METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **16**: 237–244 (2009)
Published online 27 November 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.113

A laboratory-based study of understanding of uncertainty in 5-day site-specific temperature forecasts

Mark S. Roulston[a†] and Todd R. Kaplan[b*]
[a] *Met Office, FitzRoy Road, Exeter EX1 3PB, UK*
[b] *School of Business and Economics, University of Exeter, Streatham Court, Rennes Drive, Exeter EX4 4PU, UK*

ABSTRACT: The impact of presenting uncertainty in 5-day location-specific temperature forecasts on the decision making of non-specialists was tested in an experimental economics laboratory. Undergraduate students studying a range of disciplines were asked to select which of two criteria involving temperature would be most likely to occur based on a given 5-day forecast. If they selected a criterion that was subsequently satisfied they were given a small cash reward. It was found that students presented with uncertainty information (the 50th and 90th percentile confidence intervals) in addition to the expected temperature were more likely to select the most probable criterion. This was true irrespective of the academic subject the participants were studying. Copyright © Royal Meteorological Society and Crown Copyright, 2008

KEY WORDS uncertainty; predictability; probability

Received 29 October 2007; Accepted 8 October 2008

1. Introduction

It is now widely recognized and acknowledged in the weather forecasting community that, in principle, information concerning forecast uncertainty can enhance the value of forecasts and improve the ability of forecast-users to make better decisions (AMS, 2002; NRC, 2006). An increasing number of studies have attempted to estimate the economic value of weather forecasts, and in particular the value of providing uncertainty information to users (e.g. Richardson, 2001; Thornes and Stephenson, 2001). These studies typically assume optimal decision making on the part of users. It is likely that many commercial users do indeed have well developed tools for managing other types of risk and will be able to make optimal use of risk-oriented forecasts – such as ensemble forecasts (Palmer, 2002). However, there is a greater debate over whether more casual users, such as the general public, who receive their forecasts through free-at-point-of-use channels such as TV and the web, are capable of understanding probabilistic information well enough to benefit from its inclusion in forecasts disseminated *via* these channels. It is also not clear how uncertainty information is best communicated to these users. Consequently, meteorologists' ability to provide information about forecast uncertainty now significantly exceeds the amount of such information that is communicated to the general public. This 'dissemination gap' raises the possibility that investments in forecasting tools, such as ensemble prediction systems, designed to provide risk-based information will not yield large enough socio-economic returns to justify them. The aim of this study is to assess the impact of providing graphical weather forecast uncertainty information on the decision making of these non-specialists.

Probability of precipitation (PoP) forecasts are routinely provided in public weather forecasts in the United States, as well as by commercial forecasting companies providing forecasts for media aimed at the general public. Survey-based research indicates that there is widespread confusion in the public's understanding of PoP forecasts, but that this confusion is as much attributable to forecasters not clearly defining the event class (whether the probability refers to an area, a time frame or refers to the chance of rain at a given place within a given time) than to an inherent inability of the public to understand probability (Murphy *et al.*, 1980; Gigerenzer *et al.*, 2005). More generally, this type of research highlights the importance of objectively evaluating the effectiveness with which forecast products are communicated – whether it is to forecasters or end-users.

The value of a weather forecast derives from its ability to influence decisions being made in the face of uncertainty (Murphy, 1993). The question of how people make decisions under uncertainty is one of the principal research themes of the field of experimental economics (Roth, 1993). The methods developed by experimental economists to study individual choice are

* Correspondence to: Todd R. Kaplan, School of Business and Economics, University of Exeter, Streatham Court, Rennes Drive, Exeter EX4 4PU, UK. E-mail: t.r.kaplan@ex.ac.uk
† The contribution of Mark S. Roulston was written in the course of his employment at the Met Office, UK and is published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

useful tools for objectively determining how well people understand weather forecasts. A preliminary application of laboratory-based experimental economic methods to this question suggests that, at least in some contexts, people can interpret forecast uncertainty information and use it to make better decisions (Roulston *et al*., 2006). In this paper such methods are used to test specific formats for presenting probabilistic weather information. Laboratory-based studies have also been conducted by psychologists that indicate that student forecasters can make use of probabilistic information when deciding whether to issue wind advisories (Joslyn *et al*., 2007).

