HHS Public Access

Author manuscript

Wiley Interdiscip Rev Comput Stat. Author manuscript; available in PMC 2022 March 01.

Published in final edited form as:

Wiley Interdiscip Rev Comput Stat. 2021 ; 13(2): . doi:10.1002/wics.1514.

Aggregating predictions from experts: a review of statistical methods, experiments, and applications

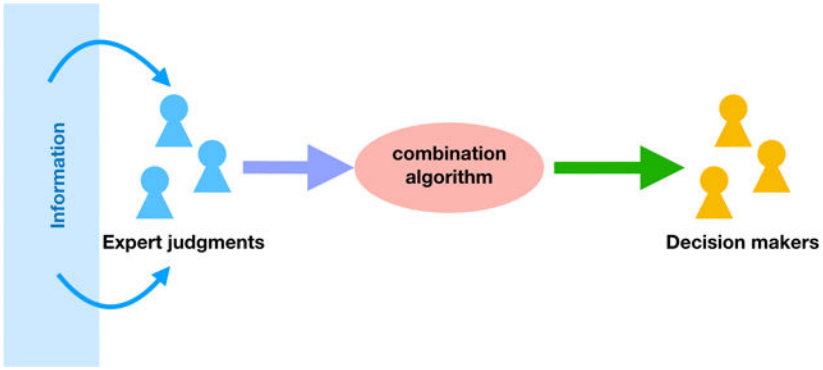
Thomas McAndrew[a,*], **Nutcha Wattanachit**[a], **Graham C. Gibson**[a], **Nicholas G. Reich**[a]

[a]Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts at Amherst, Amherst, Massachusetts, USA

Abstract

Forecasts support decision making in a variety of applications. Statistical models can produce accurate forecasts given abundant training data, but when data is sparse or rapidly changing, statistical models may not be able to make accurate predictions. Expert judgmental forecasts—models that combine expert-generated predictions into a single forecast—can make predictions when training data is limited by relying on human intuition. Researchers have proposed a wide array of algorithms to combine expert predictions into a single forecast, but there is no consensus on an optimal aggregation model. This review surveyed recent literature on aggregating expert-elicited predictions. We gathered common terminology, aggregation methods, and forecasting performance metrics, and offer guidance to strengthen future work that is growing at an accelerated pace.

Graphical Abstract



Keywords

Forecast combination; Forecast aggregation; Judgmental forecasting; Expert judgment; Consensus

*Corresponding author mcandrew@umass.edu (Thomas McAndrew).

1. Introduction

Forecasting presents decision makers with actionable information that they can use to prevent (or prepare for) economic (Mak et al. [Mak], 1996; Shin et al. [Shi], 2013; Huang et al. [Hua], 2016), engineering (Guangliang [Gua], 1996; Zio [Zio], 1996; Neves & Frangopol [Nev], 2008), ecological (Borsuk [Bor], 2004; Failing et al. [Fai], 2004; Morales-Nápoles et al. [Mor], 2017; Johnson et al. [Joh2018], 2018), social (Craig et al. [Cra], 2001; Kläs et al. [Kla], 2010; Cabello et al. [Cab], 2012), and public health burdens (Alho [Alh], 1992; Evans et al. [Eva], 1994).

Advances in computing power made statistical forecasts, models that take as input a structured data set and output a point estimate or probability distribution, a powerful tool (Al-Jarrah et al. [Jar], 2015; Kune et al. [Kun], 2016; Wang et al. [Wan], 2016). Statistical models exploit correlations between data to find patterns, but when data is rapidly changing, sparse, or missing completely, the accuracy of these models can suffer. Judgmental forecasts attempt to overcome data limitations present in statistical models by eliciting predictions from experts (Clemen & Winkler [Cle], 1986; Genest, Zidek, et al. [Gen], 1986; Clemen [Cle], 1989). Experts are able to make predictions without structured data, and instead, rely on their experience and contextual knowledge of the prediction task. Expert forecasts are most readily found in finance, business, and marketing (Petrovic et al. [Pet], 2006; Kabak & Ülengin [Kab], 2008; Franses [Fra], 2011; Seifert & Hadida [Sei], 2013; Shi, 2013; Song et al. [Son], 2013; Alvarado-Valencia et al. [Alv], 2017; Baecke et al. [Bae], 2017). These fields focus on decision makers and their ability to make predictions from data that cannot easily be collected and fed to a statistical model. Other areas of active research in expert opinion are quality assurance (Kla, 2010), politics (Graefe et al. [Gra], 2014a; Graefe et al. [Gra], 2014b; Satopää et al. [Sat], 2014; Graefe [Gra], 2015; Cai et al. [Cai], 2016; Gra, 2018; Hanea et al. [Han], 2018; Wang & Zhang [Wan], 2018), economics (Mak, 1996; Shi, 2013; Hua, 2016), engineering (**ZIO1996127**; Cra, 2001; Ren-jun & Xian-zhong [Ren], 2002; Jin et al., 2007; Tartakovsky [Tar], 2007; Nev, 2008; Wang et al. [Wan], 2008; Brito et al. [Bri], 2012; Brito & Griffiths [Bri], 2016; Hathout et al. [Hat], 2016), sports (Gu et al. [Gu], 2016), sociology (Adams et al. [Ada], 2009; Cab, 2012), meteorological (Abramson et al. [Abr], 1996), ecological (Bor, 2004; Fai, 2004; Cooke et al. [Coo], 2014a; Joh2018, 2018), environmental science (Zio & Apostolakis [Zio], 1997; Li et al. [Li], 2012; Mantyka-Pringle et al. [Man], 2014; Mor, 2017), and public health (Alh, 1992; Evans et al., 1994; Kurowicka et al. [Kur], 2010; Jana et al. [Jan], 2019). The diversity and breadth of applications underscore the importance of expert opinion in a wide variety of disciplines.

Combining expert judgment can be divided into two goals: (i) to increase forecast accuracy (Bates & Granger, 1969; Granger & Ramanathan, 1984) and/or (ii) create a predictive ‘consensus’ distribution that individual experts agree represents their collective thoughts (Stone, 1961; Gen, 1986; Cooke et al. [Coo], 1991). Models focused on combining experts to boost accuracy view experts as additional sources of information or predictive models that can be combined with other expert and statistical predictions. To evaluate forecasting accuracy, ground truth data is collected and compared against model predictions by using quantitative metrics and experimental evidence. Building a consensus distribution has a different goal—building a predictive distribution that optimally represents experts’ collective

uncertainty. Success is typically measured on a qualitative scale. For example, success may be measured by the type of data needed from experts to build a distribution and whether the algorithm used to combine expert judgment accurately represents their collective opinions. Whether to combine expert judgement for improved accuracy or consensus depends on the application. Improving accuracy will usually require ground truth data and expert-elicited predictions. In a data-sparse scenario, consensus requires the ability to solicit expert judgements and combine them in a way representative of the group without necessarily appealing to quantitative metrics.

Research combining expert opinion to produce an aggregate forecast has grown rapidly, and a diverse group of disciplines apply combination forecasting techniques. Cross-communication between different applied areas of combination forecasting is minimal, and as a result, different scientific fields are working in parallel rather than together. The same mathematical ideas in combination forecasting are given different labels depending on application. For example, the literature refers to taking an equally-weighted average of expert forecasts as: equal-weighting (Sarin [Sar], 2013; Coo, 2014a; Han, 2018), unweighted (Gra, 2015), and 50–50 weighting (even when more than two forecasts are averaged) (Alv, 2017).

This review focuses on methods for aggregating expert judgments. The aim is to survey the current state of expert combination forecasting literature, propose a single set of labels to frequently used mathematical details, look critically at how to improve expert combination forecasting research, and suggest future directions for the field.

We map key terminology used in combining expert judgemental forecasts and consolidate related definitions. A textual analysis of articles highlights how combination forecasting techniques have evolved. A prespecified list of questions was asked of every manuscript: whether point predictions or predictive densities were elicited from experts, methods of aggregating expert predictions, experimental design for evaluating combination forecasts and how forecasts were scored (evaluated). We tabulated techniques for evaluating forecasts and condensed terms referring to the same evaluative metric.

Section 2 gives a brief historical background of combination forecasting and current challenges. Section 3 describes our literature search, how articles were included in our analysis set, and our analysis. Section 4 reports results and section 5 discusses common themes, terminology, advocates for key areas that need improvement, and recommends future directions for aggregating expert predictions.

2. Background

2.1. Human judgmental forecasting

Judgmental forecasting models—predictions elicited from experts or non-expert crowds and combined into a single aggregate forecast—have a long history of making well calibrated and accurate predictions (Edmundson, 1990; Bunn & Wright, 1991; Lawrence & O’Connor, 1992; O’Connor et al., 1993). Advances in judgmental forecasting take two paths: building

sophisticated schemes for combining predictions (Cle, 1989; Clemen & Winkler, 1999a; Cle, 2008) and eliciting better quality predictions (Helmer, 1967; Ayyub [Ayy], 2001).

Initial combination schemes showed an equally-weighted average of human-generated point predictions can accurately forecast events of interest (Galton, 1907). More advanced methods take into account covariate information about the forecasting problem and about the forecasters themselves (for example weighting experts on their past performance). Compared to an equally-weighted model, advanced methods show marginal improvements in forecasting performance (Winkler, 1971; McLaughlin, 1973; Armstrong [Arm], 1985; Cle, 1989; Fischer & Harvey, 1999).

In this work we will study combinations of expert predictions. Combining non-expert predictions often falls into the domain of crowdsourcing, and crowdsourcing methods tend to focus on building a system for collecting human-generated input rather than on the aggregation method (section 2.5 discusses open challenges to a crowdsourcing approach).

Past literature suggests experts make more accurate forecasts than novices (Armstrong, 1983; Alexander Jr, 1995; Spence & Brucks, 1997; Clemen & Winkler, 1999a; Armstrong, 2001a; Lawrence et al., 2006; French, 2011), and a typical definition of an expert is someone whose forecast you would adopt because you consider your knowledge about a forecasting target a subset of the expert's knowledge (DeGroot, 1988). Several reasons could contribute to an expert's increased accuracy: domain knowledge, the ability to react to and adjust for changes in data, and the potential to make context-specific predictions in the absence of data (Armstrong, 1983; Alexander Jr, 1995; Spence & Brucks, 1997; Lawrence et al., 2006). The increased accuracy of expert opinion led some researchers to exclusively study expert forecasts (Armstrong, 2001a; French, 2011; Genre et al., 2013), however crowdsourcing—asking large volumes of novices to make predictions and using a simple aggregation scheme—rivals expert-generated combination forecasts (Howe [How], 2006; Lintott et al., 2008; Prill et al., 2011). Whether or not expert or non-expert predictions are solicited, judgmental forecasting agrees that human judgment can play an important role in forecasting.

Judgmental forecasts can have advantages over statistical forecasting models. Human intuition can overcome sparse or incomplete data issues. Given a forecasting task with little available data, people can draw on similar experiences and unstructured data to make predictions, whereas statistical models need direct examples and structured data to make predictions. When data is plentiful and structured, statistical models typically outperform human intuition (Meehl, 1954; Kleinmuntz, 1990; Yaniv & Hogarth, 1993). But whether a statistical or judgemental forecast is best depends on the circumstances.

An understanding of the type of forecasts that models can produce and a mathematical description of a combination forecast can clarify how judgmental data, number of forecasters, and the combination scheme interact.

2.2. A framework for combination forecasting

Forecasting models can be statistical, mechanistic, or judgmental. We define a forecasting model $\mathcal{M}$ as a set of probability distributions over all possible events. Each probability

distribution is typically assigned a vector (θ), called the model's parameters, that is used to differentiate one probability distribution from another $\mathcal{M} = \{P_\theta | \theta \in \Theta\}$, where P_θ is probability distribution for a specific choice of θ, and Θ are all possible choices of model parameters.

Models can produce two types of forecasts: point predictions or predictive densities. Point forecasts produce a single estimate of a future value (Bates & Granger, 1969; Granger & Ramanathan, 1984) and are frequently used because they are easier to elicit from experts and early work was dedicated to combining specifically point forecasts (Galton, 1907; Bates & Granger, 1969; Granger & Ramanathan, 1984). Probabilistic forecasts are more detailed. They provide the decision maker an estimate of uncertainty (probability distribution) over all possible future scenarios (Stone, 1961; Winkler, 1968, 1981; Gen, 1986; Dawid et al., 1995; Clemen & Winkler, 1999a; Ranjan & Gneiting, 2010; Gneiting, Ranjan, et al., 2013; Hora & Kardeş [Hor], 2015). Probabilistic densities can be thought of as more general than point forecasts. A point forecast can be derived from probabilistic forecast by taking, for example, the mean, median, or maximum a posteriori value. A probabilistic density assigning all probability mass to a single value can be considered a point forecast.

A combination forecast aggregates predictions, either point or probabilistic, from a set of models and produces a single aggregate forecast (Winkler, 1981; Gen, 1986; Clemen & Winkler, 1999a; Winkler et al., 2019). Given a set of models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_N$, a combination model $\mathcal{G} : \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_N \rightarrow \mathcal{F}$ maps the Cartesian product of all models onto a single class of suitable probability distributions (Gneiting, Ranjan, et al., 2013). The goal of combination forecasting is to find an optimal aggregation function $\mathcal{G} \in \mathcal{G}$. Typically the model $\mathcal{G}$ is parameterized $\mathcal{G} = \{G_\upsilon | \upsilon \in \Upsilon\}$ such that finding an optimal G amounts to finding the parameter vector υ that produces an optimal forecast.

There are several ways to improve a combination model's forecasting ability. Combination models can improve forecast accuracy by considering a more flexible class of aggregation functions ($\mathcal{G}$). Soliciting expert opinion (versus novices) can be thought of as improving individual forecasts $\mathcal{M}$ used as input into the combination model. Crowdsourcing takes a different approach to improve forecast accuracy (How, 2006; Abernethy & Frongillo, 2011; Brabham, 2013; "Crowdsourcing the Future: Predictions Made with a Social Network," 2014; Moran et al., 2016). These methods consider a simple class of aggregation functions $\mathcal{G}$ and collect a large number of human-generated forecasts $\mathcal{M}$. By accumulating a large set of human-generated predictions, a crowdsourcing approach can create flexible models with a simple aggregation function.

