EXPERT JUDGMENT AND CLIMATE FORECASTING: A METHODOLOGICAL CRITIQUE OF "CLIMATE CHANGE TO THE YEAR 2000"

THOMAS R. STEWART and MICHAEL H. GLANTZ
Environmental and Societal Impacts Group, National Center for Atmospheric Research,*
P.O. Box 3000, Boulder, CO 80307, U.S.A.

Abstract. The National Defense University's study of climate change to the year 2000 was based largely on the judgments of the members of two expert panels. Although the study has been widely distributed and apparently read by policy makers in the U.S. and abroad, the method of eliciting and analyzing expert judgment has not been critically reviewed. This paper uses the literature on judgment and subjective probability to evaluate the expert judgment methods used in the study.

1. Introduction

This paper analyzes the methodological underpinnings of a study of future climate change. The study, undertaken by the United States' National Defense University (NDU), was designed to elicit expert opinion about the probability of different climate futures. Its conclusions suggested that climate in the next twenty years would be similar to that of the recent past.

There is evidence to suggest that the NDU study received widespread attention in the disciplines related to atmospheric sciences and to agriculture, among the highest level of policymakers in the United States, and perhaps in other countries as well. Yet to date its methodological underpinnings, not to mention the implications of its substantive findings, have not been critically assessed. It has been reviewed benignly in several places (Kraemer, 1978; U.N. FAO, 1979; *Science News*, 1978; Sellers, 1979; McKay and Williams, 1982), with the reviews merely describing in brief the study's findings.

Because of the widespread dissemination and apparent use of the NDU study's findings by government policy planners, it is important to examine closely the basis for its conclusions. In this paper we examine the validity and, therefore, reliability of the study's findings from a methodological perspective. Specifically, we examine the method used to elicit judgments from the expert group in the first part of the NDU study and we identify and describe some method-induced biases likely to have affected the study's results. First we briefly review the historical setting of the study and summarize its results.

2. Historical Setting

The 'Green Revolution', based on the development of the high-yield varieties of grain in

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Climatic Change 7 (1985), 159–183. 0165–0009/85.15.

the 1960's, fostered new hope that technology could buffer developing as well as developed societies from the vagaries of weather. Sharply increasing productivity on land was being reinforced by expectations that parallel increases could also be derived from the sea. Not only were abundant, but yet untapped, living marine resources viewed as a reservoir of protein that could supplement protein derived from terrestrial resources, but new technologies such as factory ships were being touted as the wave of the future. It was also in the 1960s that hypotheses were being presented about ways to slow high population growth rates that were widespread throughout the Third World. Thus, by 1970 an optimistic feeling was emerging that world food problems were on the verge of being brought under control (Brown, 1970).

This feeling changed by the mid-1970s, following a period of climate-related shocks to agronomists, agricultural planners, and farmers. After decades of what in retrospect has been referred to as benign worldwide weather conditions, climate anomalies and subsequent societal reactions to them led to a revised view of food production. In the early 1970s there were droughts in the U.S.S.R., China, Central America, the West African Sahel, East Africa, and Australia. The resulting crop failures sharply increased demand for available grain stocks. The rapid depletion of grain stocks, in turn, caused a sharp increase in grain prices in the international marketplace (e.g., Hopkins and Puchala, 1978). There was a concurrent sharp decline in fish landings (this was also climate related) that generated a new skepticism about the ability of the seas to supply with new technology the hundreds of millions of metric tons of fish per year as had been suggested in earlier fish stock assessments. Some scientists insisted that living marine resources were already overexploited.

A new political perspective on American food exports was formed. Secretary of Agriculture Earl Butz, among others, began to boast that food had become an important foreign policy tool in the American diplomatic negotiating kit, the strength of which might offset the inconveniences, for example, of the threat of additional sharp increase in oil prices. Reinforcing Butz's contention were the scores of headlines referring to the politics of droughts and of starvation, the economics of grains, and so forth. The U.S. Central Intelligence Agency, too, showed a new interest in climate and produced a number of public documents related to climate and food, and the need for climatological research within the intelligence community (U.S. CIA, 1974).