This paper presents the results of laboratory-based trials of a format designed to communicate the 5-day temperature forecast for a single location in a probabilistic way. The format tested was the most popular of five formats presented to users of the U.K. Met Office's website as part of an online questionnaire that was conducted between the 13 and 19 June 2006. Examples of the format – a 'fan chart' in which the 50% and 90% confidence intervals are indicated – are shown in Figures 1(b) and 2(b). In the questionnaire 49% of 1144 respondents identified this format as the most useful of five options presented, and 44% also identified it as the easiest to understand. Another very similar format, in which only the 80% interval was given (instead of both the 50% and 90% intervals) was also popular, with 23% of participants declaring it the most useful and 32% saying that it was the easiest to understand.

While providing users with forecasts in formats that they believe to be useful and easy to understand is obviously an important aim, it does not automatically follow that users will be able to interpret these forecasts well enough to extract value from them. It is not inconceivable that survey respondents may either overestimate or underestimate their ability to understand probabilistic information. To address this issue, an experiment was designed to test how well people can use the information provided in the 5-day temperature fan charts. During the winter of 2006/2007 the experiment was conducted at the Finance

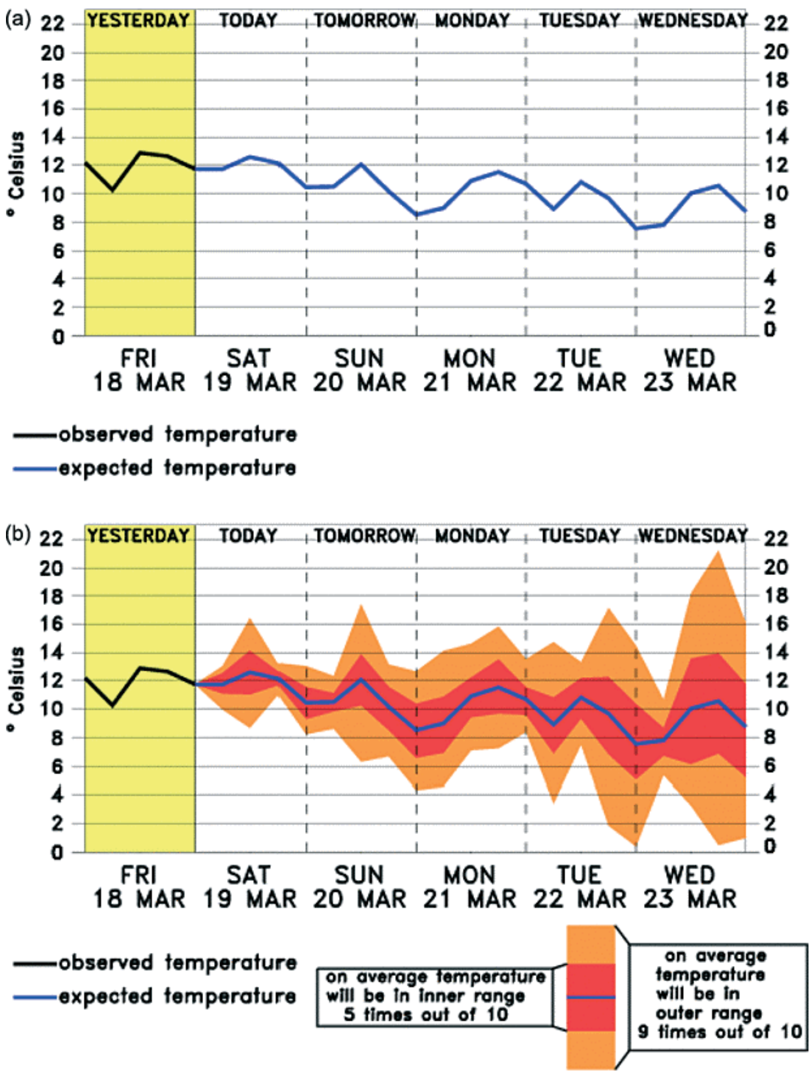

Figure 1. (a) The forecast presented to group A in round 12 of the experiment. The options in this round were to receive £0.50 if:

1. The temperature at midday on Monday is below 7 °C or if
2. The temperature at midday on Wednesday is below 6 °C.

(b) The forecast presented to group B in round 12 of the experiment. The options were the same as those presented to group A.

