ABSTRACT

Title of dissertation: EVALUATING METHODS FOR MODELING AND AGGREGATING CONTINUOUS DISTRIBUTIONS OF FORECASTER BELIEF

Joe Tidwell, Doctor of Philosophy, 2017

Dissertation directed by: Thomas Wallsten
Department of Psychology

The "Wisdom of the crowds" is the concept that the average estimate of a group of judges is often more accurate than any single judges estimate. This dissertation explores a variety of elicitation, modeling, and aggregation methods for time-based forecasting questions at both the individual and consensus levels, and shows that accurate continuous forecast distributions can be modeled from relatively few judgments from individual forecasters.

For individual forecasters, eliciting judgments with fixed versus random cut points, and modeling those judgments with least-squares methods led to the most accurate forecasts. While gamma distributions fit the empirical judgments more closely than exponential distributions, exponential fits yielded more accurate model forecasts, suggesting that the greater flexibility of the gamma distribution tended to over-fit the empirical forecasts.

For consensus forecasts, random cut points across individual forecasters yielded more accurate forecasts than fixed cut points, suggesting that across a group of

forecasters, random bins may help average over individual-level forecast errors introduced through partition dependence bias and an arbitrary set of fixed cut points. With respect to modeling methods, a mixture of forecaster distributions fit with a Bayesian Dirichlet-multinomial model performed best across a variety of metrics and yielded forecast accuracies on par with advanced discrete aggregation techniques. This model also provides a natural way to weight individual forecasters according to expertise and other factors.

Differences in forecast accuracy between modeling methods varied greatly depending on when an event occurred relative to the range over which forecaster judgments were elicited, particularly when events occurred long after the last date for which forecasters provided judgments. In these cases, the modeled forecasts depend heavily on the assumptions of the model versus the elicited judgments, and forecasts should be cautiously interpreted as representing crowd belief.

The results of this research shows that with a limited number of discrete elicited judgments, it is possible to obtain continuous aggregate models of forecaster belief that are as accurate as discrete forecast aggregation methods, but can also provide decision makers with forecasts for arbitrary partitions of the event space and can be easily integrated into a broad range of decision analyses.

EVALUATING METHODS FOR MODELING AND AGGREGATING CONTINUOUS DISTRIBUTIONS OF FORECASTER BELIEF

by

Joe Tidwell

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Michael Dougherty, Chair/Advisor
Professor Paul Hanges
Professor Robert Slevc
Professor Emeritus Thomas Wallsten, Advisor
Professor Michel Wedel, Dean's Representative

Acknowledgments

I am grateful to the the continuous elicitation team of the Good Judgment Project who provided the empirical data for this dissertation. This team was led by Donald Moore and Thomas Wallsten, and was one component of the larger Good Judgment Project directed by Barbara Mellers, Don Moore, and Philip Tetlock. I was a member of this research team and led development of forecast modeling and aggregation methods, but the credit for designing, implementing, and managing this experiment belong to the entire team. While all of the work in this dissertation developing and validation consensus models is my own, the experiment and data I validated these models on was the product of many people's diligent work, and this dissertation would not have been possible without their efforts.

I am also thankful for the help of my family, particularly my wife Sharon and children Grace, Laura, Cuatro, and Audrey. This dissertation would not have been possible without their extensive patience and support.

Table of Contents

List of Figures

List of Abbreviations

ACE	Aggregative Contingent Estimation program
CI	Confidence Interval
CRPS	Continuous Rank Probability Score
DM	Dirichlet-Multinomial
GJP	Good Judgment Project
GJP-CE	Good Judgment Project, Continuous Elicitation Team
HDI	Highest Density Interval
IARPA	Intelligence Advanced Research Projects Activity
LOP	Linear Opinion Pool
LS	Least-Squares
MAE	Mean Absolute Error
MCMC	Markov chain Monte Carlo
PIT	Probability Integral Transform
RPS	Rank Probability Score

Chapter 1: Introduction

The "Wisdom of the crowds" has been a frequent topic in popular media for several years. The concept is that the average estimate of a group of judges is often more accurate than any single judge's estimate. This idea is at least as old as Aristotle: "For it is possible that the many, though not individually good men, yet when they come together may be better, not individually but collectively" (Jowett and Davis, 1908). Remarkably, in the subsequent passages of the *Politics*, Aristotle lays out some of the most important aspects of judgment elicitation and aggregation that still apply today, including: judgment aggregation is most effective when individuals hold disjoint information; that depending on an individual judge's expertise, there is an optimal weight (vote) that each judge should receive; and that even small groups of judges, who individually may not hold any significant expertise, can make more optimal decisions than the any individual expert.

The first example of "Wisdom of the crowds" in the scientific literature is Sir Francis Galton's 1907 analysis of a competition to guess what the weight of an ox would be after it had been slaughtered and dressed (Galton, 1907b). Contestants at a stock show in Plymouth, England purchased tickets and wrote down their best guess for the weight. Galton showed that on average the 787 guesses deviated from

the true weight of 1,198 pounds (lbs) by 37 lbs. In contrast, the median of the guesses deviated from the correct value by 9 lbs and the mean was within 1 lb [1]. Galton took this as evidence that the collective judgment of groups, even of non-experts, could rival or exceed that of individuals. Considerable research since this time has shown that aggregating judgments, and in particular probabilistic forecasts, almost universally increases accuracy (Wallis, 2011; Wallsten et al., 1997).

An independent line of research begun in the mid 1950's discovered that mathematical models of human judgment could often outperform the judgment of the experts on which the models were based (Meehl, 1954) . In one example, Goldberg (1970) modeled the judgment process of 29 clinical psychologists attempting to differentiate psychotic from neurotic patients based on scores in the patients' Minnesota Multphasic Personality Inventory (MMPI). He fit a linear regression for each clinician predicting their diagnoses from the 11 MMPI scores and found that the regression models outperformed the clinicians not only in the training sample, but also on new out of sample diagnoses.

In this dissertation I combine both of these approaches, judgment aggregation and mathematical models of human judgment, to develop new methods for modeling continuous probabilistic forecasts of uncertain events. In particular, I focus on methods for forecasts of events based on time, i.e. the probability that an event will happen by some date, and for events without a well-defined reference class that can occur only one time, for example the probability of a specific conflict between

[1] Galton did not provide the mean in the original paper, but in response to a letter to the editor requesting it (Galton, 1907a).

two nation states. In the following sections I provide background on probabilistic forecasting, forecast calibration, methods for determining forecast accuracy, and different approaches for aggregating forecasts.

1.1 Background

Over four years, the Intelligence Advanced Research Projects Activity (IARPA) conducted what is likely the most extensive investigation to date into the elicitation, mathematical aggregation, and communication of probabilistic forecasts of non-repeatable uncertain events. This project, the Aggregative Contingent Estimation (ACE) program (IARPA, 2010), focused on forecasts of events of interest to the intelligence community. Nevertheless, the research and innovations obtained from this work apply more generally to any forecasting problem, and particularly to forecasts of events characterized by high epistemic uncertainty.

The ACE program focused on innovating methods for discrete forecasts of continuous phenomena. One type of these discrete questions elicited judgments for a single event partition. For example, "Before 10 June 2015, will Ukraine officially announce that it will hold a referendum on the structure of its government?" Forecasters selected either 'yes' or 'no', then provided a probability judgment corresponding to their choice. Another type of discrete question elicited judgments for multiple partitions of the event space. For example, "How many additional countries will report cases of the Ebola virus as of 9 May 2014?" Ace researchers provided forecasters with multiple options, in this case [0,1,2,3 or more], and provided prob-

ability judgments for each option.

While this approach has proven itself useful in many domains (For examples in economics, energy, and health risk, (see Croushore, 1993; Usher and Strachan, 2013; Hetes et al., 2011), it is nevertheless constrained in that it can only provide a decision maker raw or aggregate judgments with respect to the event partitions that were explicitly elicited. Take the first example of the Ukrainian referendum. This question elicits judgments for a small range of the more fundamental question "*When* will Ukraine officially announce that it will hold a referendum on the structure of its government?" The answer could be any time between when the question is elicited and infinitely far into the future, i.e. never. A probability distribution over all possible values of the forecasted event would be much more useful to decision makers than would a discrete forecast, since the continuous forecast could provide relevant forecasts for any partition of the variable and be more easily integrated into a broad range of decision analyses.

At least two research teams involved with the ACE program explored the feasibility of eliciting, modeling, and aggregating continuous forecast distributions. The most extensive work was conducted by the Continuous Elicitation Team of the Good Judgment Project (GJP-CE) led by Don Moore and Tom Wallsten, of which I was member. This work focused on producing continuous forecast methods that would be effective when using relatively few judgments from each individual forecaster. The GJP-CE method was as follows:

1. For each forecasting question, elicit interval probability judgments from a set

of forecasters. Typically, three to five judgments per forecaster were elicited.

2. Fit a two-parameter probability distribution to each forecaster's judgments via least-squares. The distribution chosen to fit these judgments depended on the bounds of the forecasted variable. Unbounded variables were fit with the normal distribution. Double-bounded variables, for example ratios or percentages, were fit with the beta distribution. Left-bounded variables, typically time-based variables, were fit with the gamma distribution.

3. Aggregate these fitted distributions into a a single parametric consensus distribution of the same distributional form as fit at the forecaster level. The primary aggregation method was to take the median of each distribution parameter at the forecaster level, and then use these values as the parameters for the consensus distribution.

GJP-CE evaluated this method, which they termed "Median-θ", with a 9-month on-line forecasting tournament. Approximately 350 forecasters participated, producing over 30,000 sets of interval probability judgments for 132 events. Though analysis of these forecasts is currently incomplete, tentative results indicate that the GJP-CE method yielded consensus forecasts that were as least as accurate as advanced discrete aggregation methods.

1.1.1 Motivation for Research

While their tentative results are promising, there are at least three reasons to believe that the Median-θ method can be improved. First, the choice of the gamma

distribution to model forecaster judgments necessarily restricts the kinds of belief that the forecaster models can represent. Second, since individual forecasters provide relatively few judgments, the flexibility of any two parameter distribution fit to these judgments raises the risk of over-fitting the forecasts. Third, constraining consensus level forecasts to follow the gamma distribution can obscure the heterogeneity of belief across individual forecasters.

One useful way of representing continuous belief for time based events is through the hazard function of a probability distribution. The hazard function is the ratio of the density function to the survival function (Eq 1.1), and is the instantaneous rate of occurrence of an event. At the forecaster level, modeling judgments with two-parameter distributions restricts the possible forms of subjective belief and may over-fit and/or mis-specify the hazard function. For example, the gamma distribution is capable of representing constant, concave and increasing, and convex and decreasing hazards (Figure 1.1-a). Since GJP-CE modeled probability judgments for time-based forecasting questions with the gamma distribution, the potential hazards their methods can capture are limited to those listed above. In terms of forecaster belief, this means that this method must produce forecaster level models that either "believe" that the chance of an event occurring is constant as time passes; that the event is getting more likely but approaches a constant maximum chance of occurrence as time passes; or that the event is getting less likely but approaching a constant minimum chance of occurrence as time passes. The gamma distribution can not model beliefs such as those in Figure 1.1-b: the rate of increase in likelihood of an event increasing as time passes (convex and increasing hazard)

or that an event is most likely to occur relatively soon, or far into the future, but unlikely to occur for some medium term (bathtub hazard).

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{1.1}$$

Another potential problem with fitting simple distributions at the forecaster level is particular to this forecasting environment, where for a given forecasting question relatively few interval probability judgments are elicited per forecaster. If one assumes that forecasters produce judgments with some error component, then as the number of elicited bins decreases, the potential to over-fit these judgments increases. For example, if one elicits three interval probability bins for a forecasting question, as long the cumulative probabilities of these judgments are strictly increasing, *any* two-parameter distribution with appropriate support will fit these observations perfectly. This implies that rather than averaging out production error, fitting two parameter distributions maximizes the influence of production error in the fitted model.

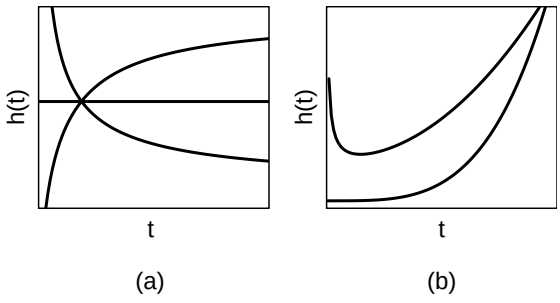

Figure 1.1: Hazard Examples

The inability of two-parameter distributions to model plausible subjective belief at the forecaster level extends to the the consensus level as well. However, at the

consensus level there is the additional complication that forecasters may have very different continuous subjective estimates of the likelihood of some event. Defining the consensus as a single two-parameter distribution may inadvertently gloss over heterogeneity of forecaster belief, and provide a false picture of certainty and/or unanimity amongst forecasters.

For questions defined primarily by aleatory uncertainty (random variability) versus epistemic uncertainty (insufficient/unknown information), it is reasonable to assume that the the goal of aggregation is to obtain a continuous probability measure that itself is an estimate of some objective and observable long-run phenomenon. Assume that a group of forecasters observes a sufficiently long sequence of rolls of a biased 6-sided die, and is then asked to define a categorical distribution of the probabilities of each side appearing on a given roll. All forecasters have equal access to information, and one should expect an equal weighted mixture of all forecaster distributions to yield a good estimate of the true distribution, assuming that each forecaster judgment is a function of the true probability perturbed by individual error and that forecasters are independent. Aggregation in this context averages out individual error, and provides an accurate consensus judgment of the true discrete distribution. In the case of forecasting problems principally defined by aleatory uncertainty, individual subjective belief is an estimate of an objective and observable true long-run probability, and aggregation attenuates error in individual forecast production.

In the case of forecasting problems defined by high epistemic uncertainty, this line of reasoning breaks down. Consider the question "At what point will Bashar

al Assad step down or otherwise be removed from the office of president of Syria?", and assume that one could perfectly model continuous subjective distributions of forecaster belief for this question. There is no well-defined reference class for this question; there is no practically definable probability model at the forecaster level. Even if we assumed that each forecaster had a covert and well defined causal model with a joint probability over many events that could yield a marginal distribution of time until Assad was no longer in office, it would be practically impossible to elicit this joint distribution and the marginal distribution shouldn't be expected to follow any simple form. Given that such a model existed, it's practically unknowable to the forecaster, much less to anyone else, and even less plausibly could be well-modeled by something like the gamma distribution.

For forecasting questions defined primarily by aleatory uncertainty, the "best" consensus distribution is the one that most accurately models an observable environmental phenomenon with a well-defined reference class. For questions defined by high epistemic uncertainty, the "best" consensus distribution is the one that most accurately models a summary of crowd belief.

1.2 Properties of Forecasts

1.2.1 Calibration

Calibration refers to the consistency between probabilistic forecasts and the frequency with which the forecasted events occur (Gneiting et al., 2007). If a forecaster is well-calibrated, whether the forecaster is an individual or a model, then the

forecasted events should occur proportionally to the probability assigned to those events. For example, if a weatherman was well-calibrated, then it would rain 50% of the days he forecasted a 50% chance of rain.

A common form of miscalibration is overprecision (Moore and Healy, 2008). A forecaster is overprecise if her forecasts obtain confidence intervals that are excessively narrow. If an economist repeatedly estimated a 50% confidence interval for the U.S. inflation rate, but only 25% of the forecasted rates fell within her 50% confidence interval, then her forecasts would be overprecise.

In line with Gneiting et al. (2007), I will evaluate the calibration of the forecasters and consensus methods with the Probability Integral Transform (PIT). Assume a strictly continuous predictive cumulative distribution function F and an observed event x. The PIT, $Z_F = F(x)$ is simply the the cumulative probability that F assigns to the outcome. Given a set of events X, one can evaluate the calibration of the predictive distribution $F(X)$ with the empirical distribution of Z_F. If $F \sim X$, i.e. if the predictive distribution is identically distributed to the distribution of outcomes, then Z_F is necessarily distributed standard uniform. F is probabilistically calibrated if $Z_F \sim U(0, 1)$ (Rosenblatt, 1952).

PIT histograms can generally classify forecasts as either neutrally dispersed, underdispersed, or overdispersed. Dispersion in this case refers to the variance of the PIT (Z_F), *not* the predictive distribution F. Since the PIT is distributed on the interval [0,1], its variance (σ_Z^2) is necessarily constrained to the interval $\left[0, \frac{1}{4}\right]$. A perfectly calibrated forecast will be distributed standard uniform with a variance $\sigma_Z^2 = \frac{1}{12}$, or about 0.083, and will be considered neutrally dispersed. When $\sigma_Z^2 > \frac{1}{12}$,

the forecast is underdispersed relative to a neutral forecast, and when $\sigma_Z^2 < \frac{1}{12}$ the forecast is overdispersed relative to a neutral forecast.

Figure 1.2 shows examples of the different dispersion patterns. For each example, I simulated 10,000 observed outcomes from $X \sim N(0, 2)$. The middle panel shows a neutrally dispersed forecaster with a continuous forecast $F \sim N(0, 2)$. The PIT is uniformly distributed, so the forecaster is calibrated. At any arbitrary confidence interval, the proportion of observations that fall within the interval match the forecasted probabilities. The left panel shows an underdispersed forecaster with forecast $F \sim N(0, 1)$. The variance of the PIT is greater than a a standard uniform distribution For any central confidence interval, the proportion of observations falling within that interval are less than the range of the interval. This pattern is consistent with an overprecise or "overconfident" forecaster. The right panel shows an overdispersed forecaster. The variance of the PIT is less than $\frac{1}{12}$ and for any central confidence interval, the proportion of observations falling within that interval are greater than the range of the interval.

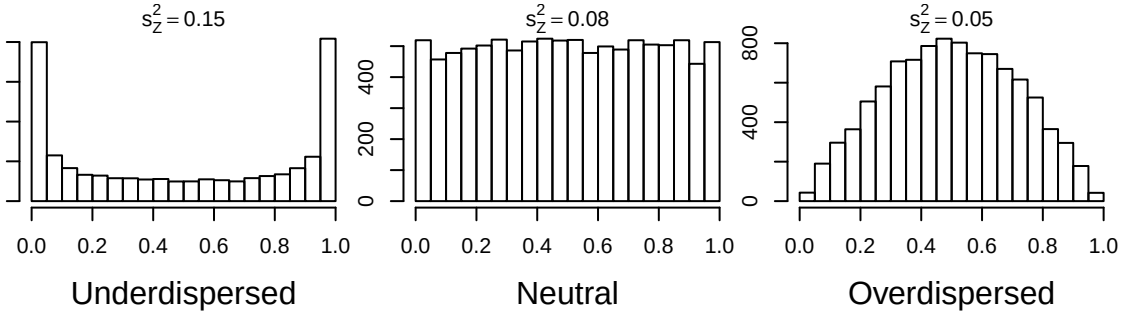

Figure 1.2: PIT histograms for underdispersed (overconfident/overprecise), neutrally dispersed (calibrated), and overdispersed (underconfident/underprecise) forecasts.