Fueling the new-found political interest in climate was the debate that was then raging in journals, meetings, conferences, and ultimately in the popular press about whether the global climate was changing. Was it becoming warmer (with the appearance of book titles like *Hothouse Earth*) or were we moving toward an Ice Age (with book titles like *The Cooling*)? Still others, seeing the issue surrounded by a high degree of scientific uncertainty, suggested strategies to cope with climate variability and change regardless of direction (e.g., *The Genesis Strategy: Climate and Global Survival*). Selective evidence to support each of these contending views could be found in the scientific literature (e.g., it was 'time' for a new ice age, or thermal pollution or increased CO_2 loading of the atmosphere would ultimately lead to a global warming). This debate commanded the attention of Congress (e.g., Congressional Research Service, 1976).

To policymakers the debate about future climates was a confusing one. Each side seemed to have good factual support for its contentions about what future climate might be like, partly because scientists tended to draw attention only to information that supported their views or that descredited opposing views. Policymakers began to question why the U.S. Department of Agriculture staff had not prepared contingency plans to cope with the impacts on U.S. agriculture of any of the various future climate scenarios that were being debated in the newspapers and scientific reports.

Out of this setting a call for a study on climate emerged. Since policy decisions could not wait for the resolution of the scientific debate, the National Defense University (NDU) initiated a study to summarize existing expert judgments with regard to future global climate. To that end the NDU published three reports based on a study of climate changes and their impacts on global agricultural productivity. These reports, issued over a span of six years, dealt with three topics in series. The first report, *Climate Change to the Year 2000* (1978), was to "define and estimate the likelihood of changes in climate during the next 25 yr, and to construct climate scenarios for the year 2000" (p. xvii). The findings of the first report served as the input into part two, *Crop Yields and Climate Change* (1981), designed to assess crop yields for the different scenarios. The findings of part two, in turn, served as the input into the third report, *The World Grain Economy and Climate Change to the Year 2000* (1983).

The NDU study was supported financially by the Department of Defense at an estimated cost of about $100 000, excluding approximately 9 person-years of contributed research by a small, multidisciplinary staff assigned by the Department of Defense, the U.S. Department of Agriculture, and the National Oceanic and Atmospheric Administration, as well as contributions by numerous experts and advisers who received nominal honoraria. The $100 000 was primarily for the Institute for the Future to assist in the development of the study design (Glantz *et al.*, 1982).

3. Results of the Study

The conclusions of all three parts of the study downplay the impact of future climate change on world food policy. The results of the first part of the study were summarized as follows: "The derived climate scenarios manifest a broad range of perceptions about possible temperature trends to the end of this century, but suggest as most likely a climate resembling the average for the past thirty years" (p. iii). It was further stated in the report's summary that "the salient finding is that the likelihood of catastrophic climatic change by the year 2000 is assessed as being small" (p. xvii).

With respect to the second part of the study, it was concluded that "the most likely climate change [according to this study], a slight global warming, . . . was found to have negligible effects on 15 'key' crops", and that "the potential crop-yield effects of technological change are judged to be severalfold larger than the effects of the posited climate changes" (NDU, 1980, p. v).

In the third part "the main conclusion . . . is that through the end of this century the world is very unlikely to face climate changes of such magnitude as to affect the world food situation significantly" (NDU, 1983, p. 4). It was also concluded that *"The significance of this study* [all three parts] *is that the United States can consider its proper role in the world food situation without great concern that climatic changes during the rest of this century will upset its calculations"* (p. 4, emphasis original). Although the NDU reports contain appropriate warnings regarding the uncertainty of their conclusions, the reports emphasize the main conclusion but not the uncertainty that surrounds it. Clearly these are potentially important findings with implications for U.S. foreign economic and political policies for the next few decades.

