A Laboratory Study of the Benefits of Including Uncertainty Information in Weather Forecasts

MARK S. ROULSTON

Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

GARY E. BOLTON

Laboratory for Economic Management and Auctions, The Pennsylvania State University, University Park, Pennsylvania

ANDREW N. KLEIT

Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

ADDISON L. SEARS-COLLINS

Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia

(Manuscript received 28 September 2004, in final form 23 May 2005)

ABSTRACT

Modern operational methods of numerical weather prediction, such as "ensemble forecasting," allow assessments of state-dependent predictability to be made. This means that forecast-specific estimates of the forecast standard errors are possible. Quantitative estimates of forecast uncertainty are often not communicated to the public as it is unclear what the value of this information will be to people who must make weather-dependent decisions. Using laboratory-based methods developed by experimental economists to study individual choice it is found that nonspecialists are able to make better decisions that increase their expected reward while reducing their exposure to risk, when provided with information about the day-to-day uncertainty associated with temperature forecasts. The experimental framework used herein may provide a useful tool for evaluating the effectiveness with which weather forecasts can be communicated to end users.

1. Introduction

Among the general public, weather forecasts are probably the most visible and commonly used type of scientific prediction. Since the 1990s, there has been a substantial research effort in meteorology to estimate the uncertainty associated with weather forecasts on a day-to-day basis. "Ensemble forecasting," in which multiple simulations of the atmosphere are made to reflect uncertainty, has become a standard tool of operational weather forecasting. This type of forecast can provide quantitative information about the impact of uncertainty in both the current state of the atmosphere and the accuracy of the simulation models themselves (Molteni et al. 1996; Houtekamer et al. 1996; Toth and Kalnay 1997; Stensrud et al. 1999; Palmer 2000). The potential value of this extra information to users has been demonstrated in many contexts (AMS 2002; Palmer 2002; Zhu et al. 2002). Despite this, most of the forecasts disseminated by the mainstream media, including those on the World Wide Web, provide little information about uncertainty, with the exception of probabilities of precipitation and a handful of other weather phenomena. One reason for this is that it is not clear whether the general public would be able to make effective use of uncertainty information. Here we provide experimental evidence that nonspecialists can make better decisions when provided with quantitative information about forecast uncertainty.

Corresponding author address: Mark S. Roulston, Met Office, FitzRoy Road, Exeter EX1 3PB, United Kingdom.
E-mail: mark.roulston@metoffice.gov.uk

Claims that including uncertainty information in weather forecasts enhances the value of the forecasts are often based on normative or prescriptive decision-making studies, where a critical assumption is that forecast users will make optimal decisions (Richardson 2001; Mylne 2002). The question of how people actually interpret and use forecast information has been addressed using surveys (Murphy and Winkler 1971; Murphy et al. 1980; Wong and Yan 2002) and descriptive case studies (Stewart 1997) that, while valuable, often lack controls and repeatability.

To assess how people will actually use uncertainty information, we adopted a laboratory approach developed by experimental economists that makes use of financial incentives to motivate participants (Davis and Holt 1992). Laboratory experiments using financial incentives have been used to investigate decision-making problems in economics (e.g., Kagel and Roth 1995) as well as psychology (e.g., Suleiman et al. 2004), business (e.g., see the special issue of *Interfaces,* 2002, vol. 32, no. 5), and anthropology (e.g., Wedekind and Milinski 2000). Laboratory experiments have aided in the design of institutions including the Federal Communications Commission's spectrum auctions and the American Medical Association's Resident Matching Program (Roth 2002). While many of the problems studied in the research cited above do not directly or exclusively involve financial goals, financial incentives are used in the laboratory to induce a preference structure so that we might then study how the decision maker pursues his or her preferences over the problem (as opposed to studying what the preferences might be). Laboratory studies often use college students as a first step in a research program. If students solve the problem at hand well, it increases confidence that the general population can do so; at the very least, it shows that there are circumstances under which people can solve the problem. If students do poorly, the data may provide insight into the difficulty. Either way, the initial experiments provide the baseline for follow-up studies that either pursue elaborations of the problem or check robustness with regard to issues such as the population sampled (e.g., Cooper et al. 1999).