DOI: 10.1002/met

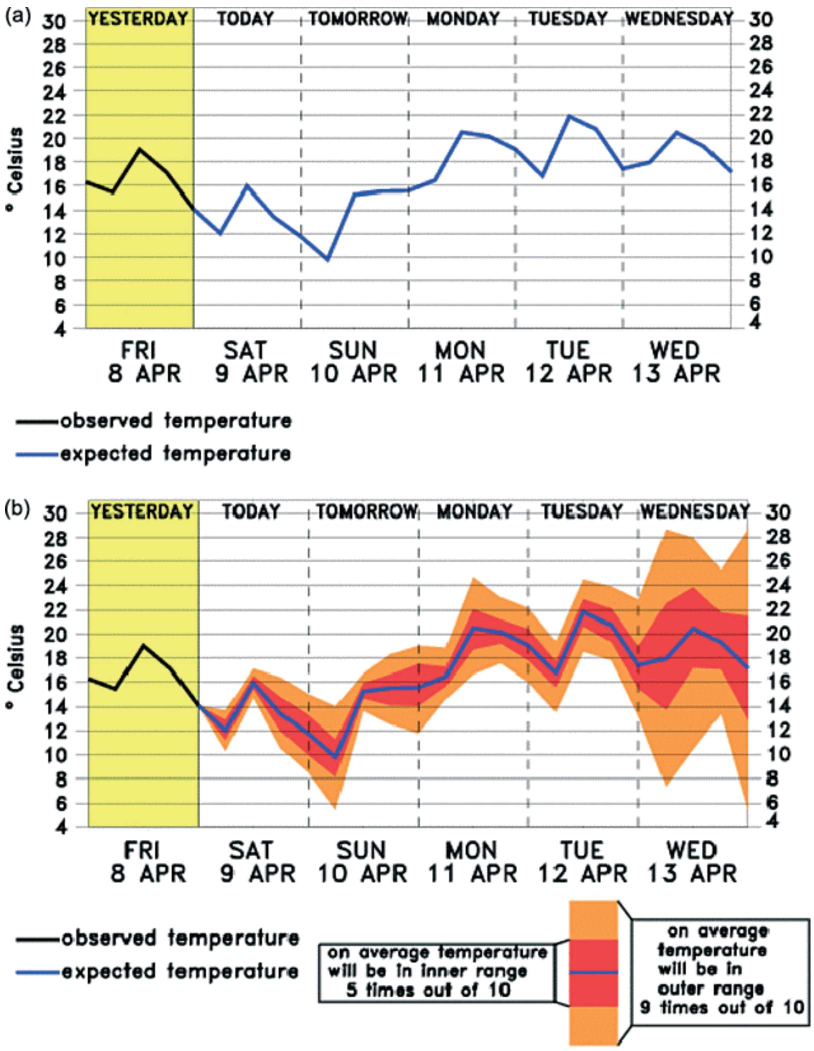

Figure 2. (a) The forecasts presented to group A in round 15 of the experiment. The options in this round were to receive £0.50 if:

1. The temperature at midday on Sunday is above 15 °C or if
2. The temperature at midday on Wednesday is above 14 °C

(b) The forecasts presented to group B in round 15 of the experiment. The options were the same as those presented to group A.

and Economic Experimental Laboratory at Exeter University (FEELE) using 153 students as subjects.

2. Method

Undergraduates at the University of Exeter, studying a variety of academic subjects, were recruited to take part in the experimental sessions. In total, 153 students participated. These participants were divided into two treatment groups: A and B. Both groups were presented with a sequence of 20 'lotteries' in which they could choose to receive £0.50 (approximately US \$0.98 at the time of the experiments) if one of two criteria concerning temperatures in the next 5 days was satisfied. To assist their decision, in each lottery, they were provided with a 5-day temperature forecast. Two examples of these lotteries, along with the different forecast information presented, are shown in Figures 1(a), (b), 2(a) and (b). The two criteria in each lottery always had the following structure:

Criterion A: The temperature at midday on Day D1 is above/below X° C

Criterion B: The temperature at midday on Day D2 is above/below Y° C.

In every lottery both criteria involved the same preposition: both were 'above' or both were 'below', there were no 'mixed' lotteries in which one criterion was above and one below.