This framework makes clear the goals of any combination forecasting model. Some focus on improving individual models $\mathcal{M}$, others focus on more flexible aggregation functions ($\mathcal{G}$). In this work we will consider combination forecasting models that include expert-elicited forecasts as their raw material and pursued building more flexible aggregations models.

2.3. A brief timeline of existing work

Francis Galton was one of the first to formally introduce the idea of combination forecasting. In the early 20th century, he showed aggregating point estimates from a crowd via an unweighted average was more accurate compared to individual crowd estimates (Galton, 1907). Galton's work was empirical, but laid the foundation for exploring how a group of individual conjectures could be combined to produce a better forecast.

Since Galton, combination forecasting was mathematically cast as an opinion pool. Work in opinion pools began with Stone (Stone, 1961) in the early 1960s. He assumed a set of experts had an agreed upon utility function related to decision making, and that experts could each generate a unique probability distribution to describe their perceived future state of nature. To build a single combined forecast, Stone proposed a convex combination of each expert's probability distribution over the future—an opinion pool. Equally weighting individual predictions would reproduce Galton's model, and so the opinion pool was a more flexible way to combine expert opinions.

In the late 1960's, Granger and Bates formalized the concept of an optimal combination forecast. In their seminal work (Bates & Granger, 1969), several methods were proposed for how to combine point predictions to reduce, as much as possible, the combined forecast's variance. Methods for combining forecasts was further advanced by Granger and Ramanathan, and framed as a regression problem (Granger & Ramanathan, 1984). Work by Granger, Bates, and later Ramanathan inspired several novel methods for combining point forecasts (Coo, 1991; Wallis [Wal], 2011; Gneiting, Ranjan, et al., 2013; Hor, 2015). Combination forecasts often produce better predictions of the future than single models.

It wasn't until the 1990's that Cooke generalized the work of Stone and others, and developed an algorithm coined Cooke's method, or the Classical Model (Cooke et al. [Coo], 1988; Coo, 1991) for combining expert judgment. Every expert was asked to provide a probability distribution over a set of possible outcomes. To assign weights to experts, a calibration score statistic compared the expert's probability distribution to an empirical distribution of observations. Experts were assigned higher weights if their predictions closely matched the empirical distribution. The calibration score was studied by Cooke and asymptotic properties were summarized based on Frequentist procedures (Coo, 1988; Cooke [Coo], 2015). Cooke's model also assigned experts a weight of 0 for poor predictive performance, and if an expert's performance was under some user-set threshold they were excluded from the opinion pool. Cooke's model garnered much attention and has influenced numerous applications of combining expert opinion for forecasting (Zio, 1996; Clemen & Winkler [Cle], 2007; Cle, 2008; Bri, 2012; Sar, 2013; Coo, 2014b; Coo, 2014a; Coo, 2015; Hor, 2015; Hat, 2016; Bolger & Houlding [Bol], 2017; Mor, 2017; Han, 2018).

Alongside Frequentist approaches to combination forecasting, Bayesian approaches began to gain popularity in the 1970's (Morris, 1974). In the Bayesian paradigm, a decision maker (called a supra Bayesian), real or fictitious, is asked to evaluate expert forecasts and combine their information into a single probability distribution (Hogarth, 1975; Keeney, 1976). The supra Bayesian starts with a prior over possible future observations and updates their state of knowledge with expert-generated predictive densities. Combination formulas can be

specified via a likelihood function ℓ meant to align expert-generated predictive densities with observed data. The difficulties introduced by a Bayesian paradigm are familiar. The choice of likelihood function and prior will affect how expert opinions are pooled. Past work proposed many different likelihood functions, and interested readers will find a plethora of examples in Genest and Zidek (Gen, 1986), and Clemen and Winkler (Cle, 1986; Cle, 1989; Clemen & Winkler, 1999a).

2.4. Recent work in combination forecasting

Recent work has shifted from combining point estimates to combining predictive densities. Rigorous mathematical theory was developed and framed the problem of combining predictive densities (Gneiting, Ranjan, et al., 2013). Work combining predictive densities showed results similar in spirit to Granger and Bates' (Bates & Granger, 1969; Granger & Ramanathan, 1984) work on combining point predictions. Ranjan and Gneiting (Ranjan & Gneiting, 2010; Gneiting, Ranjan, et al., 2013) showed a set of calibrated predictive distributions, when combined using a linear pool, necessarily leads to an overdispersed and therefore miscalibrated combined distribution. This mimics Granger and Bates' results (Bates & Granger, 1969). They showed combining unbiased point predictions can lead to a combination method that makes biased point estimates.

This work in miscalibrated linear pools inspired new methods for recalibrating forecasts made from a combination of predictive densities. To recalibrate, authors recommend transforming the aggregated forecast distribution. The Spread-adjusted Linear Pool (SLP) (Berrocal et al. [Ber], 2007; Glahn et al. [Gla], 2009; Kleiber et al. [Kle], 2011) transforms each individual distribution before combining, the Beta Linear Pool (BLP) applies a beta transform to the final combined distribution (Ranjan & Gneiting, 2010; Gneiting, Ranjan, et al., 2013), and a more flexible infinite mixture version of the BLP (Bassetti et al., 2018), mixture of Normal densities (Baran & Lerch, 2018), and empirical cumulative distribution function (Garratt et al., 2019) also aim to recalibrate forecasts made from a combination of predictive densities.

Machine learning approaches assume a broader definition of a model as any mapping that inputs a training set and outputs predictions. This allows for more general approaches to combining forecasts called: ensemble learning, meta-learning, or hypothesis-boosting in machine learning literature. Stacking and the super-learner approach are two active areas of machine learning research to combine models. Stacked generalization (stacking) (Wolpert [Wol], 1992) proposes a mapping from out-of-sample predictions made by models (called base-learners) to a single combination forecast. The function that combines these models is called a generalizer and can take the form of any regression model, so long as it maps model predictions into a final ensemble prediction. The super-learner ensemble takes a similar approach to stacking (Van der Laan et al. [Van], 2007; Polley & Van Der Laan [Pol], 2010). Like stacking, the super-learner takes as input out-of-sample predictions from a set of models. Different from stacking, the super-learner algorithm imposes a specific form for aggregating predictions, a convex combination of models, such that the weights assigned to each model minimize an arbitrary loss function that includes the super-learner predictions and true outcomes of interest. By restricting how predictions are aggregated, super-learner is

guaranteed better performance under certain conditions (Van, 2007; Pol, 2010). Stacked and super-learner models often perform better than any individual forecasts and their success has led to applying them to many different problems (Sakkis et al. [Sak], 2001; Che et al. [Che], 2011; Wang et al. [Wan], 2011; Syarif et al. [Sya], 2012), however the machine learning community is debating issues with stacked models (Ting & Witten [Tin], 1999) and how they can be improved (Džeroski & Ženko [Dze], 2004).

2.5. Open challenges in combination forecasting

Combination forecasting has three distinct challenges: data collection, choice of combination method, and how to evaluate combination forecasts.

Crowdsourcing (How, 2006; Abernethy & Frongillo, 2011; Brabham, 2013; "Crowdsourcing the Future: Predictions Made with a Social Network," 2014; Moran et al., 2016) and expert elicitation (Amara & Lipinski, 1971; O'Hagan et al., 2006; Yousuf [You], 2007) are two approaches to collecting judgemental forecasts that attempt to balance competing interests: the quantity and quality of judgemental predictions. Crowdsourcing trades expertise for a large number of contributors. Expert judgemental forecasting takes the opposite approach and focuses on a small number of independent high-quality forecasts. Both methods try to enlarge the space of potential predictions so that a combination method can create a more diverse set of predictive densities over future events (Bates & Granger, 1969; Dietterich et al., 2002).

Combination methods are faced with developing a set of distributions over events of interest that take predictions as input and produce an aggregated prediction aimed at optimizing a loss function. Major challenges are how to account for missing predictions (Capistrán & Timmermann, 2009), correlated experts (**armstrong1985long**; Bunn [Bun], 1979, 1985), and how to ensure the combination forecast remains calibrated (Ber, 2007; Gla, 2009; Ranjan & Gneiting, 2010; Kle, 2011; Gneiting, Ranjan, et al., 2013; Garratt et al., 2019).

No normative theory for how to combine expert opinions into a single consensus distribution has been established, and a lack of theory makes comparing the theoretical merits of one method versus another difficult. Instead, authors compare combination methods using metrics that measure predictive accuracy: calibration, and sharpness (Dawid, 2007; Gneiting & Raftery, 2007; Gneiting & Ranjan, 2011; Jolliffe & Stephenson, 2012; Hor, 2015). Combination methods that output point forecasts are compared by measuring the distance between a forecasted point estimate and empirical observation. Probabilistic outputs are expected to be calibrated and attempt to optimize sharpness, or the concentration of probability mass over the empirical observations (Gneiting & Raftery, 2007; Gneiting & Ranjan, 2011; Jolliffe & Stephenson, 2012; Hor, 2015).

2.6. Past Reviews on Combination forecasting

Our review underlines the digital age's impact on combination forecasting. Collecting expert opinion in the past required one-on-one meetings with experts: in person, by phone, or mailed survey, and the internet decreased the burden of eliciting expert opinion by using online platforms to ask experts for their opinion (How, 2006). Past work focused on using

statistical models to combine forecasts, but increases in computing power broadened the focus from statistical models to machine-learning techniques. Our review explores how the digital age transformed combination forecasting and is an updated look at methods used to aggregate expert forecasts.

Many excellent past reviews of combination methods exist. Genest and Zidek give a broad overview of the field and pay close attention to the axiomatic development of combination methods (Gen, 1986). Clemen and Winkler wrote three reviews of aggregating judgmental forecasts (Cle, 1986; Cle, 1989; Clemen & Winkler, 1999a). The most cited manuscript overviews behavioral and mathematical approaches to aggregating probability distributions, reviews major contributions from psychology and management science, and briefly reviews applications. These comprehensive reviews center around the theoretical developments of combination forecasting and potential future directions of the science. Our work is an updated, and more applied, look at methods for aggregating expert predictions.

3. Methods

3.1. Search algorithm

The Web of Science database was used to collect articles relevant to combining expert prediction. Expert was defined by the authors of each manuscript, and we did not impose any restrictions on how expert was defined. The search string entered into Web of Science on 2019–03-06 was **(expert* or human* or crowd*) NEAR judgement AND (forecast* or predict*) AND (combin* or assimilat*)** and articles were restricted to the English language. All articles from this search were entered into a database. Information in this article database included: the author list, title of article, year published, publishing journal, keywords, and abstract (full database can be found at https://github.com/tomcm39/AggregatingExpertElicitedDataForPrediction).

To decide if an article was related to combining expert judgement, two randomly assigned reviewers (co-authors) read the abstract and were asked if the article should be included for analysis (in-scope) or excluded (out of scope). We defined an article as in-scope if it elicited expert judgments and combined them to make a prediction about natural phenomena or a future event. An article moved to the next stage if both reviewers agreed the article was in-scope. If the two reviewers disagreed, the article was sent to a randomly assigned third reviewer to act as a tie breaker and was considered in scope if this third reviewer determined the article was in-scope.

Full texts were collected for all in-scope articles. In-scope full texts were divided at random among all reviewers for a detailed reading. Reviewers were asked to read the article and fill out a prespecified questionnaire (Table 4). The questionnaire asked reviewers to summarize: the type of target for prediction, the methodology used, the experimental setup, and terminology associated with aggregating expert opinion. If after a detailed review the article is determined to be out of scope it was excluded from analysis. The final list of articles are called analysis-set articles.

3.2. Analysis of full text articles

From all analysis-set articles, abstract text was split into individual words, we removed English stop words—a set of common words that have little lexical meaning—that matched the Natural Language Toolkit (NLTK)'s stop word repository (Loper & Bird [Lop], 2002), and the final set of non-stopwords were stemmed (Willett [Wil], 2006).

A univariate analysis: (i) counted the number of times a word w appeared in abstract text per year $n_w(t)$, (ii) the total number of words among all abstracts in that year (N_t), and (iii) the frequency a word appeared over time $N_w = \sum_t nw(t)$. If a word w did not appear in a given year it received a count of zero ($n_w(t) = 0$).

Words were sorted by N_w and a histogram was plotted of the top 5% most frequently occurring words in abstract text. Among the top 12 most frequently occurring words, we plotted the proportion ($n_w(t)/N_w$) of each word over time.

Full text articles were scanned for key terms related to aggregating expert judgments. Evaluation metrics, a preferred abbreviation, related names, whether the metric evaluated a binary or continuous target, and formula to compute the metric was included in a table (Table 3). Terms specific to aggregating judgmental data were grouped by meaning and listed in a table (Table 1) along with a single definition. If multiple terms mapped to the same concept, our preferred label was placed at the top.

Frequencies and percents were computed for 'Yes/No' prespecified questions related to analysis-set articles (Statistics are presented in Table 2 and the list of all questions can be found in Table 4). Questions with text answers were summarized in the results.

4. Results

4.1. Search results

The initial Web of Science search returned 285 articles for review. After random assignment to two reviewers, 218 articles were agreed to be out of scope. The most frequent reasons for exclusion were the lack of experts used for prediction or the use of experts to revise, rather than directly participate in generating, forecasts. The 67 in-scope articles come from 50 articles two reviewers agreed to be in-scope, and 17 out of 74 articles a randomly assigned third reviewer considered in-scope. Full text analysis determined another 14 articles out of scope, and the final number of analysis-set articles was 53 (Fig. 1).

Analysis set articles were published from 1992 to 2018. Publications steadily increase in frequency from 1992 until 2011. After 2011, publication rates rapidly increase until 2018 (Fig. 2).

Analysis-set articles were published in 34 journals, and the top publishing journals are: the *International Journal of Forecasting* (4 articles), *Reliability Engineering & System Safety* (3 articles), and *Risk Analysis* and *Decision Analysis* (2 articles each). Combination forecasting articles often emphasize the role of decision makers in forecasting, and these top-publishing journals sit at the intersection of forecasting and decision sciences.

