Use and Communication of Probabilistic Forecasts

Adrian E. Raftery*

Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA

Received 20 August 2014; revised 11 March 2015; accepted 23 September 2015
DOI:10.1002/sam.11302
Published online 23 February 2016 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Probabilistic forecasts are becoming more and more available. How should they be used and communicated? What are the obstacles to their use in practice? We review experience with five problems where probabilistic forecasting played an important role. This leads us to identify five types of potential users: low stakes users, who do not need probabilistic forecasts; general assessors, who need an overall idea of the uncertainty in the forecast; change assessors, who need to know if a change is out of line with expectations; risk avoiders, who wish to limit the risk of an adverse outcome; and decision theorists, who quantity their loss function and perform the decision-theoretic calculations. This suggests that it is important to interact with users and consider their goals. Cognitive research tells us that calibration is important for trust in probability forecasts and that it is important to match the verbal expression with the task. The cognitive load should be minimized, reducing the probabilistic forecast to a single percentile if appropriate. Probabilities of adverse events and percentiles of the predictive distribution of quantities of interest often seem to be the best way to summarize probabilistic forecasts. Formal decision theory has an important role but in a limited range of applications. © 2016 Wiley Periodicals, Inc. Statistical Analysis and Data Mining: The ASA Data Science Journal 9: 397–410, 2016

Keywords: calibration; cognitive research; decision theory; risk; uncertainty

1. INTRODUCTION

Much progress has been made over the past few decades in the development of methods for probabilistic forecasting, and probabilistic forecasts are now routinely used in several disciplines. These include finance, where risk management is a common use of probabilistic forecasts in financial asset management [1], and trading decisions that are made based on predictive distributions of assets, often using automated computer-trading programs. In marketing, predictive distributions of future sales and inventory are commonly made using statistical models, such as ARIMA models [2], and used as the basis for stocking and other decisions. In medicine, probabilistic forecasts of binary outcomes (e.g. the probability of surviving more than 5 years) are widely used [3]. However, in other areas, the development of probabilistic forecasting methods is more recent, and use of these methods in practice is at an earlier stage. How should probabilistic forecasts be used and communicated? What are the obstacles to their use in practice? Can these be overcome? Can they be presented in ways that make them more useful to possibly sceptical users?

Communicating uncertainty is inherently a challenging problem. Kahneman [4] identified people’s resistance to uncertainty as

‘a puzzling limitation of our mind: our excessive confidence in what we believe we know, and our apparent inability to acknowledge the full extent of our ignorance and the uncertainty of the world we live in. We are prone to overestimate how much we understand the world and to underestimate the role of chance in events. Overconfidence is fed by the illusory certainty of hindsight.’

There are various possible explanations for this. One is that people’s cognitive bandwidth is limited, and uncertainty information increases cognitive load. For example, adding a range to a point or ‘best’ forecast triples the cognitive load.

A more fundamental explanation is again proposed by Kahneman [4]:

‘An unbiased appreciation of uncertainty is a cornerstone of rationality, but it is not what people and organizations want. Extreme uncertainty is paralyzing under dangerous
circumstances, and the admission that one is merely guessing is especially unacceptable when the stakes are high. Acting on pretended knowledge is often the preferred solution.’

A related possible explanation arises when forecasters and decision makers are different people, as is often the case in policy-making contexts. Then the decision maker may wish to push the responsibility for the decision onto the forecaster, and when the forecaster provides a range or a probabilistic forecast, this is harder to do than when a single number is given. If things go wrong, it is easier to blame the forecaster who gave an incorrect forecast.

In this article, we will describe experience with probabilistic forecasting in five different contexts and try to draw some conclusions. These will lead us to identify five types of potential users of probabilistic forecasts: low stakes users, general assessors, change assessors, risk avoiders, and decision theorists. Each may have different needs.

Some suggestions are that it is important to interact with users and consider their goals; the ways of doing this include meetings and web surveys. This is a cognitive problem as well as a statistical one. The cognitive research tells us that calibration is important for trust in probability forecasts and that it is important to match the verbal expression with the goal. The cognitive load should be minimized to the extent possible, even reducing the probabilistic forecast to a single number if appropriate. Probabilities of adverse events and percentiles of the predictive distribution of quantities of interest often seem to be the best way to summarize probabilistic forecasts.

Formal decision theory has an important role in a limited range of applications, particularly when users are aware of their loss functions and when there is agreement on the loss function to use. This arises most clearly when costs and losses are measured in monetary terms. Decision theory is also useful in research on the use of probabilistic forecasts to analyze different possible decision rules.

In the following sections, we will describe experience with five problems where probabilistic forecasting played an important role: setting aboriginal whaling quotas, probabilistic weather forecasting, projecting the worldwide HIV/AIDS epidemic, probabilistic population projections for the United Nations, and deciding on the number of funded graduate students to admit. We will then discuss what conclusions can be drawn from this experience.

2. SETTING ABORIGINAL BOWHEAD WHALING QUOTAS

For centuries, the Western Arctic stock of bowhead whales, *Balaena mysticetus*, off the coasts of Alaska and Siberia, has been the object of small-scale subsistence hunting by the Inuit, or Eskimo, peoples of the area, for whom it is vital both nutritionally and culturally; see Fig. 1. The stock was severely depleted by commercial whaling by Yankee and European whalers in the late 19th and early 20th centuries. Commercial whaling of bowhead whales (although not other whale species) effectively ended around 1915, and the species was first legally protected from commercial whaling from 1931 by the League of Nations Convention and then by the International Whaling Commission (IWC), founded in 1946.