Group A was given a graphical deterministic (or 'point') forecast that contained no information about forecast uncertainty, whereas group B was given a format that included information concerning forecast uncertainty. The sequence of lotteries and the outcomes of the lotteries were identical for each participant. The forecasts and the outcomes were not authentic weather data but were generated synthetically. The forecasts were not labelled as referring to a particular location. The outcomes were selected from a normal distribution centred on the expected value given in both forecasts and with a width consistent with the uncertainty information

DOI: 10.1002/met

provided to group B. In other words, the uncertainty information given to group B was 'perfect' in that it accurately reflected the probability that the outcome would lie in particular intervals. The dates on the forecast on each successive round were separated by 1 week, this was done to minimize the possibility that participants would use prior views they may have about day-to-day correlations in temperature. The temperatures in the forecasts were tuned to be typical of temperatures in the United Kingdom and the uncertainties were adjusted to be broadly in line with temperature forecast errors at horizons of 1–5 days, although the frequency of highly uncertain forecasts was higher than would be expected in reality. Examples of two of the lotteries and the different forecast information given to groups A and B are shown in Figures 1(a), (b), 2(a) and (b). Notice that the predicted expected temperature is the same for both groups, the only difference in the forecasts is the extra information about uncertainty provided to group B. After the students had made their selection of temperature criterion they were told what the actual temperatures were and whether either of the criteria had been satisfied. If the criterion they had chosen was satisfied their balance was credited with £0.50. At the end of the session they were paid their cash balances in addition to a £5.00 show-up payment. The breakdown of the participants by subject of study, gender and the treatment group that they were assigned to is shown in Table I.

In 14 of the 20 lotteries, a hypothetical participant who assumed that the uncertainty associated with the deterministic forecast (group A) was the same at all forecast lead times (and hence would make a choice based on distance to the prediction line) would identify the same option as being most likely as a participant in group B with access to uncertainty information. In the remaining six lotteries the assumption of uniform uncertainty at all forecast lead times would lead to the hypothetical member of group A choosing a different option to the fully informed group B participant. These six lotteries will be referred to as 'swing questions'. Figure 1(a) and (b) are an example of a 'swing' question whereas Figure 2(a) and (b) show a 'non-swing' question.

The instructions given to participants are in Appendix A. After the lottery questions, for which cash could be earned, participants were also asked several supplementary questions. These questions were designed to gather information about where participants currently get their weather forecast information, and to measure crudely the statistical literacy of participants. These questions are listed in Appendix B.

3. Results and discussion

The average reward (not including show-up payment) for group A participants was £7.25, while the average for group B was £8.48. The percentages of participants who chose the most probable outcome, broken down by subject area, gender and treatment group are shown in Table II. On average participants in group A – who received no information about forecast uncertainty – picked the most probable outcome 68.5% of the time. Participants in group B, who had uncertainty information, selected the most probable outcome 85.2% of the time (a naïve strategy of always choosing the criterion involving the later time would result in success in 65% of the lotteries). This difference between group A and group B is approximately the same when participants are grouped by academic discipline or by gender. The difference between the performances of group A and group B participants is much larger on the swing questions. Indeed, almost all the overall difference comes from these questions.

A more detailed statistical analysis was performed by fitting the results to a probit model. The probit model predicts the probability that a participant will select the most probable outcome as a function of a set of predictor variables. It is given by:

$$\begin{aligned} F &= \Pr(\text{most probable}|\text{predictors}) \\ &= \Phi\left(\sum_{i=1} \text{coefficient}_i \times \text{predictor}_i\right) \end{aligned} \quad (1)$$

Table I. A summary of the number of participants in each treatment group, broken down according to gender and the subject area of their degree courses. Note that the normal 'order of options' was with the option referring to the earlier time given first.

Discipline	Gender	Order of options	Format A	Format B
Business/economics	Male	Normal	13	15
Business/economics	Female	Normal	7	5
Business/economics	Male	Reversed	14	12
Business/economics	Female	Reversed	3	4
Science/engineering	Male	Normal	16	15
Science/engineering	Female	Normal	4	5
Humanities	Male	Normal	9	11
Humanities	Female	Normal	11	9

Table II. A summary of the results of the experiments. A 'correct' response was one in which the participant chose option with the most probable outcome. Note that, irrespective of how participants are segmented, those with uncertainty information (format B) outperformed those without (format A).

Group	(%) Correct (format A)	(%) Correct (format B)
Business/economics	69.6	85.7
Science/engineering	68.5	85.8
Humanities	66.5	83.8
Male	69.3	85.8
Female	66.8	83.7
Swing questions	24.5	76.3
Non-swing questions	87.4	89.0
Overall	68.5	85.2

DOI: 10.1002/met

where Φ is the cumulative distribution function of the standard normal distribution. The probit model was introduced in the 1930s to model data that could only take on values from 0 to 100% (Bliss, 1934). The use of a conventional linear model to predict the probability F would be problematic as F would not be constrained to lie between 0 and 1. A list of the predictor variables is given in Table III. All the predictor variables, with the exception of 'Question Number', are dummy variables that are either equal to 0 or 1 depending upon the condition they represent. 'Question Number' was included to detect possible learning effects in which participants get better at answering as the session progresses. Table III lists the values of the coefficients in Equation (1) found by maximum likelihood estimation. Table III also gives the value of dF/dx for each predictor, this is the change (marginal effect) in the probability that the participant will choose the most likely outcome associated with that predictor. For example, if the participant is British then the probability of them selecting the most likely outcome is, on average, 11.9% higher than for non-British participants. The p-values indicating the statistical significance of each predictor are also given in Table III.