The top 10 most frequent words found in articles' abstracts are related to our initial search: "expert", "judgment", "forecast", "combin*", and "predict". Words related to modeling and methodology are also frequent: "model", "method", "approach", "predict". The word "assess" appears less frequently in abstracts and the word "accuracy" even less frequent (Fig. 3). The proportion of words: "expert", "forecast", "model", "method", and "data" appear intermittently in the 1990s and appear more consistently in the 2000s (Fig. 4). The words "probabili*" and "predict" occur in abstract text almost exclusively after the year 2000. The rise of "forecasts", "model", and "data" suggests data-driven combination forecasting schemes may be on the rise, and the uptick of "probabil*" and "predict" could be caused by an increase in aggregating expert probability distributions (rather than point forecasts).

Using expert judgement and data-driven models for forecasting are becoming more popular, and may suggest literature is focused on combining expert judgment to increase forecasting accuracy instead of building a consensus distribution. There is a potential gap in expert judgement used for probabilistic forecasting, but we did observe an increase in the frequency of the word-stem "probabil*".

4.2. Forecasting terminology

Forecasting terminology centered around six distinct categories (Table 1): frameworks for translating data and judgment into decisions (Forecasting support system, probabilistic safety assessment), broad approaches to aggregating forecasts (behavioral aggregation, mathematical combination, integrative judgment), specific ways experts can provide predictions (integrative judgment, judgemental adjustment), terms related to weighting experts (equal weighted linear pool, nominal weights), different names for classical models (Cooke's method, mixed estimation), and philosophical jargon related to combination forecasting (Laplacian principle of indifference, Brunswik lens model).

Few concepts in the literature are assigned a single label, the majority are given multiple labels. Some concepts' labels are similar enough that one can be swapped for another. For example, equal-weighted, 50–50, and unweighted all refer to assigning equal weights to expert predictive densities in a linear opinion pool. Other concepts are assigned different labels, for example forecasting support system and adaptive management, that may make it difficult to understand both terms refer to the same concept.

To condense terminology related to combination forecasting, we identified and grouped together frequently occurring terms in the literature that describe the same combination forecasting concept and suggested a single label (Table 1).

4.3. Forecasting targets

Forecasting research focused on predicting categorical variables (34%) and time-series (40%), but the majority of articles attempted to predict a continuous target (68%) (Table. 2).

The type of forecasting target depended on the application. Ecological and meteorological articles (Abr, 1996; Bor, 2004; Tar, 2007; Kur, 2010; Li, 2012; Coo, 2014a; Man, 2014; Mor, 2017; Joh2018, 2018; Wan, 2018) focused on continuous targets such as: the prevalence of

animal and microbial populations, deforestation, and climate change. Economics and managerial articles focused on targets like: the number of tourist arrivals, defects in programming code, and monthly demand of products (Fai, 2004; Kab, 2008; Shi, 2013; Son, 2013; Hua, 2016). Political articles focused on predicting presidential outcomes, a categorical target (Hurley & Lior [Hur], 2002; Gra, 2014a; Gra, 2014b; Morgan [Mor], 2014; Gra, 2015, 2018). Risk-related targets were continuous and categorical: the probability of structural damage, nuclear fallout, occupational hazards, and balancing power load (Zio, 1996; Zio, 1997; Mu & Xianming [Mu], 1999; Cra, 2001; Ren, 2002; Nev, 2008; Wan, 2008; Ada, 2009; Kla, 2010; Bri, 2012; Cab, 2012; Bri, 2016; Hat, 2016; Bae, 2017; Jan, 2019). Public health papers predicted continuous targets over time, like forecasting carcinogenic risk (Eva, 1994) and US mortality rates (Alh, 1992).

Targets were often either too far in the future to assess, for example predicting precipitation changes in the next 1 million years (Zio, 1997), or related to a difficult-to-measure quantity, such as populations of animals with little or no monitoring (Bor, 2004; Man, 2014; Joh2018, 2018). The majority of analysis-set articles placed more importance on the act of building a consensus distribution than studying the accuracy of the combined forecast (Eva, 1994; Abr, 1996; Zio, 1996; Zio, 1997; Mu, 1999; Bor, 2004; Fai, 2004; Cle, 2007; Tar, 2007; Kab, 2008; Nev, 2008; Wan, 2008; Ada, 2009; Kla, 2010; Kur, 2010; Bri, 2012; Cab, 2012; Li, 2012; Shi, 2013; Son, 2013; Baron et al. [Bar], 2014; Coo, 2014a; Man, 2014; Mor, 2014; Hor, 2015; Bri, 2016; Gu, 2016; Hat, 2016; Joh2018, 2018; Wan, 2018). It is important to note the ability of experts to build forecasts and support decision-making with very little, or any, observed data on targets of interest.

All articles defined a small number of specific forecasting targets. The majority of targets related to safety. Public health, ecology, and engineering applications focused on forecasting targets that, if left unchecked, could negatively impact human lives or the surrounding environment. For example, carcinogenic risk imposed by the concentration of chloroform found in drinking water (Eva, 1994). What differed between articles was whether ground truth data was collected on targets, or if no data was collected, that ground truth data could be collected in the near future.

4.4. Forecasting methodology

Articles were split into three groups, those that took (i) a Bayesian approach, (ii) Frequentist approach, or (iii) neither.

Articles taking a Bayesian approach accounted for 25% of analysis-set articles and emphasized how priors can compliment sparse data (Abr, 1996; Zio, 1997; Ren, 2002; Cle, 2007; Tar, 2007; Nev, 2008; Bri, 2012; Man, 2014; Bri, 2016; Hua, 2016; Bol, 2017; Wan, 2018). Many papers focused on assessing risk (Zio, 1997; Tar, 2007; Bri, 2012; Bri, 2016). For example, the risk of losing autonomous underwater vehicles was modeled using a Bayesian approach that incorporated objective environmental data and subjective probabilities of loss solicited from experts (Bri, 2012; Bri, 2016). Other papers assessed the impact of subsurface hydrology on water contamination (Tar, 2007), the risk of structural deterioration (Nev, 2008), and the economic risk associated with government expenditures (Wan, 2018).

Bayesian methods involved beta-binomial models, Bayesian decision trees, mixture distributions, or Bayesian belief networks. For example, work by (Zio, 1997) used a linear pool to combine expert judgments. The probability density over the target of interest x was defined as

$$f(\tilde{x}) = \sum_{e=1}^{E} w_e f_e(\tilde{x} \mid \mathscr{D})$$

where w_e is a weight assigned to each expert and $f_e(\tilde{x} \mid \mathscr{D})$ is the probability expert e places on $\tilde{x}$, having observed data $\mathscr{D}$. The goal is to find weights w_e that optimize a loss function. They extended this linear pool model by assuming experts could observe different data based on the target of interest x. The new linear pool was defined as

$$f(\tilde{x}) = \sum_{e=1}^{E} g(\mathscr{D}_e \mid x) f_e(\tilde{x} \mid \mathscr{D}_e)$$

where $g(\mathscr{D}_e|x)$ is the probability expert e observed data $\mathscr{D}_e$. The goal is then to accurately compute $g(\mathscr{D}_e|x)$ and use this probability density to weight experts.

Often Bayesian models involved complicated posterior computations, requiring numerical integration to compute forecast probabilities. Past work suggested a Bayesian framework could better model subjective probabilities elicited from experts (Cle, 2007), however Frequentist techniques were used in almost 50% of articles.

Frequentist models for combining forecasts (Alh, 1992; Eva, 1994; Mak, 1996; Mu, 1999; Hur, 2002; Ren, 2002; Bor, 2004; Wan, 2008; Ada, 2009; Kla, 2010; Kur, 2010; Fra, 2011; Cab, 2012; Sei, 2013; Shi, 2013; Coo, 2014a; Gra, 2014b; Baldwin [Bal], 2015; Hor, 2015; Gu, 2016; Hat, 2016; Alv, 2017; Bae, 2017; Mor, 2017; Han, 2018; Jan, 2019) were typically convex combinations of expert judgment or linear regression models that included expert judgment as a covariate. Including expert judgment as a covariate in a linear regression model is related to judgemental bootstrapping (Armstrong, 2001b) and the Brunswik lens model (Hammond & Stewart, 2001). Both techniques are mentioned in analysis-set articles and rely on a Frequentist regression that divides human judgment into predictions inferred from data and expert intuition,

$$y_e \mid x_e, \beta_0, \beta, \sigma^2 \sim \mathcal{N}\left(\beta_0 + \beta' x_e, \sigma^2\right)$$

where y represents the expert's forecast, $\mathcal{N}$ is a Normal distribution, x_e is a vector of objective information about the target of interest, β are estimated parameters, and σ^2 is argued to contain expert intuition. This model can then infer what covariates (x_e) are important to expert decision making and to what extent expert intuition (σ^2) is involved in prediction.

Articles that did not use classic regression combined statistical predictions with qualitative estimates made by experts using fuzzy logic (an extension of traditional logic where elements can partially belong to many sets). Cooke's method inspired articles to take a mixture model approach and weighted experts based on how well they performed on a set of ground-truth questions.

Articles using neither Bayesian or Frequentist models (Ren, 2002; Fai, 2004; Pet, 2006; Kab, 2008; Li, 2012; Hora et al. [Hor], 2013; Son, 2013; Bar, 2014; Gra, 2014a; Mor, 2014; Gra, 2015; Cai, 2016; Gra, 2018; Joh2018, 2018) resorted to: dynamical systems, simple averages of point estimates and quantiles from experts, and tree-based regression models. As a simplified example from (Joh2018, 2018), they aimed to predict the population of geese using a differential equation. The population of geese at time $t + 1$ (N_{t+1}) was modeled as

$$\frac{dN_t + 1}{dt} = N_t r\left[1 - \left(\frac{N_t}{K}\right)^{\theta}\right] - h_t N_t$$

where probability distributions over parameters r, the intrinsic growth rate, θ, the density dependence, and h_t, the harvest rate, could be estimated from data. To estimate the carrying capacity K, a probabilistic distribution was fit to expert judgments. A forecast of the geese population was generated by repeatedly sampling parameters (r, θ, h_t, K) and integrating the differential equation.

The majority of models were parametric. Non-parametric models included: averaging quantiles, equally weighting expert predictions, and weighting experts via decision trees. These models allowed the parameter space to grow with increasing numbers of judgmental forecasts. Parametric models included: linear regression, ARIMA, state space models, belief networks, the beta-binomial model, and neural networks. Expert judgments, when combined and used to forecast, showed positive results in both nonparametric and parametric models. Parametric Bayesian models and non-parametric models could better cope with sparse data than a parametric Frequentist model. Bayesian models used a prior to lower model variance when data was sparse and non-parametric models could combine a expert judgments without relying on a specific form for the aggregated predictive distribution.

Authors more often proposed combining expert-generated point estimates compared to predictive distributions. A diverse set of models were proposed to combine point estimates: regression models (linear regression, logistic regression, ARIMA, exponential smoothing), simple averaging, and neural networks (Mak, 1996; Ada, 2009; Cab, 2012; Bar, 2014; Gra, 2014b), and fuzzy logic (Ren, 2002; Pet, 2006; Kab, 2008; Jan, 2019). Authors that combined predictive densities focused on simpler combination models.

Most predictive distributions were built by asking experts to provide a list of values corresponding to percentiles. For example, a predictive density would be built by asking each expert to provide values corresponding to the 5%, 50% (median), and 95% percentiles. Combination methods either directly combined these percentiles by assigning weights to each expert density (Zio, 1997; Kab, 2008; Sar, 2013; Bri, 2016; Cai, 2016; Bol, 2017; Mor,

2017; Han, 2018), or built a continuous predictive distribution that fit these discrete points (Abr, 1996; Fai, 2004; Nev, 2008; Wan, 2008; Kur, 2010; Bri, 2012).

4.5. Forecasting evaluation metrics

Only 42% of articles evaluated forecast performance using a formal metric. Formal metrics used in analysis-set articles are summarized in Table 3. The articles that did not include a metric to compare forecast performance either did not compare combination forecasts to ground truth, evaluated forecasts by visual inspection, or measured success as the ability to combine expert-generated forecasts. Among articles that did evaluate forecasts, most articles focused on point estimates (68%) versus probabilistic forecasts (23%), and two articles did not focus on point or probabilistic forecasts from experts.

The most commonly used metrics to evaluate point forecasts were: the Brier score (mean squared difference between the forecast and observed outcome), mean absolute (and percentage) error, and root mean square error. Even when predictive densities were combined, the majority of articles output and evaluated point estimates.

A small number of articles combining probability distributions used metrics that evaluated aggregated forecasts based on density, not point forecasts. Expert forecasts were evaluated using relative entropy and a related metric, the calibration score (see Table 3 for details). These metrics were first introduced (and not part of the analysis set articles) by Cooke (Coo, 1988; Coo, 1991).

The logscore is one of the most cited proper scoring rules for assessing calibration and sharpness (Gneiting & Raftery, 2007; Gneiting & Ranjan, 2011; Hor, 2015) for predictive densities, but was not used in any of the analysis-set articles. Instead, analysis-set articles emphasized point estimates and used metrics to evaluate point forecasts.

Three articles conducted an experiment but did not use any formal metrics to compare the results. Two articles used no evaluation and one article visually inspected forecasts.

4.6. Experimental design

Among all analysis-set articles, 43% conducted a comparative experiment (Alh, 1992; Mak, 1996; Cra, 2001; Hur, 2002; Ren, 2002; Pet, 2006; Fra, 2011; Hor, 2013; Sei, 2013; Gra, 2014a; Gra, 2014b; Sat, 2014; Bal, 2015; Gra, 2015; Cai, 2016; Hua, 2016; Alv, 2017; Bae, 2017; Bol, 2017; Mor, 2017; Gra, 2018; Han, 2018; Jan, 2019). We defined a comparative experiment as an evaluation of two or more models using a formal metric, like those found in Table 3.

Articles compared: (i) traditional statistical models to models that combined expert judgments (Pet, 2006; Gra, 2014a; Mor, 2017; Han, 2018), (ii) individual experts' performance vs. aggregated expert performance (Gra, 2014a; Gra, 2014b; Gra, 2015; Bol, 2017; Gra, 2018), and (iii) different models for combining expert judgments (Mak, 1996; Cra, 2001; Hur, 2002; Hor, 2013; Sei, 2013; Sat, 2014; Bal, 2015; Bae, 2017; Jan, 2019). The type of statistical models compared to expert judgement depended on if the target was binary or continuous. For binary targets, logistic regression, K-nearest neighbors, neural

networks, and discriminant analysis was fit to training data and used to make predictions. For continuous targets, ARIMA, neural networks, the median and trimmed mean, and parametric regression models were fit to training data. Statistical models tended to perform worse than ensembles that included expert judgment. But when past work compared statistical models against ensembles, they built ensembles from a combination of statistical forecasts and expert judgment. It isn't clear if the ensemble of statistical models plus expert judgment is because of the added expertise or because ensembles tend to perform better than any of the individual models that are a part of the ensemble.