Our approach provides a complement to, and a bridge connecting, existing prescriptive and descriptive research. The basic model of economic agents is that, all other things being equal, they prefer more money to less, and less risk to more (risk aversion) (Arrow 1971). We therefore sought to test the hypothesis that people can use information about forecast uncertainty to make decisions that simultaneously increase their expected rewards and reduce risk. Our approach was to confront human decision makers with an idealized decision-making task. The task chosen was the simple "cost–loss" problem, which has become somewhat of a classic pedagogical example to explain and evaluate the benefits of probabilistic weather forecasts (Richardson 2003).

TABLE 1. The penalties (in tokens) that the participants in the decision-making game had to pay for failing to salt the roads on nights when the temperature fell below freezing.

Failure to salt on . . .	Penalty (tokens)
Sunday night	7000
Monday night	8000
Tuesday night	9000
Wednesday night	10 000
Thursday night	6000
Friday night	4000
Saturday night	2000

2. Experimental design

The participants in the experiment were drawn from the student population at The Pennsylvania State University (Penn State) in the summer of 2004. The students had a median age of 22, with 65% being undergraduates and 35% graduates. Two-thirds of the participants were female. The students were studying a broad range of subjects, with business (31%) and engineering (27%) accounting for over half the participants. Fewer than 4% were meteorology majors.[1]

Participants performed a computer-based task at Penn State's Laboratory for Economic Management and Auctions. The task was a game in which they had to manage a road maintenance company responsible for salting the roads. They were told that the company has a contract with the City for which it is paid 30 000 tokens a month, and salting the roads on a given night would cost the company 1000 tokens. Under the terms of the contract, if the company fails to salt on a night when the temperature falls below freezing, the City charges them a penalty designed to reflect the demand for road usage the following morning. These penalties are shown in Table 1. The game was played for 30 rounds (corresponding to a 30-day month). In each round, which represented a day–night cycle, the participants were presented with a forecast of the overnight minimum temperature. They then decided whether to salt the roads or not. After making their decisions, each participant was informed what the actual temperature was, and whether he or she had incurred any penalties (no further feedback was given).

Participants played the game three consecutive

[1] More details of the subject pool, the full instructions they were given, and the software used to conduct the experiments can be found online (http://www.meteo.psu.edu/~roulston/wx_experiment.html).