The only predictors of performance that were significant at the 1% level are 'Early Correct', 'Swing Question', 'Swing Question and Format A', 'Nationality is British' and 'Checks Weather at least every 2–3 days'. The 'Early Correct' predictor was equal to one if the earlier criterion (e.g. Sunday as opposed to Wednesday) was the most probable. The results indicate that if the option referring to the earlier day happened to be the most likely then participants were 11.6% less likely to choose it. In other words, participants were biased towards selecting the later option (two sessions were run with the order of the options reversed so that in each round the first option referred to the *later* time). If the question was a swing question this reduced the chance of participants picking the optimal option by 15.9% on average, and for participants in group A – who had no uncertainty information – their chances of making an optimal choice were reduced by an additional 47.7% on average. Also, as noted above, British participants were 11.9% more likely to select the most probable option. Since the non-British participants were from non-English speaking countries this predictor is likely to be strongly correlated with whether a participant's first language was English or not. As all the instructions and questions were in English the reduction in performance of non-native speakers is perhaps not surprising. It is interesting to note that the reduction in performance associated with not being a native English speaker was smaller than the reduction associated with not having uncertainty information. Participants who claimed to check the weather at least every 2–3 days were 4.8% less likely to choose the most probable outcomes. Whether participants obtained their forecasts from the internet or not did not make a statistically significant difference.

Two additional predictors were significant at the 10% level. Participants studying humanities were 4.1% less likely to select the most probable option. From Table II it can be seen that while the performance of humanities students was marginally lower, the difference in performance between humanities students in group A and group B was comparable with students studying other subjects. Students who answered the question about rolling a fair die twice incorrectly were 4.1% less likely to pick the most probable outcomes. The relatively small effect of this predictor might be because participants had no cash incentive to answer this question correctly. It might also indicate that even a fairly basic understanding of probability is not required to be able to use the type of uncertainty information presented to group B. Table IV shows a summary of how many participants answered each individual questions correctly. It also shows the underlying probability of each of the two statements being true and which of the statements did turn out to be true. Notice

Table III. The results of fitting a probit model to the experimental data.

Predictor	Coeff	dF/dx	P-value
Question Number	0.0061	0.0017	0.226
Reverse order	−0.0733	−0.0206	0.381
Early Correct	−0.4009	−0.1162	0.000
Early Correct & Format B	0.1542	0.4076	0.186
Format B	0.0171	0.0047	0.849
Swing Question	−0.5307	−0.1590	0.000
Swing Question & Format A	−1.3706	−0.4768	0.000
Gender is Male	0.0247	0.0068	0.698
Humanities	−0.1454	−0.0414	0.092
Science/Engineering	−0.0855	−0.0240	0.283
Nationality is British	0.3798	0.1189	0.001
Die Question Mistake	−0.1423	−0.0405	0.052
Sample Question Mistake	0.0011	0.0003	0.986
Gets Weather from Internet	0.0965	0.0274	0.208
Checks Weather at least every 2–3 days	−0.1737	−0.0484	0.003
Constant	1.0033		0.000

DOI: 10.1002/met

Table IV. A summary of how many participants answered each question correctly, the underlying probability of statements 1 and 2 being correct, which statements turned out to be correct and also whether the question was a 'swing' question and whether the outcome associated with the earlier date was the most likely event.