Other literary sources focused on comparing an aggregated expert judgment model to individual responses. The majority of results found that the ensemble performed better, but there were cases when individuals outperformed the aggregated model (Mor, 2017). Worse aggregated performance was attributed to some experts giving very poor, "outlier-like", estimates. Guarding against outlier was also the main message of (Hor, 2013). They found aggregating via the median (robust to outliers) outperformed a simple average.

Equal weighting was a common benchmark used as a comparison to a novel aggregation method. Some work compared novel aggregation methods to a Delphi technique (Cai, 2016), prediction market (Gra, 2014a; Gra, 2014b; Gra, 2018), and regression models (Mak, 1996). Novel models of aggregation, compared to equal weighting, performed better when expert predictions were simulated (Hur, 2002; Hor, 2013; Bal, 2015; Bol, 2017). Models performed similar to simpler aggregation models when trained on real expert output (Alh, 1992; Mak, 1996; Cra, 2001; Ren, 2002; Pet, 2006; Fra, 2011; Sei, 2013; Gra, 2014a; Gra, 2014b; Sat, 2014; Gra, 2015; Cai, 2016; Hua, 2016; Alv, 2017; Bae, 2017; Mor, 2017; Gra, 2018; Han, 2018; Jan, 2019).

Many articles did not evaluate their forecasting methods because no ground truth data exists. For example, articles would ask experts to give predictions for events hundreds of years in the future (Zio, 1996; Zio, 1997). Articles that didn't evaluate their combined forecast but did have ground truth data concluded that the predictive distribution they created was "close" to a true distribution. Still other articles concluded their method successful if it could be implemented at all.

4.7. Training data

Rapidly changing training data—data that could change with time—appeared in 41% of articles. Data came from finance, business, economics, and management and predicted targets like: monthly demand of products, tourist behavior, and pharmaceutical sales (Pet, 2006; Wan, 2008; Kla, 2010; Fra, 2011; Bae, 2017). In these articles, authors stress experts can add predictive power by introducing knowledge not used by statistical models, when the quality of data is suspect, and where decisions can have a major impact on outcomes. The rapidly changing environment in these articles is caused by consumer/human behavior.

Articles applied to politics stress that experts have poor accuracy when forecasting complex (and rapidly changing) systems unless they receive rapid feedback about their forecast accuracy and have contextual information about the forecasting task (Gra, 2014a; Sat, 2014;

Gra, 2015, 2018). Political experts, it is argued, receive feedback by observing the outcome of elections and often have strong contextual knowledge about both candidates.

Weather and climate systems were also considered datasets that rapidly change. The Hailfinder system relied on expert knowledge to predict severe local storms in eastern Colorado (Abr, 1996). Weather systems are rapidly changing environments, and this mathematical model of severe weather needed training examples of severe weather. Rather than wait, the Hailfinder system trained using expert input. Expert knowledge was important in saving time and money, and building a severe weather forecasting system that worked.

Ecology articles solicited expert opinion because of sparse training data, a lack of sufficient monitoring of wildlife populations, or to assign subjective risk to potential emerging biological threats (Kur, 2010; Li, 2012; Man, 2014)

Manuscripts that explicitly mention the training data describe the typical statistical model's inability to handle changing or sparse data, and suggest expert predictions may increase accuracy (Sei, 2013; Son, 2013).

4.8. Number of elicited experts and number of forecasts made

Over 50% of articles combined forecasts from less than 10 experts. (Fig. 5). Several articles describe the meticulous book-keeping and prolonged time and effort it takes to collect expert judgments. The costs needed to collect expert opinion may explain the small number of expert forecasters.

Two distinct expert elicitation projects produced articles that analyzed over 100 forecasters. The first project (Sei, 2013) asked experts from music record labels to predict the success (rank) of pop singles. Record label experts were incentivized with a summary of their predictive accuracy, and an online platform collected predictions over a period of 12 weeks.

One of the most successful expert opinion forecasting systems enrolled approximately 2000 participants and was called the Good Judgement Project (GJP) (Ungar et al. [Ung], 2012; Mellers et al. [Mel], 2014; Sat, 2014). Over a period of 2 years, an online platform was used to ask people to assign a probability to the occurrence of geo-political events and to self-assess their level of expertise on the matter. Participants were given feedback on their performance and how to improve with no additional incentives. Both projects that collected a large number of forecasters have common features. An online platform was used to facilitate data collection, and questions asked were simple, either binary (yes/no) questions or to rank pop singles. Both project incentivized participants with feedback of their forecasting performance.

Close to 80% of articles reported less than 100 total forecasts (Fig. 6) and studies reporting more than 10^4 forecasts were simulation based (except the GJP). Recruiting a small number of experts did not always result in a small number of forecasts. Authors assessing the performance of the Polly Vote system collected 452 forecasts from 17 experts (Gra, 2014a; Gra, 2015, 2018), and a project assessing the demand for products produced 638 forecasts from 31 forecasters (Alv, 2017).

One simulation study varied how many experts were aggregated and compared forecasting performance (Hor, 2013). This work studied an increasing number of experts that were (i) independent and well-calibrated to the forecasting target, and (ii) correlated to one another and overconfident about their forecast. Results showed that an optimal choice of aggregation method depended on the number of experts and whether they were correlated and calibrated to the target. Experts who are correlated and over confident (which the authors have reported is closer to what they see in practice) produce less calibrated aggregated forecasts. The choice of mean vs. median quantile aggregation showed a small improvement in calibration.

The time and energy required to collect expert opinion is reflected in the low number of forecasters. Some studies did succeed to produce many more forecasts than recruited forecasters, and they did so by using an online platform, asking simpler questions, and giving forecasters feedback about their forecast accuracy.

5. Discussion

Combining expert predictions for forecasting continues to shows promise, however rigorous experiments that compare expert to non-expert and statistical forecasts are still needed to confirm the added value of expert judgement. The most useful application in the literature appeals to a mixture of statistical models and expert prediction when data is sparse and evolving. Despite the time, effort, and cost it takes to elicit expert-generated data, the wide range of applications and new methods show the field is growing. Authors also recognize the need to include human intuition into models that inform decision makers.

In any combination forecast, built from expert or statistical predictions, there is no consensus on how to best combine individual forecasts or how to compare one forecast to another (Table 3). In addition to methodological disagreements familiar to any combination algorithm, expert judgemental forecasts have the additional burden of collecting predictions made by experts. Expertise is subjective and subject-specific, and the literature has not settled on how to define expertise. An entire field is devoted to understanding how experts differ from non-experts (Dawid et al., 1995; Rikers & Paas, 2005; Farrington-Darby & Wilson, 2006; Ericsson & Ward, 2007; de Groot, 2014). Methods for collecting data from experts that are unbiased and in the least time-consuming manner is also an area of open inquiry. An investigator must spend time designing a strategy to collect data from experts, and experts themselves must make time to complete this prediction task. There is a vast literature on proper techniques for collecting expert-generated data (Normand et al. [Nor], 1998; Ayy, 2001; Powell [Pow], 2003; Leal et al. [Lea], 2007; You, 2007; Martin et al. [Mar], 2012). Expert elicitation adds an additional burden to combination forecasting not present when aggregating purely statistical models.

We identified four key themes reiterated in combination forecasting literature: (i) the use of human intuition to aid statistical forecasts when data is sparse and rapidly changing, (ii) including experts because of their role as decision makers, (iii) using simpler aggregation models to combine predictive densities and more complicated models to combine point predictions, and (iv) the lack of experimental design and comparative metrics in many manuscripts.

Many articles introduced expert judgment into their models because the data needed to train a statistical model was unavailable, sparse, or because past data was not a strong indicator of future behavior. When training data was available, researchers typically used expert forecasts to supplement statistical models. Authors argued that experts have a broader picture of the forecasting environment than is present in empirical data. If experts produced forecasts based on uncollected data, then combining their predictions with statistical models was a way of enlarging the training data. Expert-only models were used when data on the forecasting target was unavailable. Authors argued context-specific information available to experts and routine feedback about their past forecasting accuracy meant expert-only models could make accurate forecasts. Though we feel this may not be enough to assume expert-only models can make accurate forecasts, without any training data these attributes allow experts to make forecasts when statistical models cannot.

Most articles in this review took a definition of expertise similar to Dawid and Degroot (Dawid et al., 1995). They considered experts those people with past experiences related to the target of interest, or those people the decision maker would consider more knowledgeable than themselves—subject matter experts. No analysis-set article explicitly defined expertise, but using Dawid's definition, expertise should be thought of as context-specific and tied to the forecasting target. Some forecasters may be better suited for different targets, different scenarios, or specific combinations of each.

Applications varied, but each field stressed the reason for aggregating forecasts from experts was due to decision-making under uncertainty. For example: deciding on how a company can improve their marketing strategy, what choices and actions can affect wildlife populations and our environment, deciding on the structural integrity of buildings and nuclear power plants. Numerous articles emphasized the role of decision making in these systems by naming the final aggregated forecast a decision maker.

A longer history of combining point forecasts (Galton, 1907; Bates & Granger, 1969; Granger & Ramanathan, 1984) has prompted advanced methods for building aggregated forecasts from point estimates. Simpler aggregation techniques, like linear pools, averaging quantiles, and rank statistics, were used when combining predictive densities. Besides the shorter history, simple aggregation models for predictive densities show comparable, and often, better results than more complicated techniques (Cle, 1989; Rantilla & Budescu, 1999). The reasons why simple methods work so well for combining predictive densities is mostly empirical at this time (Makridakis & Winkler, 1983; Cle, 1989; Rantilla & Budescu, 1999), but under certain scenarios, a simple average was shown to be optimal (Wallsten et al., 1997a; Wallsten et al., 1997b).

A small percentage of research took time to setup an experiment that could rigorously compare combination forecasting models. Most articles took a consensus distribution approach and measured success on whether or not the combination scheme could produce a forecast and inspected the results visually. In some cases visual inspection was used because ground truth data was not present, but in this case, a simulation study could offer insight into the forecasting performance of a novel combination method. No manuscripts compared predictions between forecasts generated by experts only, a combination of experts and

statistical models, and statistical models only. Past research is still unclear on the added value experts provide statistical forecasts, and whether expert-only models provide accurate results.

To support research invested in aggregating expert predictions and improve their rigorous evaluation, we recommend the following: (i) future work spend more time on combining probabilistic densities and understanding the theoretical reasons simple aggregation techniques outperform more complicated models, when appropriate (ii) authors define an appropriate metric to measure forecast accuracy and develop rigorous experiments to compare novel combination algorithms to existing methods. If not feasible we suggest a simulation study that enrolls a small, medium, and large number of experts to compare aggregation models. We also suggest (iii) authors carefully define expertise. Expertise should be defined in enough detail so that others can apply the same set of criteria to a novel pool of potential judgmental forecasters.

We recommend that the two goals of combining expert judgment begin to intersect. Historical models of meteorological phenomena, and more recent human-in-the-loop research point to more accurate and understandable models by involving experts/decision makers. Current research, called ‘human-in-theloop’ machine learning, integrates human participants directly into a machine learning workflow and has shown positive performance in many different applications (Cranor, 2008; Yu et al., 2015; Holzinger, 2016; Li, 2017). Participants can select models and model parameters for training, evaluate the output, and create complicated models of their own. As active participants, decision makers are able to build accurate models that, because they took part in building the model, they trust. Weather forecasting is another example of a successful forecasting system by supporting close communication between forecasting models and decision makers (Murphy & Winkler, 1974a, 1974b, 1984). Forecasts are frequent, transparent about uncertainty, and provide meteorologist’s a condensed model about the future they can rapidly compare to the truth. We expect more research will explore how to build stronger forecasting systems by blending together the two combination forecasting goals of a more accurate forecast with a trusted consensus distribution.

Aggregating expert predictions can outperform statistical ensembles when data is sparse, or rapidly evolving. By making predictions, experts can gain insight into how forecasts are made, the assumptions implicit in forecasts, and ultimately how to best use the information forecasts provide to make critical decision about the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was funded by the National Institute of General Medical Sciences (NIGMS) Grant R35GM119582. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of

the NIH or the NIGMS. The funders had no role in study design, data collection and analysis, decision to present, or preparation of the presentation.

References

Galton F. (1907). Vox populi. Nature, 75(7), 450–451. 10.1038/075450a0

Meehl PE (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. 10.1037/11281-000

Stone M. (1961). The Opinion Pool. The Annals of Mathematical Statistics, 32(4), 1339–1342. https://doi.org/www.jstor.org/stable/2237933

Helmer O. (1967, 3). Analysis of the Future: The Delphi Method (tech. rep.). The RAND Corporation. https://www.rand.org/pubs/papers/P3558.html

Winkler RL (1968). The Consensus of Subjective Probability Distributions. Management Science, 15(2), B–61. www.jstor.org/stable/2628853

Bates JM, & Granger CW (1969). The Combination of Forecasts. Journal of the Operational Research Society, 20(4), 451–468. 10.1057/jors.1969.103

Amara RC, & Lipinski AJ (1971). Some views on the use of expert judgment. Technological Forecasting and Social Change, 3, 279–289. 10.1016/S0040-1625(71)80019-7

Winkler RL (1971). Probabilistic prediction: Some experimental results. Journal of the American Statistical Association, 66(336), 675–685. 10.1080/01621459.1971.10482329

McLaughlin RL (1973). The forecasters' batting averages. Business Economics, 58–59.

Morris PA (1974). Decision Analysis Expert Use. Management Science, 20(9), 1233–1241. www.jstor.org/stable/2630184.

