indicates overconfidence, while a hump-shaped PIT chart indicates underconfidence.

3.3 Relating forecast evaluation to the economic setting

When probability forecasts are made in a decision-making problem, it would be nice if the scoring rule could be related in some manner to the problem itself. The general measures of accuracy provided by standard scoring rules and components of them such as calibration and sharpness are useful in any situation. However, a rule connected to the specific problem at hand could be even more useful, just as a loss function related to the utilities in a given problem is more appropriate than the ubiquitous quadratic loss function for point estimation.

An early note by McCarthy (1956) suggests that a scoring rule can be connected directly to a decision-making problem. Building on this idea, Savage (1971, p. 799) considers scoring rules viewed as a share in a business and states that in principle, "every such share leads to an at least weakly proper scoring rule," at the same time indicating uncertainty about the practicality of such a scheme.

In the spirit of business sharing, Johnstone et al. (2011) develop tailored scoring rules designed to align the interest of the forecaster and the decision maker. Analytical expressions for the scoring rules are developed for simple decision-making situations but it is necessary to express the rules in numerical form when the problems get at all complex. The complexities could involve the decision-making problem itself (e.g. the structure of the problem, the uncertainties, or the nature of the decision maker's utility function). For other than relatively simple decision-making problems, it may be infeasible to abandon the standard scoring rules in an attempt to develop tailored rules.

3.4 Evaluating Probability Forecasts in Competitive Settings

When Galton (1907) elicited estimates for the weight of the ox, he offered a reward to the farmer with the closest estimate to the real weight. It was a competition. Today, technology enables firms to collect forecasts from experts, from their employees, or from the public, through forecasting competitions. Platforms such as Kaggle, HeroX, and CrowdANALYTIX offer firms creative ways to set up prediction challenges, share data, and offer high rewarding prizes. The 2006 $1 Million Netflix Prize (Bennett and Lanning 2007) is perhaps the most well-known point forecasting competition in recent years. The Global Energy Forecasting Competition is an example of popular probability forecasting competition (Hong et al. 2016).

In such settings, participants submit their forecasts with the goal of winning a prize or achieving high rank recognition. When forecasters compete against each other, their motivation often becomes more about relative performance than absolute performance. This will be even more pronounced when leaderboards are made publically available for all to see. Such winner-takes-all formats imply that proper scoring rules are no longer proper. With point forecasting, Lichtendahl et al. (2013a) show that individuals who compete should exaggerate their forecasts in order to stand out and beat others. Lichtendahl and Winkler (2007) show this for probability forecasting. A competitive forecaster who wants to do better than others will report more extreme probabilities, exaggerating toward zero or one.

Lichtendahl and Winkler (2007) also develop joint scoring rules based on business sharing, showing that these scoring rules are strictly proper and overcome the forecasters' competitive instincts and behavior. Witkowski et al. (2018) suggest the Event-Lotteries Forecaster Selection Mechanism, a mechanism by which forecasting competitions can be incentive-compatible, rewarding the top performer as well as rewarding truth telling.

Prediction markets, building on the notion of efficient markets from finance and on sports betting markets, have often been proposed as an alternative to combining mechanisms and forecasting competitions. The market provides the incentive role of a scoring rule, and market dynamics take care of the aggregation of the participants' implicit probability forecasts. For example, participants in a prediction market for an election can buy "shares" in the candidates, where a share in the candidate who wins pays $1 and shares in other candidates pay $0. The price of a candidate's shares at a given time represents an aggregate probability of that candidate winning, and plots of the prices over time show the changes in the probabilities. This has been implemented, e.g., in the Iowa Electronic Markets operated by the University of Iowa (Wolfers and Zitzewitz 2004).

Bassamboo et al. (2018) demonstrate the use of prediction markets in forecasting quantities important to operational decisions, such as sales forecasts, price commodity forecasts, or product features. Atanasov et al. (2017) compare the performance of forecasting competitions to prediction markets. They show that while prediction markets are initially more accurate, forecasting competitions can improve with feedback, collaboration, and aggregation.

3.5 Summary

With greater computer power and interest in analytics, probability forecasts are encountered and used more frequently, a welcome trend. However, most of these forecasts are never evaluated formally, so an opportunity to learn from past performance is being lost. The three applications we will discuss in Section 4 are notable exceptions. For example, the U.S. National Weather Service (NWS) is a pioneer not only in making probability forecasts on a regular basis and issuing them to the general public, but also in the systematic evaluation of these forecasts (Murphy and Winkler 1984). They have used the Brier score to evaluate probabilities of precipitation for over 50 years, and the forecasters see their scores. Such scores are not only useful for decision makers to evaluate forecasters, but even more so to help forecasters learn from their good and bad scores and improve their future forecasts.

Part of the problem with lack of use of evaluations is a lack of widespread understanding of methods for evaluating probabilities. When point forecasting is taught, evaluation is commonly included, using measures like MSE and MAE. When students learn about probabilities, they seldom learn about evaluating them, and any evaluation numbers they encounter seem like they came from a black box. The wide array of potential scoring rules for different situations can be confusing, but teaching the basic scoring rules is not difficult, and it helps if they are decomposed into calibration and sharpness terms.

As the use of probability forecasts increases, there is promise for increasing use of evaluations of more of these forecasts. The starting point is greater understanding of the evaluation options by the forecasters themselves and increasing demand for evaluations from users. For example, the incentives for making better decisions under uncertainty should lead to some ex post focus on how the probabilities impacted the decision and whether probabilities could be improved in future decisions. Steps like creating leaderboards for the increasing number of forecasting competitions give exposure to evaluation metrics and further motivate the development of better forecasting techniques.

4. Applications

To illustrate recent use of some of the ideas discussed in Sections 1-3 in important situations, we will consider three applications: hurricane path prediction, macroeconomic forecasting, and forecasts of future geopolitical events. These applications illustrate the increasing use of probability forecasts and the aggregation of such forecasts. More importantly, they demonstrate the importance of probability forecasts in challenging decision-making situations. They also demonstrate the potential for more widespread consideration of such forecasts, their dissemination to the public where appropriate, and the importance of good visualization of the forecast and its uncertainty in dissemination.

4.1. Hurricane path prediction

One important forecasting application involving the aggregation of probability forecasts is to hurricanes. For any tropical cyclone that forms in the Atlantic Ocean, the U.S. National Hurricane Center (NHC) makes forecasts of its path for 12, 24, 36, 48, 72, 96, and 120 hours ahead. The NHC makes these forecasts every six hours, producing the well-known "cone of uncertainty". See Figures 1 and 2 for two high profile examples.

Figure 1 shows the cone for Hurricane Katrina, a category-5 storm that made landfall in 2005 near the city of New Orleans, killing 1,833 people. In Figure 2, we see Hurricane Maria's cone. Hurricane Maria hit the Caribbean island of Puerto Rico in 2017. This major storm, also a category-5 hurricane, is estimated to have killed 4,645 people, according to a Harvard study (Kishore et al. 2018).

The NHC's forecasts come from an "ensemble or consensus model"—a combination of up to 50 model forecasts. Meteorologists were one of the first groups of forecasters to use the term "ensemble" and to take seriously the idea that better forecasts could be produced by averaging or aggregating multiple models' forecasts.