1.2.2 Accuracy

Calibration in itself is not a sufficient measure of forecast quality. Assume that two weather forecasters predicted the chance of rain for a series of days, and that the true probability of rain on any give day was 50%. Forecaster #1 always predicted a 50% chance of rain. Forecaster #2 always predicted 100% chance of rain on days that it did rain, and 0% chance of rain on days that it did not rain. Both forecasters are perfectly calibrated, but clearly Forecaster #1's predictions are no better than flipping a coin, while Forecaster #2's predictions are perfectly diagnostic.

Scoring rules are measures (cost-functions) that simultaneously evaluate forecast diagnosticity as well as calibration. Strictly proper scoring rules uniquely obtain the best possible score if the forecaster reports their true estimated probability, i.e. there is no way to 'game' the score by reporting any probability other than true belief. Since a scoring rule is just a cost function, there are an unlimited number of possible scoring rules; however, some are much more widely adopted than others.

The Brier score (Brier, 1950) is one of the most common scoring rules for binary forecasts of the probability that an event will or will not occur. The Brier score is a quadratic the loss function in Equation 1.2, where $F(q)$ is the forecasted probability for whether the event will occur before the cut point (or have a value less than the cut point) q and $\mathbb{1}_{o \leq q}$ equals 1 if the event occurs before q, otherwise 0. Scores can range from 0 to 2, with the best possible score of 0 and worst possible score of 2.

$$S_{BS} = 2\left[F(q) - \mathbb{1}_{o \leq q}\right]^2 \tag{1.2}$$

While the weather forecasters in the example were both perfectly calibrated, the expected values of their Brier scores show that the second forecaster is more accurate. Forecaster #1 always predicts a 50% chance, so his expected score is .5. Forecaster #2 always correctly predicts the chance of rain as either 100% or 0%, so her expected score is 0.

The rank probability score (RPS) is a brier score for questions with more than two possible outcomes (Matheson and Winkler, 1976). This score divides forecasts for n ordered alternatives into a series of binary alternatives as the threshold moves up from lower to higher (1.3). A Brier score is calculated for each of the $n-1$ binary partitions and the final result is the average of the separate scores where n is the number of cut points; q_i the quantile associated with cut point i, $i = 1...n$; o the observed value; F the predictive cumulative distribution function; and $\mathbb{1}_{o \leq q_i}$ equals 1 if o is less than or equal to q_i, otherwise 0. Like the Brier score, RPS scores can range from [0,2], with 0 the best possible score and 2 the worst possible score.

$$S_R = \frac{2}{n} \sum_{i=1}^{n} \left[F(q_i) - \mathbb{1}_{o \leq q_i}\right]^2 \tag{1.3}$$

There are scoring rules for continuous forecasts just as there are for discrete. Like in the discrete case, these scores simultaneously measure calibration and resolution, though the concept is somewhat different in the continuous case and is generally referred to as sharpness (Ranjan and Gneiting, 2010).

Assume that a series of events is distributed $E_t \sim N(\mu_t, 1)$, where $\mu_t \sim N(0, 1)$. Think of t as indexing a random forecasting question, and μ_t as the state of the world at the the time of the forecast. The outcome of each event is random, but conditioned on μ_t. Unconditionally, E_t is necessarily distributed $E_t \sim N(0, \sqrt{2})$. A forecaster who always makes predictions based on the unconditional distribution with the distribution $N(0, \sqrt{2})$, and a forecaster who always conditioned on μ_t and for each event chose the distribution $N(\mu_t, 1)$ would both be perfectly calibrated (Fig 1.3), but the second forecaster would provide more accurate forecasts over a series of events, because the variance of his predictive distribution is smaller relative to the first forecaster.

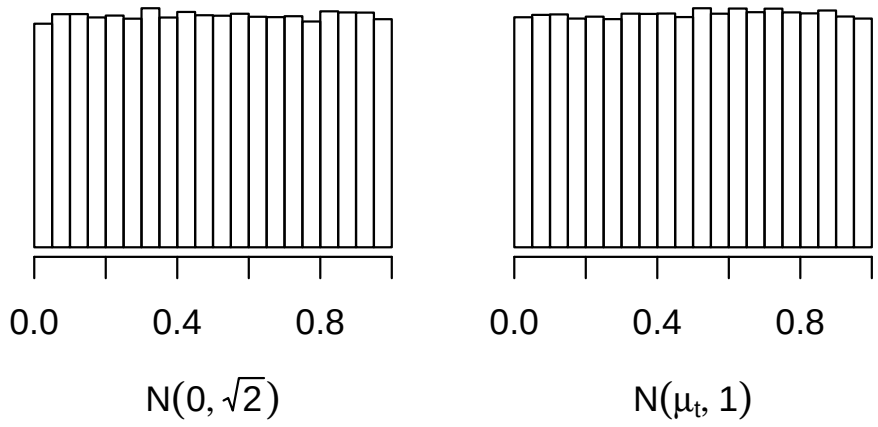

Figure 1.3: PIT histograms for two calibrated continuous forecasters, where $\mu_t \sim N(0, 1)$, and the true distribution of the event is $E \sim N(\mu_t, 1)$. Both forecasters are perfectly calibrated in expectation, but the forecaster on right, who conditions on μ_t will produce more accurate forecasts.

The continuous rank probability score (CRPS) (Eq 1.4) is a continuous scoring

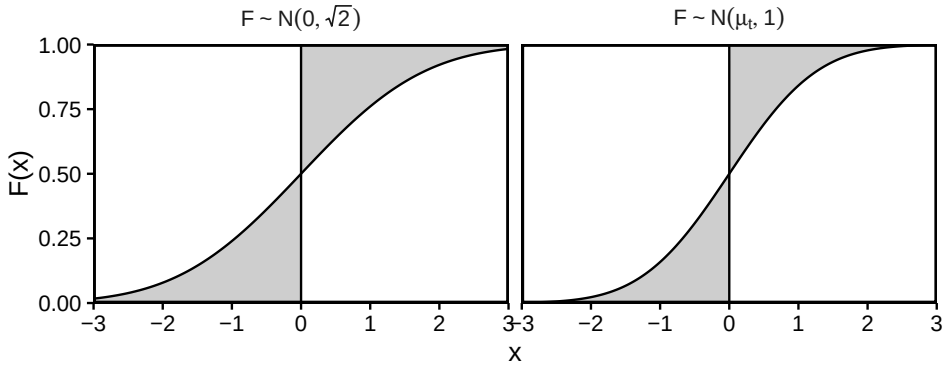

Figure 1.4: Visualization of the CRPS score. The right panel plots the CDF of $N(\mu_t, 1)$ and the left panel $N(0, \sqrt{2})$. The vertical black line is the outcome of the forecast question, 0. The CRPS equals the shaded area between the outcome and the CDFs. Though both forecasters are calibrated, the forecaster who conditions on μ_t obtains a more accurate (lower) CRPS as indicated by the smaller shaded region.

rule that scores the entire predictive distribution (Matheson and Winkler, 1976). It is a generalization of mean absolute error (MAE) and is equivalent to integrating the RPS across all possible real-valued thresholds (Matheson and Winkler, 1976). Like the RPS, lower scores reflect more accurate forecasts, but the CRPS has no upper limit.

$$S_C = \int_{-\infty}^{\infty} [F(x) - \mathbb{1}_{o \leq q_i}]^2 \tag{1.4}$$

The more probability a forecast concentrates near the outcome, the lower the CRPS score will be. Figure 1.4 shows the two example continuous forecasts when $\mu_t \sim N(0, 1)$. The right panel shows the CDF for the predictive forecast $F \sim N(\mu_t, 1)$, and the left panel the CDF for the predictive forecast $F \sim N(0, \sqrt{2})$.

The CRPS is equal to the shaded area between the step function defined by the outcome at 0, and the respective CDFs. Despite both forecasts being calibrated, The forecast with the smaller standard deviation is concentrated closer to the outcome, and therefore the area between the outcome and the CDF is smaller and yields a lower CRPS score ($S_R = .23$) compared the the score for the forecast with a standard deviation ($S_R = .33$).

1.2.3 Aggregation

One of the most common methods to aggregate continuous distributions is the linear opinion pool (LOP). The LOP is a mixture of either the density or cumulative probability functions of the continuous forecasts (Clemen and Winkler, 1999). Though common, it is also well known that linear aggregation is often suboptimal. With calibrated forecaster level distributions linear aggregation will necessarily produce uncalibrated, underconfident consensus forecasts (Ranjan and Gneiting, 2010) (Hora 2004, Ranjan 2010). To continue the previous example, though the forecaster with $F_1 \sim N(0, \sqrt{2})$ and and forecaster with $F_2 \sim N(\mu_t, 1)$ were both calibrated, the linear combination of their distributions F_3 yields an underprecise consensus distribution that scores worse than than the best forecaster F_2 (Fig 1.5).

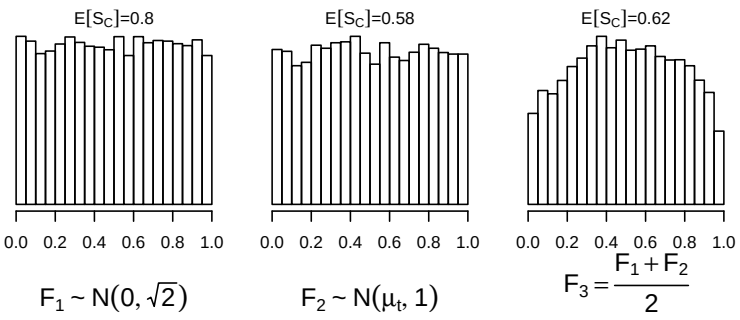

Figure 1.5: PIT histograms for three continuous forecasts, with the expected values of S_C for each forecast. The true data generating process is $F \sim N(\mu_t, 1)$, $\mu_t \sim N(0, 1)$. Although the left and middle forecasts are calibrated, the unweighted linear combination of these distributions in the right panel results in an overdispersed aggregate forecast that scores worse than F_2.

Despite the known shortcomings of LOPs, there is little agreement about what constitutes an optimal distribution aggregation method. For example, Lichtendahl et al. (2013) demonstrate that under certain conditions, aggregating over quantiles obtains better calibrated consensus distributions and sharper forecasts, and argue that sharpness should be the primary goal in aggregation. However, Ranjan and Gneiting (2010) show that a LOP re-weighted using a beta transformation mitigates the miscalibration inherent to finite mixtures of cumulative probabilities, and Hora et al. (2013) argue that under certain conditions replacing mean aggregation of cumulative probabilities with median aggregation leads to better calibration and sharper forecasts.

While the existing research into continuous aggregation is both analytically and computationally rigorous, most often this works begins with the implicit as-

sumption that the distributions over which one will aggregate are perfect representations of subjective belief and are well-calibrated. Real-world forecasts, on the other hand, tend to be underdispersed and therefore overconfident (Gneiting et al., 2007). Because of this, even though linear aggregation necessarily increases dispersion and introduces miscalibration if the source distributions are calibrated, it often improves aggregate scores since the source distributions are overdisprsed and miscalibrated to begin with.

1.3 Summary

Researchers have recently developed at least one approach, the median-θ consensus method, that may yield aggregate continuous consensus forecasts that are generally as accurate as more extensively researched discrete aggregation methods. The goal of this dissertation is to extend that continuous modeling and aggregation work and discover consensus methods that can consistently outperform median-θ. There are at least three reasons why this method may be improved: the choice of the gamma distribution to model forecaster judgments necessarily restricts the kinds of belief that the forecaster models can represent; since in the forecasting environment under study individual forecasters provide relatively few judgments, the flexibility of any two parameter distribution fit to these judgments raises the risk of over-fitting the forecasts; and constraining consensus level forecasts to follow the gamma distribution, or any particular parametric function, may obscure the heterogeneity of belief between individual forecasters. In the following chapters I detail

a set of forecast modeling and aggregation methods that may improve consensus forecast accuracy over the median-θ method, test these models against a large empirical forecast dataset, use Monte Carlo experiments to determine the potential for the modeling methods to capture forecaster hazard functions, and conclude with a general discussion of the research findings.

Chapter 2: Consensus Methods

I define a "consensus method" as the combination of three components: a forecaster model, forecaster estimates derived from the forecaster models, and an aggregation function that combines the forecaster estimates. A forecaster model uses a set of judgments for one forecasting question for one forecaster to estimate a distribution function for all possible values of the forecasted variable. For example, one could fit a parametric distribution to a forecaster's judgments via least-squares to obtain parameters that represent the forecaster's continuous subjective belief. I will fit forecaster judgments with two methods: least-squares (LS) and a Bayesian Dirichlet-Multinomial model (DM).

Forecaster estimates are the information that will be aggregated to form the consensus distribution. I will use either the forecaster cumulative distribution functions (F), or the parameters of the forecaster distributions (θ), as the forecaster estimates.

I will use the simple mean or median as the aggregation functions to combine forecaster estimates into a single consensus distribution. For example, if you had parameters for a set of forecasters you could take the average of each parameter across the forecasters to obtain a consensus level parametric distribution that represented

the aggregate belief of the forecasters.

I will identify all modeling methods with reference to these three components. For example, $\mu(F_{LS})$ denotes a consensus method that takes the mean cumulative probabilities of forecaster level distributions fit via least-squares.

2.1 Forecaster Models

I use two methods to estimate forecaster level continuous subjective belief, least-squares (LS) and a Bayesian Dirichlet-Multinomial model (DM). LS is an attractive method because it is both mathematically and computationally simple (at least for distributions with few parameters), requires relatively little programming or mathematical expertise to implement, and should provide consistent and predictable results. These are significant considerations for any real-world application. Regardless of its potential benefits, a consensus method that requires extensive monitoring and broad programming or mathematical expertise will be far less likely to be implemented outside of an academic environment.

At both the forecaster and consensus levels I developed and evaluated Bayesian analogues of the LS methods. My goal for these models was to establish whether a Bayesian approach was practical and could obtain forecast accuracies in the same range as the LS methods, not to develop *the* optimal Bayesian method. Given that Bayesian methods are practical and dependable, I believe that this approach will be more useful than LS methods because they provide a transparent way to estimate, recalibrate, and aggregate forecaster judgments within a hierarchy of clearly defined

priors and likelihoods.

2.1.1 Least Squares - LS

The least-squares (LS) model finds the best fitting distribution parameters $\hat{\theta}$ for a single forecaster's judgments for one forecasting question by minimizing the mean squared error between the elicited probabilities p_i and corresponding quantiles q_i (Eq 2.1). The objective function is simple, but can be difficult to optimize for functions with complex search spaces. This is particularly true when trying to fit distributions with many parameters. One way to improve optimization is to repeatedly fit the distribution across a grid of different starting values and take the estimated parameters that yield smallest error, though this doesn't guarantee that you will obtain the global minima. As the complexity of the search space increases, so does the number of unique starting values that should be evaluated, and therefore the necessary computational resources.

$$\arg\min_{\hat{\theta}} \sum_{i=1}^{n} \left[p_i - F(q_i|\hat{\theta}) \right]^2 \tag{2.1}$$

2.1.2 Dirichlet Multinomial - DM

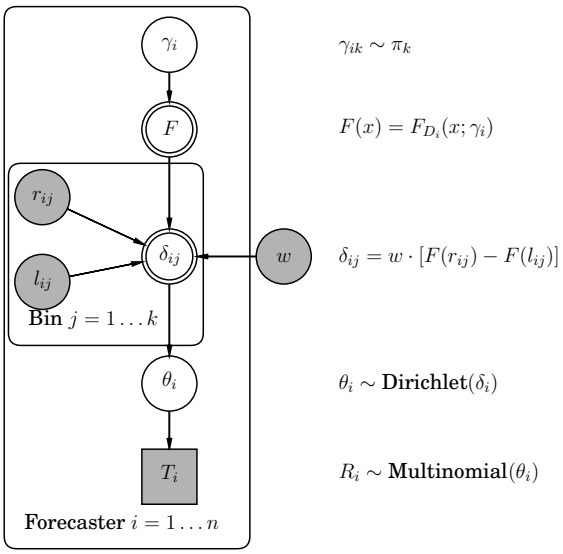

Figure 2.1: Dirichlet Multinomial model for forecaster judgments.

In the empirical dataset I evaluate in Chapter 3, forecasters distributed 100 tokens across k bins for each forecast. Instead of modeling probabilities like the LS method, the Dirichlet Multinomial (DM) method (Fig 2.1) directly models the distribution of tokens. Two potential advantages of this approach is that it directly models the empirical data, versus first transforming the tokens into probabilities, and it provides an intuitive way to think about the priors in the model.

Assume T_i is a k length vector of the observed distribution of a forecaster's 100 tokens across k bins. I model T_i as multinomial with k length parameter θ bin probabilities. Each token within one of the k bins represents a single occurrence of that interval. The k length Dirichlet prior δ_i on θ_i controls how closely the probabilities in θ_i correspond to the empirical token counts. You can consider the

δ_i as 'pseudo-counts' of observations that have already been observed. If a forecast had four bins, and $\delta_i = [5, 5, 5, 5]$, then the prior is effectively adding 5 additional tokens to each bin. If this was the top level of the model, since the Dirichlet is conjugate to the multinomial, the posterior distribution of bin probabilities would be distributed $\text{Dir}(t_1 + 5, t_2 + 5, t_3 + 5, t_4 + 5)$, where t_j was the token count in bin j for forecaster i.

In order to estimate a continuous distribution from the tokens, I add a continuous prior distribution D over the concentration parameters δ_i for each bin. The value of each δ_{ij} is the cumulative probability F_D assigns to the interval for each bin, defined by the left (l_{ij}) and right (r_{ij}) quantiles for each bin multiplied by a fixed scaling factor w. The model therefore estimates parameters for D, γ, that integrate the prior information for all the parameters and the observed data. I take the posterior estimate for D as the subjective distribution for each forecaster.

2.2 Forecaster Estimates

After producing a forecaster model, I selected the unit of forecaster belief to aggregate over, either cumulative probabilities (F), distribution parameters (θ), or avoiding forecaster models altogether, and just selected the empirical interval forecasts ($\forall$). Probabilities and parameters are subsequently passed to a linear aggregation function, while the empirical forecasts are directly fit with a parametric distribution at the consensus level.