Question	Treatment A		Treatment B		Probability		Actual	Swing question	Early date correct
	Correct	Wrong	Correct	Wrong	1	2			
1	73	4	72	4	0.85	0.49	1	N	Y
2	76	1	72	4	0.52	0.66	2	N	N
3	65	12	58	18	0.85	0.49	Both	N	Y
4	5	72	44	32	0.11	0.25	2	Y	N
5	69	8	66	10	0.57	0.86	2	N	N
6	66	11	71	5	0.69	0.96	Both	N	N
7	70	7	69	7	0.54	0.9	Both	N	N
8	67	10	63	13	0.59	0.94	Both	N	N
9	12	65	57	19	0.91	0.61	1	Y	Y
10	16	61	59	17	0.9	0.66	Both	Y	Y
11	65	12	72	4	0.04	0.27	2	N	N
12	46	31	69	7	0.05	0.25	Neither	Y	N
13	56	21	50	26	0.81	0.47	1	N	Y
14	28	49	60	16	0.03	0.23	2	Y	N
15	73	4	70	6	0.62	0.9	2	N	N
16	6	71	59	17	0.03	0.2	2	Y	N
17	49	28	67	9	0.81	0.59	1	N	Y
18	72	5	76	0	0.04	0.24	2	N	N
19	71	6	68	8	0.54	0.76	2	N	N
20	70	7	73	3	0.49	0.26	Both	N	Y

that the most probable of the two statements turned out to be true in every lottery except for question number 12 in which neither of the two statements actually occurred.

To evaluate the overall ability of the model in Equation (1) to predict whether a participant would choose the most likely option an ROC (receiver or relative operating characteristic) curve was plotted. This curve is shown in Figure 3. ROC curves are frequently used in meteorology to evaluate forecast skill (Mason, 2003) as well as in psychology (Swets, 1973). The ROC curve is parameterized by probability threshold. The ideal curve would have a zero false alarm rate and a hit rate of one – placing it at the top left of the diagram. The diagonal line is the 'no skill' line. Curves to the top left of this line indicate skill. The area between the ROC curve and the diagonal line thus provides a measure of predictive skill. This area would be 0.5 for perfect forecasts. For the curve in Figure 3 the area is 0.31 indicating that the model does have predictive skill, with the caveat that this ROC curve was calculated for 'in-sample' predictions, that is the data being predicted were the same as the data used to fit the model. Because of this the ROC curve should be viewed more as a measure of 'goodness-of-fit' than a verification of predictive skill.

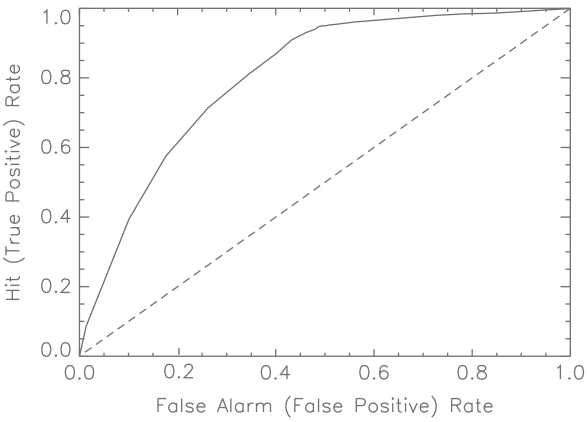

Figure 3. The receiver or relative operating characteristic (ROC) curve of the model given by Equation (1). The curve is parameterized by probability threshold. The dashed diagonal line is the 'no skill' line. The area between the curve and the no skill line is 0.31. A perfect forecasting model would have a hit rate of 1.0 and a false alarm rate of 0.0, placing it in the top left corner of the graph. The area between such a curve and the diagonal would be 0.5.

Figure 4 is a graph of the time it took participants to answer and review each question under each of the treatments. It can be seen that as the experiment progressed response times fell – falling by over 50% between the beginning and the end of the session. Also, there was little difference between the response times of those participants who had uncertainty information and those who did not. Overall, it took 21.7 s on average to answer each question and 5.8 s to review the results for each lottery.

The probit regression was run breaking the question number into two separate variables: one that equalled the question number if the format was A and zero otherwise, and one that equalled the question number if the format was B and zero otherwise. For the latter variable, Coeff = 0.0187251, dF/dx = 0.0051612, and P-value = 0.011. For the former variable, the result was not significant (P-value = 0.438). All other numbers in

DOI: 10.1002/met

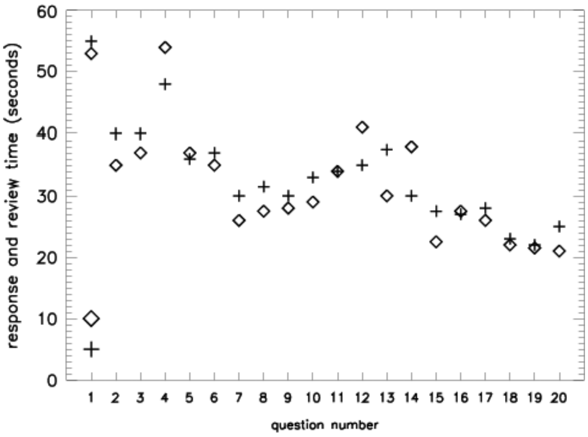

Figure 4. The average time it took for participants to answer and review each question. The diamonds are the average times for group A, who did not receive uncertainty information, and the crosses are the times for group B, who received uncertainty information.