Some models included in the NHC's ensemble are dynamical, while some are statistical. Other models used by the NHC are hybrids of these two types of models. Dynamical models make forecasts by solving the physical equations of motion that govern the atmosphere. These models are complex and require a number of hours to run on a supercomputer. The statistical models, on the other hand, rely on "historical relationships between storm behavior and storm-specific details such as location and date".[1]

[1] NHC Track and Intensity Models, *U.S. National Hurricane Center*, accessed July 19, 2018 at https://www.nhc.noaa.gov/modelsummary.shtml.

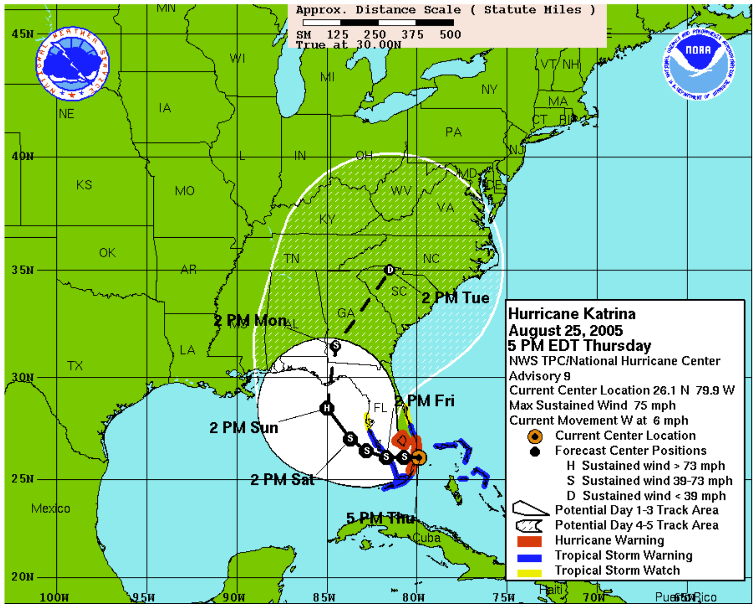


Figure 1. Cone of uncertainty at the time Hurricane Katrina first became a hurricane in 2005.[2]

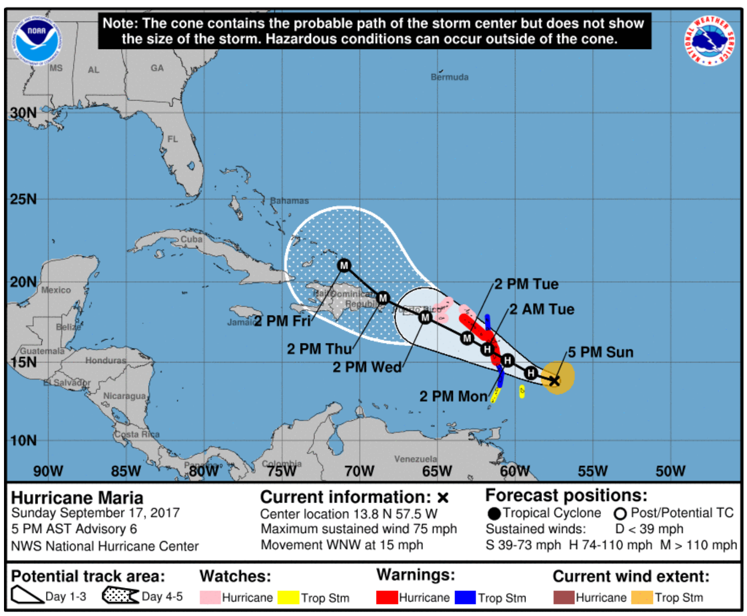


Figure 2. Cone of uncertainty at the time Hurricane Maria first became a hurricane in 2017.[3]

One thing to notice about these cones is that Hurricane Maria's is much narrower. Each storm's cone is the probable track of the center of the storm, along with a set of prediction circles. A cone's area is swept out by a set of 2/3 probability circles around the storm's most likely path. These probabilities are set

[2] KATRINA Graphics Archive, *U.S. National Hurricane Center*, accessed July 19, 2018 at https://www.nhc.noaa.gov/archive/2005/KATRINA_graphics.shtml.

[3] MARIA Graphics Archive, *U.S. National Hurricane Center*, accessed July 19, 2018 at https://www.nhc.noaa.gov/archive/2017/MARIA_graphics.php?product=5day_cone_with_line_and_wind.

so that 2/3 of the last five year's annual average forecast errors fall within the circle (U,S. National Hurricane Center 2017). We note that the cone is formed by a set of circles, rather than the set of intervals we typically see with time series, because the storm's location at a point in time is described by two dimensions—its latitude and longitude on the map.

In 2005, the annual average forecast error at 48 hours ahead was 101.2 nautical miles (1 nautical mile = 1.15 miles). By 2017, the annual average forecast error at 48-hours ahead had dropped to 52.8 nautical miles. Figure 3 depicts the dramatic improvements the NHC has achieved in the accuracy of its forecasts.

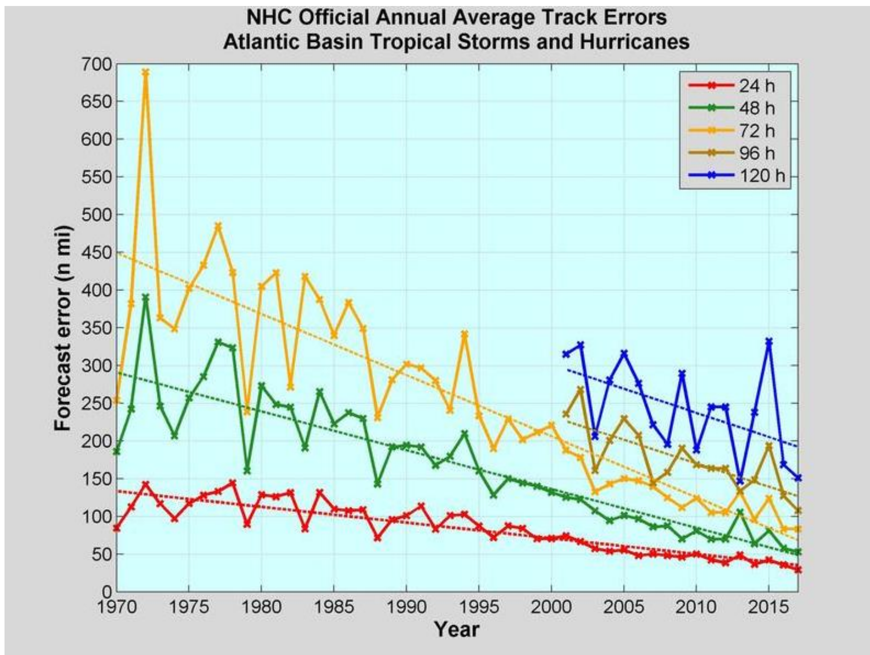


Figure 3. Annual average forecast errors (1970-2017). [4]

Since 2010, the cones of uncertainty have shrunk by 36%. The NHC attributes these improvements to "remarkable advances in science". Researchers at NHC have improved their models of atmospheric processes involving radiation and clouds. Computers run at higher resolutions, and satellites beam down clearer images of cloud tops. Narrower cones can have big impact on society. According to Jeff Masters, co-founder of Weather Underground, "Substantially slimmer cones mean fewer watches and warnings along coastlines … Since it costs roughly $1 million per mile of coast evacuated, this will lead to considerable savings, not only in dollars, but in mental anguish." (Miller 2018)

4.2 Macroeconomic forecasting

Since 1968, the U.S. Survey of Professional Forecasters (SPF) has asked many private-sector

[4] National Hurricane Center Forecast Verification: Official Error Trends, *U.S. National Hurricane Center*, accessed July 19, 2018 at https://www.nhc.noaa.gov/verification/verify5.shtml.

economists and academics to forecast macroeconomic quantities such as the growth in gross domestic product (GDP), the unemployment rate, and the inflation rate. Started by the American Statistical Association and the National Bureau of Economic Research (NBER), the survey has been conducted by the Federal Reserve Bank of Philadelphia since 1990. On the survey, the panelists are asked to make both point forecasts and probability forecasts (Croushore 1993).