This left the question of whether and how to regulate aboriginal whaling by the Inuit. It was generally recognized that it would be unfair to penalize the Inuit people for a problem that was not of their making as they had been whaling sustainably for centuries and to ban aboriginal whaling would damage their livelihood and culture. This led to a tension between two conflicting goals: on one hand, to protect the bowhead whale stock and allow it to

![Fig. 1](a) Bowhead whale, *Balaena mysticetus*. (b) Community celebration after Inuit bowhead whale hunt. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Statistical Analysis and Data Mining: The ASA Data Science Journal DOI:10.1002/sam
recover to its pre-commercial whaling levels and on the other hand, to satisfy the subsistence and cultural needs of the Inuit people.

The IWC’s solution was to allow continued limited aboriginal subsistence whaling but with a quota to be set at a level low enough to allow the stock to recover. A key quantity for setting the quota in a given future year was the replacement yield (RY) in that year, namely the greatest number of whales that could be taken without the population decreasing. This is unknown and is subject to considerable uncertainty. Because it is important that the quota not exceed this unknown value, a conservative value or ‘lower bound’ is sought, which should take account of all non-negligible sources of uncertainty.

The future RY had traditionally been forecast using a deterministic population dynamics model in which births are added and natural deaths and kills are subtracted. This requires age-specific fertility and natural mortality rates as inputs and outputs the population for each future year, broken down by age and sex. The inputs are unknown and subject to considerable uncertainty.

Until 1991, the lower bound was set by doing several runs of the model with different scenarios or variants, consisting of combinations of ‘central’, ‘high’, and ‘low’ values of the inputs. The range of values of the RY output was then treated as a rough prediction interval. In 1991, however, the IWC Scientific Committee rejected this approach as statistically invalid, noting that it had no probabilistic interpretation and could lead to, for example, decisions that were riskier than they seemed. They recommended that statistically principled methods be developed.

In response to this, we developed the Bayesian melding method to make inferences about outputs from the population dynamics model, taking account of all known substantial uncertainties about the inputs [5–7]. This yielded a posterior predictive distribution of RY for future years; an example is shown in Fig. 2.

Once this was available, the IWC Scientific Committee recommended that the fifth percentile of this distribution be taken as a precautionary lower bound on the RY and thus as an upper bound on the allowed hunting quota. The recommendation was accepted by the International Whaling Commission itself (consisting mostly of politicians and senior civil servants, such as Fisheries ministers and officials from the then 40 IWC member countries). Taking account of this lower bound, the desire to allow a margin for future recovery of the stock, and the Inuit subsistence and cultural needs, the Commission set a quota slightly below the lower bound.

This approach was used successfully for the following 10 years. Over that period, the bowhead whale stock prospered, increasing substantially, while the Inuit whale hunt continued and the related Inuit culture was preserved. The basic statistical ideas have since been used for other wildlife management problems [8–10].

The fifth percentile of the posterior predictive distribution effectively became the ‘best forecast’ for this problem. It was necessary to compute the full posterior distribution in order to calculate it. However, once the fifth percentile had been calculated and agreed as valid by the IWC Scientific Committee, most of the policy-making attention focused on it, and the rest of the distribution (including measures of its center such as the median or mode) was largely ignored. Thus, the cognitive load was no larger than for a single ‘best’ forecast.

The responsibility for making a single best forecast had been met by the forecasters (in this case, the IWC Scientific Committee), only in this case, it was a lower bound rather than a predictive median or mode or a deterministic point forecast. Probabilistic forecasts were important in this application because the first priority was to limit the risk of an adverse outcome, namely a decrease of the whale stock.

Note that formal decision theory was not used in this problem. The IWC Scientific Committee has considered using formal decision theory for such problems but, in general, has not done so because they considered that reaching an agreement on the relative costs involved was not feasible. For example, what is the ratio of the cost to the stock of killing a whale to the benefit to the Inuit community? Consensus on the answers to questions like these would be hard to achieve [11–13].

Instead, the preferred approach was to set the quota so that the risk of the stock decreasing as a result would be no more than 5%. This eventually commanded broad agreement, even in a body where debates have often been contentious because of the environmental sensitivities associated with whaling. Value at risk in finance is another
3. PROBABILISTIC WEATHER FORECASTING

3.1. Methods and Probcast Website

Probabilistic weather forecasts consist of predictive probability distributions of future weather quantities. In particular, they yield probabilities of future adverse weather events, such as freezing temperatures, high rainfall, or wind storms. Since 1992, probabilistic weather forecasts have been produced by major weather forecasting agencies using ensembles of deterministic numerical weather predictions [14]. However, these have been little used as the basis for public forecasts because they are typically poorly calibrated.

In response to this situation, methods for postprocessing ensembles to produce calibrated probabilistic weather forecasts have been developed based on statistical methods, including ensemble Bayesian model averaging [15] and ensemble model output statistics (EMOS) [16]. In addition to temperature, methods were developed for precipitation [17], wind speeds [18], wind directions [19], wind vectors [20], and visibility [21].

Based on these forecasts, we set up a prototype real-time probabilistic weather forecasting website for the general public in the North American Pacific Northwest at www.probcast.com [22]; see Fig. 3. Its design and content were based on extensive cognitive experiments and ethnographic studies of forecasters and end users [23–27].