Murphy AH, & Winkler RL (1974a). probability forecasts: a survey of National Weather Service forecasters. Bulletin of the American Meteorological Society, 55(12), 1449–1453. https://www.jstor.org/stable/26255518

Murphy AH, & Winkler RL (1974b). Subjective probability forecasting experiments in meteorology: Some preliminary results. Bulletin of the American Meteorological Society, 55(10), 1206–1216. 10.1175/1520-0477(1974)055×1206:SPFEIMy2.0.CO;2

Hogarth RM (1975). Cognitive Processes and the Assessment of Subjective Probability Distributions. Journal of the American Statistical Association, 70(350), 271–289. 10.2307/2285808

Keeney RL (1976). A Group Preference Axiomatization with Cardinal Utility. Management Science, 23(2), 140–145. https://www.jstor.org/stable/2629821

Bunn DW (1979). The Synthesis of Predictive Models in Marketing Research. 10.2307/3150692

Winkler RL (1981). Combining Probability Distributions from Dependent Information Sources. Management Science, 27(4), 479–488. 10.1287/mnsc.27.4.479

Armstrong JS (1983). Relative Accuracy of Judgemental and Extrapolative Methods in Forecasting Annual Earnings. Journal of Forecasting, 2(4), 437–447. 10.1002/for.3980020411

Makridakis S, & Winkler RL (1983). Averages of Forecasts: Some Empirical Results. Management Science, 29(9), 987–996. https://www.jstor.org/stable/2630927

Granger CW, & Ramanathan R. (1984). Improved methods of combining forecasts. Journal of Forecasting, 3(2), 197–204. 10.1002/for.3980030207

Murphy AH, & Winkler RL (1984). Probability Forecasting in Meteorology. Journal of the American Statistical Association, 79(387), 489–500. 10.2307/2288395

Armstrong JS (1985). Long-range forecasting: From Crystal Ball to Computer. Wiley.

Bunn DW (1985). Statistical efficiency in the linear combination of forecasts. International Journal of Forecasting, 1(2), 151–163. 10.1016/0169-2070(85)90020-2

Clemen RT, & Winkler RL (1986). Combining Economic Forecasts. Journal of Business & Economic Statistics, 4(1), 39–46. 10.2307/1391385

Genest C, Zidek JV et al. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. Statistical Science, 1(1), 114–135. 10.1214/ss/1177013825

Cooke R, Mendel M, & Thijs W. (1988). Calibration and information in expert resolution; a classical approach. Automatica, 24(1), 87–93. 10.1016/0005-1098(88)90011-8

DeGroot MH (1988). A Bayesian view of assessing uncertainty and comparing expert opinion. Journal of Statistical Planning and Inference, 20(3), 295–306. 10.1016/0378-3758(88)90094-8

Clemen RT (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5(4), 559–583. 10.1016/0169-2070(89)90012-5

Edmundson R. (1990). Decomposition; a strategy for judgemental forecasting. Journal of Forecasting, 9(4), 305–314. 10.1002/for.3980090403

Kleinmuntz B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. Psychological Bulletin, 107(3), 296. 10.1037/0033-2909.107.3.296 [PubMed: 2190252]

Bunn D, & Wright G. (1991). Interaction of Judgemental and Statistical Forecasting Methods: Issues & Analysis. Management Science, 37(5), 501–518. 10.1287/mnsc.37.5.501

Cooke R et al. (1991). Experts in uncertainty: Opinion and subjective probability in science. Oxford University Press on Demand.

Alho JM (1992). Estimating the strength of expert judgement: The case of US mortality forecasts. Journal of Forecasting, 11(2), 157–167. 10.1002/for.3980110206

Lawrence M, & O'Connor M. (1992). Exploring judgemental forecasting. International Journal of Forecasting, 8(1), 15–26. 10.1016/0169-2070(92)90004-S

Wolpert DH (1992). Stacked generalization. Neural Networks, 5(2), 241–259. 10.1016/S0893-6080(05)80023-1

O'Connor M, Remus W, & Griggs K. (1993). Judgemental forecasting in times of change. International Journal of Forecasting, 9(2), 163–172. 10.1016/0169-2070(93)90002-5

Yaniv I, & Hogarth RM (1993). Judgmental Versus Statistical Prediction: Information Asymmetry and Combination Rules. Psychological Science, 4(1), 58–62. 10.1111/j.1467-9280.1993.tb00558.x

Evans JS, Gray GM, Sielken RL, Smith AE, Valdezflores C, & Graham JD (1994). Use of probabilistic expert judgment in uncertainty analysis of carcinogenic potency. Regulatory Toxicology and Pharmacology, 20(1), 15–36. 10.1006/rtph.1994.1034 [PubMed: 7838990]

Evans J, Gray G, Sielken R, Smith A, Valdezflores C, & Graham J. (1994). Use of probabilistic expert judgment in uncertainty analysis of carcinogenic potency. Regulatory Toxicology and Pharmacology, 20(1), 15–36. 10.1006/rtph.1994.1034 [PubMed: 7838990]

Alexander JC Jr (1995). Refining the Degree of Earnings Surprise: A Comparison of Statistical and Analysts' Forecasts. Financial Review, 30(3), 469–506. 10.1111/j.1540-6288.1995.tb00842.x

Dawid A, DeGroot M, Mortera J, Cooke R, French S, Genest C, Schervish M, Lindley D, McConway K, & Winkler R. (1995). Coherent combination of experts' opinions. Test, 4(2), 263–313. 10.1007/BF02562628

Abramson B, Brown J, Edwards W, Murphy A, & Winkler RL (1996). Hailfinder: A Bayesian system for forecasting severe weather. International Journal of Forecasting, 12(1), 57–71. 10.1016/0169-2070(95)00664-8

Guangliang S. (1996). A Multi-hierarchical Comprehensive Evaluation Model and its Application. Systems Engineering, 2.

Mak B, Bui T, & Blanning R. (1996). Aggregating and updating experts' knowledge: An experimental evaluation of five classification techniques. Expert Systems with Applications, 10(2), 233–241. 10.1016/0957-4174(95)00049-6

Zio – Zio E. (1996). On the use of the analytic hierarchy process in the aggregation of expert judgments. Reliability Engineering & System Safety, 53(2), 127–138. 10.1016/0951-8320(96)00060-9

Spence MT, & Brucks M. (1997). The Moderating Effects of Problem Characteristics on Experts' and Novices' Judgments. Journal of Marketing Research, 34(2), 233–247. 10.2307/3151861

Wallsten TS, Budescu DV, Erev I, & Diederich A. (1997a). Evaluating and Combining Subjective Probability Estimates. Journal of Behavioral Decision Making, 10(3), 243–268. 10.1002/(SICI)1099-0771(199709)10:3×243::AID-BDM268y3.0.CO;2-M

Wallsten TS, Budescu DV, & Tsao CJ (1997b). Combining linguistic probabilities. Psychologische Beitrage.

Zio E, & Apostolakis G. (1997). Accounting for expert-to-expert variability: A potential source of bias in performance assessments of high-level radioactive waste repositories. Annals of Nuclear Energy, 24(10), 751–762. 10.1016/S0306-4549(96)00052-7

Normand S-LT, McNeil BJ, Peterson LE, & Palmer RH (1998). Eliciting expert opinion using the Delphi technique: Identifying performance indicators for cardiovascular disease. International Journal for Quality in Health Care, 10(3), 247–260. 10.1093/intqhc/10.3.247 [PubMed: 9661064]

Clemen RT, & Winkler RL (1999a). Combining Probability Distributions From Experts in Risk Analysis. Risk Analysis, 19(2), 187–203. 10.1111/j.1539-6924.1999.tb00399.x

Fischer I, & Harvey N. (1999). Combining forecasts: What information do judges need to outperform the simple average? International Journal of Forecasting, 15(3), 227–246. https://doi.org/0.1016/S0169-2070(98)00073-9

Mu L, & Xianming W. (1999). Multi-hierarchical durability assessment of existing reinforced-concrete structures. In Proceedings of the 8th International Conference on Durability of Building Materials and Components. https://www.irbnet.de/daten/iconda/CIB1808.pdf

Rantilla AK, & Budescu DV (1999). Aggregation of Expert Opinions. In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. IEEE. 10.1109/HICSS.1999.772751

Ting KM, & Witten IH (1999). Issues in Stacked Generalization. Journal of Artificial Intelligence Research, 10, 271–289. 10.1613/jair.594

Armstrong JS Combining Forecasts. In: In Principles of Forecasting. International Series in Operations Research and Management Science. Springer, 2001, pp. 417–439. 10.1007/978-0-306-47630-319.

Armstrong JS Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting. In: In Principles of Forecasting. International Series in Operations Research and Management Science. Springer, 2001, pp. 171–192. 10.1007/978-0-306-47630-39.

Ayyub BM (2001). Elicitation of Expert Opinions for Uncertainty and Risks. CRC press.

Craig PS, Goldstein M, Rougier JC, & Seheult AH (2001). Bayesian Forecasting for Complex Systems Using Computer Simulators. Journal of the American Statistical Association, 96(454), 717–729. 10.1198/016214501753168370

Hammond KR, & Stewart TR (2001). The Essential Brunswik: Beginnings, Explications, Applications. Oxford University Press. https://psycnet.apa.org/record/2001-18779-000

Sakkis G, Androutsopoulos I, Paliouras G, Karkaletsis V, Spyropoulos CD, & Stamatopoulos P. (2001). Stacking Classifiers for AntiSpam Filtering of E-Mail. In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. https://www.aclweb.org/anthology/W01-0506

Dietterich TG et al. (2002). Ensemble learning. The Handbook of Brain Theory and Neural Networks, 2, 110–125.

Hurley W, & Lior D. (2002). Combining expert judgment: On the performance of trimmed mean vote aggregation procedures in the presence of strategic voting. European Journal of Operational Research, 140(1), 142–147. 10.1016/S0377-2217(01)00226-0

Loper E, & Bird S. (2002). NLTK: The Natural Language Toolkit. arXiv:cs/0205028

Ren – Ren-jun Z, & Xian-zhong D. (2002). Optimal combined load forecast based on the improved analytic hierarchy process. In Proceedings. International Conference on Power System Technology. IEEE. 10.1109/ICPST.2002.1047570

Powell C. (2003). The Delphi technique: Myths and realities. Journal of Advanced Nursing, 41(4), 376–382. 10.1046/j.1365-2648.2003.02537.x [PubMed: 12581103]

Borsuk ME (2004). Predictive Assessment of Fish Health and Fish Kills in the Neuse River Estuary Using Elicited Expert Judgment. Human and Ecological Risk Assessment, 10(2), 415–434. 10.1080/10807030490438454

Džeroski S, & Ženko B. (2004). Is combining classifiers with stacking better than selecting the best one? Machine learning, 54(3), 255–273. https://doi.org/0.1023/B:MACH.0000015881.36452.6

Failing L, Horn G, & Higgins P. (2004). Using Expert Judgment and Stakeholder Values to Evaluate Adaptive Management Options. Ecology and Society, 9(1). http://www.ecologyandsociety.org/vol9/iss1/art13/

Rikers RM, & Paas F. (2005). Recent advances in expertise research. Applied Cognitive Psychology, 19(2), 145–149. 10.1002/acp.1108

Farrington-Darby T, & Wilson JR (2006). The nature of expertise: A review. Applied Ergonomics, 37(1), 17–32. 10.1016/j.apergo.2005.09.001 [PubMed: 16256934]

Howe J. (2006). The Rise of Crowdsourcing. Wired magazine, 14(6), 1–4. https://www.wired.com/2006/06/crowds/

Lawrence M, Goodwin P, O'Connor M, & Önkal D. (2006). Judgmental Forecasting: A Review of Progress over the Last 25 years. International Journal of Forecasting, 22(3), 493–518. 10.1016/j.ijforecast.2006.03.007

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, & Rakow T. (2006). Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley & Sons. 10.1002/0470033312

Petrovic D, Xie Y, & Burnham K. (2006). Fuzzy decision support system for demand forecasting with a learning mechanism. Fuzzy Sets and Systems, 157(12), 1713–1725. 10.1016/j.fss.2006.03.011

Willett P. (2006). The Porter stemming algorithm: then and now. Program: electronic library and information systems, 40(3), 219–223. 10.1108/00330330610681295

Berrocal VJ, Raftery AE, & Gneiting T. (2007). Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecasts. Monthly Weather Review, 135(4), 1386–1402. 10.1175/MWR3341.1

Clemen RT, & Winkler R. l. Aggregating Probability Distributions. In: In Advances in Decision Analysis: From Foundations to Applications. 2007. 10.1017/CBO9780511611308.010.

Dawid AP (2007). The geometry of proper scoring rules. Annals of the Institute of Statistical Mathematics, 59(1), 77–93. 10.1007/s10463-006-0099-8

Ericsson KA, & Ward P. (2007). Capturing the Naturally Occurring Superior Performance of Experts in the Laboratory: Toward a Science of Expert and Exceptional Performance. Current Directions in Psychological Science, 16(6), 346–350. 10.1111/j.1467-8721.2007.00533.x

Gneiting T, & Raftery AE (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102(477), 359–378. 10.1198/016214506000001437

Jin W, Lu Q, & Gan W. (2007). Research progress on the durability design and life prediction of concrete structures. Journal of Building Structures, 28(1), 7–13.

Leal J, Wordsworth S, Legood R, & Blair E. (2007). Eliciting expert opinion for economic models: an applied example. Value in Health, 10(3), 195–203. 10.1111/j.1524-4733.2007.00169.x [PubMed: 17532812]

Tartakovsky DM (2007). Probabilistic risk analysis in subsurface hydrology. Geophysical Research Letters, 34(5). 10.1029/2007GL029245

Van der Laan MJ, Polley EC, & Hubbard AE (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1). 10.2202/1544-6115.1309

Yousuf MI (2007). Using Experts' Opinions Through Delphi Technique. Practical Assessment, Research & Evaluation, 12(4), 1–8. 10.7275/rrph-t210

Clemen RT (2008). Comment on Cooke's classical method. Reliability Engineering & System Safety, 93(5), 760–765. 10.1016/j.ress.2008.02.003

Cranor LF (2008). A Framework for Reasoning about the Human in the Loop. In Proceedings of the 1st Conference on Usability, Psychology, and Security. https://www.usenix.org/legacy/event/upsec08/tech/fullpapers/cranor/cranor.pdf

Kabak Ö, & Ülengin F Aggregating forecasts to obtain fuzzy demands. In: In Computational Intelligence In Decision and Control. World Scientific, 2008, pp. 73–78. 10.1142/97898127994700012.