A widely followed forecast from the survey is real GDP growth for the next five quarters ahead. The Philadelphia Fed aggregates the point forecasts made by the panelists and reports their median to the public shortly after it receives all the forecasts for the quarter. Let's take an example from the 2018:Q2 survey (the survey taken in the second quarter of 2018, collecting forecasts for 2018:Q2 and the following four quarters).

For example, the survey was sent to 36 panelists on April 27, 2018, and all forecasts were received on or before May 8, 2018. On May 11, 2018, the survey's results were released. The median forecasts of real GDP growth were 3.0%, 3.0%, 2.8%, 2.4%, and 2.6% for the next five quarters, respectively.[5] Based on these forecasts and the survey's historical errors, the "fan chart" in Figure 4 was created, showing forecasted quarter-to-quarter growth rates in real GDP.

These fan charts are not all that different in principle from the NHC's cones of uncertainty. Both are based on historical forecast errors, but the fan charts and cones are constructed differently. The Philadelphia Fed's fan is generated by overlaying central prediction intervals, covering from 25% probability up to 80% probability. These probabilities come from a normal distribution with mean equal to the median panelists' forecast and variance equal to the mean squared error of past forecasts (at the same horizon) over the period from 1985:Q1 to 2016:04.[6]

Another closely watched forecast is the distribution comprised of "mean probabilities" for real GDP growth. Panelists are asked to give probabilities over 11 pre-determined bins for annual real GDP growth in the next four years (including the current year). To aggregate these probabilities, the Philadelphia Fed averages the panelists' probabilities in each bin. In other words, they form a linear opinion pool. This opinion pool communicates information similar to the fan chart, but instead of using past point forecasting errors to describe the uncertainty in real GDP growth, the panelists' own forward-looking uncertainties are

[5] Survey of Professional Forecasters: Second Quarter 2018, Federal Reserve Bank of Philadelphia, accessed July 30, 2018 at https://www.philadelphiafed.org/-/media/research-and-data/real-time-center/survey-of-professional-forecasters/2018/spfq218.pdf?la=en.

[6] "Error Statistics for the Survey of Professional Forecasters for Real GNP/GDP", Federal Reserve Bank of Philadelphia, accessed July 30, 2018 at https://www.philadelphiafed.org/-/media/research-and-data/real-time-center/survey-of-professional-forecasters/data-files/rgdp/spf_error_statistics_rgdp_3_aic.pdf?la=en.

used. See Figure 5 for an example from the 2018:Q2 survey.[7]

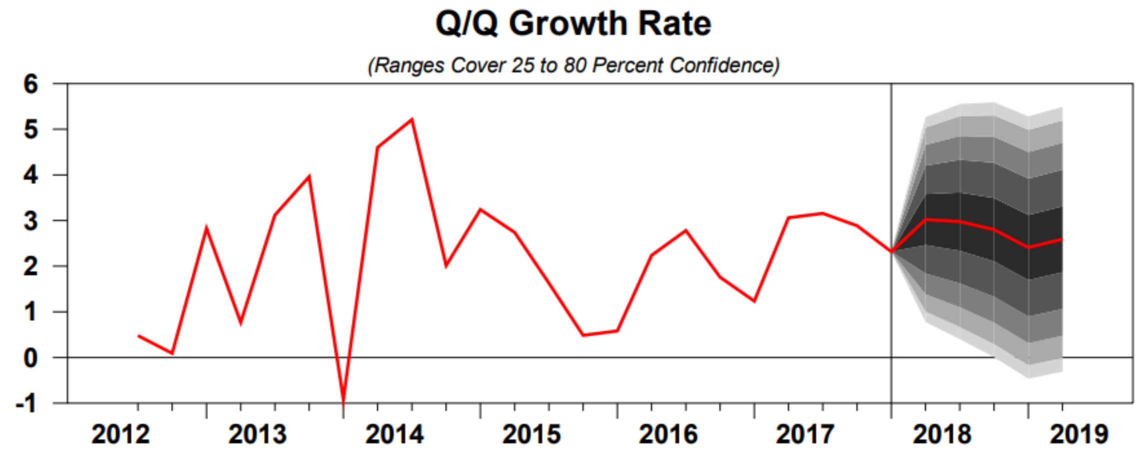


Figure 4. Fan chart of real GDP quarter-to-quarter growth rate, as of Quarter 2, 2018.

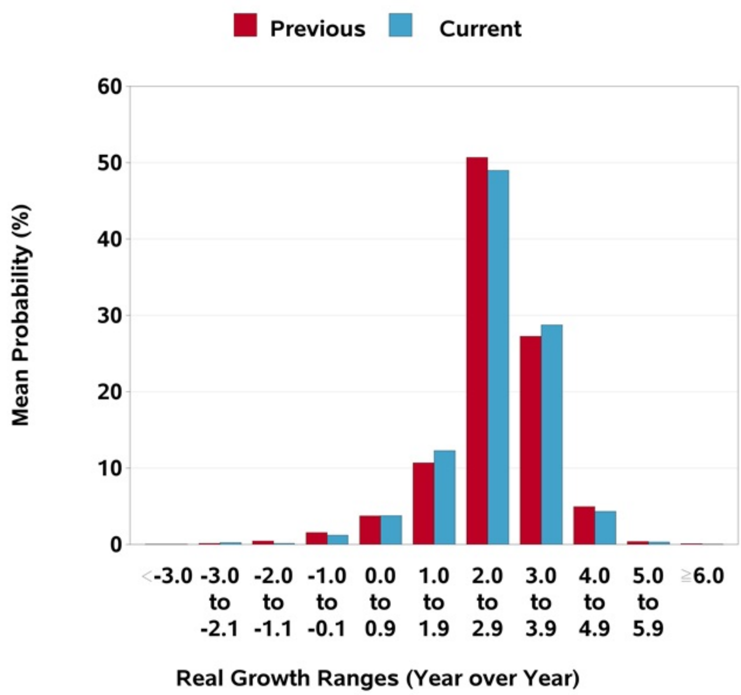


Figure 5. Mean Probabilities for Real GDP Growth in 2018, as of Quarter 2, 2018.