The importance of NDU's study and its impact on policymakers can only be surmised indirectly from the scientific and popular literature. For example, the first part of the NDU study was used extensively in the climate sections of *The Global 2000 Report to the President*, prepared for President Carter (Council on Environmental Quality, 1980) and, according to the introduction to the third part of the NDU study, was used by policymaking groups in the Reagan Administration. During the World Climate Conference held in Geneva in 1979, the NDU study was one of the few documents available in great quantity to conference participants. One could also assume that *The Global 2000 Report to the President* has been read by policymakers in several countries. According to the State Department (Tipton, 1984, personal communication), more than 5000 copies were distributed to governments with which the United States has diplomatic relations. It has also been translated by German and Spanish publishing houses.

4. Summary of the Method

The first and second parts of the study relied mainly on expert judgments obtained through mailed questionnaires. Although the content of the questionnaires used in the two parts differed, as did the composition of the expert groups, the methods used to elicit judgments and analyze the data were the same in most important respects. We will examine in detail only the method used in the first part because (a) the results of the first part were the most widely distributed, (b) those results were used in the second and third parts, and (c) our major comments on the first expert judgment study apply equally to the second.

The justification for the expert judgment approach used in the first and second studies was described by the NDU study authors as follows:

> The causes of global climate change remain in dispute. Existing theories of climate, atmospheric models, and actuarial experience are inadequate to meet the needs of policymakers for information about future climate. In the long run, research may lead to reliable forecasts of climate. For the present, however, policymakers have no recourse but to heed expert judgments – subjective and contradictory though they may be – about future world climate and its effects on agriculture and other sectors of the economy. Informed, expert judgments on the likelihood of change, or the odds for a repetition of some event, are useful to the decisionmaker weighing the costs, benefits, and risks of alternative policies (NDU, 1978, p. ix).

The method developed by the Institute for the Future for the NDU study, as described in the 1978 report, is summarized below. All quoted material is from the report.

(1) A panel of experts was selected by the research team with the assistance of an Advisory Group consisting of representatives of government, universities, and other research institutions. The experts "were selected both for their competence in the field of climatology[1] and for the diversity of views which they represented" (p. 1).

(2) Questionnaires were constructed and mailed to 28 panelists, 24 were returned, and 21 contained quantitative information" (p. 1). Questions were asked on the following topics: average global temperature, average latitudinal temperature, carbon dioxide and turbidity, precipitation change, precipiation variability, mid-latitude drought, outlook for the 1977 crop year, Asian monsoons, Sahel drought, and length of growing season. With one exception (carbon dioxide and turbidity), the questions were designed to elicit subjective probability distributions.

(3) Judgments of the 19 panelists who answered question 1 (global mean temperature), were combined into an overall probability distribution by means of a weighted average. The weights were based on expertise ratings of the individual panelists.

(4) Respondents were assigned to 'global mean temperature categories' based on their responses to question 1. The categories, which were selected to make the probabilities symmetric, are described in Table I.

TABLE I: Global mean temperature categories[a].

Category	Change in temperature from the present by the year 2000[b]	Probability	Number of panelists assigned to category
Large cooling	0.3° to 1.2° colder	0.10	1
Moderate cooling	0.05° to 0.3° colder	0.25	3
Same as last 30 yr	0.05° colder to 0.25° warmer	0.30	10
Moderate warming	0.25° to 0.6° warmer	0.25	4
Large warming	0.6° to 1.8° warmer	0.10	1

[a] Based on Tables 1–3 and 1–4 in NDU (1978)
[b] All temperatures are in °C. Present temperature is defined as the mean temperature for 0–80° N between 1965 and 1969.

(5) For the other questions (except those about precipitation), responses of panelists within a category were combined. The results were presented in narrative and tabular form as global mean temperature 'scenarios'. The description of these scenarios constitutes the bulk of the report. Precipitation questions were treated in a different way because panelists were asked to judge precipitation effects separately for specified temperature ranges.