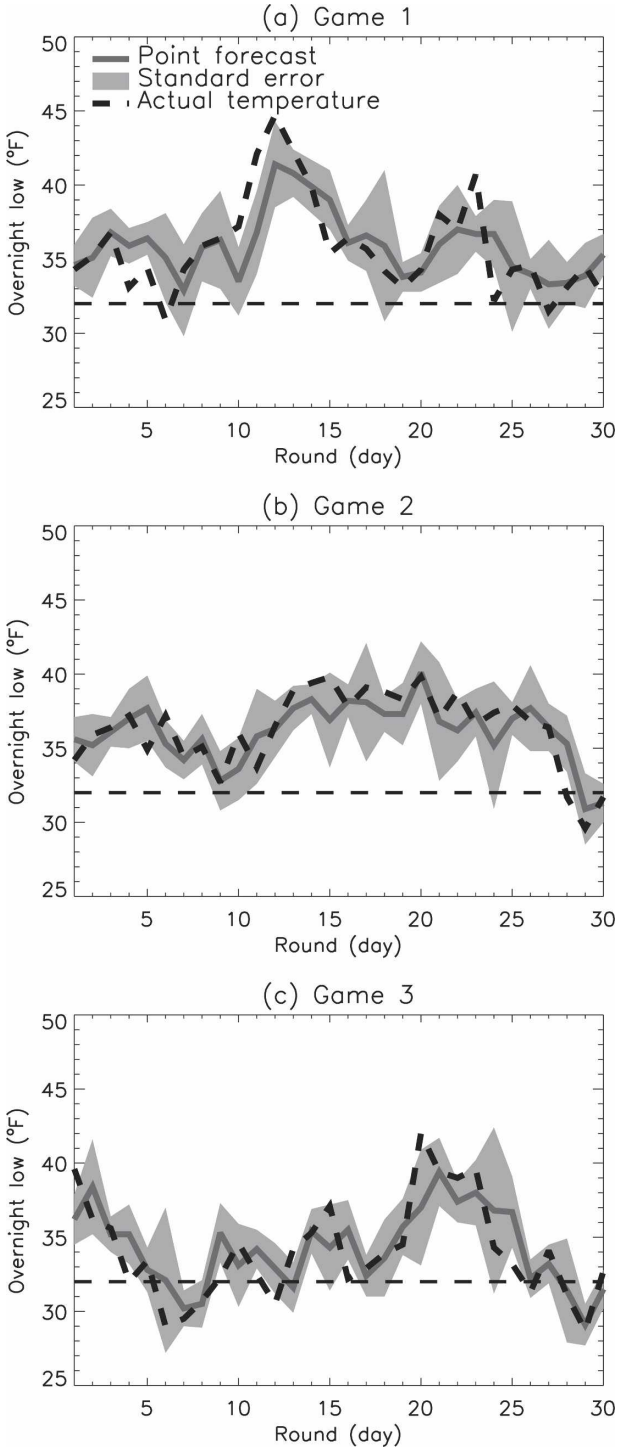

FIG. 1. The forecasts and subsequent overnight low temperatures used in the three games. The forecasts were generated using a simple stochastic model, while the actual temperatures were drawn from normal distributions with mean equal to the point forecast and a standard deviation equal to the standard error that was provided to participants in groups B and C. Participants in group A were only provided with the point forecast and were told that, in general, the average forecast error was about 2°F. Group B members were given the standard error in addition to the point forecast, while those in group C were also explicitly told the probability that the actual temperature would fall below freezing (32°F).

times. For each game, they were given a new stake of 30 000 (earnings from the prior game were not carried over; no additional feedback was given between games). At the end of the session, their remaining tokens for a randomly selected game were converted to U.S. dollars, at a rate of one dollar for every 1000 tokens. Sessions lasted less than 90 min.

The forecast information given depended on the group a participant was assigned to. Forecasts given to group A (17 participants) consisted only of a "point" forecast. Group A participants were told that the error on the forecasts would vary from day to day and on average the error would be about 2°F (1°F = 0.55°C; the United States uses Fahrenheit in most weather forecasts), and that there was a two-thirds chance of the actual temperature falling within the average error, but they were given no other information about forecast errors. This average error was provided to group A on the reasoning that most people are aware that traditional "point" temperature forecasts are not perfectly accurate and have some experience of the magnitude of the likely error. This experience, however, may depend on where the participant comes from as there is geographical variation in the average accuracy of weather forecasts. Explicitly providing group A with the average forecast error was an attempt to control for this effect. Group B (19 participants) was provided with the point forecast, and in addition these participants were given the standard error and told that there would be a two-thirds chance that the actual overnight low would fall inside the range given by the point forecast ± standard error. Group C (15 participants) was given the point forecast, the standard error, and was also explicitly told the probability that the temperature would be below freezing (32°F).

The forecasts, and the subsequent temperature realizations, were synthetically generated using a stochastic model designed to reproduce realistic sequences of forecasts and realizations. For each round the temperature realization was drawn from a normal distribution with mean equal to the point forecast, and a standard deviation equal to the error given to groups B and C. The probabilistic forecasts were therefore "perfect" in the sense that they were reliable, accurately reflecting the uncertainty in the forecast. The standard errors for each round were independent of each other. Figure 1 shows the forecasts and temperature realizations used

in the three games. All three groups were exposed to the same forecasts and realizations in the same order. The only difference was the information content of the forecasts.

Of course, real road managers must make more complex decisions, dependent on more factors than whether it will freeze. The decision-making task we examine is more straightforward, with a sharper focus on the question of whether the participants were able to understand and productively utilize the forecast information presented to them.

To gauge the quality of the decisions made during the experiment, we calculate the expected profit and expected variance of profit for each decision maker, for each game. Expected profit is a better gauge of quality than realized profit in that the former better reflects the value of a decision at the time it was taken—the realized profit is but one possible consequence and is selected by chance. That said, different decisions run different risks and a risk-averse decision maker will be concerned with this too. We measure the amount of risk taken at the time of the decision by the expected variance of the decision. The variance of profit is used widely as a measure of risk (e.g., in finance applications). Moreover, under certain technical conditions, all who are risk averse should prefer a gamble with lower expected variance to one with higher expected variance, so long as the expected payoff of the former is at least as high as the latter (e.g., Rothschild and Stiglitz 1970).

Expected profit is maximized by salting the road whenever $pL > C$, that is, whenever the expected loss associated with not salting exceeds the cost of salting. To pursue such a strategy the participant would have to have an estimate of the value of p. In discussions of binary cost–loss scenarios found in the meteorological literature it is generally assumed that, given a value of p, forecast users would adopt this loss-minimizing strategy. (Observe also that, by varying the value of L, as our experiment does, we are better able to separate those following the optimal rule from those following a simpler rule, as for instance salting when the forecast minus the standard error is below 32°F.)