A related question asked on the survey is the probability of a decline in real GDP. According to the 2018:Q2 survey, the mean probability of a decline in real GDP was 0.053, 0.86, 0.111, 0.144, and 0.156 for quarters 2018:Q2 through 2019:Q2, respectively. Thus, as of early in 2018:Q2, the panel sees an

[7] Mean Probabilities for Real GDP Growth in 2018 (chart), Federal Reserve Bank of Philadelphia, accessed July 30, 2018 at https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/2018/survq218.

increasing chance of a decline in economic growth in the U.S over the next five quarters. The Philadelphia Fed refers to this forecast, namely the probability of a decline in real GDP in the quarter after a survey is taken, as the anxious index. In Q2 of 2018, the anxious index was 8.6 percent[8].

Figure 6, which is published by the Philadelphia Fed on their website, shows the anxious index over time. The shaded regions mark periods of recession as called by NBER. The index tends to increase before recessions, peaking during and declining after these periods.

An interesting point here is that the Philadelphia Fed uses the median (an extreme case of a trimmed mean) to aggregate point forecasts, whereas it uses a simple mean to aggregate probability distributions. As noted in Section 2.5, the use of trimmed means to average probability forecasts may lead to some improvements in accuracy when probability forecasts are evaluated with a proper scoring rule.

Another survey of business, financial, and academic economists is conducted monthly by the Wall Street Journal (WSJ). The survey asks for point and probability forecasts, using a simple mean to aggregate both types of forecasts. For example, in their survey of 57 economists conducted August 3-7, 2018, the average probability of a recession beginning in the next 12 months was 18%, the probability of a NAFTA pullout was 29%, and the probability of tariffs on autos was 31% (Zumbrun 2018).

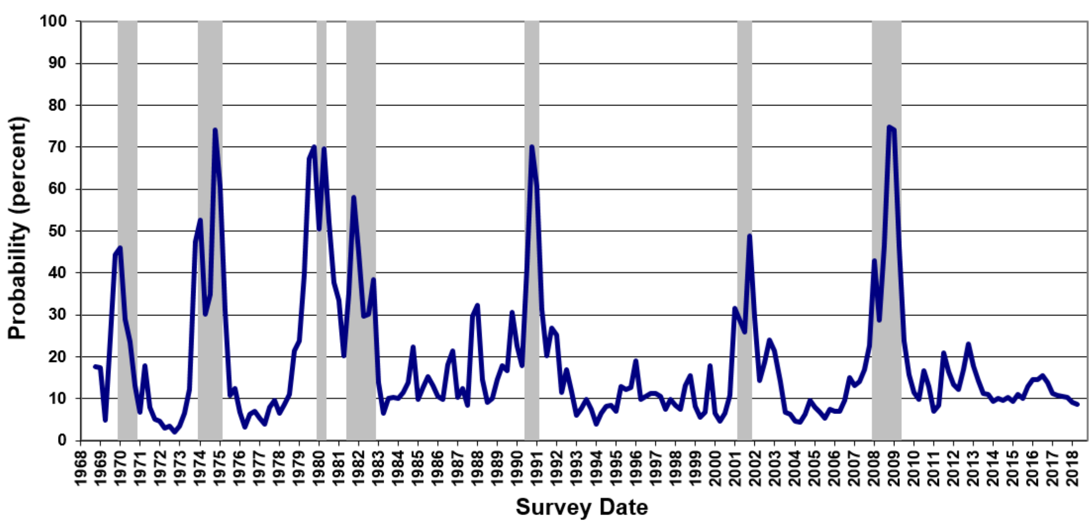


Figure 6. The SFP's Anxious Index 1968:Q4 – 2018:Q2.

4.3 Forecasts of future geopolitical events

In 2010, the U.S. Intelligence Advanced Research Projects Activity (IARPA) announced the start of a new research project, the Aggregative Contingent Estimation (ACE) Program. The focus of ACE was

[8] "The Anxious Index", accessed July 30, 2018 at https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/anxious-index.

on the development of innovative research related to efficient elicitation and aggregation of probability judgments and the effective communication of the aggregated probabilistic forecasts (IARPA 2010). IARPA's interest in this project was due to its heavy reliance on information, such as the likelihood of future geopolitical events, elicited from intelligence experts. Inspired by wisdom-of-crowds research, the agency was hoping that the accuracy of judgment-based forecasts could be improved by cleverly combining independent judgments.

The ACE program ran as a tournament and involved testing forecasting accuracy for real-time occurring events. Research teams from different institutions could test their elicitation and aggregation approaches against each other. The Good Judgment Project, a team based at the University of Pennsylvania and the University of California, Berkeley, was one of five research teams selected by IARPA to compete in ACE. The team, led by Philip Tetlock, Barbara Mellers, and Don Moore, officially began soliciting forecasts from participants in September of 2011. The Good Judgment Project was the ACE forecasting tournament winner, outperforming all other teams by forming more accurate forecasts by more than 50%. The tournament concluded in 2015 (Tetlock and Gardner 2015).

Throughout the competition, thousands of volunteers participated in predicting world events. Over 20 research papers were inspired by the data[9], hundreds of popular press pieces were published, bestselling books were authored, and the data from the project was made available in order to encourage further development of aggregation techniques[10]. The main development coming out of the Good Judgment Project is the idea of "superforecasting", which includes four elements: (1) identifying relative skill of the forecasts by tracking their performance ("talent-spotting"); (2) offering training to the participants in order to improve their forecasting accuracy, including learning about proper scoring rules; (3) creating diverse teams of forecasters; and (4) aggregating the forecasts while giving more weight to talented forecasters.

In 2015, the Good Judgment Project led to a commercial spinoff, Good Judgment Inc. Good Judgment Inc. offers firms access to its platform, enabling firms to crowdsource forecasts important to their business. The firm also publishes reports and indices compiled from forecasters made by a panel of professional superforecasters. In addition, Good Judgment Inc. runs workshops and training to help improve forecasting capabilities. Finally, Good Judgment Open is part of the Good Judgment Inc's website that is open to the public to participate in forecasting tournaments. Anyone interested can participate in forecasting geopolitical and worldwide events, such as entertainment and sports. Figure 7 presents forecasting challenges available to the public, providing a sense of the types of topics that are typical of Good Judgment Open. Figure 8 presents the consensus trend, which is the median of the most recent 40% of the forecasts. This type of feedback to forecasters is an example of good visualization. Figure 9 illustrates a leaderboard

[9] https://goodjudgment.com/science.html
[10] https://dataverse.harvard.edu/dataverse/gjp

maintained by the site to track and rank forecasters' performance, including feedback on Brier scores.