Table III were not substantially different. This indicates a potential learning effect under format B only. The purpose of the current study was to assess the impact of uncertainty information. A study focusing on how rapidly users improve in their use of uncertainty information could be conducted by reversing or randomizing the order of the lotteries.

4. Conclusions

The results of the experiments described in this paper indicate that, within the context studied, undergraduates studying a range of academic subjects who were provided with graphical information about uncertainty in 5-day temperature forecasts were able to make better decisions, on average, than undergraduates who were not provided with uncertainty information. The improvement in performance associated with receiving uncertainty information was independent of what subjects the students were studying, or their gender. Whether the students had demonstrated an understanding of probability theory by correctly answering a question about dice rolling was a weak predictor of how much the provision of forecast uncertainty information improved decision-making. Furthermore, the provision of uncertainty information did not significantly slow down the decision making of participants. It is speculated that this is because participants without this information were generating their own hypotheses about what the uncertainties of the forecasts were, and that this process was as time-consuming as attempting to interpret explicit uncertainty information. Determining exactly how participants were using the uncertainty information is beyond the scope of this study. They may have been consciously interpreting the information as probabilistic or they have constructed heuristics based on the width of the coloured bands.

In the experiment described in this paper the provision of uncertainty information mainly improved decision-making for certain lotteries – the ‘swing’ questions. These questions were the ones in which assuming uniform uncertainty across forecast horizons would not give the correct answer. For the ‘non-swing’ questions assuming uniform uncertainty leads to the correct answer and so access to uncertainty information is of less value. Another situation in which providing uncertainty information might not improve decision-making is when the uncertainty of a forecast at a given horizon is constant (e.g. if the expected error on a day ahead forecast was *always* 2 °C and the expected error on a 2 day forecast was *always* 3 °C). In this situation, once a decision maker had learned the expected error, providing explicit uncertainty information would not be of value. It is now appreciated in meteorology, however, that expected forecast errors depend on the current state of the atmosphere and thus vary on a day-to-day basis (see Palmer, 2006 and references therein). Forecast errors are certainly not uniform at all lead times and furthermore, the growth rate of forecast errors varies from 1 day to the next. Therefore the forecast conditions that define the ‘swing’ questions will occur in real weather forecasts.

The current results concern one particular way of presenting uncertainty information (graphical representation of error frequencies) and one particular decision-making task. A previous study using the methods of experimental economics used different formats to represent uncertainty (standard errors and event probabilities) and a cost-loss based decision-making task (Roulston *et al.*, 2006). That study also found that participants provided with uncertainty information made better decisions. From these studies it is not possible to make general claims about the ability of people to understand uncertainty information and apply that information to any decision-making task. These experimental studies, however, do indicate that there exist contexts and formats for which the decision-making of non-specialists is improved by providing information about forecast. Providing information about forecast uncertainty, therefore, can be a means to improve the value of forecasts to users. These studies also illustrate how the ability of users to extract value from a particular type of forecast might be objectively tested. Ensuring that users can extract value from forecasts is arguably as important as verifying the skillfulness of forecasts, a task now regarded as integral to the forecasting process.

Acknowledgements

We wish to thank David Budescu, Naomi Feldman and Ken Mylne for useful comments, Tim Miller for research assistance, and participants at the ESA World Meetings in Rome and ‘Meteorology Meets Social Science: Risk, Forecast and Decision’ in Exeter. This research was funded by the Public Weather Service of the Met Office.

Appendix A: Experimental procedure

The experiments were conducted in the Finance and Economics Experimental Laboratory at Exeter University (FEELE). The sessions occurred on four separate

DOI: 10.1002/met

days – with two sessions being held on each day. The sessions were advertised as being 90 min in length, although all the participants completed the assigned tasks in less than half this time. Participants were paid in cash on leaving the laboratory. Participants were only allowed to take part in one of the eight sessions. The instructions given were:

'You are about to participate in an experiment involving the interpretation of weather forecasts. If you follow the instructions carefully and make wise decisions, you may earn a significant amount of money. Your earnings will depend on your decisions. Participants in this experiment do not interact with one another, so your earnings do NOT depend on the decisions of the other participants. All of your decisions will remain anonymous and will be collected through a computer network. Your decisions are to be made at the computer at which you are seated. Your total earnings from the experiment will be paid to you, in cash, at the end of the experiment.