2.2.1 Probabilities - F

Aggregating probabilities has three potential advantages: preserving consensus among forecaster models, relaxing parametric requirements at the consensus level, and reducing overdispersion. Linear aggregation guarantees that if a set of subjective distributions agree on the mean or modal value of a distribution, or on the width of any given interval, that this agreement is preserved through aggregation (Hora, 2004). For example, if all forecaster models agreed that there was a 60% chance that an event would occur in the next 30 days, then the linear combination of their judgments would also predict a 60% chance of the event occurring in the next 30 days. However, such consistency may not always be ideal, even if it at first it seems intuitive. (Wallsten and Diederich, 2001) show that for conditionally independent judgments, as the number of judges increases, the average probability judgment becomes increasingly diagnostic of the outcome. In the limit, if the average of all judgments exceeds 50%, the conditional probability of the event converges to 1. If the average judgment falls below 50%, the conditional probability converges to 0. This implies that if conditional independence holds, preserving consensus among the forecaster level models may under/over estimate the true conditional probability of the event, given the agreement among the forecaster models.

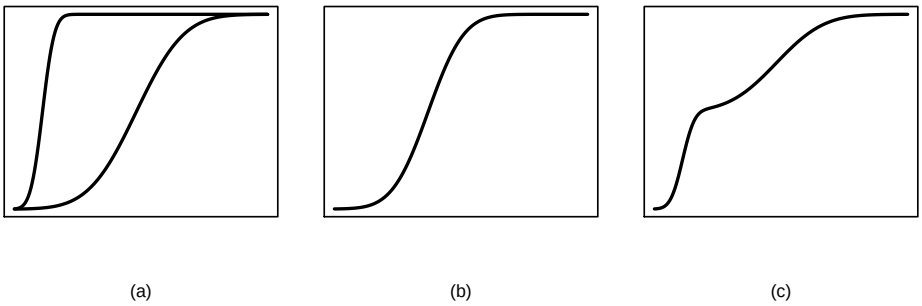

Figure 2.2: Examples of different aggregation methods. Panel (a) shows two forecaster level forecast CDFs, with different means and variances. Panel (b) shows the CDF that results from averaging the parameters of the distributions in panel (a). Panel (c) shows the linear combination of the CDFs in panel (a).

Aggregating over probabilities will yield a mixture consensus distribution that is potentially much more flexible than constraining the consensus to a single parametric form, and therefore capable of representing a much wider range of aggregate belief. Panel (a) in Figure 2.2 shows the CDFs for two continuous forecaster distributions. Clearly one distribution has a lower mean and smaller variance than the other. If you constrain the consensus distribution to follow one simple parametric form, then you'll obtain something similar to the CDF in panel (b), which obscures the differences in belief between the two forecaster distributions. Panel (c) is a linear combination of the two forecast distributions. While this consensus CDF lacks a simple form, it better represents the hetereogenity of belief at the forecaster level.

Linear aggregation over probabilities may also improve consensus accuracy since linear combination tends to reduce dispersion, and probability judgments tends to be overdispersed, i.e. "overconfident". In a sense, this is two wrongs making a

right. If forecaster models are well calibrated, then linear aggregation of the distribution functions will lead to underdispersed consensus forecasts, and may decrease consensus accuracy relative to other methods (Ranjan and Gneiting, 2010).

2.2.2 Parameters - θ

While there have been a few studies that fit continuous distributions directly to interval or cumulative probability judgments (Abbas et al., 2008; Wallsten et al., 2015), there are no published studies I am aware of that establish the properties of aggregating over forecaster level distribution parameters. Unpublished research by the Good Judgment Project shows that this can be a useful method (Tidwell et al., 2015), but until their research is more developed this evidence isn't conclusive. Nevertheless, there are at least two reasons that aggregating over parameters may yield accurate and useful consensus forecasts: it is a much more constrained aggregation goal than aggregating probabilities, and it is a computationally and mathematically simple operation that could be easily implemented by almost anyone.

To optimally aggregate probabilities, the forecaster models need to accurately reflect true subjective belief for the entire region of interest of the support of the forecasted variable, and the aggregation function needs to optimally combine the probabilities. If individual forecast production has a significant error component, and between forecaster belief is highly variable, then it may be difficult for probability aggregation to effectively filter signal from a very noisy channel. Aggregating distribution parameters, on the other hand, only needs to optimize the relevant

parameters for the consensus model. It is possible that relative to other methods, estimating and aggregating parameters will be more robust to extreme forecaster models and provide more consistent consensus forecasts.

The greatest potential benefit of aggregating parameters is it's simplicity. This method would require only very limited computational resources, and the aggregation steps are simple enough that the modeling could be done in a spreadsheet.

2.2.3 Empirical (no modeling) - $\forall$

Aggregating both probabilities and parameters requires continuous forecaster models. Given that these models are often fit to relatively few judgments for any individual forecaster, and given that forecasters tend to produce overdispersed judgments and/or introduce error at the time of forecast production, it is likely that continuous parametric distributions will often overfit the observed judgments and yield a distorted model of forecaster belief. One way to mitigate this potential distortion is to refrain from fitting forecaster level models at all, and instead treat the judgments from all forecasters as a single sample and estimate a consensus model directly from the empirical judgments. However, this method necessarily eliminates the distinction individual forecaster and consensus level error or variability, and imposes very constrained possible forms of belief at the consensus level.

2.3 Forecast Aggregation

I will aggregate forecaster model probabilities and parameters with the the simple mean and median. While there are more sophisticated methods to weight individual forecasts, given that consensus aggregation is two steps removed from the empirical judgments and that it is hard to predict how many of the candidate methods will behave with real-world data, I believe the best option is to keep the aggregation functions simple. Mean aggregation has proved very successful in many environments, but it can be very sensitive to extreme forecasts. In these cases, median aggregation should yield more robust consensus distributions (Lichtendahl et al., 2013). For the empirical judgments, I will fit the LS and DM models directly to all forecasts and bypass estimating forecaster level models.

2.4 Summary

Table 2.1 lists all combinations of forecaster models, estimates, and aggregation functions that I will test in this dissertation. I selected each of these to address some combination the potential consensus limitations described above. Aggregating probabilities will allow flexible consensus functions that reflect heterogeneity of belief between forecasters and potentially attenuate overfit of parametric distributions at the forecaster level. Aggregating fitted distribution parameters may be robust to extreme forecasts and, if accurate, could be easily implemented in real-world environments. If forecaster level models distort true subjective belief, then fitting

consensus distributions simultaneously to all judgments may yield more accurate forecasts. Finally, though the Bayesian and least-squares methods are similar, the Bayesian approach provides transparent and intuitive mechanisms to recalibrate forecasts and introduce additional information to the model outside of the empirical judgments.

Consensus Symbol	Aggregation Function	Forecaster Estimates	Forecaster Model
$M(\theta_{LS})$	Median	Distribution Parameters	Least-squares
$\mu(\theta_{LS})$	Mean	Distribution Parameters	Least-squares
$M(F_{LS})$	Median	Cumulative Probabilities	Least-squares
$\mu(F_{LS})$	Mean	Cumulative Probabilities	Least-squares
$M(F_{DM})$	Median	Cumulative Probabilities	Dirichlet-Multinomial
$\mu(F_{DM})$	Mean	Cumulative Probabilities	Dirichlet-Multinomial
$\theta_{LS}(\forall)$	Least-squares	Empirical Judgments	None
$\theta_{DM}(\forall)$	Dirichlet-Multinomial	Empirical Judgments	None

Table 2.1: Nomenclature and brief description of all candidate consensus methods.

Chapter 3: Analysis of Empirical Data

I validated the candidate consensus methods against a large forecasting data set collected by the continuous elicitation (CE) team of the Good Judgment Project. This team was led by Donald Moore and Thomas Wallsten, and was one component of the larger Good Judgment Project directed by Barbara Mellers, Don Moore, and Philip Tetlock. I was a member of this research team and led development of forecast modeling and aggregation methods, but the credit for designing, implementing, and managing this experiment belong to the entire CE team. The experiment and data I describe below are the product of many people's diligent effort, and were not directly collected for this dissertation; however, all of the work in this dissertation developing and validation consensus models is my own.

The experiment was a 10-month on-line forecasting tournament. We recruited 382 forecasters from a list of subjects who had volunteered to participate in the main Good Judgment Project (GJP) forecasting tournament, but who were not selected because the GJP had met it's quota for participants. We provided forecasters the opportunity to provide judgments for 127 forecasting questions. Forecasters could respond to as many, or as few, questions as they wanted, and they could revise their forecasts as many times as they chose to as long as a forecasted event had not yet

occurred.

All question topics were created by a team of political scientists from the University of Pennsylvania and focused on socio-political events of interest to the intelligence community. Topics included the likelihood of conflict between foreign nations, prices of commodities, foreign elections, and events related to disease and world health. I retained 47 of the 127 forecasting questions for consensus method evaluation. This was the total number of questions that were about time-based events, where the events had occurred so the forecasts could be scored, and were elicited in similar forms in the GJP and GJP-CE forecasting tournaments (Table 3.1).

Condition	Forecasts	IFPs
fixed	5931	47
random	4392	47

Table 3.1: Number of IFPs and forecasts by condition

Forecasters provided judgments through a web-based forecasting platform developed by Lumenogic, a forecasting and business solutions consultancy company. To make a forecast, forecasters moved tokens into or out of the bins assigned to them for that question (Figure 3.1). The platform required forecasters to exhaustively distribute all 100 tokens.

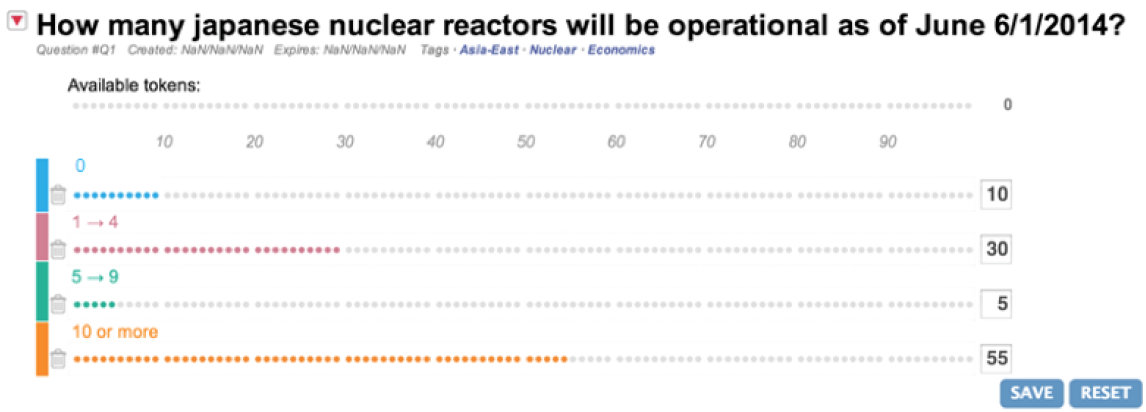

Figure 3.1: Screen shot of interval probability elicitation using Lumenogic's Fullwalk platform.

This analysis uses data taken from one GJP-CE experiment that was run during this tournament. In this experiment, we randomly assigned 382 forecasters to 3 different experimental conditions defined by 2 manipulations: how the range over which bins was set, and how the bins within this range were determined. We labeled the conditions Expert-Set:Fixed, Expert-Set:Random, and Self-Set:Proportional.

For each forecasting question, the computer divided the continuum for the question into four or five mutually exclusive and collectively exhaustive binned intervals, defined by three or four cut points. For the Expert-Set:Fixed and Expert-set:Random conditions, for each forecasting question a group of experts provided bounds on a subjective central 96% confidence interval (CI) around the correct answer, i.e., estimates of the 2nd and 98th quantiles that the true value might attain. We denote these two values L and H respectively. Focusing on the Fixed condition first, we used the L and H values to generate cut points in the following manner.

Let $R = H - L$ denote the range on the interval, C the number of cut points (i.e. number of bins - 1) and $S = R/C$ a unit value for setting the size of the interior bins. For questions assigned three cut points (four bins) we set cut-points at $L + 0.5S$, $L + 1.5S$ and $L + 2.5S$. For questions assigned four cut points (five bins), the first three cut points were as described and the fourth was set at $L + 3.5S$.

In the Random manipulation, we randomly generated three or four cut points between L and H for each question. In order to prevent the resulting intervals from being excessively small, we imposed restrictions ensuring that each interval was at least 1/10th (in the case of five bins) or 1/8th (in the case of four bins) of the total 96% confidence interval, and that the lowest and highest cut points were at least that minimum interval from L and H respectively.

For this analysis, I only retained forecasts from the Expert-Set:Fixed and Expert-Set:Random conditions, for a total of 484 forecasters, 11,420 forecast sets of 4 to 5 interval probability judgments, for 47 forecasting questions that elicited interval probability judgments for date-based events.

3.1 Methods

3.1.1 Modeling Forecasters

Each unique forecast consisted of three to five interval judgments for that forecasting question, with 100 tokens exhaustively distributed across the elicited intervals. I fit continuous distributions to these judgments with the least squares (LS) and Dirichlet multinomial (DM) forecaster methods. I used both methods to fit

exponential, gamma, and generalized gamma distributions to the observed forecasts.

Least Squares (LS)

For each forecast, I first converted the cut points from dates to the relative number of days from the date of the forecast. Assume a forecaster responded to a question on January 6th and that the cut points for the question were January 10th, January 15th, and January 20th. The platform would display the question with judgments for four intervals: before January 10th, January 10th to January 14th, January 15th to January 19th, and January 20th or later. The number of days relative to the forecast date for these cut points would then become $q_i = \{4, 9, 14\}$, where q is the quantile for the distribution that will be fit to the judgments for forecast i.

I divided the number of tokens in each elicited bin by 100 to obtain a vector of pseudo probabilities p_i. In order to fit any of the distributions to a forecast set, there must be at least as many non-zero bins as there are parameters in the distribution, otherwise no solution exists. Approximately 39% of forecasted bins contained no tokens, and consequently 0% probability. Since this would have substantially reduced the number of forecasts that could be fit with distributions, I assigned .05% probability to any 0 value in p_i, and then renormalized the the vector to sum to 1.

$$\arg\min_{\hat{\theta}} \sum_{i=1}^{n} \left[p_i - F(q_i|\hat{\theta}) \right]^2 \tag{3.1}$$

For each p_i, I searched for distribution parameters than minimized Eq 3.1

across a grid of different starting values in order to minimize the chance of a single optimization run converging on a local minima, and retained the fit that minimized the objective function. I conducted all optimization using the `optim` function in `R` (R Core Team, 2015).

For exponential distributions, I used simulated annealing to optimize the mean of the distribution (Eq 3.2), $\lambda = \frac{1}{\mu}$, with starting values of $\mu = \{1, 10, 25, 50, 100, 200, 400\}$. For gamma distributions (Eq 3.3) I used Broyden, Fletcher, Goldfarb and Shanno (BFGS) optimization, with the factorial combination of $\alpha = \{.5, 1, 2, 4, 8, 10\}$ and $\beta = \frac{\alpha}{\mu}$ as starting values.

$$\lambda e^{-\lambda x} \tag{3.2}$$

$$\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{3.3}$$

$$\frac{\alpha \beta^{-\alpha c}}{\Gamma(c)} t^{\alpha c - 1} \exp\left[\frac{-t}{\beta}^{\alpha}\right] \tag{3.4}$$

Dirichlet Multinomial (DM)

As with the LS method, I converted the cut points from dates to the relative number of days from the date of the forecast. Unlike the LS method, there was no need to transform the token judgments to pseudo probabilities, since the DM method directly models the distribution of tokens across the bins (Fig 3.2).

Each forecast i consisted of k bins with 100 tokens exhaustively distributed

across the bins. The distribution of these tokens T_i was modeled as multinomial with probability parameter θ_i. I placed a Dirichlet prior over this parameter. The vector of concentration parameters for the Dirichlet, δ_i was fully determined by a weight w and the cumulative probability of a continuous hyper-prior F_D and the left and and right bounds of each bin (l_{ij}, r_{ij}). For example, if the bounds for the first bin were 0 and 10, then δ_{i1} would be $w[F_D(10|\gamma) - F_D(0|\gamma)]$, where γ is the set of parameters for the hyper-prior.

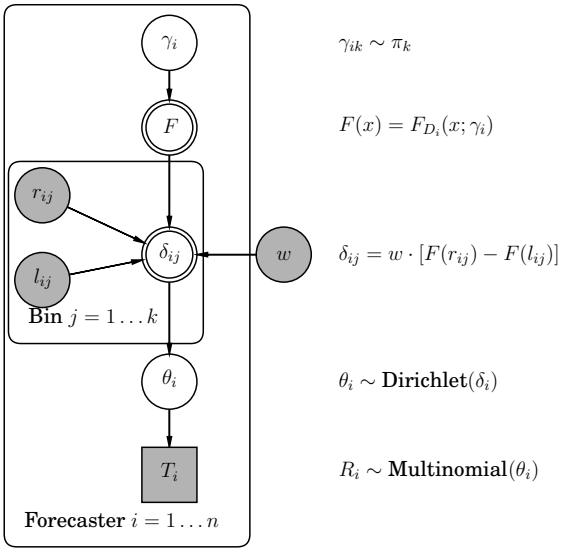

Figure 3.2: Dirichlet Multinomial model for forecaster judgments.

I held $w = 100$ constant across all forecasts. This was an entirely practical decision. The relatively large value of w helped the models to more consistently converge in a reasonable number of samples. However, this also amplified the influence of F_D relative to the observed counts. I estimated each forecast with F_D distributed as an exponential, gamma, and generalized gamma distribution. For exponentials, I set the prior on the rate parameter as $\lambda \sim Exp(\frac{1}{10})$. For gamma distributions, I

set the prior shape as $\alpha \sim Gamma(1, \frac{1}{4})$ and prior rate $\beta \sim Exp(\frac{1}{10})$. I estimated all forecaster DM models with with the Stan Markov Chain Monte Carlo (MCMC) sampler (Stan Development Team, 2015) in `R`, with 4 chains, 3,000 samples, and I discarded the first 1,500 samples from each chain as burn-in after sampling completed.

3.1.2 Consensus Models

Each of the eight consensus methods I validated on this data combined a forecaster level model (LS,DM), a unit to aggregate over, either parameters (θ) or probabilities (F), and either a mean (μ) or median M aggregation function. Each method is identified by the combination of these components. For example, $\mu(F_{LS})$ denotes the method that takes the mean of forecaster level cumulative distribution functions fit with least squares.

Each consensus method had to generate a forecast for every day that a forecasting question was open. Since the number of forecasters available to aggregate over each day depended on how many forecasts had been made, and since the left bound 'Day 0' for any consensus forecast was the day the forecast was made, I calculated a new consensus distribution for every day, for every method, for every forecasting question.

All consensus methods followed the same forecast selection rules. Many forecasting questions remained open over long intervals, up to nine months, and it is likely that very old forecasts would reduce the accuracy of consensus models. For a

given question and day, I selected all forecaster level models for forecasts made on or before that day. Of those forecasts models, I retained the greater of either 50% of the forecasts or the number of forecasts made in the previous three days. The resulting set of forecast models was then passed on to the consensus method specific modeling functions.