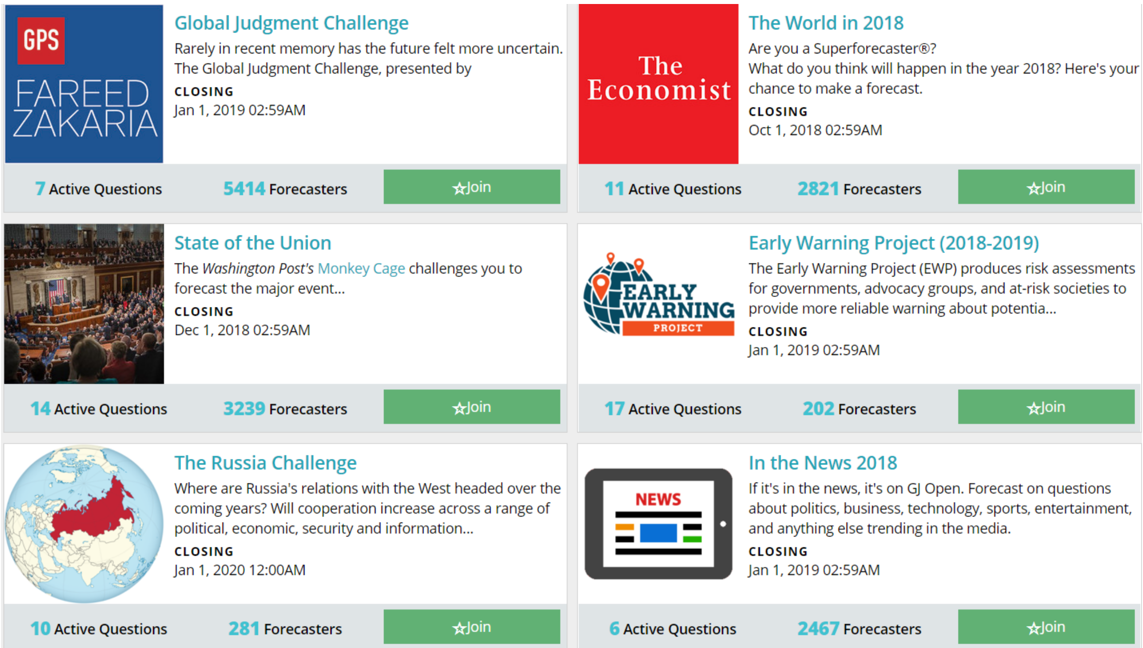


Figure 7. Current Forecasting Challenges on www.gjopen.com

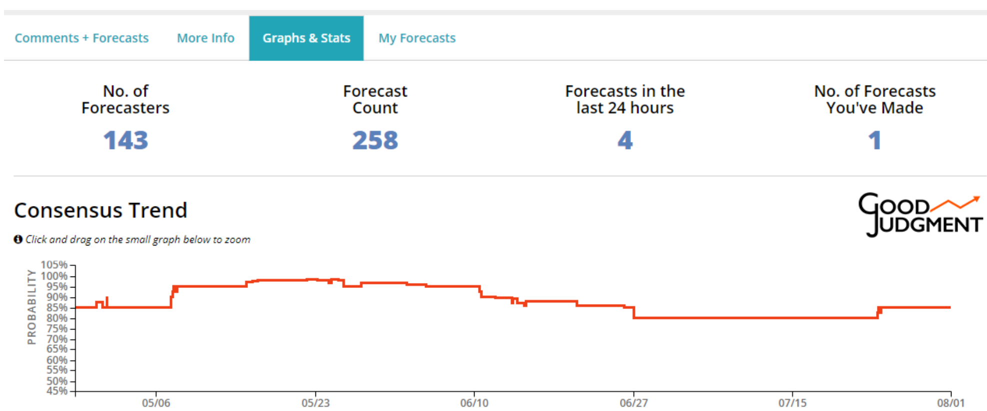
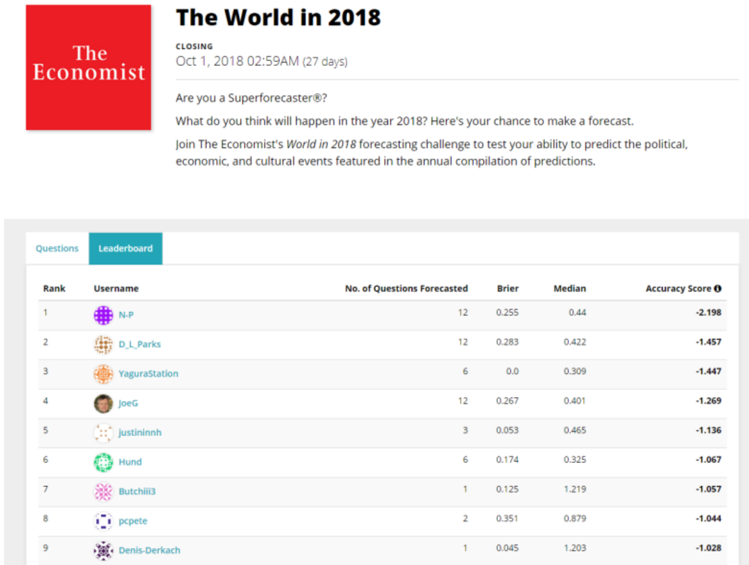


Figure 8. Consensus trend for the probability that Brazil's Workers' Party nominates a candidate other than Luiz Inacio Lula da Silva for president.



Rank	Username	No. of Questions Forecasted	Brier	Median	Accuracy Score
1	N-P	12	0.255	0.44	**-2.198**
2	D_L_Parks	12	0.283	0.422	**-1.457**
3	YaguraStation	6	0.0	0.309	**-1.447**
4	JoeG	12	0.267	0.401	**-1.269**
5	justininnh	3	0.053	0.465	**-1.136**
6	Hund	6	0.174	0.325	**-1.067**
7	Butchiii3	1	0.125	1.219	**-1.057**
8	pcpete	2	0.351	0.879	**-1.044**
9	Denis-Derkach	1	0.045	1.203	**-1.028**

Figure 9. Leaderboard for The World in 2018 Forecasting Challenge.

4.4 Summary

The three applications described above all involve multiple probability forecasts and the aggregation of those forecasts for situations of interest. Also, they all demonstrate the importance of generating effective visualization. However, the nature of the forecasts differ. First, we looked at forecasts of the path of a severe storm with uncertainty about how where it will move on the two-dimensional grid, how quickly it will move, and how strong it will be at different points along its path. Next, we considered probability distributions of macroeconomic quantities at fixed points of time in the future with updates. Finally, we described a project involving probabilities of important geopolitical events. Each of these applications provides probability forecasts that are very important for decision making, and each shows the increasing interest in probability forecasts to represent uncertainty.

The three applications also differ somewhat in how the forecasts are created. For weather forecasting, the human forecasters of the NWS have access to model-generated forecasts but can adjust those forecasts based on other inputs and their subjective judgments:

> The NWS keeps two different sets of books: one that shows how well the computers are doing by themselves and another that accounts for how much value the humans are contributing. According to the agency's statistics, humans improve the accuracy of precipitation forecasts by about 25% over the computer guidance alone, and temperature forecast by about 10%. Moreover, … these ratios have been relatively constant over time: as much progress as the computers have made, (the) forecasters continue to add value on top of it. Vision accounts for a lot. (Silver 2012, p. 125)

This sort of process also seems to be common among the panelists in the Philadelphia Fed's survey even though, unlike the weather forecasters, they are "independent contractors" and most likely do not all use the same models.