In a round in which a participant decides to salt, their expected loss is the cost of salting, $C = 1000$ tokens. The expected variance of this loss is zero. If they decided not to salt, and the probability of freezing is p, then their expected loss is pL, where L is the value of that night's penalty. The expected variance of the loss is given by

$$\text{Variance of Loss} = p(L - pL)^2 + (1 - p)(0 - pL)^2 = L^2 p(1 - p). \quad (1)$$

This variance is a measure of the risk that the participant exposed themselves to by choosing not to salt.

3. Results and discussion

Figure 2 displays the results for all the participants, for all three games. Table 2 summarizes mean expected profits and risk exposures, and mean realized profits. Inspection of Table 2 indicates that participants in groups B and C made decisions which both increased their expected profit, and reduced the expected variance of these profits, relative to participants in group A. At the same time, there appears to be little improvement from the additional information provided to group C relative to group B.

Statistical analysis bears the observations from Table 2 out: Regression analysis taking an individual total expected earnings as the dependent variable and an indicator variable for standard error forecast information (hence for groups B and C) as the independent variable yields a positive coefficient that is strongly statistically significant (two-tailed p value) for game 1 ($p < 0.001$, $R^2 = 0.45$), game 2 ($p = 0.025$, $R^2 = 0.10$), and all games ($p < 0.001$, $R^2 = 0.26$) and that is weakly statistically significant for game 3 ($p = 0.076$, $R^2 = 0.06$). Similarly, regressing individual expected variance on an indicator variable for standard error forecast information yields a negative coefficient that is strongly statistically significant for game 1 ($p < 0.001$, $R^2 = 0.46$), game 2 ($p < 0.001$, $R^2 = 0.35$), and all games ($p = 0.002$, $R^2 = 0.18$) but not statistically significant for game 3 ($p = 0.179$, $R^2 = 0.04$). In no case, for either expected profit or variance, does an added indicator variable for probability information (hence group C) produce a coefficient that is statistically significant at any standard level.[2]

In sum, groups B and C's ability to both boost their expected reward, while simultaneously decreasing their exposure to risk, indicates that participants who were

[2] The reported regressions are based on 51 observations (one per participant). The results are identical to those obtained for a general linear model Analysis of Variance (ANOVA) with nested factors. Pairwise t tests do not exploit the combined information about standard error forecast information in groups B and C but nevertheless tell a similar story: Comparing mean expected rewards (two-tailed tests assuming unequal variance) in group A to either group B and C shows improvement with information for game 1 and all games at the 5% level, at the 10% level for game 2, and not significant at standard levels for game 3 or for any comparisons between B and C. The same statements holds for pairwise t tests on mean expected variances save now game 2 comparisons of A with either B or C are significant at the 5% level.