Probability Methods: $\mu(F_{LS})$**,** $M(F_{LS})$**,** $\mu(F_{DM})$**,** $M(F_{DM})$

$$F_i^* = \frac{F_i(t - t_0)}{S_i(t_0)} \tag{3.5}$$

The first step for each consensus method was to set all forecaster models on the same scale. Assume that a there were three forecasts for a question, elicited on January 1st, January 5th, and January 10th. I would first fit a continuous distribution to the forecaster judgments (In practice, I fit each forecaster distribution only once and then referenced the fitted parameters for any consensus method that required them). Since the left bound of day 0 each fitted forecast F_i is different (1st, 5th, 10th), the forecast models can not be directly aggregated. For example, for a consensus forecast on January 15th, the cumulative probability for each distribution for the event occurring on January 20th would be $F_1(19)$, $F_2(15)$, and $F_3(10)$.

To set the forecaster models on the same scale, I took conditional cumulative probabilities of the forecaster distributions (Eq 3.5). If t_0 is the date of the consensus forecast and t the date to forecast a probability, the a forecaster's conditional distribution $F_i^*(t)$ scales the probability of the event happening between date t_0 and

t by the forecaster's survival function $S_i(t_0)$, the probability that the event occurs after t_0.

The median and mean methods took the respective average of the forecaster conditional distributions at any date t as the value of the consensus CDF. The $\mu(F_{LS})$ and $M(F_{LS})$ methods used conditional distributions based on the least squares fits, and the $\mu(F_{DM})$ and $M(F_{DM})$ methods used conditional distributions based on the Dirichlet Multinomial fits.

Parameter Methods: $\mu(\theta_{LS})$, $M(\theta_{LS})$

In order to average over forecaster parameters, I required parameters for every forecaster for every day a forecasting question was open, and also parameters that were based on the same day 0. Assume the same set of forecasts in the previous example. For the θ consensus methods I divided the forecaster conditional distributions F_i^* by 20 equidistant cut points between t and $t + 180$. This yielded 20, 9-day, intervals and one open interval $t > 180$. I then re-fit a continuous distribution to these intervals with the LS method to obtain an updated set of parameters.

The independent average of the forecaster parameters became the consensus distribution parameters. For example, if it was a gamma distributed $M(\theta_{LS})$ consensus model, the consensus method would be distributed Gamma$(\bar{\alpha}, \bar{\beta})$, where α and β were vectors of all the forecaster shape and rate parameters.

Empirical Methods: $\theta_{LS}(\forall)$, $\theta_{DM}(\forall)$

Both θ methods directly modeled the forecaster judgments instead of aggregating over forecaster level models, effectively treating each set of forecasts for a consensus forecast day as a single sample. However, like the previous methods, forecasts first had to be equated on the same Day 0.

Rather than refit distributions, these methods reallocated tokens. I proportionally redistributed tokens from bins that occurred prior to the consensus forecast day, or for which the consensus forecast day fell inside the bin. Assume a forecast made on January 5th assigned 20 tokens to the interval [Jan 5th - Jan 14th) and the remaining tokens to subsequent bins. For a consensus forecast on January 10th, I would remove 10 tokens from the first bin, and divide the vector tokens for all bins by .9, the remaining percentage of tokens left.

For the $\theta_{LS}(\forall)$ I then fit exponential, gamma, and generalized gamma distributions as I did at the forecaster level, except instead of fitting one forecaster's pseudo probabilities, I simultaneously fit all the forecasters' judgments. For the the $\theta_{DM}(\forall)$ method, I similarly fit all bins of tokens for all forecasters.

3.2 Results

3.2.1 Model Quality

Many forecasts did not include enough non-zero bins to fit the gamma or generalized gamma distributions, and some forecasters assigned all 100 tokens to

the top open interval, which prevented even fitting an exponential distribution. Although I redistributed a small amount of probability to the zero bins as detailed above, I was still unable to successfully fit some of the forecasts.

$$R^2 = 1 - \frac{\sum_i (p_i - \hat{F}(q_i))^2}{\sum_i (p_i - \bar{p})^2} \tag{3.6}$$

For LS fits, I defined a successful fit as any model that the optimization routine converged and the resulting coefficient of determination R^2 (Eq 3.6) was greater than 0. R^2 measures how closely a model corresponds to the observed data, where $\hat{F}$ is the estimated cumulative distribution function, p_i the cumulative observed probability judgments, and q_i the cut points used to elicit the forecaster's judgments. If a model perfectly fits the data, $R^2 = 1$. Any model that fits the data worse than the mean will yield $R^2 < 0$. This statistic only measures how closely the fitted cumulative distribution corresponds to the observed judgments, and is not any measure of how well the fitted model represents a forecaster's latent continuous subjective belief.

For DM fits, I defined a successful fit as any model that yielded with $R^2 > 0$ and a Gelman-Rubin convergence diagnostic $\hat{R} < 1.2$ (Gelman and Rubin, 1992). $\hat{R}$ assesses convergence by comparing the between chain and within chain variances for each estimated parameter. If all parameters obtain an $\hat{R} < 1.2$ the model has demonstrated strong convergence evidence (Brooks and Gelman, 1998).

Table 3.2 contains the proportion of successful fits by forecaster modeling method and distribution. On one end the exponential distribution fit virtually all forecasts, while neither the LS or DM methods successfully fit a generalized

gamma distribution to a large proportion of the forecasts. For the DM method, less than half of the forecasts were fit successfully with the generalized gamma distribution. I judged that the difference in the number of forecasts available for aggregation, and therefore the amount of potential signal to extract from the crowd, between the few generalized fits and the other distributions was too great to draw any meaningful inferences, no matter the results. I therefore exclude the generalized gamma distribution from all analyses other than fit quality.

	Distribution			
Condition	*exp*	*gamma*	*gengamma*	Method
Fixed	0.99	0.84	0.78	LS
Random	0.99	0.77	0.71	LS
Fixed	0.98	0.83	0.48	DM
Random	0.98	0.83	0.45	DM

Table 3.2: Proportion of 'successful' fits by forecaster model method and distribution. For LS models, a successful fit was defined as any fitted distribution for which the optimization routine converged and the resulting $R^2 > 0$. For DM fits a successful fit was any model with an $\hat{R} < 1.2$ and $R^2 > 0$ for a distribution parameterized with the means of the posterior distribution of parameters.

For forecasts that were fit successfully, fit quality was uniformly high (Table 3.3). The gamma and generalized gamma distributions, whether LM or DM models, obtained nearly perfect fits. As expected, the exponential distribution fits were .15

to .20 lower than the other distributions. Since many of the forecasts contained only 3 bins, the gamma and generalized gamma distributions are flexible enough to fit almost any response pattern. The single parameter exponential distribution, on the other hand, is far less flexible. Unless judgments happened to be distributed perfectly exponential, this distribution could not obtain fits as close to the observed judgments as the other distributions.

	Distribution			
Condition	*exp*	*gamma*	*gengamma*	Method
Fixed	0.87	1.00	1.00	LS
Random	0.77	0.99	1.00	LS
Fixed	0.89	0.99	1.00	DM
Random	0.79	0.98	0.99	DM

Table 3.3: Mean coefficient of determination R^2 for all forecaster model fit methods and distributions, only including LS models that obtained $R^2 > 0$ and converged, and DM models that obtained $R^2 > 0$ and $\hat{R} < 1.2$

3.2.2 Calibration

Forcasters

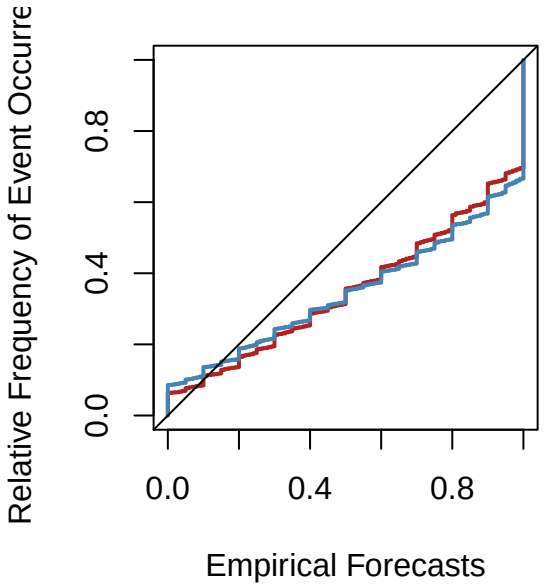

Figure 3.3: Calibration curves for empirical forecasts for the fixed (red) and random (blue) elicitation conditions. The black identity line represents perfect calibration. Both conditions show severe overconfidence.

Empirical forecasts were substantially overconfident except for low probability judgments (Fig 3.3). I assessed calibration of the forecaster models with the variation and dispersion of the PIT histogram. I first obtained a PIT variance for each forecaster by calculating the cumulative probability for the outcome of each of their forecasts using each combination of fitting method (LS, DM), and fitted distribution type (exponential, gamma). This provided four variances per forecaster, one for each combination of methods, and treats each combination of subject and

specific forecast modeling method as a unique forecasting unit.

The primary difference in calibration for forecasters was by fitted distribution. Figure 3.4 shows that gamma distributed fitted distributions were consistently more underdispersed (PIT variance $> .083$) than the exponential fits ($\mu_{diff} = 0.035, SD = .003$).

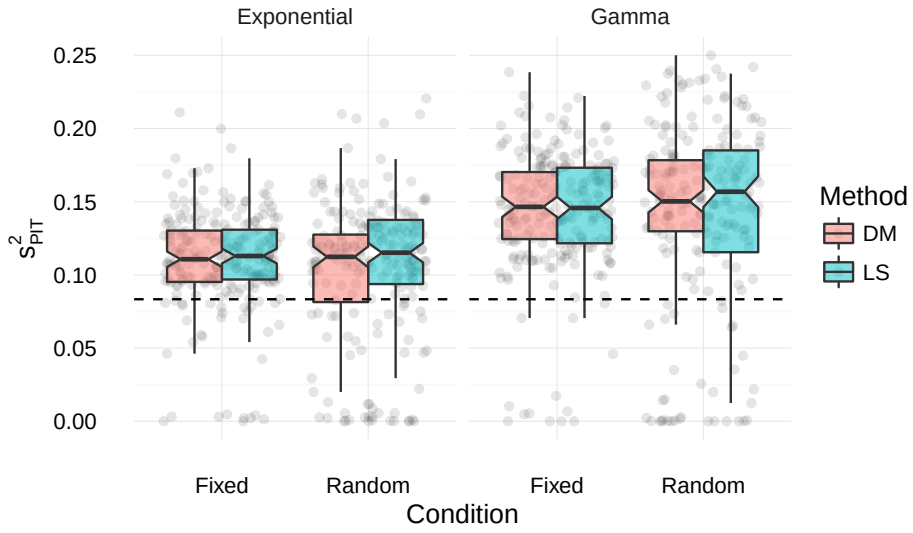

Figure 3.4: Distribution of forecaster PIT variances (S^2_{PIT}) by elicitation format, modeling method, and fitted distribution. Boxplot notches show the asymptotic standard error of the median. Individual values are jittered points.

I evaluated the calibration of each experimental condition and method by considering each combination of condition and method as a forecasting unit. For each experimental condition (fixed bins, random bins), I calculated the PIT across every forecast for each of the combinations of fitting method (LS, DM), and fitted distribution type (exponential, gamma). Each forecast from each forecaster contributed fours scores, one for each combination of fitting methods. The PIT histograms

(Fig 3.5) show that all conditions and manipulations were underdispersed, with all variances well above the neutral threshold of .083 (Table 3.4).

The left skew of the histograms also reveals that forecasts were biased to predict events occurring far sooner than they actually did. The relative height of the last decile / bar each plot indicates that the probability the fitted distributions assigned to many events was close to 100%, and therefore that the events occurred far into the right tail of the distributions.

Since the PIT method yields only a single histogram/variance for each combination of methods and conditions, I estimated marginal differences in dispersion with an independent non-parametric bootstrap for each of the comparisons. For elicitation condition, fitting method, and distribution type, I calculated the difference in PIT variance between the levels (e.g. fixed - random) on 1,000 resamples of the data. The entire dataset (10,323 forecasts) was sampled with replacement each iteration. Random elicitation was consistently more underdispersed than fixed elicitation (95% CI[.010, .014]), least-squares fitting more underdispersed than Dirichlet-multinomial (95% CI[.004, .008]), and the gamma distributions were more underdispersed than the exponential (95% CI[.025, .029]).

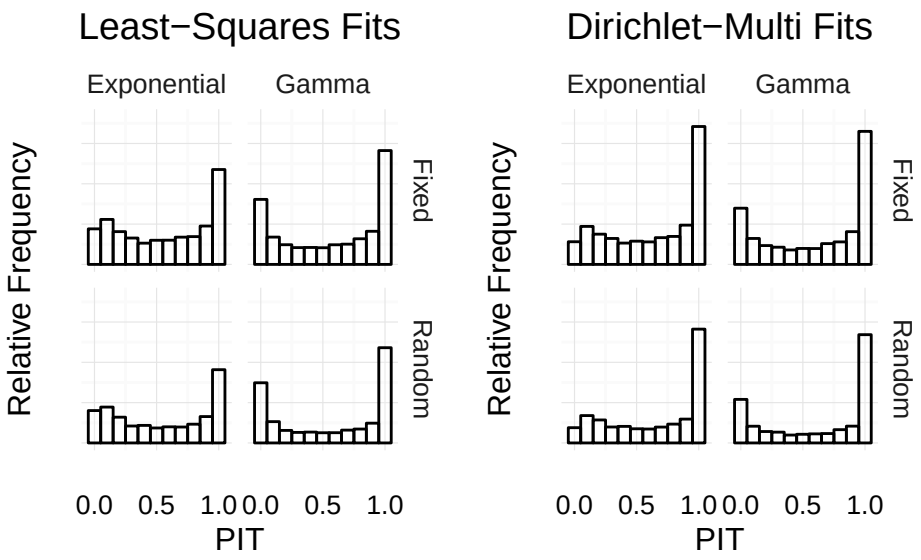

Figure 3.5: PIT histograms for each elicitation condition (Fixed, Random), forecast fitting method (Least-Squares, Bayes), and fitted distribution (Exponential, Gamma)

Distribution	Condition	Least-Squares	Dir-Mult
Gamma	Random	0.17	0.16
Gamma	Fixed	0.15	0.15
Exponential	Random	0.14	0.13
Exponential	Fixed	0.13	0.12

Table 3.4: PIT variance across forecasters for each experimental condition, fitting method, fitted distribution

Consensus

Every consensus method produced a forecast for every forecasting question, for every day that the question was open and that there was at least one forecast

level forecast. In total, there were 138,611 unique consensus forecasts across all the consensus methods. In general, the consensus methods tended to mitigate the severe underdispersion of the forecaster level distributions, though there are some cases where underdispersion increased relative to individual forecasters. For example, the combination of the $\mu(\theta_{LS})$ consensus method and gamma distributions yielded not only underdispersed, but extremely biased forecasts (Fig 3.6).

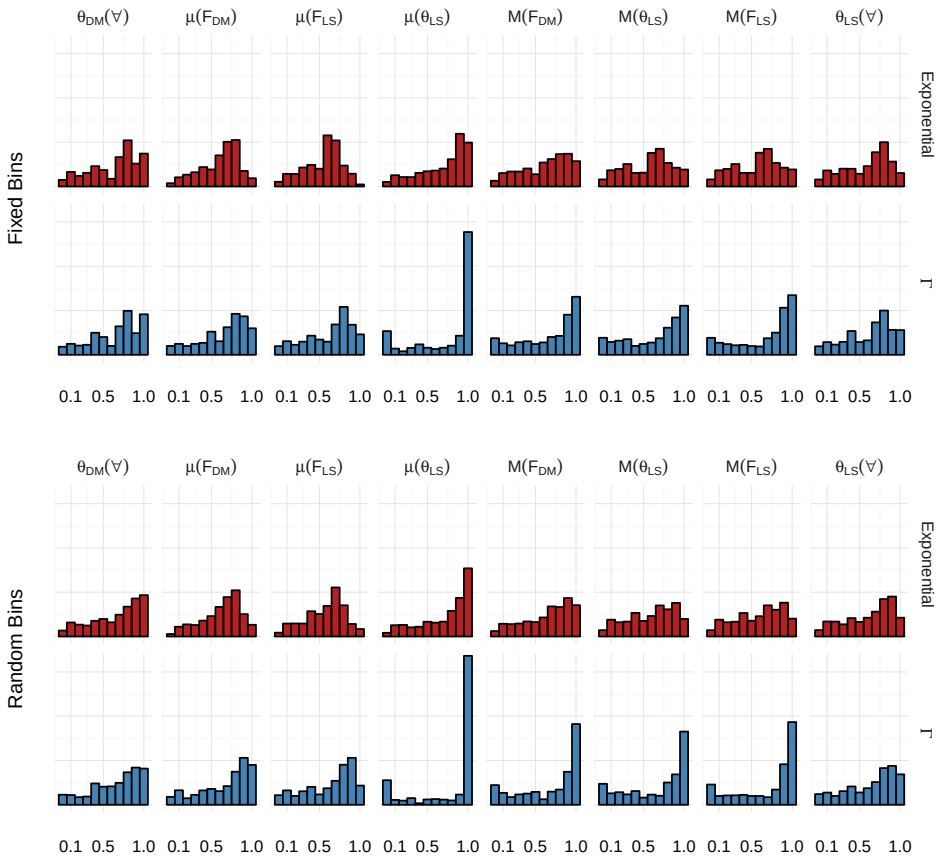

Figure 3.6: PIT histograms for every consensus method, by elicitation format and distribution type.

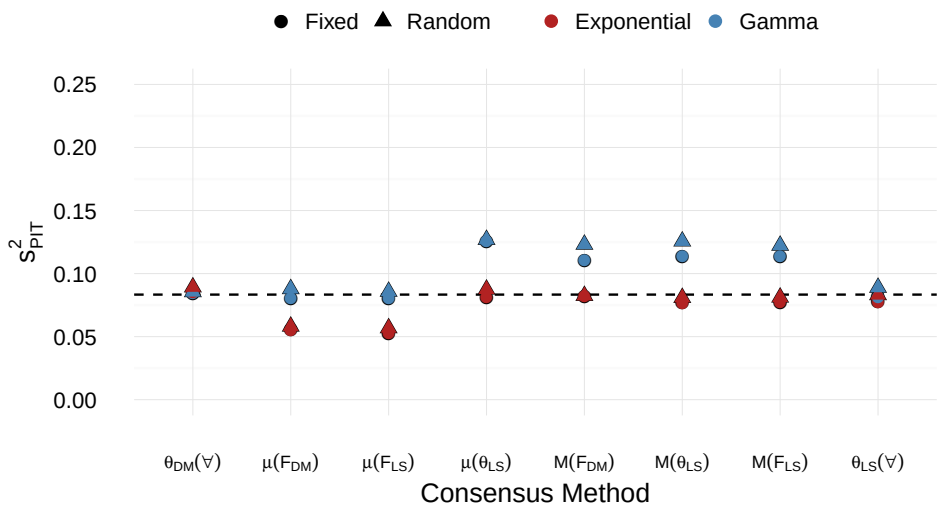

Figure 3.7: PIT variances by consensus method, condition, and distribution.