> In an optional last section of the special survey, we asked the panelists about their use of mathematical models in generating their projections, and how their forecast methods change, if at all, with the forecast horizon. … Overwhelmingly, the panelists reported using mathematical models to form their projections. However, we also found that the panelists apply subjective adjustments to their pure-model forecasts. The relative role of mathematical models changes with the forecast horizon. (Stark 2013, p. 2)

ACE and the Good Judgment Project present different types of situations, less amenable to mathematical modeling. The forecasters can use any means available to formulate their forecasts, but ultimately the forecasts are subjective because they typically involve one-off events. Moreover, although the forecasts in the first two applications are primarily model based, they too are ultimately subjective. The forecasters can and often do adjust the model-generated forecasts that are available, and the building of these models depends on subjective choices for methods and parameters in the first place.

In 2016 IARPA announced their follow up study to ACE, the Hybrid Forecasting Competition (HFC). Similar to ACE, the HFC focused on forecasting geopolitical events. This time, however, IARPA was interested in studying the performance of hybrid forecasting models, combining human and machine, or model-based, forecasting. According to the study announcement:

> Human-generated forecasts may be subject to cognitive biases and/or scalability limits. Machine-generated forecasting approaches may be more scalable and data-driven, but are often ill-suited to render forecasts for idiosyncratic or newly emerging geopolitical issues. Hybrid approaches hold promise for combining the strengths of these two approaches while mitigating their individual weaknesses. (IARPA 2016, p. 5)

5. Where Are We Headed? Prescriptions and Future Directions

Although the focus in this paper is on averaging probability forecasts, increasing the quality and use of such averaging is dependent in part on increasing the quality, use, and understanding of probability forecasts in general. The use of probability forecasts and their aggregation has been on the rise across many domains, driven to a great extent by the growth and availability of data, computing power, and methods from analytics and data science. In some arenas, such as the hurricane forecasting discussed in Section 4.1, all of these factors have helped to increase understanding and modeling of physical systems, which in turn has led to improved probability forecasts. Moreover, probability forecasts are increasingly communicated to the public and used as inputs in decision making.

This is illustrated by the three applications in Section 4 and by Nate Silver, who has 3.13 million followers on Twitter and runs the popular fivethirtyeight.com website, which routinely reports on all sorts of probabilities related to politics, sports, science and health, economics, and culture. FiveThirtyEight's focus is squarely on probability forecasts, gathering lots of data from different sources and using sophisticated methods to analyze that data and blend different types of data, accounting for the uncertainty in forecasts. From Silver's overview of their forecasting principles before giving details of their model for the 2018 U.S. House of Representatives election: "Our models are probabilistic in nature; we do a *lot* of thinking about these probabilities, and the goal is to develop probabilistic estimates that hold up well under real-world conditions." (Silver 2018)

The interest in analytics and data science, paired with today's computing power, has enabled the development of more sophisticated forecasting models. Some of the more successful models have drawn on multiple disciplines, such as statistics and computer science. On the statistics side, advances in Bayesian methods, which are inherently probabilistic, are valuable in probability forecasting. For instance, discussions on Andrew Gelman's blog at andrewgelman.com involve some cutting-edge statistical modeling techniques, such as Stan. In terms of computer science, machine learning is making great strides in developing models with methods like quantile regression using the gradient boosting machine (Friedman 2001, Ridgeway 2017) and quantile regression forests (Meinshausen 2006) to produce accurate probability forecasts.

Aggregation methods and hybrid approaches using both statistical modeling and machine learning are being developed. In the recent M4-competition on time series forecasting, such hybrids have been shown to perform better than approaches using only statistical modeling or only machine learning (Makridakis et al. 2018). Although the M4-competition focused mainly on point forecasts, it considered uncertainty by asking for 95% prediction intervals, the end points of which are quantiles. The top two methods for point forecasts (a hybrid method first and an aggregation method second, both involving statistical modeling and machine learning) were also first and second for the 95% intervals. These approaches using both statistical modeling and machine learning are relatively new but early results suggest that they have great potential. More broadly, the surge of work on improving model-based forecasts and their aggregation bodes well for the future.

The Good Judgment Project discussed in Section 4.3, with its focus on probability forecasts for important one-off geopolitical events that tend to be less suitable for mathematical modeling, necessitates more reliance on subjective judgments. This brings in the consideration of notions from psychology, specifically behavioral decision making. IARPA's HFC study is looking at the performance of hybrid forecasting models that combine subjective and model-based or machine-based forecasts. Like the recent model-based work, the path-breaking work initiated by IARPA is young, has led to successful probability forecasts, and still has a great upside.

As should be clear by now, many of the recent developments in probability forecasts have incorporated the aggregation of information and forecasts from multiple sources. Often the forecasts are aggregated via a simple, robust method. The Philadelphia Fed uses a simple average when aggregating probability distributions and a different robust method, the median, when aggregating point forecasts. The

Good Judgment Project aggregates probabilities with a slightly less robust method, a weighted average. FiveThirtyEight's forecasts for the 2018 U.S. House of Representatives election illustrate how complex things can get in forecasting situations, with different types of information and different levels of aggregation. For example, within each House district probability forecasts are first aggregated using a weighted average with calibration adjustments for individual polls and then combined this with other factors. Then final forecasts for the districts are aggregated to obtain forecasts for the overall makeup of the House, taking into account dependence among forecast errors in different districts and other adjustments (Silver 2018).

We are encouraged by the increased use of aggregation of probability forecasts and especially the complex types of aggregation exemplified by the FiveThirtyEight house forecasts. However, with forecasts consisting of probability distributions, the probabilities or densities are generally aggregated. Viable and potentially more useful alternative options have been proposed, e.g., aggregating quantiles instead of probabilities (Lichtendahl et al. 2013b) or generating trimmed pools.

After any aggregation, when final probability forecasts have been formulated, a very important step is the communication of such forecasts. This communication can be to the general public or to specific decision makers for whom the forecasts could be very helpful. When communicating to the public, it is important to realize that probabilities can be difficult to understand for the lay person. Even though understanding is improving as people are exposed to more and more probabilities, probability statements in the media are often misinterpreted given their technical nature and the fact that multiple realizations are needed in order to determine the value of the forecasts.

The lay person may anchor on wanting a "correct forecast" with little understanding of what that means in terms of probability forecasts. A casual observer may hear a reported probability of rain of 0.20 and think that is low enough that it wouldn't rain (implicitly rounding the 0.20 to zero). Then if it actually rains, the observer concludes that such probabilities are useless. From election outcomes and climate change to economic outlook, the popular press routinely reports on how misinterpretation of (and perhaps skepticism about) probability forecasts has led decision makers astray. For example, Leonhardt (2017) writes:

> "The rise of big data means that probabilities are becoming a larger part of life. And our misunderstandings have real costs. Obama administration officials, to take one example, might have treated Russian interference more seriously if they hadn't rounded Trump's victory odds down to almost zero. Alas, unlike a dice roll, the election is not an event we get to try again."

Other numerical information can sometimes be confused with probabilities. For example, when a poll reports that 55% of the voters in an election poll said they would vote for Candidate A and 45% for Candidate B, some might interpret those as the probabilities of the candidates winning the election, which is not correct. Another point that is often overlooked is that there is sometimes confusion about the event or variable and not the probabilities. For complex issues like climate change, the events associated with any probability need to be defined very carefully to combat misinterpretation by the forecasters or later by recipients of the forecast. Even for a seemingly simple event such as the probability of rain, there can be confusion between the probability of rain at a given point in the area (the NWS definition), the probability of rain somewhere in the area (which is often larger), or yet some other interpretation.