The website contains three kinds of information. First are percentiles of decision-critical weather quantities, namely temperature and the amount of precipitation. The 10th, 50th,
3.2. Cognitive Findings

An important part of the probabilistic weather forecasting project consisted of carrying out cognitive experiments to determine how best to convey the uncertainty information. There is a long tradition in psychology of assessing people’s understanding of probability and uncertainty by offering them simple gambles [28], but there has been less research on how best to communicate uncertainty about complex real-life outcomes.

Calibration of the probability forecast (e.g. 80% prediction intervals contain the truth 80% of the time on average) is an important requirement for probabilistic forecasts [29]. One series of experiments showed that providing calibrated probability forecasts improves weather-related decision making and increases trust in the forecast [30–32]. This is good news for probabilistic forecasting, showing that ordinary people can understand and use probabilities to improve their decision making.

Joslyn and Savelli [27] found that users of standard (deterministic) weather forecasts have well-formed uncertainty expectations, suggesting that they are prepared to understand explicit uncertainty forecasts for a wide range of parameters. Explicit uncertainty estimates may be necessary to overcome some of the anticipated forecast biases that may be affecting the usefulness of existing weather forecasts. Despite the fact that these bias expectations are largely unjustified, they could lead to the adjustment of forecasts that could in turn have serious negative consequences for users, especially with respect to extreme weather warnings.

Joslyn et al. [24] reported on a series of experiments to investigate the effects of various aspects of how probability forecasts are presented: framing (positive versus negative), format (frequency versus probability), probability (low vs. high), and compatibility between the presentation and the decision task. They showed that the key factor is the match between the verbal expression and the task goal. The other three factors (framing, format, and probability) made a much smaller difference. In one experiment, people were asked to decide whether or not to post a wind advisory for winds above 20 knots and were given probability information. When people were told the probability that wind speed would be above 20 knots, they made few errors (where an error would be, for example, to post an advisory when winds above 20 knots were unlikely). However, when they were told the probability that wind speed would be below 20 knots, they made far more errors, even though the information is mathematically equivalent. This indicates that when the verbal expression and the task were mismatched, more errors were made.

Another series of experiments was carried out to assess whether it was better to present probability forecasts in terms of probability (e.g. 10% chance) or frequency (e.g.
one time in 10). It has been argued that uncertainty presented as frequency is easier for people to understand [33,34]. However, Joslyn and Nichols [35] found that people better understood the forecast when it was presented in a probability format rather than a frequency format, in contrast with earlier research. This is more good news for probabilistic forecasting, indicating again that ordinary people can understand probabilities.

3.3. Assessment

Overall, the Probcast website has been reasonably successful, attracting about 2 million unique visits since it was set up in 2005 [36]. Public probabilistic weather forecasting (beyond probability of precipitation, which has been issued by the U.S. National Weather Service for about 40 years) is now being considered and evaluated by several national and other weather agencies, and Probcast provides both a methodology for producing calibrated probabilistic forecasts and a model of how they might be communicated to the public. It has also been cited by the National Research Council [37] as a possible model for communication of uncertainty in weather forecasting.

While specialists sometimes argue that the public does not understand probabilities and so there is little point in issuing probabilistic forecasts, the research results from the Probcast project suggest otherwise. The cognitive results indicate that users are ready for explicit uncertainty statements in forecasts and that including them can improve decision making and increase trust in the forecast. The fairly wide public use of the Probcast website, in spite of its lack of substantial institutional backing and its narrow geographical range (the North American Pacific Northwest), suggest that the public is ready for probabilistic forecasts on a broader scale, although of course only a portion of the public would actively use them (notably those with higher-stakes weather-related decisions to make).

The cognitive experiments carried out as part of our project by Susan Joslyn’s research group at the University of Washington suggest that probabilities of particular adverse weather events (e.g. freezing temperatures, precipitation, heavy precipitation, high winds) and percentiles (10th, 50th, 90th) of the predictive distribution of continuous weather quantities of interest (e.g. temperature, amount of precipitation, wind speed) are useful quantities to provide to users [38]. The work suggests that both are understandable to people and that they make better decisions when they have this information.

A common prescription is that probabilities should be used in decision making using decision theory [39]. This says that each possible outcome imposes a loss on the decision maker and that the decision made should minimize the expected overall loss. In this case, the expectation would be taken over possible future weather outcomes, and the losses might relate, for example, to the costs of issuing a high wind warning if no high winds occur and to the damage that high winds would cause in the absence of a warning. This seems to be a very useful framework when the utilities associated with different outcomes can be quantified on the same scale, typically money. The clearest weather example that we know of is decision making by wind energy companies that have to bid for contracts to provide specified amounts of energy at given prices and with specified penalties for failing to fulfil the contract in the presence of great uncertainty about future wind speeds [40].

However, we did not incorporate decision-theoretic concepts explicitly into the Probcast website. The Probcast website is used by many users with, presumably, quite different costs and losses. Thus, decision theory would involve providing solutions for individual users or a wide range of solutions for users with different utility functions. This might be done, for example, by asking the user to submit his or her utility function online and then performing the decision-theoretic calculations and providing the solution to the user. This would be ambitious and beyond the scope of the Probcast website.

It seems that most people are unaware of their utility functions and may even be unwilling to specify them when the losses involved are on different scales (e.g. money vs. possible loss of life). Thus people may find it easier to use probabilistic forecasts to make decisions that limit the risk of adverse outcomes to acceptable levels, rather than carrying out a full decision-theoretic analysis.