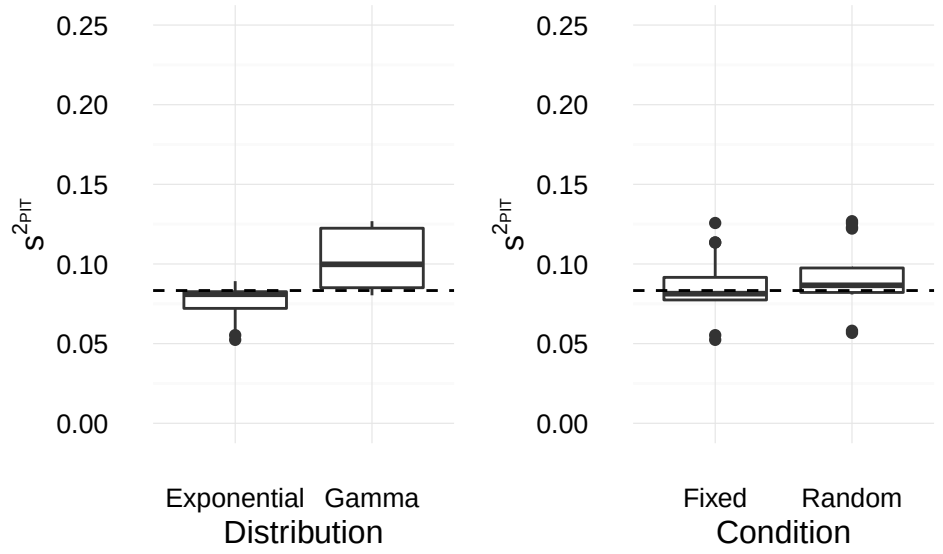

Figure 3.8: Distribution of PIT variances by condition and distribution.

While there were only minor differences in dispersion between methods and distributions at the forecaster level, there were clear differences in the calibration (Fig 3.6) and PIT variances (Fig 3.7) at the consensus level. Across consensus

methods, the gamma distributions tended to be underdispersed while the exponential distributions were slightly overdispersed (Fig 3.8).

Given the extreme bias of the forecaster level models, and the attenuated yet still considerable bias at the consensus level, the variance of the PIT may be less informative than visual distribution of the histograms. As a simpler alternative to PIT variances, I calculated the central 50% and 90% confidence intervals for each consensus method, condition, and distribution (Table 3.5). A well calibrated forecast method would capture the same proportion of events as the central confidence interval. Across methods the gamma distribution models tended to exhibit greater overprecision than the random distribution models (Fig 3.9). You could infer these intervals directly from the PIT histograms, but the table and boxplot format makes it easier to compare intervals across methods and conditions. While the pattern of intervals is similar across methods, the mean-theta method for the random condition obtained extremely narrow confidence intervals.

		Exponential		Gamma	
Method	Condition	50%	90%	50%	90%
$\theta_{DM}(\forall)$	Fixed	40	74	39	72
	Random	36	66	40	68
$\theta_{LS}(\forall)$	Fixed	47	83	43	76
	Random	40	77	38	70
$\mu(F_{DM})$	Fixed	57	90	39	72
	Random	53	87	33	64
$M(F_{DM})$	Fixed	44	76	30	55
	Random	41	69	24	42
$\mu(F_{LS})$	Fixed	70	90	41	76
	Random	63	90	37	73
$M(F_{LS})$	Fixed	55	82	29	59
	Random	48	80	23	48
$\mu(\theta_{LS})$	Fixed	33	64	17	27
	Random	28	55	11	18
$M(\theta_{LS})$	Fixed	55	81	24	47
	Random	48	80	20	39

Table 3.5: Width of consensus method central 50% and 90% confidence intervals by experimental condition, fitting method, and distribution.

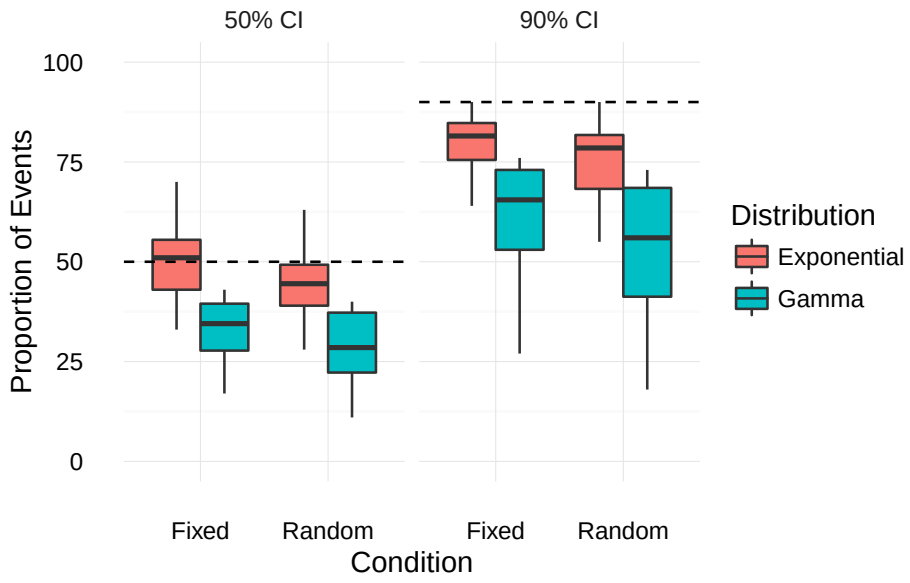

Figure 3.9: Proportion of events occurring within the 50% and 90% central confidence intervals for exponential and gamma distributions, collapsed across consensus methods.

3.2.3 Accuracy

I scored discrete forecasts with the rank probability score (RPS) in Equation 3.7 and continuous forecasts with the continuous rank probability score in Equation 3.8. The RPS is a Brier score for questions with at least two categories. This score divides forecasts for n ordered alternatives into a series of binary alternatives as the threshold moves up from lower to higher. The Brier score is calculated for each of the $n-1$ binary partitions and the final result is the average of the separate scores where n is the number of cut points; q_i the quantile associated with cut point i, $i = 1...n$; o the observed value; F the predictive cumulative distribution function; and $\mathbb{1}_{o \leq q_i}$ equals 1 if o is less than or equal to q_i, otherwise 0. In the form in Equation

3.7, the scores can range from [0,2], with 0 the best possible score and 2 the worst possible score.

$$S_R = \frac{2}{n} \sum_{i=1}^{n} \left[F(q_i) - \mathbb{1}_{o \leq q_i} \right]^2 \tag{3.7}$$

Since S_R is dependent on the location of partitions, I could not score the cut points used to originally elicit the forecasts. Although the fixed and random elicitation conditions had identical minimum and maximum values in which the cut points could be distributed, the location of the cut points differed by condition. The cut points were identical for every forecaster in the fixed condition, and effectively different for every forecaster in the random condition. For example, assume a continuous forecast distributed $N(0, 1)$ and an outcome $o = 1$. If a forecasting question had one cut point at 0, then the forecast would score $2\left[F(x < 0) - \mathbb{1}_{1 \leq 0}\right]^2 = 2(.5 - 0)^2 = .5$. However, if the cut point was at 1.64, then the forecast would score $2\left[F(x < 1.64) - \mathbb{1}_{1 \leq 1.64}\right]^2 = 2(.95 - 1)^2 = .005$. Although the only difference in examples was the location of the cut point, the second forecast appears much more accurate than the first.

I scored continuous forecasts, i.e. the fitted continuous forecaster and consensus distributions, with the Continuous Rank Probability Score (CRPS). The CRPS (Eq 3.8) is a generalization of mean absolute error (MAE) and is equivalent to integrating the RPS across all possible real-valued thresholds (Matheson and Winkler, 1976). It is a generalization of MAE in the sense that since the score integrates the area between the predictive CDF and the step function for the observed outcome

$(P(x < o) = 0, P(x \geq o = 1))$ over the support of the variable, the score will be in units of forecasted event. The more probability density that a predictive distribution places close to the true outcome, the lower the score will be.

$$S_C = \int_{-\infty}^{\infty} [F(x) - \mathbb{1}_{o \leq q_i}]^2 \tag{3.8}$$

3.2.3.1 Forecaster Level

RPS (S_R)

Both the modeled and empirical forecaster RPS scores were quite variable (Fig 3.10). I evaluated the modeled forecasts with a varying-intercepts beta regression using `R` (R Core Team, 2015) and the `rstan` package (Stan Development Team, 2015) in order to estimate the effects of the conditions and forecast modeling methods on the RPS scores.

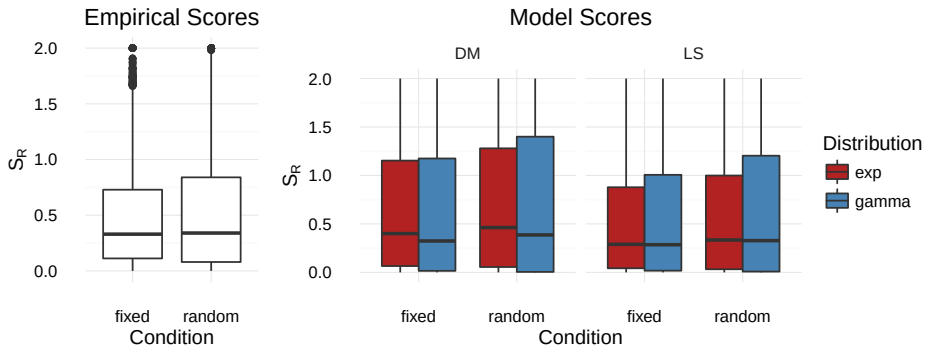

Figure 3.10: Distribution of S_R scores. The left panel displays scores for the empirical forecasts, scored against the cut points used to elicit the judgments. The right panel displays scores for the forecaster level continuous models by modeling method, distribution, and condition, scored against the GJP cut points.

		Method	
Condition	Distribution	DM	LS
Fixed	Exp	0.66 (0.67)	0.54 (0.59)
	Gamma	0.64 (0.70)	0.59 (0.66)
Random	Exp	0.71 (0.71)	0.59 (0.64)
	Gamma	0.71 (0.76)	0.65 (0.72)

Table 3.6: Mean (SD) S_R for all conditions and forecast modeling methods at the forecaster level.

I modeled the scores as the linear combination of regression coefficients β and intercept-free model matrix X of the full interaction of condition (fixed, random), distribution (exponential, gamma) and model type (Dirichlet-multinomial, least-squares). Each predicted score was offset by a varying intercept α_j for the corresponding forecasting question for the score to account for differences in the

forecasting questions that could skew the results. I set a $N(logit(.25), 1)$ prior on the constant α_0 to regularize the scores towards .25, the RPS score for a forecaster who always guesses 50% probability for a dichotomous event. The remaining priors are included in Figure 3.11.

$$
\begin{aligned}
S_{R_i} &\sim \mathcal{B}(\mu_i\phi, (1-\mu)\phi) \\
\text{logit}(\mu_i) &= \alpha_{j[i]} + \beta_0 + X_i\beta \\
\alpha_0 &\sim N(\text{logit}(.25), 1) \\
\beta &= 10 \cdot \theta \\
\theta &\sim N(0, 1) \\
\phi &\sim \text{Cauchy}(0, 5) \\
\alpha &\sim N(0, \sigma_\alpha) \\
\sigma_\alpha &\sim |N(0, 1)|
\end{aligned}
$$

Figure 3.11: Hierarchical beta regression model for forecaster S_R scores, with varying intercepts α_j for each forecasting question j; model matrix X of predictors for experimental condition, fitted distribution, and forecaster model; and regression coefficients β.

I sampled 4 chains, with 3,000 iterations each, and discarded the first 1,500 samples in each chain as a burn-in. I assume an adequate posterior sample based on visual inspection of the predicted scores and distribution of the Gelman-Rubin $\hat{R}$ convergence diagnostic (Gelman et al., 2013) (Fig 3.12). All of the regression

coefficients, excluding the varying intercepts, are included in the caterpillar plot in Figure 3.13. Parameters for condition, distribution, and modeling method imply a comparison to the unnamed reference level, e.g. the reference level for the 'random' parameter is the the fixed condition.

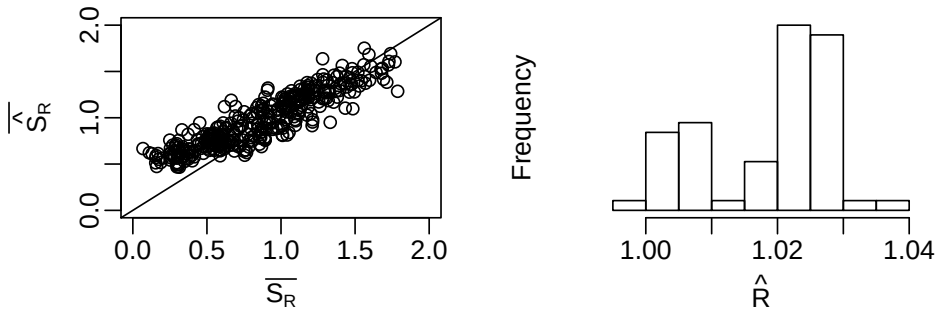

Figure 3.12: Diagnostic plots for Bayesian varying-intercepts beta regression model for forecaster level RPS scores. The left panel plots the mean empirical S_R and posterior $\hat{S}_R$ scores by condition, method, distribution, and forecasting question. The right panel shows the distribution of the Gelman-Rubin convergence diagnostic $\hat{R}$.

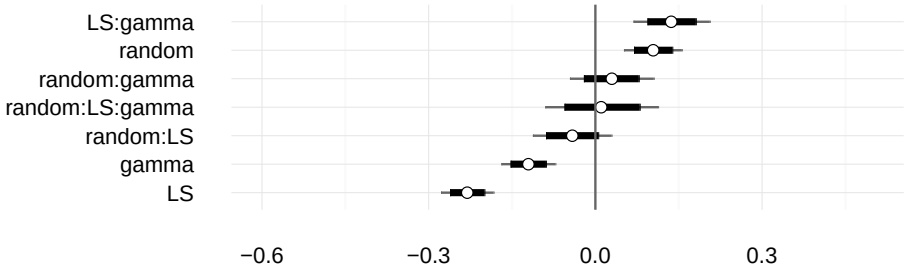

Figure 3.13: HDI and posterior median for each non-varying coefficient for the varying-intercepts beta regression model for forecaster level RPS scores. The thick black lines represent the posterior 50% HDI. The thin lines represent the 95% HDI.

Though it's easy to discern the credible intervals of the coefficients in Figure 3.13, it can be somewhat more difficult to mentally transform coefficients to direct differences in manipulations and conditions. Figure 3.14 plots the relevant marginal differences in means for the method, distribution, and elicitation format. Forecaster model had the largest effect ($\mu = -0.08$, 95% HDI [-0.10, -0.06]), though the gamma distribution scored reliably lower than the exponential and fixed bins scored lower than random. The LS forecaster model with fixed bins, both for the gamma and exponential distributions scored lower than any other combination of method and condition (Table 3.7).

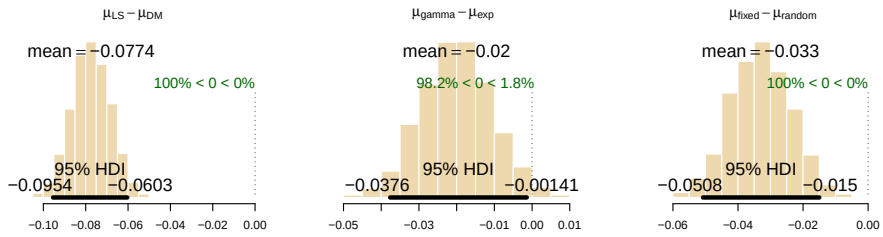

Figure 3.14: Posterior distribution of simple effects for the varying-intercepts beta regression model for forecaster level RPS scores. Solid horizontal black lines display the width of the 95% HDI, vertical green lines with text show what percentage of the distribution falls above/below 0.

		Method	
Distribution	Condition	DM	LS
Exponential	Fixed	0.39 [0.38, 0.40]	0.34 [0.33, 0.35]
	Random	0.41 [0.40, 0.42]	0.35 [0.34, 0.36]
Gamma	Fixed	0.36 [0.35, 0.38]	0.34 [0.33, 0.35]
	Random	0.39 [0.37, 0.40]	0.36 [0.35, 0.37]

Table 3.7: Posterior 'cell' medians (95% HDI) for each combination of experimental condition, distribution, and forecaster model for the varying-intercepts beta regression model for forecaster level RPS scores.

CRPS (S_C)

The CRPS can be difficult to compare across forecasting questions because it is sensitive to distance and can be interpreted in units of the forecasted variable. Assume two continuous forecasts for two independent events. The first event occurs 10 days after the forecast, and the second event occurs 100 days after the forecast, and both forecasts yield scores of $S_C = 1$. In terms of the magnitude of the scores, the forecasts are equivalently accurate, about one day off from the true value. The forecast for the event that occured 100 days from the forecast is clearly more impressive, but without transforming the score, they appear to equally accurate.

$$S_\gamma = \frac{S_C}{\text{Resolution} - \text{Date}} \tag{3.9}$$

In order to make the scores for each forecasting more directly comparable, I scaled all scores by the distance in days from the forecast to the event resolution (Eq 3.9). I interpret this transformed score as the relative absolute error of the forecast as a proportion of the distance in days between forecast production and event resolution. The scores for the previous example would then become $\frac{1}{10}$ and $\frac{1}{100}$, with the score for the further event becoming ten times as relatively accurate as the the more immediate event.

I was also interested in whether where we set the elicitation cut points relative to the event resolution affected the forecast scores. To evaluate this, I included a covariate in all CRPS analyses that was an index of this relative distance between event resolution and the elicitation interval. GJP-CE assigned a maximum and minimum date within which cut points could be assigned for each forecasting question. For the fixed elicitation condition, the minimum cut point delineated the first elicitation bin, and the maximum cut point delineated the minimum bound of the last open-interval bin. For the random elicitation condition, all cut points were randomly distributed between the maximum and minimum. I defined a measure d as the ratio of the distance in days between the event resolution and minimum date (Range_{low}) and the width in days of interval over which cut points were distributed (Eq 3.10).

$$d = \frac{\text{Resolution} - \text{Range}_{low}}{\text{Range}_{high} - \text{Range}_{low}} \tag{3.10}$$

I modeled the scores as the linear combination of regression coefficients β

and intercept-free model matrix X of the full interaction of condition (fixed, random), distribution (exponential, gamma) and model type (Dirichlet-multinomial, least-squares), and the continuous predictor d. Each predicted score was offset by a varying intercept α_j for the corresponding forecasting question for the score to account for differences in the forecasting questions that could skew the results. I used mildly regularizing priors on all parameters (Fig 3.15).

$$log(S_{C_i}) \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha_{j[i]} + X_i\beta$$

$$\sigma \sim |N(0, 10)|$$

$$\alpha \sim N(0, 5)$$

$$\beta = 10 \cdot \theta$$

$$\theta \sim N(0, 1)$$

$$\sigma_\alpha \sim |N(0, 10)|$$

Figure 3.15: Hierarchical regression model for forecaster S_C scores, with varying intercepts α_j for each forecasting question j; model matrix X of predictors for experimental condition, fitted distribution, forecaster model, and distance d; and regression coefficients β.