Visualization can be very helpful in increasing understanding of probability forecasts, a point brought home by the idiom that a picture is worth a thousand words. Visualization ranges from standard bar charts to fancier displays of probabilities over space or time, often animated. Some situations lend themselves better to visualization than others. The hurricane forecasts in Section 4.1 are good examples, particularly Figures 1 and 2, which pack a lot of useful information into relatively easy-to-understand visuals. The fan chart for GDP growth rate in Figure 4 is also helpful, as are graphs of probabilities over time such as the anxious index in Figure 6 and the consensus trend for the nomination of a candidate in Figure 8. Creativity is often needed in coming up with a good visualization, and specialized software makes it easier to implement visualization. The improvement in visualization tools such as Tableau or Power BI allow more people to develop the skill necessary for creating useful graphics.

If probability forecasts are intended for specific decision makers, care should be taken to give a probability that best matches the needs of the decision makers and stakeholders. For example, the NWS provides a wide variety of accurate probability forecasts to the public, but decision makers often have problems that require weather forecasts with probabilities that are more tailored to their needs. Increasingly, they turn to the private sector for help. A 2006 survey by the American Meteorological Society (AMS) showed that the private industry earned revenues in excess of $1.8 billion (Mandel and Noyes 2013).

Larger global firms like The Weather Company (acquired by IBM in 2016) and AccuWeather have advantages of scale and sophistication to deal with larger clients and complex decisions. Smaller services have the leverage of geographical proximity and familiarity, enabling them to provide tailored forecasts for specific points on short notice in dealing with smaller local clients and smaller decisions.

> "Clients need to get data and forecasts that are directly relevant to their business operations and costs. They need forecasters to be willing to explain how these data and forecasts translate into the decisions they need to make. And they need forecasters who are willing to speak in terms of probabilities, not unattainable certainties. In (the Superintendent of a small airport in New Hampshire's) words, 'Part of the value of the service is knowing levels of probability. In my work, de-icing the airfield may cost me $60,000-$80,000, so I need to understand the likelihood of any particular weather occurrence.'" (Mandel and Noyes 2013, pp. 16-17)

Firms in this industry benefit from free access to NWS data and forecasts and from the boom in analytics and data science. "In forecasting, many companies are adopting machine learning and advanced statistical techniques to post-process model output from the NWS ... to improve forecasts at a range of time and space scales ... 'We do a lot to make forecasts better, including using machine learning for multi-model ensembles and bias correction.' – Weather service provider" (National Weather Service 2017, p. 14). The NWS attributes the growth in demand for weather information to three factors: increased costs associated with increases in large storms, greater sophistication in companies to take advantage of weather data as they invest in analytics and data science skills, and increasing use of weather data for decision making. This growth is expected to continue due to climate change and further increases in the sophistication of forecasts.

The formulation, aggregation, and communication of probability forecasts, as well as their use in decision making, naturally occur before the event or variable being forecast occurs. That leaves us with the important step of evaluating the probability forecast after observing what occurs. Unfortunately, the increase in exposure to and use of probability forecasts has not been accompanied by a comparable increase in the frequency of evaluations. This is due in part to limited exposure to and understanding of scoring rules and their decompositions, which is understandable for those without formal training in or experience with probability forecasting. However, even for those with some training or experience, there may be little or no exposure to scoring rules and their decompositions.

In practice, there should be more emphasis on the use of scoring rules. Scoring rules like the Brier score have been used extensively in weather forecasting for over 50 years, and the scores are communicated to the forecasters. Weather forecasting is an ideal area for such evaluations, because forecasts are made often and the time between forecast and realization is often short. Thus, large numbers of forecast-realization pairs can be accumulated quickly, facilitating evaluations. Scoring rules are also used by FiveThirtyEight and in competitions such as Kaggle that offer large prizes.

In many cases, especially with probability forecasts reported in the media, not enough forecast-realization pairs are obtained to be able to report reliable evaluations. For longer-range forecasts, evaluations cannot be made for a long time, by which time the forecasts may have been discarded or forgotten. Also, probability forecasts that are proprietary may be evaluated, but those probabilities and evaluations are never released. This is the case in many business settings, such as forecasts for demand of new products, project completion times, and costs.

Although scoring rules can be used to evaluate forecasts and, by extension, the forecasters who made them or the models that created them, their most valuable role is to provide feedback to the forecasters to help them improve their future forecasts. Constant feedback on the quality of the forecasts leads to learning (Regnier 2018). This learning, along with improvements in weather data and models, has contributed to the steady increase in the accuracy of probability forecasts of weather events over time.

Ideally the feedback should include a decomposition of overall scores into separate scores for calibration and sharpness. It is easy to see how forecasters can learn from calibration feedback for repetitive events. Consider all of the times a forecaster's probability for an event such as "rain tomorrow" was 0.3 and suppose that the relative frequency of rain over those occasions was 0.5. This suggests to the forecaster that an adjustment should be made to increase probability forecasts of 0.3. When multiple forecasts of the same events or variables are available, as is often the case given the increased aggregation of multiple forecasts, comparisons of the sharpness of forecasts from different sources can stimulate forecasters to try to improve in terms of sharpness. With continued feedback over time, a forecaster can learn, leading to improved forecasts.

We expect all of the above developments to continue to grow and improve rapidly. There are still many situations for which point forecasts are used instead of probability forecasts. Whenever possible, those forecasts should be supplemented or replaced by probability forecasts. We cannot stress too strongly the importance of increasing the use of probability forecasts to convey the uncertainty associated with the event or variable of interest. Dissemination of the probability forecasts is desirable to increase the exposure of the public to probabilities. Increased evaluation of the forecasts when possible, even if only provided to the forecasters, can be valuable in leading to improved future probabilities.

In conclusion, we feel that probability forecasting and the aggregation of probability forecasts have made great strides in recent years and have a promising future. We expect to see greater use of and exposure to probability forecasts, which in turn will increase understanding of probabilities and how they can contribute to better decisions.

References

Armstrong, S. J. 2001. Combining forecasts. In *Principles of Forecasting: A Handbook for Researchers and Practitioners*. S. J. Armstrong, ed. Kluwer Academic Publishers, Norwell, MA, 417–439.

Atanasov, P., P. Rescober, E. Stone, S.A. Swift, E. Servan-Schreiber, P. Tetlock, L. Ungar, B. Mellers. 2017. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*. **63**(3): 691-706.

Bassamboo, A., R. Cui, A. Moreno. 2018. Wisdom of crowds: Forecasting using prediction markets. Working paper, Kellogg School of Management, Northwestern University, Evanston, IL.

Bates, J. M., C. W. Granger. 1969. The combination of forecasts. *Journal of the Operational Research Society*. **20**(4): 451-468.

Belleflamme, P., T. Lambert, A. Schwienbacher. 2014. Crowdfunding: Tapping the right crowd. *Journal of Business Venturing*. **29**(5): 585-609.

Bennett, J., S. Lanning. 2007. The Netflix prize. In *Proceedings of KDD cup and workshop 2007.* San Jose, CA, August 12, 2007.