5. Problems of Elicitation: The Questionnaire

A major problem to be addressed in any expert judgment study is how to elicit judgments

[1] Although the NDU report refers to the panelists as climatologists, in fact, most were not trained in climatology.

from the experts. Should a questionnaire be sent by mail or should the expert panel be interviewed in person or by telephone? Should the approach be highly structured in order to assure that the judgments of experts can be compared, or should it be open-ended to allow the experts the maximum flexibility in expressing their opinions? How should the questions be formulated in order to avoid introducing bias or errors into the experts' judgments? Is a single elicitation round sufficient, or should the experts be given feedback about the judgments of other members of the group and allowed to revise their judgments? Should communication among the experts be encouraged or discouraged? If communication is allowed, how should it be structured and organized? The best choices depend upon the context of the study and its purpose. The NDU study relied on a mailed questionnaire with no feedback to respondents or opportunities for revision, and communication among experts was not encouraged.

Although the use of a mailed questionnaire is an inexpensive way to elicit expert judgment, face-to-face interviews conducted by interviewers who are familiar with probability assessment methods are preferable (Spetzler and Stael von Holstein, 1975). Interviews permit checks on the clarity of instructions, provide the opportunity to check the consistency of responses, and can allow the respondent to examine and revise his or her answers. An experienced interviewer can detect sources of error that go unnoticed if a questionnaire is used.

If a questionnaire is used, then the structure and form of the questions included are critically important. Indeed, the form of the questions may largely dictate the results of the study. One of the most pervasive results in survey research and psychological research is that changes in the wording or structure of a question can have a major impact on the answers obtained (see Schuman and Presser, 1981, for a summary, and Tversky and Kahneman, 1981, for some relevant examples). This means that the expertise of the questionnaire designers could influence the outcome of the study as much as or more than the expertise of the respondents.

Since the question on global mean temperature (see Figure 1) was the basis for constructing the climate scenarios of the first report which were also used in the second and third reports, that question will be examined in detail. Many of the comments that are made about this question would apply equally well to the others in the questionnaire. In the following sections we examine six possible sources of bias or error in the questionnaire responses: response mode, calibration, anchoring, instructions, format, and cognitive limitations.

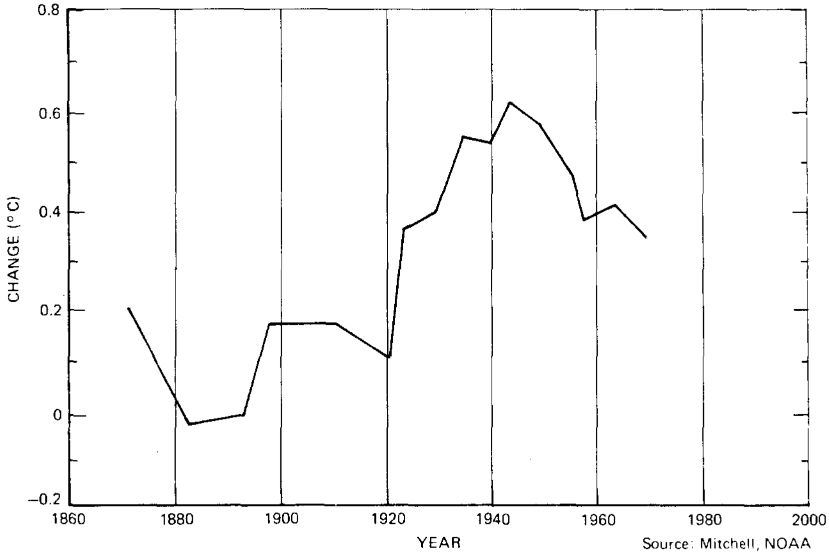
5.1. Response Mode

Information about subjective probabilities may be elicited from respondents in different formats, or 'response modes'. The respondents may be asked to supply numerical information about their subjective probability distributions *directly*, or their distributions may be inferred *indirectly* from other types of responses, e.g., their choices among certain carefully constructed bets (Spetzler and Stael von Holstein, 1975). If the response mode is direct, then the respondent may be asked to estimate the probability of a specified level

I. GLOBAL TEMPERATURES

Shown below is a historical record of changes in the annual mean temperature during the past century for the latitude band, 0-80°N.