Nevertheless, Joslyn and Leclerc [32] showed that while people do not match the optimal decision-making standard, they are closer to it when they have probabilistic information when costs and benefits are on the same scale (e.g. money). Joslyn and Leclerc [32] also found that if people were given decision advice based on optimal decision-theoretic calculations, they followed the advice only if they were also given the probabilities.

4. PROJECTING THE HIV/AIDS EPIDEMIC

The Joint United Nations Programme on HIV/AIDS (UNAIDS) publishes updated estimates and projections of the number of people living with HIV/AIDS in the countries with generalized epidemics every 2 years. Generalized epidemics are defined by overall prevalence being above 1% and the epidemic not being confined to particular subgroups; there are about 38 such countries [41]. UNAIDS projections are typically provided for no more than 5 years into the future. As part of this, statements of uncertainty are also provided.
This exercise has two main goals. The first is to develop estimation and projection methods and software for use by country health officials for planning, for example to meet future medication needs. Statements of uncertainty may be used, for example, for determining the amount of medication needed to be reasonably sure of having enough to meet the need; this would correspond to an upper percentile of the predictive distribution.

The second goal is to contribute to the basis for the UNAIDS annual reports [42]. Uncertainty statements about estimates are routine in the UNAIDS reports, perhaps because UNAIDS is a newer agency, established in 1996, by which time it had become the norm to include uncertainty measures of some kind with estimates of uncertain quantities. See Fig. 5(a) for an example. While the uncertainty statements do not feature prominently in the published report for the broad public, they underlie assessments in the report, such as the following:

‘The annual number of new HIV infections among adults and adolescents decreased by 50% or more in 26 countries between 2001 and 2012. However, other countries are not on track to halve sexual HIV transmission, which underscores the importance of intensifying prevention efforts.’

The phrase ‘not on track’ reflects conclusions drawn in part from probabilistic projections.

We developed methods for assessing uncertainty about estimates and projections using Bayesian melding [43–45]. These were reasonably well calibrated in out-of-sample prediction experiences; see, for example, ref. [43]. One example of the output is shown in Fig. 5(b). Figures such as these typically do not make their way into the most visible public reports; instead, they provide background support for the conclusions presented in these reports.

There seem to be two main kinds of use for the probabilistic estimates and projections developed by UNAIDS. The first is to provide a general assessment of the estimates and projections and how accurate they are likely to be.

The second kind of use is to assess changes. For example, reported HIV prevalence might increase in a given year, but the question that then arises is whether the increase is out of the range of normal expectations, perhaps warranting some new policy interventions. Probabilistic forecasts such as those summarized by the uncertainty bands in Fig. 5(b) can be useful in this context. For example, if the new estimated prevalence is inside the range of the projection (even if there is an increase), then there is little evidence that what is happening is beyond what could be expected in the normal run of things, and the chosen policy could be to continue as before while monitoring the situation. On the other hand, if the new estimate is outside the projected range, there may be grounds for concern and for an intervention.

5. PROBABILISTIC POPULATION PROJECTIONS FOR THE UNITED NATIONS

The United Nations (UN) publishes projections of the populations of all countries broken down by age and sex, updated every 2 years in a publication called the World
Population Prospects (WPP). It is the only organization to do so. These projections are used by researchers, international organizations, and governments, particularly those of countries with less developed statistical systems and researchers. They are used for planning, social and health research, monitoring development goals, and as inputs to other forecasting models such as those used for predicting climate change and its impacts [46,47]. They are the de facto standard [48].

Like almost all the other population projections, the UN’s projections are produced using the standard cohort-component projection method [49–51]. This is a deterministic method based on an age-structured version of the basic demographic identity that the number of people in a country at time $t + 1$ is equal to the number at time $t$ plus the number of births minus the number of deaths plus the number of immigrants minus the number of emigrants.

The UN projections are based on assumptions about future fertility, mortality and international migration rates; given these rates, the UN produces the ‘medium’ projection, a single value of each future population number with no statement of uncertainty. The UN also produces ‘low’ and ‘high’ projections using total fertility rates (the average number of children per woman) that are, respectively, half a child lower and half a child higher than the medium projections. These are alternative scenarios that have no probabilistic interpretation.

Like the UN up to 2008, most national statistical offices, including the U.S. Census Bureau and the U.K. Office of National Statistics, use assumptions about future fertility, mortality, and migration rates from experts, either internal experts or panels of outside experts. Expert knowledge is an essential part of the population projection process, and experts are generally agreed to be good at assembling and reviewing the underlying science as well as assessing the actual forecasts.

However, evidence has been mounting over the past 60 years that experts in several domains are not as good at producing forecasts from scratch themselves. Meehl [52] found that very simple statistical models beat expert human forecasters in a range of clinical disciplines, and this finding has been replicated in many subsequent studies [53]. Oeppen and Vaupel [54] showed in 2002 that expert forecasts of life expectancy at birth, both by leading demographers and forecasting organizations, had performed poorly over the previous 70 years. Forecasters generally tended to project that the future would be like the present and, in particular, that a limit to life expectancy would be reached soon, whereas life expectancy continued to increase throughout the period.

Tetlock [55] evaluated the quality of about 3000 forecasts of political events and outcomes by experts, many distinguished, and found their performance to be startlingly poor. He memorably concluded that many of the experts would have been beaten by a ‘dart-throwing chimpanzee’. In a rare counterexample, Mandel and Barnes [56] found that analysts in a Canadian intelligence agency provided calibrated forecasts of good quality.