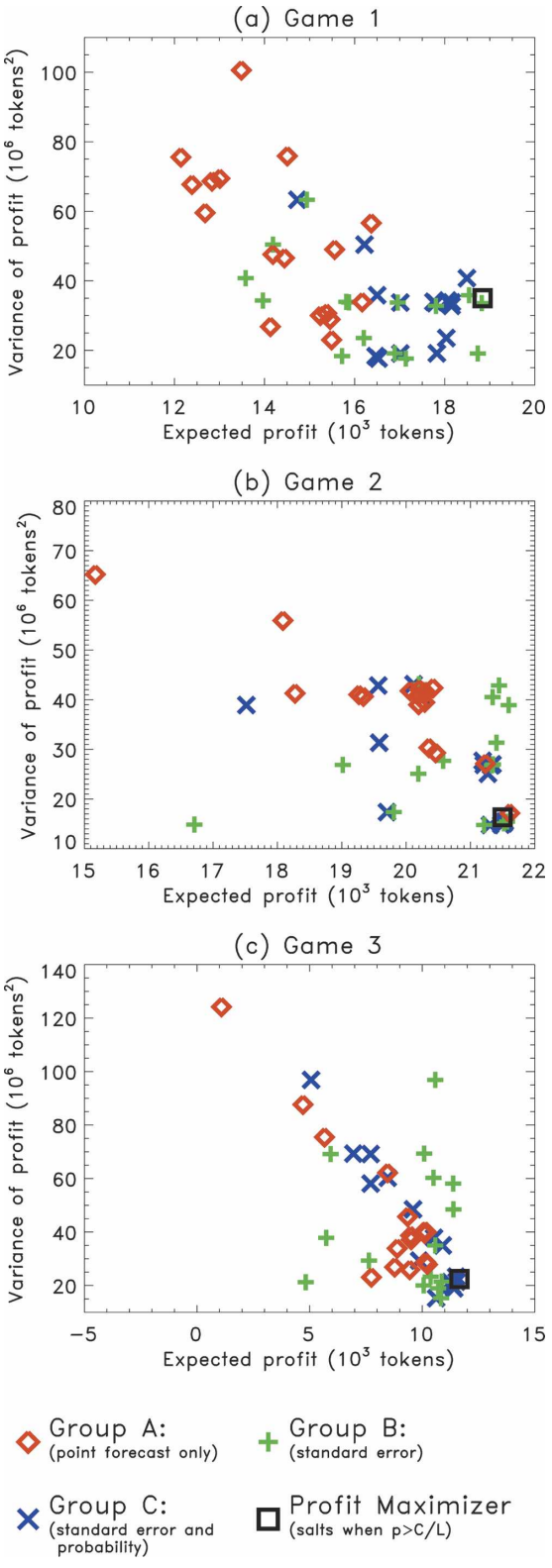

FIG. 2. The total expected profits and the total expected variance of the profit (risk exposure) calculated for the decisions of each participant in each game. Group A only received the point forecast, group B was also given the standard error, and group C had the forecast, the error, and were explicitly told the probability that it would freeze. Points have been slightly displaced where they lie directly on top of other points for clarity. For reference the results that would be obtained by a "profit maximizer" have been included. A profit maximizer would opt to salt whenever the probability of freezing exceeded the "cost–loss" ratio (cost of salting–penalty for not salting).

provided with information about forecast uncertainty were able to make decisions that increased their utility. There is, however, little evidence that giving the probability of freezing, in addition to the standard error of the forecast, improves decisions.

The reduction in the differentiation between groups in game 3 is partly a consequence of the forecasts with which the participants were presented. A larger proportion of forecasts that are well above or below freezing means that there are fewer rounds in which participants who are actively using uncertainty information will act differently from those who either do not have such information, or who are not making use of it. For example, a participant adopting the optimal strategy of salting whenever the probability of freezing exceeds the cost–loss ratio and another participant who merely salts whenever the point forecast falls within 2°F of freezing would make different decisions in 6 of the 30 rounds in the first game, 2 rounds in the second game, and 3 rounds in the final game. The fact that there are more occasions in the first game when participants with different information can be expected to act differently explains the greater statistical significance of the differentiation between groups in this game. The differences in group differentiation in the different games, however, may also be partly attributed to a learning effect as the experiment progressed. The impact of learning from experience (from repeated exposure to the decision problem) could be estimated by running the three games in a different order, although this was not done in this experiment. It is also possible that participant fatigue played a role, although such was not evident to the experiment's monitors.

Since group C was provided with the standard error and the probability of it freezing, it was not possible to ascertain the benefit of providing the probability alone, without the standard error. In a practical application, however, it would generally be easier to provide the standard error, since this is a property of the forecast, than the probability of some application-specific criterion occurring, since the relevant application would be user dependent. The question of the value of the probability of it freezing alone to decision makers could be

TABLE 2. A summary of the mean expected profits, the mean expected variances of these profits, and the mean realized profits for each group, in each of the three games. Group A was given only given a point forecast, group B was also given the standard error, while group C was given the point forecast, the standard error, and the probability of freezing.

Mean expected profit (000s of tokens) (with std dev)					
Group	Participants	Game 1	Game 2	Game 3	All games
A	17	14.33 (1.35)	19.75 (1.41)	8.46 (2.46)	14.18 (4.98)
B	19	16.55 (1.63)	20.60 (1.20)	9.61 (2.04)	15.59 (4.86)
C	15	17.26 (1.04)	20.61 (1.13)	9.63 (2.03)	15.84 (4.86)
Mean expected variance of profit (millions of tokens) (with std dev)					
Group	Participants	Game 1	Game 2	Game 3	All games
A	17	52.38 (21.75)	39.93 (10.75)	47.02 (26.50)	46.44 (20.97)
B	19	35.39 (17.00)	25.02 (11.12)	35.61 (24.44)	32.01 (18.69)
C	15	32.45 (12.78)	27.16 (9.72)	41.65 (24.38)	33.76 (17.55)
Mean realized profit (000s of tokens) (with std dev)					
Group	Participants	Game 1	Game 2	Game 3	All games
A	17	17.71 (2.31)	23.18 (1.29)	5.53 (2.94)	14.53 (7.78)
B	19	18.47 (2.09)	22.05 (1.93)	5.89 (4.03)	14.53 (7.53)
C	15	19.93 (1.75)	22.33 (1.80)	3.67 (2.82)	14.69 (8.65)

tested by a similar experiment in which one group only receives this information.