Please turn off your mobile phone and do NOT attempt to communicate with the other participants. If you have any questions, please RAISE YOUR HAND and someone will come and help you. It is important that you understand the instructions. Misunderstandings may result in lower earnings.

The experiment consists of 20 repeated rounds. In each round you will be shown a graph of the predicted temperature over the course of the next few days, similar to the one below. You will also see two statement about the future weather, called Statement A and Statement B, for example:

Statement A – The temperature at midday on Saturday is above 8 °C
Statement B – The temperature at midday on Tuesday is below 2 °C

Statements A and B may or may not be true. In other words, neither statement may be true, both statements may be true or only one of the two may be true. The statements relate to the ACTUAL temperature. Your task is to study the graph, which shows the FORECAST temperature, and work out which statement is MORE LIKELY to be true.

You will begin each round by looking at the graph and the statements and then choosing ONE of the two statements, either Statement A or Statement B. After you have chosen, you will be told the actual temperatures at the times in question and therefore whether or not each statement is true. If you chosen statement is true, you will receive a payoff of 1 Feele token, otherwise you will not receive a payoff. Feele tokens will be converted into cash at the end of the experiment, at a rate of 50 pence per token.

There are NO trial rounds, so when you start 'Round 1', you will be playing for real money. Before you do so, however, you will be asked to answer some multiple choice test questions. The answers you give to these questions do NOT affect your payment; the idea is for you to get some practice reading graphs.

If you have any question, either now or later on, please raise your hand and someone will come and help you.'

Appendix B: Supplementary questions

1. What is your gender?
   (F) female; (M) male
2. Which of the following sources of information do you consult when you want to find out the weather forecast?
   (a) Internet; (b) TV; (c) Radio; (d) Newspaper; (e) Ask someone else
3. How frequently do you look at the weather forecast?
   (a) Never or hardly ever; (b) Weekly; (c) Every 2 or 3 days; (d) Daily; (e) More than once a day
4. If a fair die is rolled twice, what is the probability that a six will appear on both occasions?
   (a) 0; (b) 1/6; (c) 1/12; (d) 1/36; (e) None of the above

References

AMS. 2002. Enhancing weather information with probability forecasts. *Bulletin of the American Meteorological Society* **83**: 450–452.

Bliss CI. 1934. The method of probits. *Science* **79**: 38–39.

Gigerenzer G, Hertwig R, van den Broek E, Fasola B, Katsikopoulos KV. 2005. "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis* **25**: 623–629.

Joslyn S, Pak K, Jones D, Hunt E. 2007. The effect of probabilistic information on threshold forecasts. *Weather and Forecasting* **22**: 804–812.

Kagel JH, Roth AE (eds). 1995. *Handbook of Experimental Economics*. Princeton University Press: Princeton, NJ, USA; 721.

Mason IB. 2003. Binary events. In *Forecast Verification A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB (eds). John Wiley and Sons: Chichester; 37–76.

Murphy AH. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**: 281–293.

Murphy AH, Lichtenstein S, Fischhoff B, Winkler RL. 1980. Misinterpretations of precipitation probability forecasts. *Bulletin of the American Meteorological Society* **61**: 695–701.

NRC. 2006. *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. National Academies Press: Washington, DC, USA; 124.

Palmer TN. 2002. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society* **128**: 747–774.

Palmer TN. 2006. Predictability of weather and climate: from theory to practice. In *Predictability of Weather and Climate*, Palmer TN, Hagedorn R (eds). Cambridge University Press: Cambridge; 1–29.

Richardson DS. 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society* **127**: 2473–2489.

Roth AE. 1993. On the early history of experimental economics. *Journal of the History of Economic Thought* **15**: 184–209.

Roulston MS, Bolton GE, Kleit AN, Sears-Collins AL. 2006. A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting* **21**: 116–122.

Swets JA. 1973. The relative operating characteristic in psychology. *Science* **182**: 990–1000.

Thornes JE, Stephenson DB. 2001. How to judge the quality and value of weather forecast products. *Meteorological Applications* **8**: 307–314.

DOI: 10.1002/met