Lintott CJ, Schawinski K, Slosar A, Land K, Bamford S, Thomas D, Raddick MJ, Nichol RC, Szalay A, Andreescu D, et al. (2008). Galaxy zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 389(3), 1179–1189. 10.1111/j.1365-2966.2008.13689.x

Neves L, & Frangopol D. (2008). Life-cycle performance of structures: Combining expert judgment and results of inspection. In Proceedings of the 1st International Symposium on Life-Cycle Civil Engineering.

Wang X, Du C, & Cao Z. (2008). Probabilistic inversion techniques in quantitative risk assessment for power system load forecasting. In 2008 International Conference on Information and Automation. IEEE. 10.1109/ICINFA.2008.4608092

Adams R, White A, & Ceylan E. (2009). An Acceptability Predictor for Websites. In International Conference on Universal Access in Human-Computer Interaction. Springer. 10.1007/978-3-642-02713-066

Capistrán C, & Timmermann A. (2009). Forecast Combination With Entry and Exit of Experts. Journal of Business & Economic Statistics, 27(4), 428–440. 10.1198/jbes.2009.07211

Glahn B, Peroutka M, Wiedenfeld J, Wagner J, Zylstra G, Schuknecht B, & Jackson B. (2009). MOS Uncertainty Estimates in an Ensemble Framework. Monthly Weather Review, 137(1), 246–268. 10.1175/2008MWR2569.1

Kläs M, Nakao H, Elberzhager F, & Münch J. (2010). Support planning and controlling of early quality assurance by combining expert judgment and defect data—a case study. Empirical Software Engineering, 15(4), 423–454. 10.1007/s10664-009-9112-1

Kurowicka D, Bucura C, Cooke R, & Havelaar A. (2010). Probabilistic Inversion in Priority Setting of Emerging Zoonoses. Risk Analysis, 30(5), 715–723. 10.1111/j.1539-6924.2010.01378.x [PubMed: 20345579]

Polley EC, & Van Der Laan MJ (2010). Super Learner In Prediction.

Ranjan R, & Gneiting T. (2010). Combining probability forecasts. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(1), 71–91. 10.1111/j.1467-9868.2009.00726.x

Abernethy JD, & Frongillo RM A Collaborative Mechanism for Crowdsourcing Prediction Problems (Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, & Weinberger KQ, Eds.). In: In Advances in neural information processing systems 24 (Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, & Weinberger KQ, Eds.). Ed. by Shaw-eTaylor J, Zemel RS, Bartlett PL, Pereira F, & Weinberger KQ Curran Associates, Inc., 2011, pp. 2600–2608. http://papers.nips.cc/paper/4382-a-collaborative-mechanism-for-crowdsourcing-prediction-problems.pdf

Che D, Liu Q, Rasheed K, & Tao X Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics. In: In Software Tools and Algorithms for Biological Systems. Springer, 2011, pp. 191–199. 10.1007/978-1-4419-7046-619.

Franses PH (2011). Averaging Model Forecasts and Expert Forecasts: Why Does It Work? Interfaces, 41(2), 177–181. 10.1287/inte.1100.0554

French S. (2011). Aggregating expert judgement. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas, 105(1), 181–206. 10.1007/s13398-011-0018-6

Gneiting T, & Ranjan R. (2011). Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. Journal of Business & Economic Statistics, 29(3), 411–422. 10.1198/jbes.2010.08110

Kleiber W, Raftery AE, Baars J, Gneiting T, Mass CF, & Grimit E. (2011). Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model Averaging. Monthly Weather Review, 139(8), 2630–2649. 10.1175/2010MWR3511.1

Prill RJ, Saez-Rodriguez J, Alexopoulos LG, Sorger PK, & Stolovitzky G. (2011). Crowdsourcing Network Inference: The DREAM Predictive Signaling Network Challenge. Science Signaling, 4. 10.1126/scisignal.2002212

Wallis KF (2011). Combining forecasts–forty years later. Applied Financial Economics, 21(1–2), 33–41. 10.1080/09603107.2011.523179

Wang G, Hao J, Ma J, & Jiang H. (2011). A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications, 38(1), 223–230. 10.1016/j.eswa.2010.06.048

Brito M, Griffiths G, Ferguson J, Hopkin D, Mills R, Pederson R, & MacNeil E. (2012). A Behavioral Probabilistic Risk Assessment Framework for Managing Autonomous Underwater Vehicle Deployments. Journal of Atmospheric and Oceanic Technology, 29(11), 1689–1703. 10.1175/JTECH-D-12-00005.1

Cabello E, Conde C, Diego I, Moguerza J, & Redchuk A. (2012). Combination and Selection of Traffic Safety Expert Judgments for the Prevention of Driving Risks. Sensors, 12(11), 14711–14729. 10.3390/s121114711 [PubMed: 23202184]

Jolliffe IT, & Stephenson DB (2012). Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley & Sons. 10.1002/9781119960003

Li W, Liu Y. j., & Yang Z. (2012). Preliminary Strategic Environmental Assessment of the Great Western Development Strategy: Safeguarding Ecological Security for a New Western China. Environmental Management, 49(2), 483–501. 10.1007/s00267-011-9794-1 [PubMed: 22190169]

Martin TG, Burgman MA, Fidler F, Kuhnert PM, Low-Choy S, McBride M, & Mengersen K. (2012). Eliciting Expert Knowledge in Conservation Science. Conservation Biology, 26(1), 29–38. 10.1111/j.1523-1739.2011.01806.x. [PubMed: 22280323]

Sya – Syarif I, Zaluska E, Prugel-Bennett A, & Wills G. (2012). Application of Bagging, Boosting and Stacking to Intrusion Detection. In International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer. 10.1007/978-3-642-31537-4_46

Ung – Ungar LH, Mellers BA, Satopaa V, Tetlock P, & Baron J. (2012). The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions. In AAAI Fall Symposium: Machine Aggregation of Human Judgment. https://www.aaai.org/ocs/index.php/FSS/FSS12/paper/view/5570/5871

Brabham DC (2013). Crowdsourcing. MIT Press.

Genre V, Kenny G, Meyler A, & Timmermann A. (2013). Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting, 29(1), 108–121. 10.1016/j.ijforecast.2012.06.004

Gneiting T, Ranjan R et al. (2013). Combining predictive distributions. Electronic Journal of Statistics, 7, 1747–1782. 10.1214/13-EJS823

Hora SC, Fransen BR, Hawkins N, & Susel I. (2013). Median Aggregation of Distribution Functions. Decision Analysis, 10(4), 279–291. 10.1287/deca.2013.0282

Sarin RK (2013). Median Aggregation, Scoring Rules, Expert Forecasts, Choices with Binary Attributes, Portfolio with Dependent Projects, and Information Security. Decision Analysis, 10(4), 277–278.

Seifert M, & Hadida AL (2013). On the relative importance of linear model and human judge(s) in combined forecasting. Organizational Behavior and Human Decision Processes, 120(1), 24–36. 10.1016/j.obhdp.2012.08.003

Shin J, Coh B-Y, & Lee C. (2013). Robust future-oriented technology portfolios: Black–Litterman approach. R&D Management, 43(5), 409–419. 10.1111/radm.12022

Song H, Gao BZ, & Lin VS (2013). Combining statistical and judgmental forecasts via a web-based tourism demand forecasting system. International Journal of Forecasting, 29(2), 295–310. 10.1016/j.ijforecast.2011.12.003

Baron J, Mellers BA, Tetlock PE, Stone E, & Ungar LH (2014). Two Reasons to Make Aggregated Probability Forecasts More Extreme. Decision Analysis, 11(2), 133–145. 10.1287/deca.2014.0293

Cooke RM Validating Expert Judgment with the Classical Model. In: In Experts and consensus in social science. Springer, 2014, pp. 191–212. 10.1007/978-3-319-08551-7_10.

Cooke RM, Wittmann ME, Lodge DM, Rothlisberger JD, Rutherford ES, Zhang H, & Mason DM (2014a). Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie. Integrated Environmental Assessment and Management, 10(4), 522–528. 10.1002/ieam.1559 [PubMed: 25044130]

Crowdsourcing the Future: Predictions Made with a Social Network. (2014). In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 10.1145/2556288.2556967

de Groot AD (2014). Thought and Choice in Chess (Vol. 4). Walter de Gruyter GmbH & Co KG. https://www.jstor.org/stable/j.ctt46n0r2

Graefe A, Armstrong JS, Jones RJ, & Cuzán AG (2014a). Accuracy of Combined Forecasts for the 2012 Presidential Election: The PollyVote. PS: Political Science & Politics, 47(2), 427–431. 10.1017/S1049096514000341

Graefe A, Armstrong JS, Jones RJ Jr, & Cuzán AG (2014b). Combining forecasts: An application to elections. International Journal of Forecasting, 30(1), 43–54. 10.1016/j.ijforecast.2013.02.005

Mantyka-Pringle CS, Martin TG, Moffatt DB, Linke S, & Rhodes JR (2014). Understanding and predicting the combined effects of climate change and land-use change on freshwater

macroinvertebrates and fish. Journal of Applied Ecology, 51(3), 572–581. 10.1111/1365-2664.12236

Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Scott SE, Moore D, Atanasov P, Swift SA, et al. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. Psychological Science, 25(5), 1106–1115. 10.1177/0956797614524255 [PubMed: 24659192]

Morgan MG (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. Proceedings of the National academy of Sciences, 111(20), 7176–7184. 10.1073/pnas.1319946111

Satopää VA, Jensen ST, Mellers BA, Tetlock PE, Ungar LH, et al. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. The Annals of Applied Statistics, 8(2), 1256–1280. 10.1214/14-AOAS739

Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, & Taha K. (2015). Efficient Machine Learning for Big Data: A Review. Big Data Research, 2(3), 87–93. 10.1016/j.bdr.2015.04.001

Baldwin P. (2015). Weighting Components of a Composite Score using Näıve Expert Judgments About Their Relative Importance. Applied Psychological Measurement, 39(7), 539–550. 10.1177/0146621615584703 [PubMed: 29881025]

Cooke RM (2015). The Aggregation of Expert Judgment: Do Good Things Come to Those Who Weight? Risk Analysis, 35(1), 12–15. 10.1111/risa.12353 [PubMed: 25648183]

Graefe A. (2015). Accuracy gains of adding vote expectation surveys to a combined forecast of US presidential election outcomes. Research & Politics, 2(1), 1–5. 10.1177/2053168015570416

Hora SC, & Karde s, E. (2015). Calibration, sharpness and the weighting of experts in a linear opinion pool. Annals of Operations Research, 229(1), 429–450. 10.1007/s10479-015-1846-0

Yu F, Seff A, Zhang Y, Song S, Funkhouser T, & Xiao J. (2015). LSUN: Construction of a Large-scale Image Dataset Using Deep Learning with Humans in the Loop. ArXiv, abs/1506.03365.

Brito M, & Griffiths G. (2016). A Bayesian approach for predicting risk of autonomous underwater vehicle loss during their missions. Reliability Engineering & System Safety, 146, 55–67. 10.1016/j.ress.2015.10.004

Cai M, Lin Y, Han B, Liu C, & Zhang W. (2016). On a Simple and Efficient Approach to Probability Distribution Function Aggregation. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 47(9), 2444–2453. 10.1109/TSMC.2016.2531647

Gu W, Saaty TL, & Whitaker R. (2016). Expert System for Ice Hockey Game Prediction: Data Mining with Human Judgment. International Journal of Information Technology & Decision Making, 15(04), 763–789. 10.1142/S0219622016400022

Hathout M, Vuillet M, Peyras L, Carvajal C, & Diab Y. (2016). Uncertainty and Expert Assessment for Supporting Evaluation of Levees Safety. In 3rd European Conference on Flood Risk Management (FLOODrisk 2016). 10.1051/e3sconf/20160703019

Holzinger A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Informatics, 3(2), 119–131. 10.1007/s40708-016-0042-6 [PubMed: 27747607]

Huang A, Qiao H, Wang S, & Liu J. (2016). Improving Forecasting Performance by Exploiting Expert Knowledge: Evidence from Guangzhou Port. International Journal of Information Technology & Decision Making, 15(02), 387–401. 10.1142/S0219622016500085

Kune R, Konugurthi PK, Agarwal A, Chillarige RR, & Buyya R. (2016). The Anatomy of Big Data Computing. Softw. Pract. Exper, 46(1), 79–105. 10.1002/spe.2374

Moran KR, Fairchild G, Generous N, Hickmann KS, Osthus D, Priedhorsky R, Hyman JM, & Valle SYD (2016). Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast. The Journal of infectious diseases, 214, S404–S408. 10.1093/infdis/jiw375 [PubMed: 28830111]

Wang C, Chen M-H, Schifano E, Wu J, & Yan J. (2016). Statistical methods and computing for big data. Statistics and its interface, 9(4), 399. 10.4310/SII.2016.v9.n4.a1 [PubMed: 27695593]

Alvarado-Valencia J, Barrero LH, Önkal D, & Dennerlein JT (2017). Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. International Journal of Forecasting, 33(1), 298–313. 10.1016/j.ijforecast.2015.12.010

Baecke P, De Baets S, & Vanderheyden K. (2017). Investigating the added value of integrating human judgement into statistical demand forecasting systems. International Journal of Production Economics, 191, 85–96. 10.1016/j.ijpe.2017.05.016

Bolger D, & Houlding B. (2017). Deriving the probability of a linear opinion pooling method being superior to a set of alternatives. Reliability Engineering & System Safety, 158, 41–49.

Li G. (2017). Human-in-the-Loop Data Integration. Proc. VLDB Endow, 10(12), 2006–2017. 10.14778/3137765.3137833

Morales-Nápoles O, Paprotny D, Worm D, Abspoel-Bukman L, & Courage W. (2017). Characterization of Precipitation through Copulas and Expert Judgement for Risk Assessment of Infrastructure. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 3(4), 04017012. 10.1061/AJRUA6.0000914

Baran S, & Lerch S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. International Journal of Forecasting, 34(3), 477–496. 10.1016/j.ijforecast.2018.01.005

Bassetti F, Casarin R, & Ravazzolo F. (2018). Bayesian Nonparametric Calibration and Combination of Predictive Distributions. Journal of the American Statistical Association, 113(522), 675–685. 10.1080/01621459.2016.1273117

Graefe A. (2018). Predicting elections: Experts, polls, and fundamentals. Judgment and Decision Making, 13(4), 334.