I sampled 4 chains, with 3,000 iterations each, and discarded the first 1,500 samples in each chain as a burn-in. I assume an adequate posterior sample based on visual inspection of the predicted scores and distribution of the Gelman-Rubin

$\hat{R}$ convergence diagnostic. All of the regression coefficients, excluding the varying intercepts, are included in the caterpillar plot in Figure 3.17. Parameters for condition, distribution, and modeling method imply a comparison to the unnamed reference level, e.g. the reference level for the 'random' parameter is the the fixed condition.

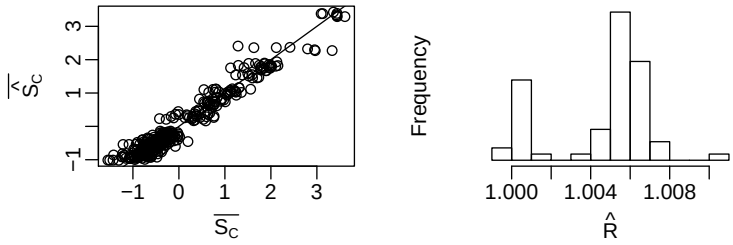

Figure 3.16: Diagnostic plots for Bayesian varying-intercepts beta regression model for forecaster level CRPS scores. The left panel plots the mean empirical $log(S_C)$ and posterior $log(\hat{S}_C)$ scores by condition, method, distribution, and forecasting question. The right panel shows the distribution of the Gelman-Rubin convergence diagnostic $\hat{R}$.

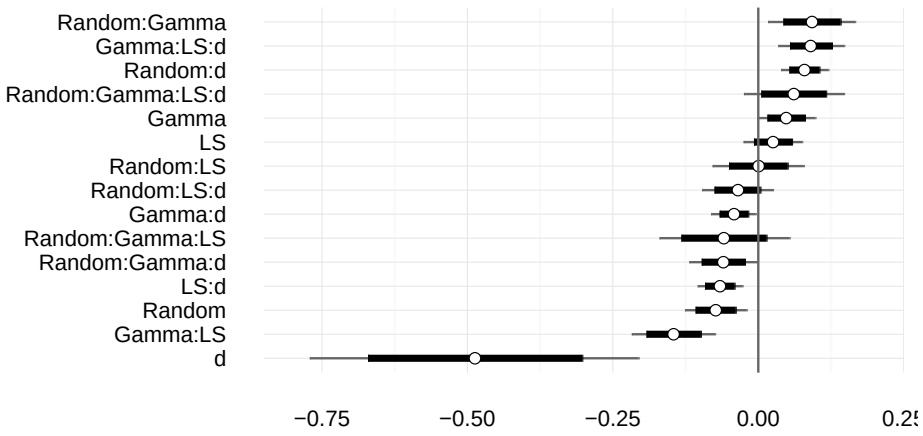

Figure 3.17: HDI and posterior median for each non-varying coefficient for the varying-intercepts beta regression model for forecaster level CRPS scores. The thick black lines represent the posterior 50% HDI. The thin lines represent the 95% HDI.

The coefficient means and HDIs in Figure 3.17 show that nearly every predictor led to credible differences in forecast scores. However, the interactions, log scale of the dependent variable, and influence of the continuous predictor d make these coefficients difficult to interpret. The remainder of analyses and visualizations in this section will focus on the posterior distribution of scores transformed back to the original S_γ values. I transformed scores back to S_γ units via $exp(\hat{y}_{ij} + \frac{\sigma_j^2}{2})$, where $\hat{y}_{ij}$ is score i from MCMC sample j, and σ_j^2 is the regression variance for sample j.

As resolution distance d increased, S_γ decreased (Fig 3.18). While this was the largest "effect" in the model, by itself it is is not particularly informative above demonstrating that the relative magnitude of the S_γ scores to the elicitation interval decreased as the resolution distance increased, i.e. that unscaled scores increased as d increased, but increased at a slower rate than d. However, differences in forecast

accuracy between the conditions interacted with d (Fig 3.19).

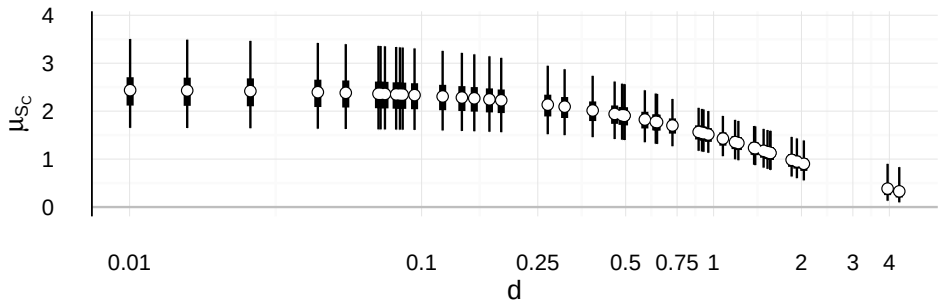

Figure 3.18: HDIs and posterior median for S_γ by resolution distance d for the varying-intercepts beta regression model for forecaster level CRPS scores. The thick black lines represent the 50% HDI. The thin lines represent the 95% HDI. d is log-scaled to make it easier to distinguish values between 0 and 1.

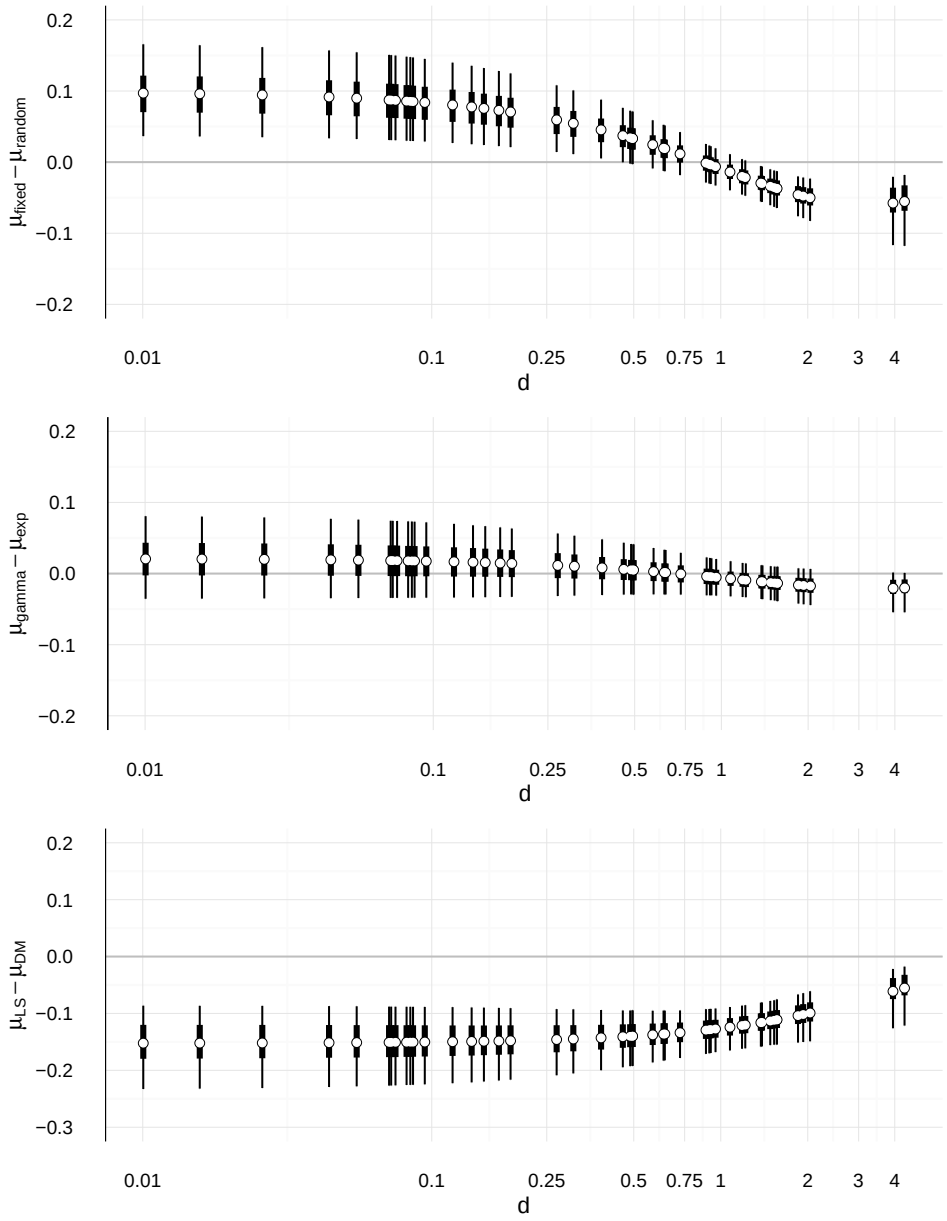

Figure 3.19: HDIs and posterior median for difference in S_γ scores by elicitation format (top panel), distribution (middle panel), and forecaster model (bottom panel). The thick black lines represent the 50% HDI. The thin lines represent the 95% HDI. d is log-scaled to make it easier to distinguish values between 0 and 1.

Fixed bin elicitation tended to score higher (worse) than random bins when

events resolved early within the elicitation interval (for $d < .05$, $\mu_{diff} = 1.21$, 95% HDI [0.09, 2.65]), but scored lower (better) than random elicitation when events resolved well after the elicitation interval (for $d > 1$, $\mu_{diff} = -0.59$, 95% HDI [-1.42, 0.02]). Overall, LS models models obtained lower scores than DM ($\mu_{diff} = -0.13$, 95% HDI [-0.22, -0.05]), though this effect was driven largely by the conditional difference between LS and DM for gamma distributions ($\mu_{diff} = -0.23$, 95% HDI [-0.48, -0.00]) (Fig 3.20).

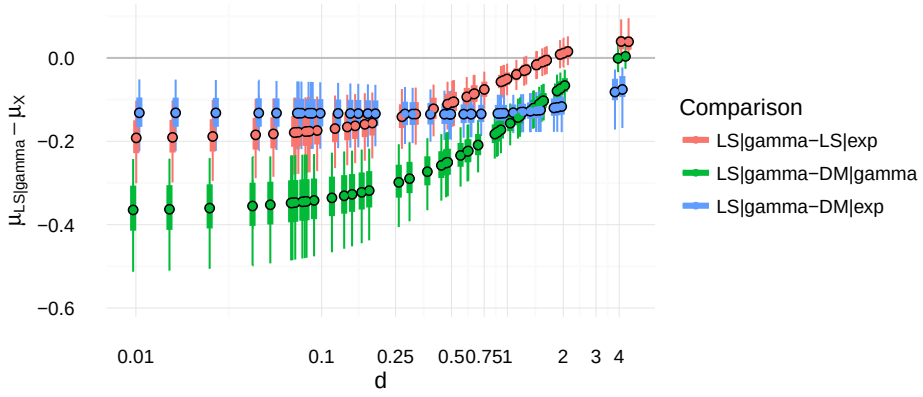

Figure 3.20: HDIs and posterior median for differences in S_γ, by distance d, between the different combinations of forecaster model (LS,DM) and distribution (exp, gamma). The thick lines represent the 50% HDI. The thin lines represent the 95% HDI. d is log-scaled to make it easier to distinguish values between 0 and 1.

Even though there are several credible interactions between the different conditions and resolution distance, there are still clear differences between the modeling methods. The left panel of Figure 3.21 plots the posterior differences in conditional means between each combination of distribution, elicitation format, and forecaster model. The right panel makes the same comparisons, but collapsed across elicitation

format. In this format, it's easier to see that the the LS forecaster model credibly outperformed all other methods.

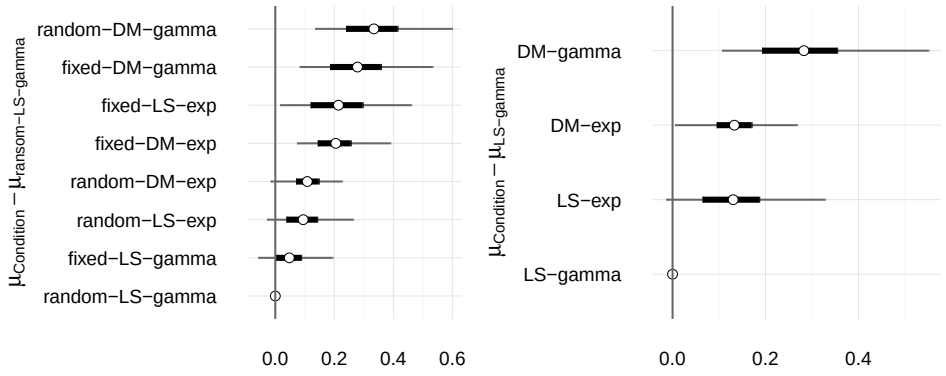

Figure 3.21: Excluding questions with $d > 1.25$, HDIs and posterior median for differences in S_γ between conditional distributions of means for elicitation format, distribution, and forecaster model, and the mean score for gamma and exponential least-squares fits. The thick lines represent the 50% HDI. The thin lines represent the 95% HDI. d is log-scaled to make it easier to distinguish values between 0 and 1.

3.2.3.2 Consensus Level

RPS (S_R)

I fit the consensus level RPS scores with the same Bayesian varying-intercepts beta regression as the forecaster level (Fig 3.11), but included indicator predictors for the different consensus methods. The model converged as indexed by all $\hat{R} < 1.2$, and the correspondence between the predicted and empirical scores indicated that the model adequately represented the data generating process (Fig 3.22).

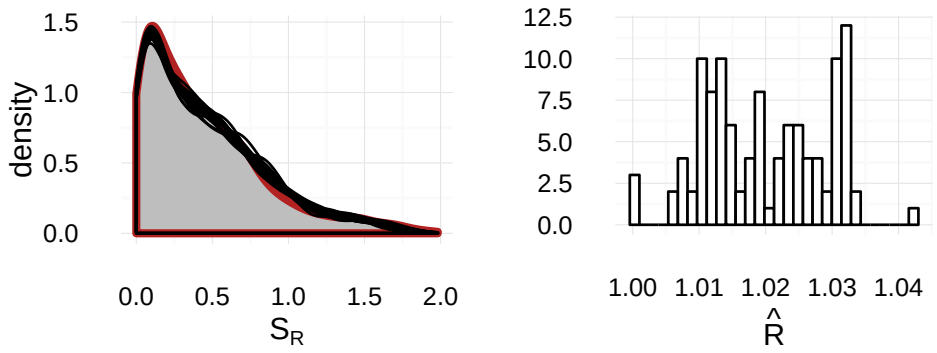

Figure 3.22: Diagnostic plots for varying-intercepts beta regression model for consensus level S_R scores. The left panel plots the empirical density of scores (red) with 20 posterior sample densities overlayed with black lines. The right panel shows the distribution of the Gelman-Rubin convergence diagnostic $\hat{R}$.

Figure 3.23 displays the posterior estimates for the regression coefficients for the effects of condition, distribution, and consensus method. Since the full interaction model includes 71 parameters, excluding the varying intercepts, I only show coefficients where at least the 50% HDI excluded 0. DM-All is the reference level for all consensus method coefficients.

There were very few credible differences between the consensus methods, distributions, and elicitation formats. Exponential methods were credibly more accurate than gamma methods ($\mu_{diff} = -0.005$, HDI[0.000, 0.010]), but the effect was minuscule.

Even though there is no clearly superior method for discrete forecasts, consensus aggregation did improve forecast accuracy. For example, the $M(F_{LS})$ consensus method with exponential fits averaged $M = -0.16$ (IQR[.06,.26]) points lower than the best forecaster model, least-squares exponential fits to fixed bin forecasts.

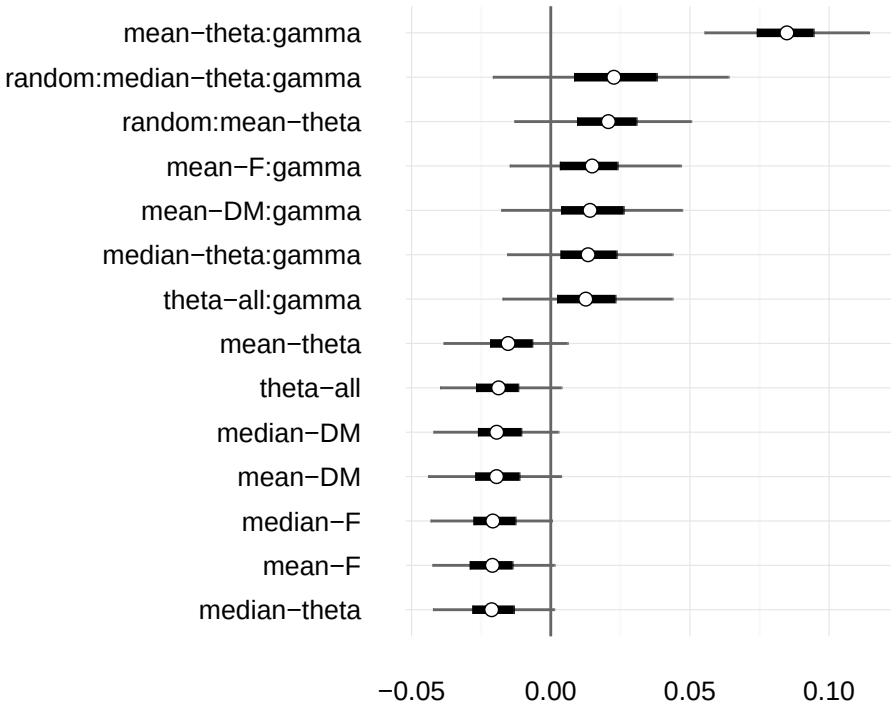

Figure 3.23: HDIs and posterior median for each non-varying coefficient where at least the 50% HDI excluded 0, for varying-intercepts beta regression model for consensus level S_R scores. The thick black lines represent the posterior 50% HDI. The thin lines represent the 95% HDI.