In collaboration with the UN Population Division, we developed new statistical methods for projecting future fertility and mortality rates probabilistically and translating these into probabilistic population projections for all countries [57–62].

An experimental version of the new probabilistic projections was issued by the UN in November 2012 at http://esa.un.org/unpd/ppp. This release was accompanied by no fanfare, but the experimental probabilistic projections have still had about 10 000 downloads per month. Official UN probabilistic population projections for all countries were issued for the first time on the same website on July 11, 2014 (World Population Day). An article analyzing the projections [63] attracted a great deal of media attention, much of which focused on the probabilistic aspects, notably the bounds of the forecast intervals.

There are other indications of the beginning of a paradigm shift from deterministic population projections based on expert assumptions to probabilistic population projections based on statistical models. For example, Statistics New Zealand changed its official population projection method to probabilistic projections in 2012 [64]. However, these releases are recent, and it remains to be seen how and to what extent ultimate users, such as policy makers and planners, make use of them. One possible use is in setting future international goals, similar to the Millenium Development Goals for 2015, for variables like child and maternal mortality. It is desirable to set goals that are ambitious but also realistic, and probabilistic projections could be useful in indicating what is realistic, suggesting setting goals that are towards the ‘good’ end of the probability distribution [65].

A possible use of probabilistic population projections is in making decisions about policy issues that depend directly on future population numbers, such as school and hospital infrastructure. One such decision is whether or not to close schools. These decisions are often based on deterministic population projections, which can have a spurious air of certainty. It is not desirable to close a school unless the probability of having to reopen it or find other premises in the future is small [66]. Thus, school closures based on deterministic projections may be unnecessary or at least premature. This suggests that there may be too many school closures in the United States, where these decisions are made locally.

Even if a deterministic population projection points to declining school enrollments, there can still be a substantial probability of them staying essentially constant or even
increasing, in which case closing the school would typically not be a good idea. Basing such decisions on reasonable upper percentiles of future school enrollments (such as the 90th percentile), rather than a deterministic projection or a predictive mean or median, could be a reasonable approach.

6. CONDITIONAL PROBABILISTIC FORECASTS: HOW MANY GRADUATE STUDENTS TO ADMIT?

Like most U.S. academic departments with Ph. D. programs, the Department of Statistics at the University of Washington, of which the author is a faculty member, faces the problem of deciding how many potential incoming graduate students to make funded offers to for the next academic year. Offers are made in December for entry the following September, 9 months later, and are binding on the department.

Incoming graduate students are funded by a mix of teaching assistantships, research assistantships, and fellowships. There are several major uncertainties to deal with in making this decision. The number of research assistantships available depends on the outcome of faculty research grant applications, which are often unknown 9 months ahead of time. Not all students accept our offers, and we do not know ahead of time how many will. We also do not know exactly how many current students will leave in the next 9 months through graduation or dropout.

Up to 2009, the departmental practice was to make a number of offers based on expected numbers of students graduating and grants and on an assumed acceptance rate. However, these calculations were based only on expectations and were not probabilistic and also did not incorporate past data in a systematic way.

This often led, in practice, to too few acceptances relative to the number of positions available, with the result that teaching assistants for Statistics courses had to be recruited from among non-Statistics graduate students. This was undesirable in that statistics teaching was not being performed by optimally qualified people, departmental teaching assistantships were ‘lost’ to other departments (in the sense that the department recruited and paid teaching assistants who were students in other departments for its courses when it could have paid its own students to do this had enough of them been available), and the pool of future potential research assistants was depleted. There are currently more jobs available for Ph.D. statisticians than graduates, so increasing the number of incoming graduate students is desirable from the labor market point of view as well. In the 5 years up to 2009, about four teaching assistants were being ‘lost’ to the department every year compared with a typical incoming class size of about ten graduate students.

The downside is that if students accept and there is no identified funding for them, the department has to scramble to find funding. This is difficult but possible within the university because many non-Statistics departments have research and teaching needs for statistically qualified people that they find hard to meet from within their own pool of students.

In 2010, the departmental faculty decided to base the decision about the number of students to admit on a probabilistic calculation instead of the then current expectation-based approach, and we took on the task of developing the appropriate method. For each possible number of offers, we computed the predictive probability distribution of the number of teaching assistant (TA) positions lost to the department as this seemed to be the key quantity for decision making. Ideally, this would be equal to zero.

With perfect knowledge, the number of TA positions lost conditional on a given number of offers is equal to

$$Y = T + R_1 + R_2 + G + L - C - A,$$

where

- $Y =$ Number of TA positions lost to department
- $T =$ Number of TA positions available
- $R_1 =$ Number of research assistant (RA) positions available within the department
- $R_2 =$ Number of RA positions available outside the department
- $G =$ Number of students graduating by September
- $L =$ Number of students dropping the program by September
- $C =$ Number of current students
- $A =$ Number of acceptances.

$T$ and $C$ are taken as known exactly, but the other quantities in Eq. (1) are uncertain at the time when the decision has to be made. Note that $Y$ can be negative, in which case the department recruits more students than it can fund with current resources and so has to find additional sources of funding.