Hanea AM, McBride MF, Burgman MA, & Wintle BC (2018). The Value of Performance Weights and Discussion in Aggregated Expert Judgments. Risk Analysis, 38(9), 1781–1794. 10.1111/risa.12992 [PubMed: 29665625]

Johnson FA, Alhainen M, Fox AD, Madsen J, & Guillemain M. (2018). Making do with less: Must sparse data preclude informed harvest strategies for European waterbirds? Ecological Applications, 28(2), 427–441. 10.1002/eap.1659 [PubMed: 29205644]

Wang L, & Zhang X. (2018). Bayesian analytics for estimating risk probability in PPP waste-to-energy projects. Journal of Management in Engineering, 34(6), 04018047.

Garratt A, Henckel T, & Vahey SP (2019). Empirically-transformed linear opinion pools.

Jana DK, Pramanik S, Sahoo P, & Mukherjee A. (2019). Interval type-2 fuzzy logic and its application to occupational safety risk performance in industries. Soft Computing, 23(2), 557–567. 10.1007/s00500-017-2860-8

Winkler RL, Grushka-Cockayne Y, Lichtendahl KC Jr, & Jose VRR (2019). Probability Forecasts and Their Combination: A Research Perspective. Decision Analysis, 16(4), 239–260. 10.1287/deca.2019.0391

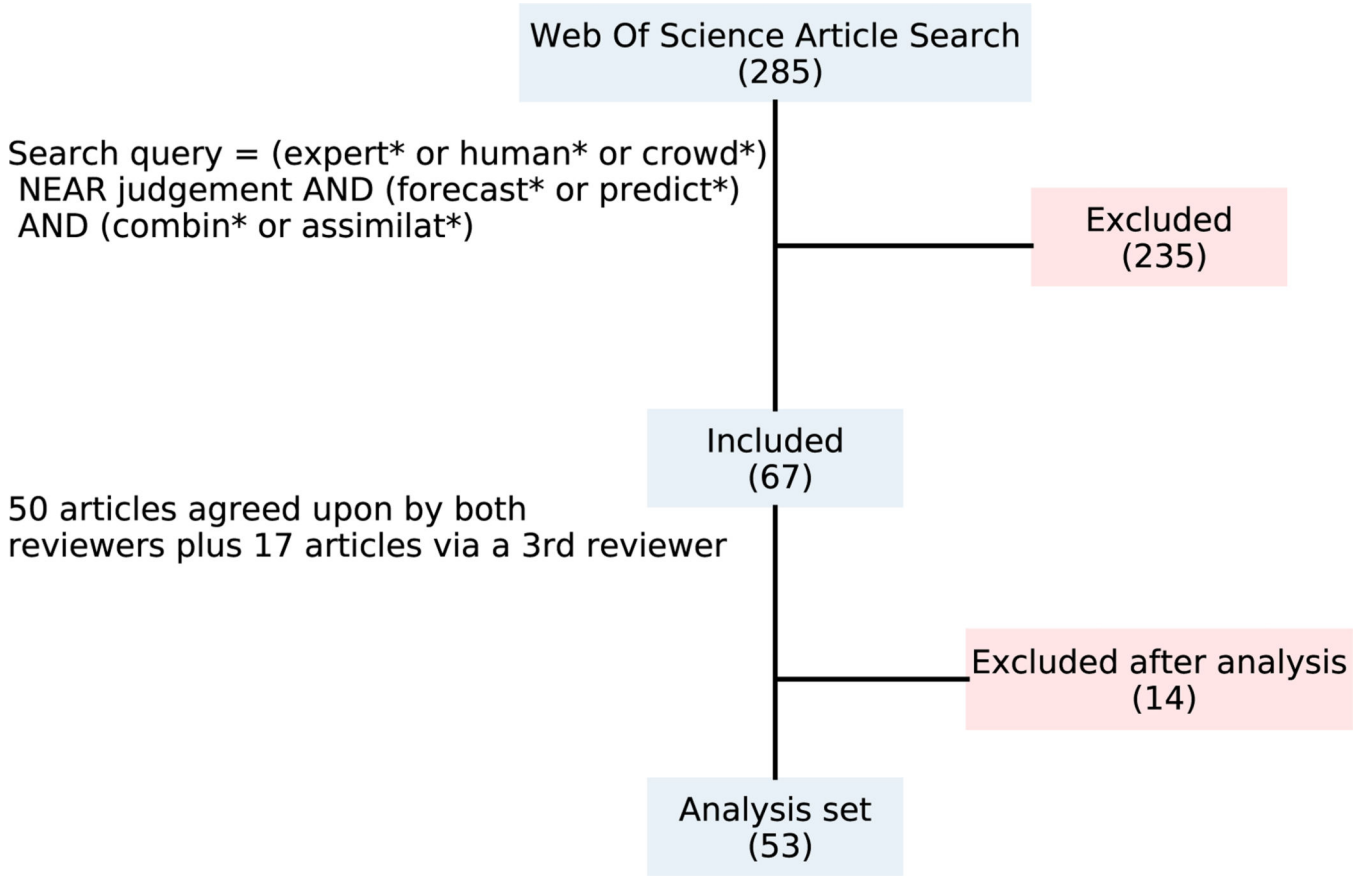


Figure 1:

A consort diagram that describes the path from collected to analysis-set article. The search term used to collect the initial set of articles is reported and all intermediate steps between initial and analysis-set articles.

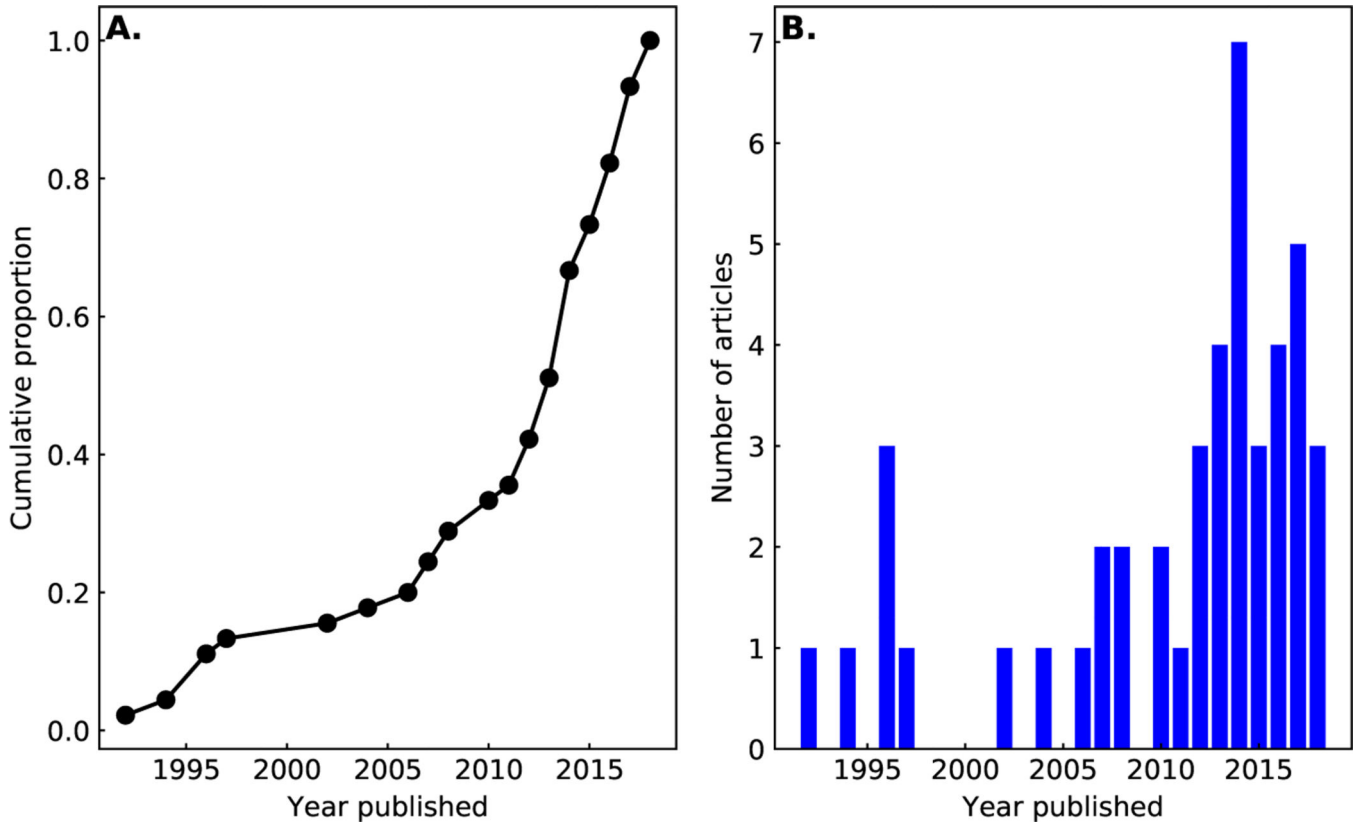


Figure 2:

The cumulative proportion (A.) and individual number (B.) of articles published per year. The earliest analysis-set article was published in 1992 and most recent in 2018. A sharp increase in publication occurred at or near 2010.

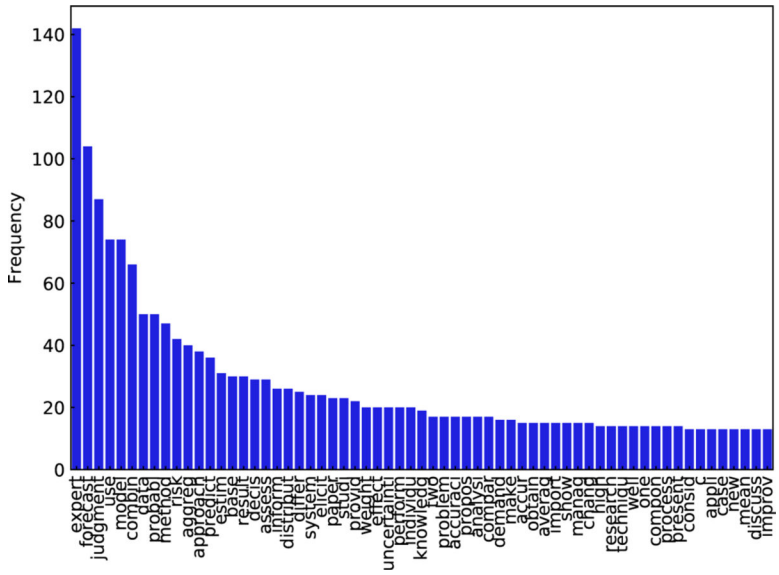


Figure 3:

The top 5 percent most frequent words used in all analysis-set abstracts. Expert, forecast, and judgment are the most frequent and likely related to the search words used to collect these articles.

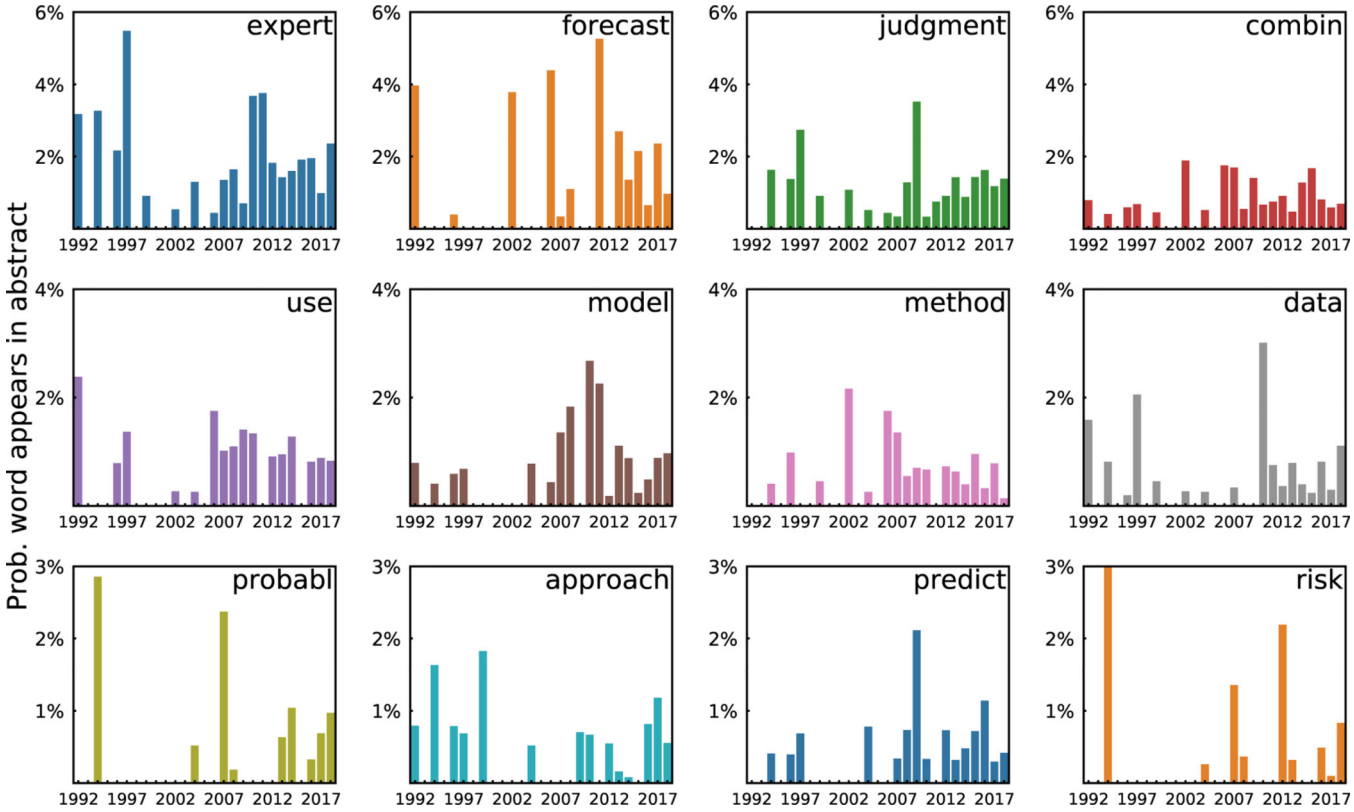


Figure 4:

The annual proportion of the top 12 most prevalent word stems among all abstract text (Note: the words probability and probabilities were stemmed to probabl). For each year, word *w* frequency was divided by the frequency of all words present in all abstracts.