CHANGE (°C) IN ANNUAL MEAN TEMPERATURE, 0-80 °N. LATITUDE



Source: Mitchell, NOAA

On the graph shown above, indicate your estimate of the general future course of the change in mean annual temperature (for 0-80° N.Lat.) to the year 2000 by:

- drawing a temperature change path to the year 2000 so that you estimate only 1 chance in 10 that the path could be even lower
- drawing a change path to the year 2000 so that you estimate an even chance that the path could be either lower or higher
- drawing a change path to the year 2000 so that you estimate 1 chance in 10 that the path could be higher

Fig. 1. Question 1 from the NDU Climate Questionnaire: Global Temperatures (Source: NDU, 1978, p. 61).

of a variable (the fixed-value method) or to estimate the level of a variable corresponding to a specified probability (the fixed-probability method). The fixed-probability direct response mode was used in the NDU study. The respondents were asked to estimate directly temperature changes corresponding to specified probabilities (e.g., one chance in ten). If the fixed-value method had been used, temperature changes would have been specified and the respondents would have been asked to estimate the probabilities of those changes.

Different response modes often produce different results (Slovic and Lichtenstein, 1971). The effects of the response-mode chosen are highly task-specific and the validity of different response modes under various conditions has not been adequately investigated

(Wallsten and Budescu, 1983, p. 167). Therefore, probability assessors should use more than one response mode and then reconcile the discrepancies between the different modes (Spetzler and Stael von Holstein, 1975; Hogarth, 1980). Since this was not done in the NDU studies, *the extent of bias introduced by response mode is unknown.*

5.2. Calibration

One of the most intensively studied properties of subjective probabilites is called 'calibration'. If a person is perfectly calibrated, then the events to which he or she assigns probability p will, in fact, occur a proportion p of the time. For example, if a well calibrated weather forecaster predicts a 20% chance of precipitation over a number of days, it will in fact rain or snow on approximately 20% of those days. Substantial differences between relative frequencies and subjective probabilities indicate poor calibration.

A pervasive finding in studies of calibration is that people are overconfident; that is, they report more certainty than they should (Lichtenstein *et al.*, 1982). Overconfidence has been found in both expert and lay judges, and in predictions of future events as well as general knowledge questions (Fischhoff and MacGregor, 1982). Overconfidence is greatest for extremely difficult tasks (Lichtenstein *et al.*, 1982).

Furthermore, attempts to 'debias' or reduce overconfidence have met with little success (Fischhoff, 1982). Four studies have, however, demonstrated some success in overcoming overconfidence. Seaver *et al.* (1978) found less overconfidence with the fixed-value response mode than with the fixed-probability mode (which was used in the NDU study). Tversky and Kahneman (1974) report a similar result. Koriat *et al.* (1980) found that calibration improved when respondents were asked to write reasons that contradicted their chosen answer. Fischhoff and MacGregor (1982) found a similar, though weaker and more ambiguous effect. In the NDU study, respondents were asked only to state their lines of reasoning for their answers. Similar instructions in the Koriat *et al.* (1980) study produced no improvement in calibration.

Although the bulk of the research on calibration indicates that overconfidence is pervasive and difficult to overcome, it is not universal. Weather forecasters, in particular, can make probabilistic forecasts that are exceptionally well calibrated (Murphy and Winkler, 1974; Winkler and Murphy, 1979). This is probably due to their training and experience and to the feedback they receive during years of making forecasts and observing results. When predicting unfamiliar events using a novel response method, calibration of weather forecasters is somewhat impaired (Wallsten and Budescu, 1983, p. 163). Dutch weather forecasters who were not accustomed to making probability forecasts made poorly calibrated forecasts during the first year of a probability forecasting experiment in The Netherlands (Daan and Murphy, 1982). After detailed feedback regarding their performance during the first year of the experience, however, the calibration of forecasts improved markedly (Murphy and Daan, 1984).