The predictive distributions of $R_1$, $R_2$, $G$, $L$, and $A$ are derived from past data and elicited information. They are treated as independent in order to derive a joint distribution. The predictive distribution of $A$ depends on the number of offers, $O$, and is modeled as binomial $(O, \pi)$, where $\pi$ is estimated from historical data. The predictive distribution

Statistical Analysis and Data Mining: The ASA Data Science Journal DOI:10.1002/sam
Fig. 6 Conditional probabilistic forecasts of the number of lost teaching assistant positions given different numbers of graduate student offers with funding. Negative values indicate the number of students that would not be funded from current funding sources.

The predictive distribution of $R_1$ is obtained by polling departmental faculty to elicit from each of them a predictive distribution of the number of research assistantships they will have available in the next academic year. The distribution of $R_1$ is then the distribution of the sum of the numbers from the faculty, obtained by convolving the elicited distributions. The predictive distributions of $R_2$ and $L$ are based on historical data on these quantities; empirical rather than model-based distributions are used. The predictive distribution of $G$ is based on current information about student progress and is typically quite tight.

The predictive distribution of $Y$, the number of lost TA positions, which is the primary quantity for decision making, is then obtained by Monte Carlo. A large number of values of each of $R_1$, $R_2$, $G$, $L$, and $A$ are simulated from their predictive distributions, and the corresponding simulated values of $Y$ are found from Eq. (1).

Fig. 6 shows conditional predictive distributions of the number of lost TA positions given several possible number of offers, and Table 1 shows percentiles of these distributions. Note that negative numbers correspond to the number of students that could not be funded from current funding sources. In these cases, alternative funding sources would be sought, such as research or teaching assistantships in departments that currently fund few or no statistics graduate students.

The verbal descriptions in Table 1 characterize how aggressive a decision is relative to the uncertainty. For example, 20 offers is the break-even point because with that number the department is equally likely to lose TA
It would be possible to relax this assumption. A second round of offers is sometimes made, depending on initial responses to the first round of offers. It would be possible to extend the model to include the second round, about which decisions are currently made without similar quantitative analysis. However, the method seems developed enough to provide useful guidance to the decision makers, and there has not yet been a strong demand for further methodological refinement.

### 7. DISCUSSION

We have described five cases in which probabilistic forecasts have been used with a certain degree of success. These lead us to identify five types of potential users of probabilistic forecasts (where the five cases do not map exactly onto the five types of users):

1. **Low stakes user:** This is a user for whom the stakes are low and/or the losses from over- and under-predicting are similar. An example might be someone deciding whether to wear a sweater or a short-sleeved shirt based on temperature; a single ‘best’ temperature forecast will often be enough in this case.

2. **General assessor:** This is a user for whom the probabilistic forecast provides a general assessment of the likely quality of the forecast. The UNAIDS annual report is a possible example. This is also important for the process of forecast improvement. The goal of forecast development should be to improve forecast accuracy and hence to reduce the uncertainty around the forecast [67]. It is hard to guide this process without an accurate assessment of forecast uncertainty.

3. **Change assessor:** For this kind of user, the probabilistic forecast provides a way of assessing whether a change in some measurement is in line with expectations or is instead a source of concern warranting action. An example might be the probabilistic forecasts of HIV prevalence produced by UNAIDS, where some changes (including increases) are to be expected, but larger increases that are ‘significant’ would sound an alarm. One-number forecasts provide no way of making this kind of assessment.

4. **Risk avoider:** Here, the goal includes keeping the risk of an adverse outcome to an acceptable level. The IWC bowhead whale quota is a good example of this, in which the risk of possible damage to the stock from aboriginal whaling was
to be kept to a low level. Note that this did lead to a 'one number' forecast, but the forecast was not the 'best' or 'central' forecast but rather a lower percentile of the predictive distribution, in this case the fifth percentile. Deciding how many graduate students to admit is another example of this, and here, again, a quantile of the distribution (for each possible decision) was the basis for decision making.

5. Decision theorist: This user has an explicit loss function and is able to quantify it. He or she uses the probabilistic forecast to explicitly minimize expected loss, as advocated by formal decision theory. This did not arise in any of the cases described here and seems most likely when the different kinds of possible loss being traded off are on the same scale, typically money. One example would be a wind energy company, which needs to bid on a contract to supply a given amount of energy, with specified penalties if the contract is not fulfilled [40].

While definite conclusions cannot be drawn from a small number of case studies, this discussion suggests that different types of users may need different types of probabilistic forecasts. The low stakes user would seem to need only a point forecast of a continuous prediction and/or the probability of a binary-valued outcome. The general assessor and change assessor need interval forecasts. The risk avoider needs a quantile forecast, while the decision theorist needs the full predictive distribution.

The fact that there are different types of users and the use of probabilistic forecasts suggests that it is important for developers of probabilistic forecasts to interact with users and consider their goals. While this may seem obvious, it is often not done. Interaction can take the form of direct contact (meetings, phone, Email, and so on) between developers and users. This can be in the context of an established scientific advisory committee with regular meetings and an official membership (as used by the IWC), or a small, less formal reference group with rotating members (as used by UNAIDS) or expert group meetings, which are effectively workshops lasting several days (as used by the UN Population Division). If the probabilistic forecasts are delivered to the general public using a website (as in the case of probabilistic weather forecasting), the interaction can take the form of a survey [27].

It is important for trust in the forecast that the probabilistic statements be at least approximately calibrated so that, for example, events given a predictive probability of 80% happen about 80% of the time on average. For the forecast to be useful, it is also important that forecast intervals be narrow, or sharp, enough to provide a basis for action. Gneiting [29] defined the key design principle of probabilistic forecasting as maximizing sharpness subject to calibration, and this has been widely accepted.