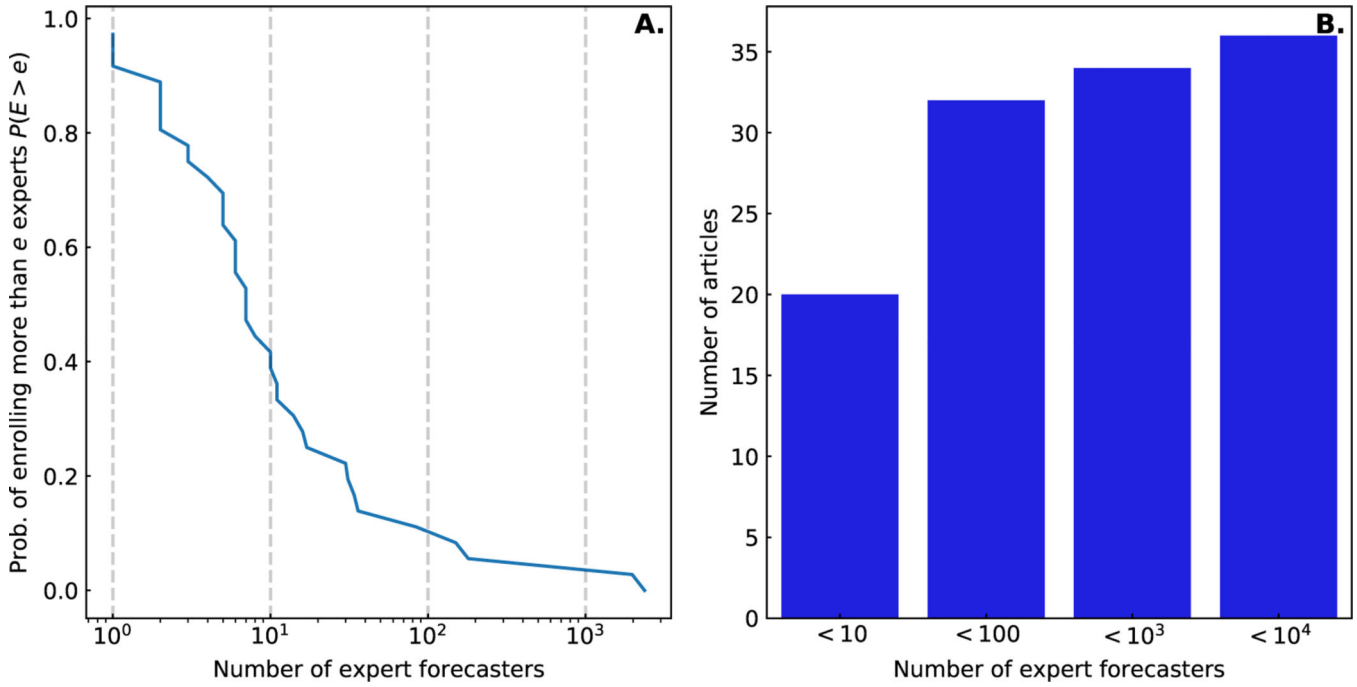


Figure 5:

The complimentary cumulative distribution (CCDF) of the number of experts elicited per article (A.). The proportion of articles enrolling less than 10, less than 100, less than 10^3, and less than 10^4 expert forecasters (B.). The small number of articles enrolling more than 10^3 were simulation studies.

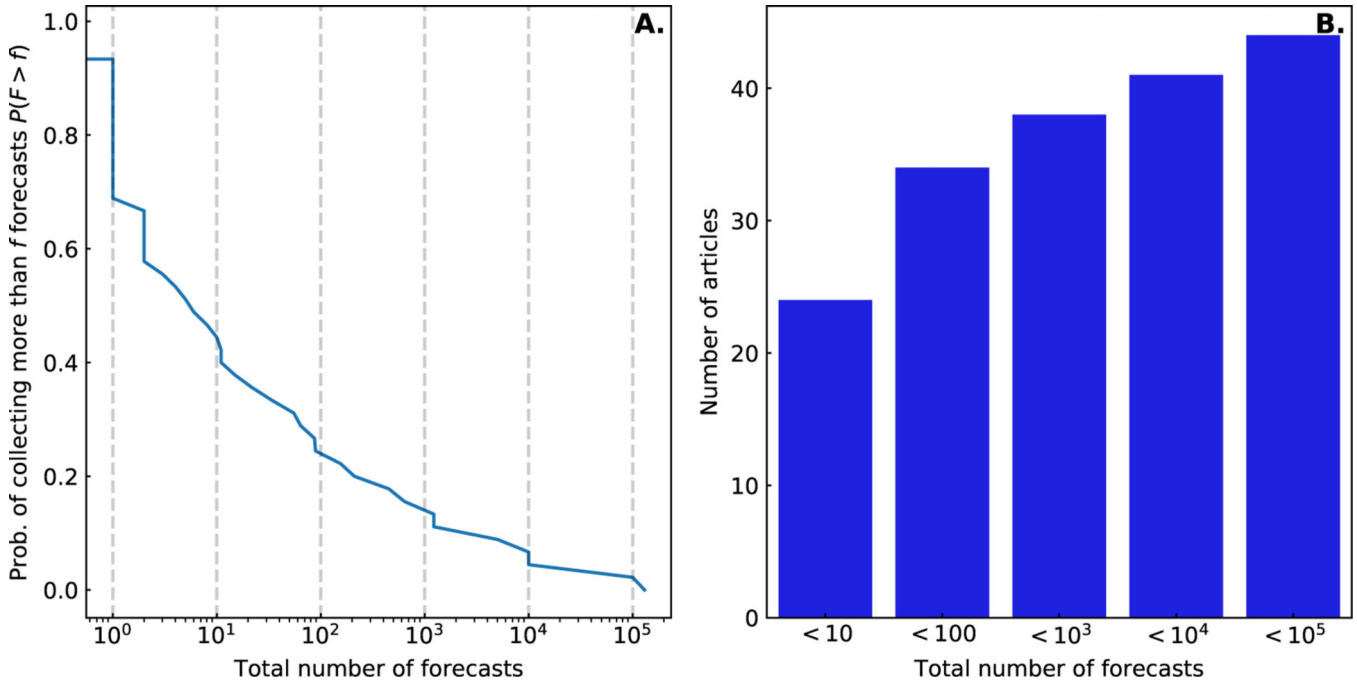


Figure 6:

Complimentary cumulative distribution of the total number of forecasts made per article (A.), and the proportion of articles eliciting less than 10, 100, 10^3, 10^4, and 10^5 forecasts. Articles collecting more than 10^4 forecasts were simulations.

Table 1:

Terminology from analysis-set articles was collected and grouped by meaning. For each definition, the preferred terms is placed on top of all related terms. Definitions and preferred terminology were agreed upon by all coauthors.

Related terms	Definition	Citations
Forecasting support system Adaptive management	A framework for transforming data and forecasts into decisions.	(Fai, 2004; Son, 2013; Alv, 2017; Bae, 2017; Joh2018, 2018)
(Probabilistic) Safety Assessment (Probabilistic) Risk Assessment	A framework for investigating the safety of a system	Zio (1996), Zio (1997), Bor (2004), Cle (2007), Tar (2007), Kla (2010), Kur (2010), Bri (2012), Coo (2014a), Hat (2016), Mor (2017), Han (2018), Wan (2018), Jan (2019)
Information set Knowledge-base	Data available to an expert, group of experts, or statistical model used for forecasting.	(Abr, 1996; Mak, 1996; Bor, 2004; Gra, 2014a; Bri, 2016; Alv, 2017)
Ill-structured tasks	When changes to an environment impact the probabilistic links between cues an expert receives and their effect (how these cues should should be interpreted).	(Sei, 2013; Hua, 2016)
Behavioral aggregation Behavioral combination Structured elicitation	The support of expert discussion until they arrive at an agreed upon consensus distribution.	(Cle, 2007; Bri, 2012; Han, 2018)
Mathematical combination Mechanical integration	The use of mathematical techniques to transform independent expert judgments into a single consensus distribution.	(Pet, 2006; Cle, 2007)
Judgmental adjustment Voluntary integration	Allowing experts to observe statistical forecasts, and provide their forecast as an adjustment to a present statistical forecast.	(Son, 2013; Hua, 2016; Alv, 2017; Bae, 2017)
Integrative judgment Knowledge-aggregation	Forecasts from experts are incorporated into a forecasting model as a predictive variable.	(Mak, 1996; Bae, 2017)
Equal weighted 50–50 Weighting Unweighted	Assigning equal weights to all experts in a combination method.	(Sar, 2013; Coo, 2014a; Gra, 2015; Alv, 2017; Han, 2018)
Nominal weights	Weights obtained by assessing experts performance on a set of calibration questions, or on observed data.	(Bal, 2015)
Cooke's method Classical model	Combining expert opinion via a linear pool where weights depend on expert's answers to calibration questions with a known answer.	(Zio, 1996; Cle, 2007; Bri, 2012; Sar, 2013; Coo, 2014a; Hor, 2015; Hat, 2016; Bol, 2017; Mor, 2017; Han, 2018)
Mixed estimation Theil-Goldeberger mixed estimation	A method for combining expert and statistical forecasts, stacking statistical and expert point predictions into a single vector and fitting a linear regression model.	(Alh, 1992; Shi, 2013)
Laplacian principle of indifference Principle of indifference	In the context of expert combination, having no evidence related to expert forecasting performance, models should weight experts equally.	(Bol, 2017)
Brunswik lens model	A framework for relating a set of criteria (or indicators), expert's judgment, and the "correct" judgment.	(Fra, 2011; Sei, 2013)

Table 2:

A prespecified list of questions was asked when reviewing all in-scope articles (See supplemental material for a table of categories and the corresponding citations, and a separate spreadsheet of the individual analysis set articles and how they were categorized). Frequencies and percentages were recorded for all binary questions. Questions a reviewer could not answer are defined as missing, causing some questions to have fewer than 53 total answers. Answers to questions are on the article level and categories are not mutually exclusive. For example, an article could explore both a Frequentist and Bayesian model.

Question	Yes N (%)	Total answers N
The primary target was		
categorical	18 (34)	53
continuous	36 (68)	53
from a time series	22 (42)	53
A novel method/model was developed	25 (47)	53
The authors implemented a		
Bayesian technique	13 (25)	52
Frequentist technique	26 (49)	53
The model was		
nonparametric	13 (25)	52
parametric	38 (73)	52
The model combined		
point estimates	30 (56)	53
probabilistic distributions	19 (36)	53
Experts depended on data that could be updated, revised, or rapidly change	21 (41)	51

Table 3:

Metrics that in-scope articles used to evaluate both point and density forecasts. A preferred term is listed (metric column), given an abbreviation and related names reported. Whether the evaluative metric operates on a continuous or binary variable is stated and the computational formula presented.Question

Metric	Abbreviation	Other names	Binary or Continuous target	Formula
Absolute Loss	AS	-	Categorical	$\lvert P(F_i) - O_i \rvert$
Quadratic Loss	QS	-	Categorical	$[P(F_i) - O_i]^2$
Prediction Accuracy	PA	-	Categorical	$N^{-1} \sum_{i=1}^{N} \sum 1(F_i = O_i)$
Brier Score	BS	-	Categorical	$N^{-1} \sum_{i=1}^{N} [P(F_i) - O_i]^2$
Mean Error	ME	-	Continuous	$N^{-1} \sum_{i=1}^{N} (F_i - O_i)$
Mean Absolute Error	MAE	Mean Absolute Deviation (MAD)	Continuous	$N^{-1} \sum_{i=1}^{N} \lvert (O_i - F_i) \rvert$
Mean Absolute Percent Error	MAPE	Mean Percent Error (MPE) Average percentage error (APE)	Continuous	$N^{-1} \sum_{i=1}^{N} \lvert (F_i / O_i - 1) \rvert$
Mean Squared Error	MSE	-	Continuous	$N^{-1} \sum_{i=1}^{N} (F_i - O_i)^2$
Root mean squared error	RMSE	Root mean squared prediction error (RMSPE)	Continuous	$\sqrt{N^{-1} \sum_{i=1}^{N} (F_i - O_i)^2}$
Proportion higher density	PHD		Continuous	$N^{-1} \sum_{i=1}^{N} \mathbb{1}\{P[F(x_i)] > P[G(x_i)]\}$
95% Coverage probability	CP	-	Continuous	$N^{1} \sum_{i=1}^{N} \mathbb{1}(F_{2.5} < O_i < F_{97.5})$
Judgemental Adjustment	JA	-	Continuous	$(F_i - G_i)/G_i$
Forecast Improvement	FCIMP	-	Continuous	$(\lvert O_i - F_i \rvert - \lvert O_i - G_i \rvert)/O_i$

Table 4:

List of close-ended questions asked of each full-text article. Questions focus on the forecasting target, model, analysis data, and experimental design.

Question	Possible answers
Forecasting target	
Identify the primary predictive target?	predictive target
The primary target was categorical	Y/N
The primary target was continuous	Y/N
The primary target was from a time series	Y/N
Experts were given data related to the forecasting target?	Y/N
Terminology	
List terms specific to aggregating crowdsourced data and quoted definition	term,def;term,def
Model	
What models were used in forecasting?	model1, model2, …, model*n*
Please list covariates included in any model	cov1, cov2, …, cov*n*
A novel model/method was developed	Y/N
Did the authors implement a Bayesian technique?	Y/N
Did the authors use a Frequentist technique?	Y/N
Did the model account for correlation among experts?	Y/N
The model combined point estimates	Y/N
The model combined probabilistic distributions	Y/N
The model was parametric	Y/N
The model was nonparametric	Y/N
Analysis data	
Experts depended on data that could be updated, revised, or rapidly change?	Y/N
Experimental design	
A comparative experiment was conducted	Y/N
How many expert forecasters were included?	integer
How many total forecasts were made?	integer
What evaluation metrics were used?	metric1, metric2, …, metric*n*