Breiman, L. 2001. Random forests. *Machine Learning*. **45**(1): 5-32.

Brier, G.W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. ***78***(1): 1-3.

Budescu, D. V., E. Chen. 2015. Identifying expertise to extract the wisdom of crowds. *Management*

Science. **61**(2): 267-280.

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

Clemen, R. T., R. L. Winkler. 1985. Limits for the precision and value of information from dependent sources. *Operations Research*, **33**(2): 427-442.

Clemen, R. T., R. L. Winkler. 1999. Combining probability distributions from experts in risk analysis. *Risk analysis*. **19**(2): 187-203.

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

Croushore, D. D. 1993. Introducing: the survey of professional forecasters. *Business Review-Federal Reserve Bank of Philadelphia*. *6*, 3.

Dawson, C. 2017. USC ISI leads IAPA contract for developing hybrid forecasting systems.

Friedman, J.H. 2001 Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. **29**(5): 1189-1232.

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

Galton, F. 1907. Vox populi. *Nature*. **75**: 450–451.

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

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

Good, I.J., 1952. Rational decisions. *Journal of the Royal Statistical Society, Series B (Methodological)*. **14**(1): 107-114.

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

Grushka-Cockayne, Y., K.C. Lichtendahl Jr., V.R.R. Jose, R.L. Winkler. 2017b. Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Operations Research*. **65**(3): 712-728.

Hong, T., P. Pinson, S. Fan, H. Zareipour, A. Troccoli, R.J. Hyndman. 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.

Howe, J. 2006. The rise of crowdsourcing. *Wired*. 14(6): 1-4.

IARPA. 2010. Aggregative Contingent Estimation (ACE) Program Broad Agency Announcement (BAA)

Solicitation Number: IARPA-BAA-10-05. Office of the Director of National Intelligence.

IARPA. 2016. Hybrid Forecasting Competition (HFC) Program Broad Agency Announcement (BAA) Solicitation Number: IARPA-BAA-16-02. Office of the Director of National Intelligence.

Johnstone, D.J., V.R.R. Jose, R.L. Winkler. 2011. Tailored scoring rules for probabilities. *Decision Analysis*. **8**(4): 256-268.

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

Jose, V.R.R., R.L. Winkler. 2008. Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*. **24**(1): 163-169.

Jose, V.R.R., R.L. Winkler. 2009. Evaluating quantile assessments. *Operations Research*. **57**(5): 1287-1297.

Kishore N., D. Marqués, A. Mahmud, M.V. Kiang, I. Rodriguez, A. Fuller, P. Ebner, C. Sorensen, F. Racy, J. Lemery, L. Maas. 2018. Mortality in Puerto Rico after Hurricane Maria. *New England Journal of Medicine*. May 29.

Lamberson, P. J., S. E. Page. 2012. Optimal forecasting groups. *Management Science*, **58**(4): 805-810.

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

Leonhardt, D. J. 2017. Opinion: What I was wrong about this year. New York Times. December 25, 2017, p. A21. https://www.nytimes.com/2017/12/24/opinion/2017-wrong-numbers.html.

Lichtendahl Jr, K. C., Y. Grushka-Cockayne, P.E. Pfeifer. 2013a. The wisdom of competitive crowds. *Operations Research*. **61**(6): 1383-1398.

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

Lichtendahl Jr, K. C., R.L. Winkler. 2007. Probability elicitation, scoring rules, and competition among forecasters. *Management Science*. **53**(11): 1745-1755.

Makridakis, S., E. Spiliotis, V. Assimakopoulous. 2018. The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*. in press, https://doi.org/10.1016/j.ijforecast.2018.06.001.

Mandel, R., E. Noyes. 2013. Beyond the NWS: Inside the thriving private weather forecasting industry. *Weatherwise*. **66**(1): 12-19.

Matheson, J.E., R.L. Winkler. 1976. Scoring rules for continuous probability distributions. *Management Science*. **22**(10): 1087-1096.

McCarthy, J. 1956. Measures of the value of information. *Proceedings of the National Academy of Sciences*. **42**: 654-655.

Meinshausen, N. 2006. Quantile regression forecasts. *Journal of Machine Learning Research*. **7**(6): 983-999.

Miller, K. 2018. Hurricane season 2018: Slimmer cone reflects forecasting improvements. *Palm Beach Post*. May 12.

Murphy, A.H., R.L. Winkler. 1984. Probability forecasting in meteorology. *Journal of the American Statistical Association*. **79**(387): 489-500.

National Weather Service. 2017. *National Weather Service Enterprise Analysis Report: Findings on changes in the private weather industry*. https://www.weather.gov/media/about/Final_NWS%20Enterprise%20Analysis%20Report_June%202017.pdf

O'Hagan A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, T. Rakow. 2006. *Uncertain judgements: Eliciting experts' probabilities*. Wiley, New York.

Ranjan, R., T. Gneiting. 2010. Combining probability forecasts. *Journal of the Royal Statistical Society, Series B*. **72**(1): 71-91.

Regnier, E. 2018. Probability forecasts made at multiple lead times. *Management Science*. **64**(5): 2407-2426.

Ridgeway, G. 2017. gbm: Generalized Boosted Regression Models. R Package version 2.1.3.

Savage, L.J. 1954. *The Foundations of Statistics*. Wiley, New York.

Savage, L.J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*. **66**(336): 783-801.

Silver, N. 2012. *The Signal and the Noise*. Penguin Press, New York.

Silver, N. 2018. How FiveThirtyEight's house model works. https://fivethirtyeight.com/features/2018-house-forecast-methodology/. August 16, 12:30 pm.

Stark, T. 2013. SPF panelists' forecasting methods: A note on the aggregate results of a November 2009 special survey.

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

Surowiecki, J. 2005. *The Wisdom of Crowds*. Anchor Books, New York.

Tetlock, P.E., D. Gardner. 2015. *Superforecasting: The art and science of prediction*. Crown, New York.

Turner, B.M., M., Steyvers, E.C., Merkle, D.V., Budescu, T.S., Wallsten. 2014. Forecast aggregation via recalibration. *Machine learning*. **95**(3): 261-289.

U.S. National Hurricane Center. 2017. National Hurricane Center Product Description Document: A User's Guide to Hurricane Products. https://www.nhc.noaa.gov/pdf/NHC_Product_Description.pdf.

Winkler, R.L. 1994. Evaluating probabilities: Asymmetric scoring rules. *Management Science*. **40**(11):

1395-1405.

Winkler, R.L. 1996. Scoring rules and the evaluation of probabilities. *Test*. **5**(1): 1-60.

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

Witkowski, J., R. Freeman, J. Wortman Vaughan, D.M. Pennock, A. Krause. 2018. *Incentive-Compatible Forecasting Competitions*.

Wolfers, J., E. Zitzewitz. 2004. Prediction markets. *Journal of Economic Perspectives*. **18**(2): 107-126.

Zumbrun, J. 2018. Growth seen hitting 3% in 2018, but risks to outlook mount after this year. https://www.wsj.com/articles/growth-seen-hitting-3-in-2018-but-risks-to-outlook-mount-after-this-year-1533823205?redirect=amp#click=https://t.co/4bHOhQERi5. August 9, 6:05 pm.