The experience we have described suggests that formal decision theory, much advocated in theory by statisticians and economists, may have less practical application than is sometimes claimed. One reason may be that people are often not aware of their loss functions. Another may be that using formal decision theory greatly increases the cognitive load in that one's loss function has to be assessed and then the decision-theoretic calculations performed. One also needs to be careful because in practice, people tend to attribute different values to equivalent losses and gains, contrary to decision theory, a finding referred to as 'prospect theory' [68,69]. Nevertheless, a recent result suggests that the scope of decision theory may be wider than we have conceded. Gneiting [70] showed that if the loss function is generalized piecewise linear as a function of the quantity being predicted probabilistically, then the optimal point forecast is a quantile of the predictive distribution.

An important special case of this is when the cost of an overestimate is a fixed multiple of the cost of an underestimate. This will often be at least roughly true, and it may be much easier to elicit that multiple than the full loss function. People may be able to say, at least approximately, how much worse an overestimate is than an underestimate, or vice versa. This also greatly simplifies the practical use of decision theory, reducing it to the calculation of a predictive quantile, so that the cognitive load is little greater than that of probabilistic forecasting by itself.

One overarching conclusion is that people can use and understand probabilities and probabilistic forecasts, even if they do not have advanced training in statistics. Cognitive research shows that probabilistic forecasts lead to better decision making than deterministic ones and also to increased trust in the forecast by users. Experience with probabilistic weather forecasting and probabilistic population projection websites has shown that there is considerable public interest in probabilistic forecasts, even in the absence of much publicity. This suggests that once probabilistic forecasts become available in a domain, they will be used: build it, and they will come.

ACKNOWLEDGMENTS

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Development through grants nos. R01 HD054511 and R01 HD070936, and by a Science Foundation Ireland E. T. S. Walton visitor award, grant reference 11/II/12079. The author is grateful to Geof Givens, Susan Joslyn, Giampaolo Lanzieri, and

Statistical Analysis and Data Mining: The ASA Data Science Journal DOI:10.1080/10618600.2016.1186234
REFERENCES

[1] P. Jorion, Value at Risk: The New Benchmark for Managing Financial Risk (3rd ed.), New York, McGraw-Hill, 2007.
[2] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, Time Series Analysis, Forecasting and Control (4th ed.), New York, Wiley, 2008.
[3] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation and Updating, New York, Springer, 2009.
[4] D. Kahneman, Thinking, Fast and Slow, New York, Farrar, Strauss and Giroux, 2011.
[5] A. E. Raftery, G. H. Givens, and J. E. Zeh, Inference from a deterministic population dynamics model about bowhead whale, balaena mysticetus, replacement yield, Department of Statistics, University of Washington, Technical Report 232, 1992.
[6] A. E. Raftery, G. H. Givens, and J. E. Zeh, Inference from a deterministic population dynamics model for bowhead whales (with discussion), J Am Stat Assoc 90 (1995), 402–430.
[7] D. Poole and A. E. Raftery, Inference for deterministic simulation models: the Bayesian melding approach, J Am Stat Assoc 95 (2000), 1244–1255.
[8] M. R. Powell, M. Tamplin, B. Marks, and D. T. Compos, Bayesian synthesis of a pathogen growth model: listeria monocytogenes under competition, Int J Food Microb 109 (2006), 34–46.
[9] J. R. Brandon, J. M. Breiwick, A. E. Punt, and P. R. Wade, Constructing a coherent joint prior while respecting biological realism: application to marine mammal stock assessments, ICES J Marine Sci 64 (2007), 1085–1100.
[10] M. G. Falk, R. J. Denham, and K. L. Mengersen, Estimating uncertainty in the revised universal soil loss equation via Bayesian melding, J Agric Biol Environ Stat 15 (2010), 20–37.
[11] A. Punt and G. Donovan, Developing management procedures that are robust to uncertainty: lessons from the International Whaling Commission, ICES J Marine Sci 64 (2007), 603–612.
[12] J. G. Cooke, R. Leaper, and V. Papastavrou, Science should not be abandoned in a bid to resolve whaling disputes, Biol Lett 5 (2009), 614–616.
[13] J. G. Cooke, R. Leaper, and V. Papastavrou, Whaling: ways to agree on quotas, Nature 482 (2012), 308.
[14] T. Gneiting and A. E. Raftery, Weather forecasting with ensemble methods, Science 310 (2005), 248–249.
[15] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles, Mon Weather Rev 133 (2005), 1155–1174.
[16] T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon Weather Rev 133 (2005), 1098–1118.
[17] J. M. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley, Probabilistic quantitative precipitation forecasting using Bayesian model averaging, Mon Weather Rev 135 (2007), 3209–3220.
[18] J. M. Sloughter, T. Gneiting, and A. E. Raftery, Probabilistic wind speed forecasting using ensembles and Bayesian model averaging, J Am Stat Assoc 105 (2010), 25–35.
[19] L. Bao, T. Gneiting, E. P. Grimit, P. Guttorp, and A. E. Raftery, Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction, Mon Weather Rev 138 (2010), 1811–1821.
[20] J. M. Sloughter, T. Gneiting, and A. E. Raftery, Probabilistic wind vector forecasting using ensembles and Bayesian model averaging, Mon Weather Rev 141 (2013), 2107–2119.
[21] R. M. Chmielecki and A. E. Raftery, Probabilistic visibility forecasting using Bayesian model averaging, Mon Weather Rev, 139 (2011), 1626–1636.
[22] C. F. Mass, S. L. Joslyn, J. Pyle, P. Tewson, T. Gneiting, A. E. Raftery, J. Baars, J. M. Sloughter, D. W. Jones, and C. Fraley, PROBCAST: a web-based portal to mesoscale probabilistic forecasts, Bull Am Meteorol Soc 90 (2009), 1009–1014.
[23] S. L. Joslyn and D. W. Jones, Strategies in naturalistic decision-making: a cognitive task analysis of naval weather forecasting. In Naturalistic Decision Making and Macroognition, J. M. Schraagen, L. Militello, T. Ormerod, and R. Lipshitz, eds. Aldershot, UK, Ashgate Publishing, 2008, 183–202.
[24] S. Joslyn, L. Nadav-Greenberg, and M. U. Taing, The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision, Appl Cogni Psychol 23 (2008), 55–72.
[25] L. Nadav-Greenberg, S. L. Joslyn, and M. U. Taing, The effect of weather forecast uncertainty visualization on decision making, J Cogni Eng Decis Mak 2 (2008), 24–47.
[26] S. Joslyn, L. Nadav-Greenberg, and R. M. Nichols, Probability of precipitation: assessment and enhancement of end-user understanding, Bull Am Meteorol Soc 90 (2009), 185–193.
[27] S. L. Joslyn and S. Savelli, Communicating forecast uncertainty: public perception of weather forecast uncertainty, Meteorol Appl 17 (2010), 180–195.
[28] D. Kahneman and A. Tversky, Choices, values and frames, Am Psychol 39 (1984), 341–350.
[29] T. Gneiting, F. Balabdaoui, and A. E. Raftery, Probabilistic forecasts, calibration and sharpness, J Roy Stat Soc B 69 (2007), 243–268.
[30] S. L. Joslyn, K. Pak, D. W. Jones, J. Pyles, and E. Hunt, The effect of probabilistic information on threshold forecasts, Wea Forecast 22 (2007), 804–812.
[31] L. Nadav-Greenberg and S. L. Joslyn, Uncertainty forecasts improve decision-making among nonexperts, J Cogni Eng Dec Mak 2 (2009), 24–47.
[32] S. L. Joslyn and J. E. LeClerc, Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error, J Experim Psychol Appl 18 (2012), 126–140.
[33] K. Fiedler, The dependence of the conjunction fallacy on subtle linguistic factors, Psychol Res 50 (1988), 123–129.
[34] A. R. Hertwig and G. Gigerenzer, The “conjunction fallacy” revisited: How intelligent inferences look like reasoning errors, J Behav Decis Mak 12 (1999), 275–305.