CRPS (S_C)

I fit the consensus level CRPS scores with the same varying-intercepts log-linear model as the forecaster level (Fig 3.15), but included indicator predictors for the different consensus methods. The model converged as indexed by all $\hat{R} < 1.2$, and the correspondence between the predicted and empirical scores indicated that the model adequately represented the data generating process (Fig 3.24).

Figure 3.25 displays the posterior estimates for the regression coefficients for

the effects of condition, distribution, and consensus method. I only display coefficients where at least the 50% HDI excluded 0.

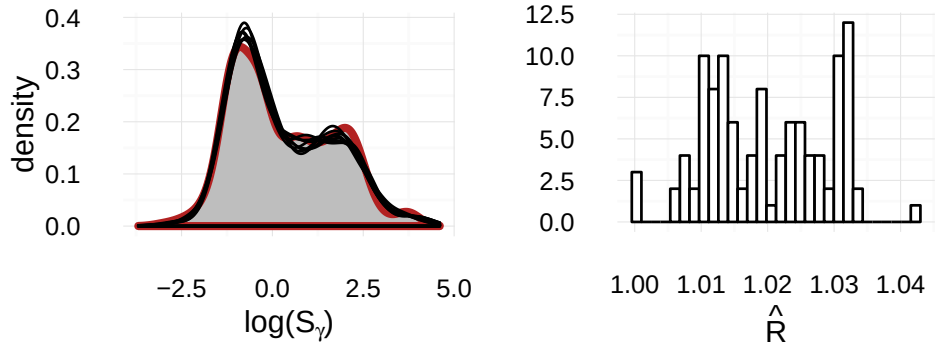

Figure 3.24: Diagnostic plots for varying-intercepts log-linear regression model for consensus level S_C scores. The left panel plots the empirical density of scores (red) with 20 posterior sample densities overlayed with black lines. The right panel shows the distribution of the Gelman-Rubin convergence diagnostic $\hat{R}$.

Similar to the forecaster level scores, as resolution distance d increased, S_γ decreased, and the random elicitation format tended to obtain lower (more accurate) scores when $d < .5$ (Fig 3.26). Unlike the forecaster level, the consensus methods based on the gamma marginally outperformed exponential methods. However, neither of these trends were consistent across all methods, due to interactions with d.

To get a better picture of how the relative accuracy of the different methods changed as a function of of the resolution distance I sampled the predictive posterior distribution of the fitted model. Instead of using the empirical values for d, I created a new model matrix with the same combinations of conditions, but 20 new d values equally spaced between 0.01 and 4.0. Figure 3.2.3.2 compares the median posterior

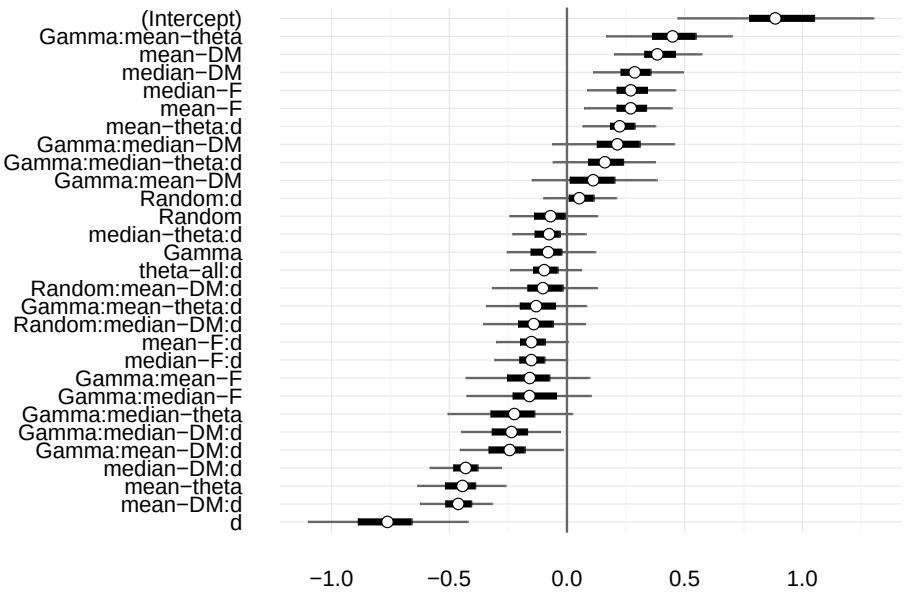

Figure 3.25: Posterior parameter estimates for varying-intercepts log-linear regression model for consensus level S_C scores, for the effects of condition, distribution, and consensus method

predictive S_γ scores for these data. For all conditions, the $M(F_{DM})$ and $\mu(F_{DM})$ methods yield more accurate forecasts for approximately $d > .5$, but as d approaches 0, they become the worst scoring methods.

Resolution distance is a property of the event, and not the forecast or question. While it is unlikely that it can be well controlled for, given that a forecast is a known distance into the elicited range there exists a minimum bound for d. For example, if a forecaster provided judgments that covered a 30 day interval from the day of the forecast, and the forecast was made 15 days ago, the minimum value d could obtain is .5. If some consensus methods perform better than others as a function of d, then this could provide a way to select the most optimal consensus method throughout

the life of a forecasted question.

I sampled the conditional posterior predictive means for all conditions and for three ranges of d ($d < .5$, $d > .5$, $.1 < d < .9$), and calculated the differences from the $M(\theta_L S)$ consensus method fit with the gamma distribution (Table 3.2.3.2). Relative to this method, there were no consistent differences in consensus accuracy given $.1 < d < .9$ or $d < .5$. For $d > .5$, the $\mu(F_{DM})$ and $M(F_{DM})$ methods were more accurate than $M(\theta_{LS})$, particularly for random set bins and gamma distributions. For these conditions $\mu(F_{DM})$ averaged M=-0.25 95% HDI[-0.52,-0.04] $S|\gamma$ points lower than $M(\theta_{LS})$. In the case that an event occurred at the end of the elicited range, the $\mu(F_{DM})$ method would then on average score $.25 \cdot$ range CRPS (S_C) points lower than $M(\theta_{LS})$.

3.3 Comparison to GJP scores

One of the advantages of this dataset is that the forecasting questions we used were concurrently issued on the main Good Judgment Project (GJP) forecasting platform. This was an independent experiment, with different forecasters, and a different web based forecasting application. The question format was slightly different as well. While all of our questions elicited three or more interval judgments, the majority of the GJP questions elicited only one or two bins, and the cut points they used to set those those bins were independent of ours. GJP aggregated discrete forecaster judgments with 38 different methods. Once a forecasting question resolved, then each of these methods would be scored and receive a mean daily Brier/RPS

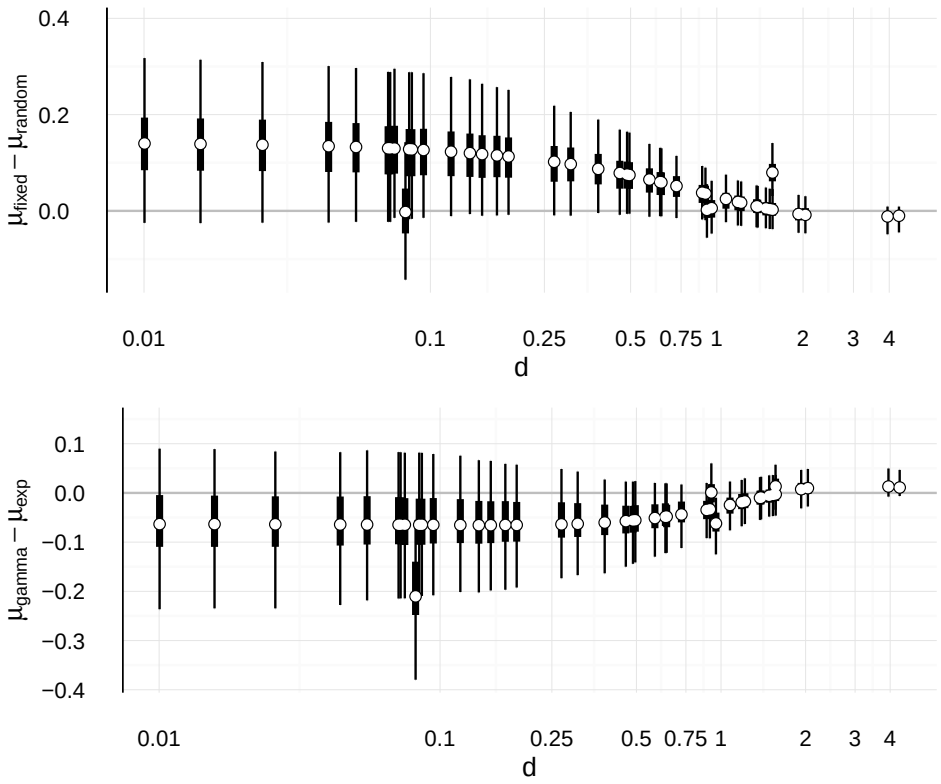

Figure 3.26: HDIs and posterior median for difference in mean daily $S\gamma$ scores by elicitation format (top panel) and distribution (middle panel), and forecaster model (bottom panel). The thick black lines represent the 50% HDI. The thin lines represent the 95% HDI. d is log-scaled to make it easier to distinguish values between 0 and 1.

score for the question.

One of the advantages of continuous forecast models is that you can impute a probability for a value that was never elicited. Using GJP's cut points, I calculated a mean daily RPS score for every question for every consensus method and compared the consensus forecast scores to those of GJP discrete aggregation methods (Fig 3.28). The worst performing consensus method on these scores was $\mu(\theta_{LS})$ with

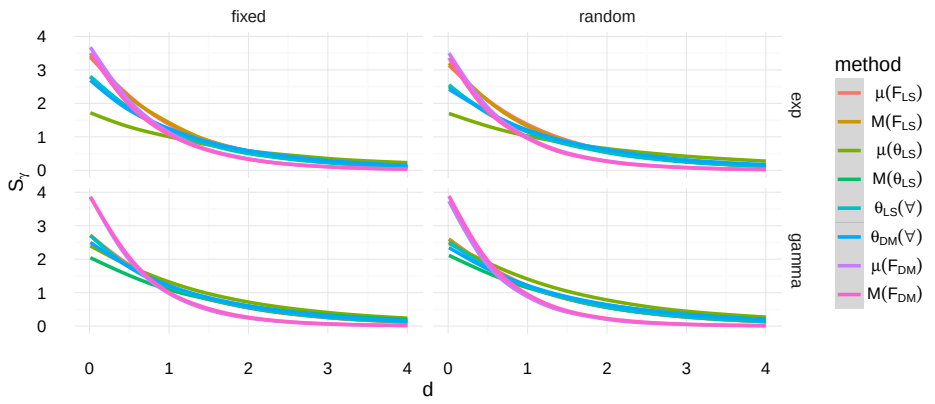

Figure 3.27: Median posterior predictive S_γ scores for $d = .01, .022, .043, \ldots, 4.0$, by elicitation format, distribution, and consensus method.

gamma distributions on random bin forecasts ($M_{S_R} = 0.59$, IQR[0.06,1.46]) scored better than 37% of GJP discrete methods. The best performing consensus method, $M(F_{DM})$ (M_{S_R}=.27, IQR[0.15,0.52]) scored better than 42% of GJP discrete methods.

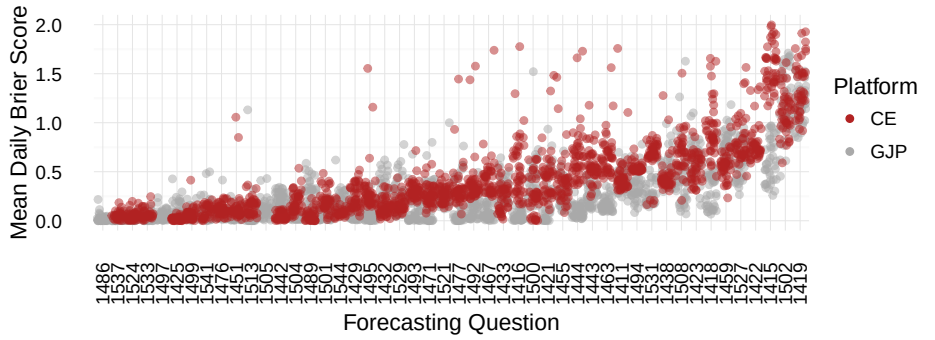

Figure 3.28: Mean Daily Brier scores for consensus methods compared to Good Judgment Project discrete methods.

3.4 Summary

Overall, the consensus methods performed quite well, particularly given the variability in empirical forecaster judgments. Consensus aggregation improved calibration, improved forecast accuracy for discrete judgments, and improved the accuracy of continuous models.

For discrete forecasts, the RPS for the $M(F_{LS})$ consensus method with exponential fits averaged $M = 0.16$ (IQR[.06,.26]) less than than the best forecaster model, least-squares exponential fits to fixed bin forecasts. The $M(F_{LS})$ RPS scores were M=39% (IQR[2%, 61%]) lower than LS fit exponential forecaster models. For continuous scores, $M(F_{LS})$ for gamma fits to random bins averaged S_γ scores M=24% (IQR[-1%, 44%]) lower than the best forecaster continuous method, exponential LS fits to fixed bin forecasts. That the consensus methods performed so well relative to GJP discrete aggregation on their own forecasts and cut points provides compelling evidence that consensus accuracy was not just a function of the particular sample of forecasts used to generate the models.

Across methods, consensus models attenuated overdispersion and improved calibration. Forecasts tended to be overdispersed at both the forecaster and consensus levels, but overdispersion was much more severe for forecaster models. Both consensus methods that directly modeled forecaster judgments instead of aggregating over forecaster models, $\theta_{DM}(\forall)$ and $\theta_{LS}(\forall)$ obtained close to neutrally dispersed PIT variances of .083. On the one hand this is evidence that the consensus methods improved forecasts.

At the forecaster level, there were a few conditions that clearly outperformed the others. For discrete forecasts, the combination of fixed bin elicitation, gamma distributions, and LS forecaster models outperformed all others. However, for continuous forecasts, random elicitation and exponential fits were marginally better. It isn't surprising that different methods would be more or less accurate depending on whether RPS or CRPS scores were evaluated. For the discrete forecasts, the probability assigned to each interval is the only influence on the RPS score. For the continuous forecasts, the CRPS is heavily influenced by the magnitude of the outcome, so questions that resolve later will tend to have higher scores. The CRPS also accounts for how much probability is concentrated around near the outcome. If that outcome occurs far beyond the elicitation interval, then the score is primarily a function of the tails of the model fit to the forecasts.

At the consensus level there were few clear differences in accuracy between the methods. While across all the methods exponential distributions yielded better RPS scores ($\mu_{diff} = -0.005$, HDI[0.000, 0.010]), the effect was small in a practical sense. For example, for a single cut-point question forecasting whether some event will occur before date q, a difference in scores of .005 is equivalent to a forecaster assigning $\mathrm{P}(x < q) = .5025$ versus $\mathrm{P}(x < q) = .5000$, or one quarter of a percentage point of probability.

Continuous consensus forecast performance depended heavily on when the event occurred relative to the range over which judgments were elicited. When events occurred early within that range ($d < .5$) no method yielded consistently more accurate scores than another, and between forecaster scores were highly variable.

When events occurred relatively later ($d > .5$) the F_{DM} consensus methods obtained the most accurate scores. Although method differences were highly conditional, random bin elicitation did consistently provide more accurate forecasts than fixed bins, though this advantage attenuated as d increased, and largely disappeared for $d > 1$.

Method	Distribution	Condition	$d > .5$	$d < .5$	$.1 < d < .9$
$\mu(F_{LS})$	Exp	Fixed	0.04 [-0.2, 0.25]	0.99 [-0.32, 2.72]	0.65 [-0.32, 1.72]
		Random	0.04 [-0.2, 0.27]	0.79 [-0.57, 2.39]	0.54 [-0.34, 1.67]
	Gamma	Fixed	-0.01 [-0.23, 0.21]	0.46 [-0.72, 1.87]	0.30 [-0.54, 1.27]
		Random	-0.01 [-0.23, 0.21]	0.39 [-0.95, 1.63]	0.25 [-0.61, 1.17]
$M(F_{LS})$	Exp	Fixed	0.04 [-0.19, 0.27]	1.00 [-0.34, 2.67]	0.66 [-0.33, 1.74]
		Random	0.04 [-0.19, 0.27]	0.81 [-0.5, 2.44]	0.54 [-0.43, 1.59]
	Gamma	Fixed	-0.02 [-0.24, 0.21]	0.47 [-0.69, 1.92]	0.29 [-0.59, 1.21]
		Random	-0.01 [-0.23, 0.21]	0.39 [-0.73, 1.77]	0.25 [-0.62, 1.15]
$\mu(\theta_{LS})$	Exp	Fixed	-0.02 [-0.25, 0.19]	-0.27 [-1.29, 0.8]	-0.2 [-0.95, 0.56]
		Random	0.03 [-0.19, 0.26]	-0.26 [-1.37, 0.69]	-0.16 [-0.9, 0.6]
	Gamma	Fixed	0.12 [-0.11, 0.39]	0.32 [-0.88, 1.55]	0.27 [-0.57, 1.22]
		Random	0.17 [-0.05, 0.46]	0.42 [-0.81, 1.79]	0.37 [-0.49, 1.37]
$M(\theta_{LS})$	Exp	Fixed	-0.02 [-0.25, 0.19]	0.51 [-0.75, 1.89]	0.31 [-0.53, 1.3]
		Random	-0.02 [-0.25, 0.19]	0.36 [-0.79, 1.65]	0.23 [-0.62, 1.11]
	Gamma	Fixed	-	-	-
		Random	0.03 [-0.19, 0.26]	0.08 [-0.99, 1.26]	0.07 [-0.71, 0.91]
$\theta_{LS}(\forall)$	Exp	Fixed	-0.04 [-0.26, 0.19]	0.49 [-0.75, 1.87]	0.3 [-0.55, 1.25]
		Random	-0.04 [-0.26, 0.17]	0.31 [-0.95, 1.56]	0.18 [-0.65, 1.04]
	Gamma	Fixed	-0.01 [-0.23, 0.2]	0.42 [-0.77, 1.81]	0.27 [-0.6, 1.18]
		Random	-0.01 [-0.22, 0.21]	0.34 [-0.91, 1.63]	0.22 [-0.6, 1.12]
$\theta_{DM}(\forall)$	Exp	Fixed	0.02 [-0.2, 0.24]	0.42 [-0.77, 1.84]	0.28 [-0.55, 1.27]
		Random	0.03 [-0.2, 0.25]	0.28 [-0.84, 1.66]	0.21 [-0.61, 1.12]
	Gamma	Fixed	0.01 [-0.2, 0.23]	0.29 [-0.9, 1.54]	0.2 [-0.63, 1.1]
		Random	0.05 [-0.16, 0.3]	0.27 [-0.95, 1.48]	0.2 [-0.65, 1.09]
$\mu(F_{DM})$	Exp	Fixed	-0.13 [-0.38, 0.08]	1.14 [-0.23, 2.9]	0.56 [-0.31, 1.7]
		Random	-0.19 [-0.46, 0.01]	0.86 [-0.46, 2.53]	0.36 [-0.53, 1.3]
	Gamma	Fixed	-0.22 [-0.47, 0]	1.12 [-0.25, 2.94]	0.46 [-0.44, 1.49]
		Random	-0.25 [-0.52, -0.04]	0.97 [-0.44, 2.65]	0.33 [-0.5, 1.35]
$M(F_{DM})$	Exp	Fixed	-0.15 [-0.4, 0.06]	0.89 [-0.43, 2.54]	0.42 [-0.45, 1.47]
		Random	-0.21 [-0.47, -0.01]	0.77 [-0.6, 2.26]	0.29 [-0.64, 1.19]
	Gamma	Fixed	-0.2 [-0.45, 0.03]	1.17 [-0.27, 3.02]	0.5[-0.41, 1.53]
		Random	-0.23 [-0.49, 0.00]	1.11 [-0.28, 2.90]	0.44 [-0.46, 1.50]

Table 3.8: Median and 95% HDI for conditional posterior predictive means.