4. Summary

We have presented a laboratory study that illustrates how methods of experimental economics may be used to objectively assess forecast communication effectiveness.

The results of our study indicate that, when making decisions in the face of uncertainty, people provided with information about uncertainty improved their decision making relative to people who lacked this information: they increased their expected profits while decreasing their risk exposure. In the context of the decision-making task used in this experiment, providing the standard error yielded most of the improvement in decision making. Stating the probability most relevant to the task—the probability of freezing—in addition to the standard error produced did not yield a statistically significant improvement over providing the standard error alone. The study illustrates that techniques developed by economists to study individual choice in the laboratory may provide a useful framework for objectively evaluating the effectiveness with which weather forecasts are communicated to users. Such evaluation is crucial as the socioeconomic value of a skillful weather forecast can be diminished by ineffective communication to decision makers.

While this illustrative study used students, larger samples of actual forecast users could be used to help in the design of forecast formats. The decision-making task could also be designed to test users' understanding of more complex forecasts, such as spatially and temporally distributed information. Laboratory experiments could also compare the value users place on the uncertainty information (by charging them for it) and the theoretical value of the information.

Acknowledgments. This experiment was conducted at the Laboratory for Economic Management and Auctions, at PSU, with funding from a Wilson Research Initiation Grant from the College of Earth and Mineral Sciences. One of the authors (ALSC) was funded by PSU's Summer Research Opportunities Program.

REFERENCES

AMS, 2002: Enhancing weather information with probability forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 450–452.

Arrow, K. J., 1971: *Essays in the Theory of Risk-Bearing*. Markham Publishing, 278 pp.

Cooper, D., J. H. Kagel, and W. Lo, 1999: Gaming against managers in incentive systems: Experimental results with Chinese students and Chinese managers. *Amer. Econ. Rev.,* **89,** 781–804.

Davis, D. D., and C. A. Holt, Eds., 1992: *Experimental Economics*. Princeton University Press, 572 pp.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Kagel, J. H., and A. E. Roth, Eds., 1995: *Handbook of Experimental Economics.* Princeton University Press, 721 pp.

Molteni, F., R. Buizza, and T. N. Palmer, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Murphy, A. H., and R. L. Winkler, 1971: Forecasters and prob-

ability forecasts: Some current problems. *Bull. Amer. Meteor. Soc.,* **52,** 239–247.

——, S. Lichtenstein, B. Fischhoff, and R. L. Winkler, 1980: Misinterpretations of precipitation probability forecasts. *Bull. Amer. Meteor. Soc.,* **61,** 695–701.

Mylne, K. R., 2002: Decision-making from probability forecasts based on forecast value. *Meteor. Appl.,* **9,** 307–315.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.,* **63,** 71–116.

——, 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.,* **128,** 747–774.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.,* **127,** 2473–2489.

——, 2003: Economic value and skill. *Forecast Verification: A Practioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley, 165–188.

Roth, A. E., 2002: The economist as engineer: Game theory, experimental economics and computation as tools of design economics. *Econometrica,* **70,** 1341–1378.

Rothschild, M., and J. E. Stiglitz, 1970: Increasing risk: 1. A definition. *J. Econ. Theory,* **2,** 225–243.

Stensrud, D. J., H. E. Brooks, J. Du, E. Rogers, and M. S. Tracton, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

Stewart, T. R., 1997: Forecast value: Descriptive decision studies. *Economic Value of Weather and Climate Forecasts*, R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 147–182.

Suleiman, R., D. V. Budescu, I. Fischer, and D. M. Messick, Eds., 2004: *Contemporary Psychological Research on Social Dilemmas.* Cambridge University Press, 422 pp.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Wedekind, C., and M. Milinski, 2000: Cooperation through image scoring in humans. *Science,* **288,** 850–852.

Wong, T. F., and Y. Y. Yan, 2002: Perceptions of severe weather warnings in Hong Kong. *Meteor. Appl.,* **9,** 377–382.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 73–83.