Statistical Analysis and Data Mining: The ASA Data Science Journal DOI:10.1002/sam
[35] S. L. Joslyn and R. M. Nichols, Probability or frequency? expressing forecast uncertainty in public weather forecasts, Meterol Appl 16 (2009), 309–314.

[36] D. W. Jones, Probcast: a tool for conveying probabilistic weather information to the public, Presented to the NUOPC Conference, June 2011, 2011.

[37] National Research Council. Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts, Washington, D.C., National Academies Press, 2006.

[38] S. Savelli and S. Joslyn, The advantages of 80% predictive interval forecasts for non-experts and the impact of visualizations, Appl Cognit Psychol 27 (2013), 572–541.

[39] J. von Neumann and O. Morgenstern, Theory of Games and Economic Behavior (2nd ed.), Princeton, NJ, Princeton University Press, 1947.

[40] P. Pinson, C. Chevallier, and G. N. Kariniotakis, Trading wind generation with short-term probabilistic forecasts of wind power, IEEE Trans Power Syst 22 (2007), 1148–1156.

[41] P. D. Ghys, T. Brown, N. C. Grassly, G. Garnett, K. A. Stanecki, J. Stover, and N. Walker, The UNAIDS estimation and projection package: a software package to estimate and project national HIV epidemics, Sexual Transm Infect 80 (2004), i5–i9.

[42] UNAIDS, Global Report: UNAIDS Report on the Global HIV/AIDS Epidemic, 2013, UNAIDS, Geneva, 2013.

[43] L. Alkema, A. E. Raftery, and S. J. Clark, Probabilistic projections of the total fertility rate for all countries, Demography 48 (2011), 815–839.

[44] D. W. Jones, Probcast: a tool for conveying probabilistic weather information to the public, Presented to the NUOPC Conference, June 2011, 2011.

[45] Intergovernmental Panel on Climate Change. Climate Change 2007: Synthesis Report, IPCC, Geneva, Switzerland, 2007.

[46] K. C. Seto, B. Güneral, and L. R. Hutyra, Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools, Proc Natl Acad Sci 109 (2012), 16083–16088.

[47] W. Lutz and K. C. Samir, Dimensions of global population projections: what do we know about future population trends and structures? Philosoph Trans Roy Soc B 365 (2010), 2779–2791.

[48] P. K. Whelpton, An empirical method for calculating future population, J Am Stat Assoc 31 (1936), 457–473.

[49] P. H. Leslie, On the use of matrices in certain population dynamics, Biometrika 33 (1945), 183–212.

[50] S. H. Preston, P. Heuveline, and M. Guillot, Demography: Measuring and Modeling Population Processes, Malden, MA, Blackwell, 2001.