Chapter 4: General Discussion

An accurate probability distribution over all possible values of the forecasted event is much more useful to a decision maker than a discrete forecast because it can provide a judgment for any partition of the event, independent of what values were elicited from forecasters, and can be easily integrated into a broad range of decision analyses. However, eliciting complete distributions from an individual judge is impractical in many contexts. Previous research from the Good Judgment Project (GJP) has shown that fitting gamma distributions to relatively few interval probability judgments, and then aggregating those distributions by taking the median of the distribution parameters across all forecasters, could yield accurate consensus models (Tidwell et al., 2015).

This dissertation evaluated the performance of potential methods to extend and improve the GJP research for forecasts of unique events that occur only once. Continuous models of forecaster judgments were fit with exponential, gamma, and generalized gamma distributions to determine whether more or less flexible functions would improve forecast accuracy. Aggregate consensus forecasts were obtained from the linear combination of forecaster level probabilities or distribution parameters, and by fitting distributions to sets of forecaster judgments instead of at the individ-

ual level.

The previous chapter detailed analytic results specific to various experimental and computational manipulations; however, these results can only inform what constitutes an optimal forecasting system subject to the goals of the forecasting system. For example, if the goal of a forecasting system is to achieve the most well-calibrated forecaster level judgments, then the combination of methods that achieve this would be the most appropriate; however, these methods may not necessarily yield the most accurate continuous forecasts at either the forecaster or consensus levels. Any continuous forecasting system will likely have a many goals, including evaluating individual forecaster performance and achieving accurate consensus models.

I divide the methods I evaluated in this dissertation broadly into elicitation and computation. Elicitation methods include anything used to elicit empirical judgments from forecasters. Computational methods include all of the analytic steps used to manipulate the empirical judgments in some way. Since the empirical judgments are fixed once elicited, the optimal elicitation methods are those that provide the most accurate forecasts across a range of forecasting goals. Computational methods are not similarly constrained. Given a set of empirical judgments, one can apply whatever computational method is best suited for a particular forecasting goal. For example, if the goals of a forecasting system were to both compare individual forecaster accuracy and obtain the most accurate consensus forecasts, then one could use different computational methods to achieve each goal, but clearly only one elicitation method.

4.0.1 What is the best elicitation method?

With the exception of forecast level discrete scores, random-set bin elicitation yielded the most accurate forecasts (i.e. lowest RPS and CRPS scores). This implies that a forecasting system should use random bins to maximize forecast accuracy across the widest number of possible forecasting goals and computational methods. There is the trade-off of potentially less accurate discrete forecaster level forecasts, but this seems like a reasonable cost given the benefits of continuous forecasts at both the forecaster and consensus levels.

A likely explanation for the superior performance of random bins are the same reasons that motivated including it as an experimental condition: to attenuate the effects of partition dependence and to elicit judgments across a greater range of values of the forecasted variable. Partition dependence is the tendency for forecasters to anchor and adjust from $\frac{1}{n}$ probability judgments for a variable partitioned into mutually exclusive and exhaustive intervals (Fox and Clemen, 2005). Fixed bins concentrate this bias at the same values of the elicited variable across forecasters, and therefore likely lead to similarly biased forecaster level models, and consequently biased consensus models. By eliciting random bins, the bias introduced by partition dependence at the forecaster level can be attenuated in aggregation at the consensus level.

One should also expect random bins to increase forecast accuracy at the consensus level because the consensus methods will aggregate over more distinct information than with fixed bins. Both fixed and random elicitation are, by definition,

interval censored at the forecaster level. However, aggregating over fixed bins carries this censoring through to the aggregate model, while random bins approximate a random sample of judgments across the elicitation range.

4.0.2 What are the best computational methods?

4.0.2.1 Forecaster Accuracy

Fitting exponential distributions via least-squares yielded the best continuous forecast accuracy. Though exponential distributions yielded poorer fits to empirical judgments, they were better calibrated and scored nearly as well as gamma fits for discrete forecasts and slightly better for continuous forecasts. This suggests that there is little benefit, and possibly a cost, to modeling small sets of judgments with flexible multi-parameter distributions. When a forecaster provides only three or four estimates, a higher parameter distribution will necessarily obtain a better fit to those judgments, but the fit between the model and the probabilities doesn't necessarily correspond to how well the distribution represents true subjective belief. Even with a large number of judgments, more flexible distributions may not necessarily do a better job of modeling true subjective belief. Assuming forecasters produce overt judgments with some error component, then the closer a model fits the observed judgments the better the model represents the true belief plus error distribution, not true belief.

Least-squares models were considerably more accurate than the Dirichlet-multinomial for forecaster judgments across elicitation format and distribution, and

should be the modeling choice if the goal is to maximize the accuracy of forecaster level continuous models. However, this consistent difference is somewhat surprising given that the LS and DM models are effectively conducting the same task: finding distribution parameters that fit the observed judgments as closely as possible. The primary difference in these approaches, aside from the token/probability distinction, is that the DM model incorporates the relatively strong weight $w = 100$ on the prior distribution for the forecasters' continuous subjective distribution. Since I needed to fit over 100,000 DM models, it was impractical to monitor and adapt the models for individual forecasters, and this weight helped ensure that the models converged and produced consistent results. Clearly the benefits of obtaining quick and consistent convergence came at the cost of model accuracy. Future work should be able to refine the DM model and its constituent priors to yield consistently tractable, as well as accurate, forecaster level models.

4.0.2.2 Consensus Accuracy

Overall, the M(F_{DM}) consensus method using gamma distributions yielded the most accurate consensus forecasts, and obtained the most accurate imputed RPS scores for the discrete GJP cut points. This combination of results suggests that this method is, in general, a good choice when the forecasting goal is to obtain the most accurate continuous consensus forecasts, and an improvement over M(θ_{LS}) However, the accuracy of this method varied greatly with the relative distance of the question resolution from the range over which forecaster judgments were elicited, d.

For $d < .5$, that is when questions resolved in the first half of the elicitation range, $M(F_{DM})$ performed worse than other methods, particularly M(θ_{LS}).

The limited success of the F_{DM} methods is promising. All of the probability averaging methods are simply mixtures of forecaster distributions. Unlike with the $M/\mu(\theta)$ or $\theta(\forall)$ methods, once a forecaster model is fit to the original judgments, it never needs to recomputed for daily aggregation. For any forecast day, the consensus is the mixture of the conditional forecaster distributions, given the current date. This is a much simpler, and less computationally intense, daily forecast process than the other methods, and should be easier to implement. Another promising aspect of the F_{DM} methods is that the Dirichlet-multinomial models provide an intuitive and relatively simple way to include external information into the model, to calibrate forecasters, and to weight forecasters in the mixture.

The result that relative model accuracy varies as a function of d is a novel finding and has implications for anyone developing a continuous forecasting system. First, where possible the range over which judgments are elicited should include the question resolution value. From a modeling perspective this is intuitive, i.e. only predict where the model is fit to empirical data. Of course, *a priori* there is no way to know when the event will happen. One potential solution is to issue a new forecasting question with an updated range as the current date approaches the end of the elicitation range.

Second, it may be useful to combine consensus methods based on the elicitation range. For example, though the resolution can not be known *a priori*, the value of d given any resolution value is known. It may be useful to model forecasts for $d < .5$

with a method that performs well in that range, e.g. $\mathrm{M}(\theta_{LS})$, and forecasts for $d > .5$ with $\mathrm{M}(F_{DM})$. If one uses the CDF as the prediction function, then different consensus methods can be combined with little additional computational effort.

4.0.3 Calibration

The discussion above intentionally omits calibration as a forecasting goal, though there are large differences in calibration between some elicitation and computational methods. For example, nearly every consensus method dramatically reduced the underdispersion of the forecaster models, with exponential consensus models closest to neutral dispersion for most methods, and even overdispersed for the mean probability averaging methods $\mu(F_{LS})$ and $\mu(F_{DM})$. This is a well known effect of linear combinations of forecast distributions (Ranjan and Gneiting, 2010; Hora, 2004; Gneiting et al., 2007). If the forecasters in the experiment had been better calibrated, it's possible that some of the aggregation methods would have introduced severe miscalibration instead of attenuating it.

Despite these results, it's unclear how appropriate the concept of calibration is for these kinds of forecasting questions, and whether or not it is an appropriate forecasting goal. Previous researchers have argued whether calibration, and overconfidence in particular, can be explained all or in part by: biased item selection (Gigerenzer et al., 1991), individuals conflating sample with population characteristics (Juslin et al., 2007), elicitation format (Winman et al., 2004), or a statistical artifact of unbiased error prone judgments (Erev et al., 1994; Dougherty, 2001),

among many other reasons. This is an even more complicated issue in this type of forecasting context, where forecasting questions are often difficult by design and where it is extremely difficult to define what constitutes a class of events.

Since calibration is the correspondence between the frequency of events and the probability assigned to those events, it can only be evaluated given a set of random outcomes for an event or class of events, but what class of events do questions like "What is the probability Assad will leave office before 1 January 2017" belong to? And even given that there was some definable class of events, it is unlikely that questions would be sampled from this class in an unbiased way. Most likely, questions will be selected based on what interests the decision makers that the forecasting system informs.

If forecasting questions are selected in a way that biases the distribution of outcomes, then even a perfectly calibrated forecaster will appear miscalibrated. Assume some class of events is distributed $E \sim N(\mu, 1)$, where $\mu \sim N(0, 1)$. Consider μ a distribution of potential forecasting questions and E the probability distribution over the outcome for any particular event. A forecaster who always conditioned on μ and forecasted the probability of an event with the distribution function $F \sim N(\mu, \sqrt{2})$ would be perfectly calibrated, but *appear* miscalibrated. For example, if this calibrated forecaster only forecasted questions where $(\mu < -1)$, then her PIT histogram would look biased like the middle panel of Figure 4.1. If she only forecasted events where $(-1 < \mu < 1)$, then she would appear underconfident like in the right panel of Figure 4.1.

This potential problem is particularly relevant for forecasts of socio-political

events like the ones in this research. A group of experts chose the questions based on many criteria, but almost certainly none of these criteria was to try and get an unbiased sample of a random process. Similarly, in an applied forecasting environment the questions will most likely be selected to meet information requirements of a decision maker, rather than to obtain a representative sample. If the forecasting questions do not reflect a random process, or do not belong to a well defined class of events, then you can not dissociate forecaster calibration from question selection.

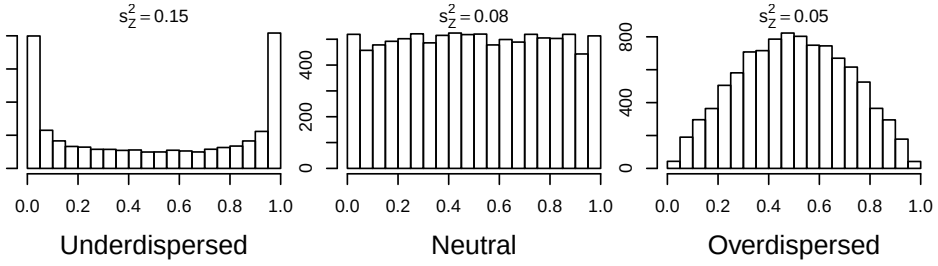

Figure 4.1: PIT histograms given non-random samples of events. Even though the underlying distribution of events is identical across the three scenarios, the observed calibration for each panel varies greatly depending on how particular events are selected.

4.1 Future Research

The results of this dissertation suggest three main areas for future research that could improve consensus forecast methods: developing more psychologically plausible models, improving the efficiency and dependability of computational methods, and understanding forecasters' subjective interpretation of time-based questions. All of the consensus methods in some way assumed that forecaster belief can be

described by common probability distributions. While it seems possible that there are cases where this could be true, it may be unreasonable to expect that something as complex as continuous subjective belief will generally follow common parametric forms. Restricting models to distributions like the exponential and gamma therefore limit how well forecaster judgment can be modeled, and consequently how well these models truly reflect both forecaster and aggregate crowd belief. The mixture models introduced in this dissertation were one step towards reducing the dependence on parametric forms. Though forecasters were still fit with exponential and gamma distributions, the mixture of their forecaster models can represent much more complex models of group belief. A logical next step is to further remove the dependence on parametric forms and model judgments with Bayesian nonparametric models like Gaussian or Dirichlet processes.

Computational efficiency and tractability are closely related to the psychological plausibility of forecast models. Compared to the Bayesian methods in this dissertation, fitting simple distributions via least-squares is a simpler, more dependable, and generally more resource efficient method to generate forecast models. For example, the Bayesian models in this dissertation included relatively strong priors to help ensure consistent and timely convergence. However, these priors also likely biased the posterior distributions, particularly for the forecaster level models where there was very little observed information to integrate into the models. Future work should focus on developing models that are not only theoretically attractive, but that also can be dependably implemented in a real-world forecasting environment.

Another potential line of research is to establish how forecasters interpret

interval judgments for time based questions for events that can only happen once. Consider this question set: "What is the probability that Assad will leave office between today and January 1st?", and "What is the probability that Assad will leave office between January 2nd and February 1st?" The second question is only interpretable given that you assume Assad is still in power on January 1st, i.e. that it is a conditional probability, yet we typically model judgments from questions like these as intervals from an unconditional distribution. If we understand how people interpret these questions and produce their judgments, we could improve the models we make from those judgments.

4.2 Summary

This dissertation showed that accurate consensus forecast distributions can be modeled from relatively few judgments from individual forecasters. With respect to elicitation format, random bins yielded more accurate forecasts than fixed bin elicitation. Random bins provide more information across the range of the forecasted variable, and in aggregation are more likely to mitigate the potential effects of partition dependence and random error in individual forecaster judgments.

At the consensus level, a mixture of forecaster distributions fit with a Bayesian Dirichlet-multinomial model and gamma distributions outperformed median parameter aggregation and obtained forecast accuracies on par with advanced discrete aggregation techniques. This model provides an intuitive way to weight and calibrate forecasters, and to include external information, that would be much more

difficult with least-squares approaches. Consequently, this model not only offers the best current option for consensus forecasts, but also holds much potential for future development and to integrate proven discrete aggregation techniques.

Unlike the consensus level, fitting with least-squares was the most accurate modeling method for individual forecasters. This is likely due to the combination of strong priors in the Bayesian models and few judgments from each forecaster for any given question. Compared to gamma distributions, exponential distributions yielded poorer fits to the empirical judgments, yet better forecast accuracy as measured by proper scoring rules. This suggests that at the forecaster level more flexible distributions may overfit error in the observed judgments. If the forecasting goal is to obtain the most accurate forecaster level models, then then the best option is to fit exponential distributions via least squares.

One unexpected result was the effect of resolution distance on forecast accuracy. Differences between modeling methods varied greatly as a function of when an event occurred relative to the range over which forecaster judgments were elicited, particularly when events occurred long after the last date for which forecasters provided judgments. With the benefit of hindsight, this is reasonable. When an event occurs long after any forecaster judgments have been collected, then a model's predictions are almost entirely a function of the assumptions of the model rather than the judgments of the forecaster.

The success of the consensus methods explored in this dissertation implies that with little additional data or forecaster effort, it is possible to obtain continuous aggregate models of forecaster belief that are as accurate as discrete forecast

aggregation methods, but can also provide decision makers with forecasts for any partition of the event over which judgments were elicited and can be easily integrated into a broad range of decision analyses.

Bibliography

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Brooks, S. and Gelman, A. (1998). Some issues for monitoring convergence of iterative simulations. *Computing Science and Statistics*, pages 30–36.

Clemen, R. and Winkler, R. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187–203.

Croushore, D. (1993). Introducing: The survey of professional forecasters. *Federal Reserve Bank of Philadelphia Business Review*, 3(13).

Dougherty, M. R. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General*, 130(4):579.

Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological review*, 101(3):519.

Fox, C. R. and Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, 51(9):1417–1432.

Galton, F. (1907a). Letter to the editor. *Nature*, 75(1952).

Galton, F. (1907b). Vox populi (the wisdom of crowds). *Nature*, 75(1949):450–451.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: A brunswikian theory of confidence. *Psychological review*, 98(4):506.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological bulletin*, 73(6):422.

Hetes, B. et al. (2011). Epa expert elicitation task force white paper. *Environmental Protection Agency, Washington DC*, 7:2013.

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

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

IARPA (2010). Aggregative contingent estimation (ace) program. http://www.iarpa.gov/index.php/research-programs/ace.

Jowett, B. and Davis, H. (1908). *Aristotle's Politics*. Oxford translation series. Clarendon Press.

Juslin, P., Winman, A., and Hansson, P. (2007). The naïve intuitive statistician: a naïve sampling model of intuitive confidence intervals. *Psychological review*, 114(3):678–703.

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

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

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472.

Stan Development Team (2015). Stan: A c++ library for probability and sampling, version 2.10.0.

Tidwell, J. W., Wallsten, T. S., and Moore, D. A. (2015). Eliciting and modeling probability forecasts of continuous quantities. (Unpublished).

Usher, W. and Strachan, N. (2013). An expert elicitation of climate, energy and economic uncertainties. *Energy Policy*, 61:811–821.

Wallis, K. F. (2011). Combining forecasts–forty years later. *Applied Financial Economics*, 21(1-2):33–41.

Wallsten, T. S., Budescu, D. V., Erev, I., and Adele, D. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10(3):243–268.

Wallsten, T. S. and Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, 41(1):1–18.

Wallsten, T. S., Shlomi, Y., Nataf, C., and Tomlinson, T. (2015). Efficiently encoding and modeling subjective probability distributions for quantitative variables.

Winman, A., Hansson, P., and Juslin, P. (2004). Subjective probability intervals: how to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1167.