Even though predicting climate changes has some of the characteristics of weather forecasting which are believed to make good calibration possible (see Murphy and Brown, 1984, for a discussion of the characteristics of weather forecasts), there are some important differences between climate change forecasts and weather forecasts. In particular, long-range climate forecasts are not made routinely and there is little opportunity for feedback about the quality of such forecasts. Furthermore, the response mode used in the NDU study was not one that was familiar to the panelists. There is little reason to believe, therefore, that the responses of the NDU panelists are calibrated better, or worse, than the responses of other people.

Calibration depends on characteristics of the judge and of the particular quantity or event being predicted, and it can be affected by training, experience, response mode, and instructions. Too little is known about the factors that affect calibration, and how those factors interact, to predict how well or how poorly calibrated a particular person's responses to a particular task will be. If we were to venture a specific prediction of how well calibrated the NDU panelists' predictions of global mean temperature were, we too would be guilty of overconfidence.

Since it is not possible to measure the calibration of probabilistic forecasts of a unique event such as global mean temperature, the effects of calibration on the results of the NDU study cannot be determined. However, *the panelists were given no instructions with regard to the calibration problem and received no training or other aid in calibrating their responses. Furthermore, the report contains no warning about the possibility that the results were affected by panelists' overconfidence which, if present, would produce probability distributions that are narrower than they should be; that is, the probabilities of extreme temperature changes could be greater than those reported in the study.*

5.3. Anchoring

'Anchoring' is another well known psychological phenomenon that may have affected the responses to question 1. People often use some value as a starting point which they then adjust by an appropriate amount in order to arrive at their judgment. For example, a person might arrive at an estimate of the population of the United States in the year 2000 by starting from an estimate of the current population and adjusting it upward. Research has shown that such starting points, or anchors, can have a strong influence on the final judgment because people typically do not adjust their response far enough away from the anchor (Tversky and Kahneman, 1974). Tversky and Kahneman (1974) suggest that anchoring may be one of the causes of the overconfidence exhibited in calibration studies. They predict that the fixed-probability response mode which was used in the NDU study encourages the anchoring that produces overconfidence, and they present experimental results supporting their prediction.

There are two possible anchors evident in question 1. First, the respondent is instructed to draw a temperature change path beginning at the end point of the annual mean temperature graph provided. Thus, the end point of the graph becomes an important anchor. Second, the respondents' first responses act as anchors for later responses. If the respondent drew change paths in the order requested, then the low change path could act as an anchor for the median change path and the low and median change paths would be available for the high change path. *The result of anchoring in question 1* *could be a bias toward a narrowing of the probability distribution.*

5.4. *Instructions*

Questions involving probabilities are particularly susceptible to effects of question format because most people lack experience and training in quantifying their uncertainty in terms of subjective probability. While the scientists on the expert panel may have been more familiar with probability than most people, they were not experienced in reporting *subjective* probability. It has been found that even experienced researchers who routinely base conclusions about their research results on probability theory are susceptible to biases (Tversky and Kahneman, 1971). Although biases are persistent and cannot always be prevented in a questionnaire, no matter how detailed the instructions (Hogarth, 1975), *the absence of any instructions regarding subjective probability is a serious flaw in the questionnaire.*

5.5. *Format*

Since the response format for question 1 was graphic, rather than numeric (another distinction that could affect results), the responses are effectively bounded by the range of the vertical temperature scale. *The choice of a range for the vertical temperature scale from –0.2 °C to 0.8 °C, relative to the 1880–1884 zero reference base, could therefore have limited the range of responses obtained.* The report contains no discussion or justification for the decision to limit responses to a 1 °C range centered on 0.3 °C.

There are other possible changes to the format of question 1 that could have significant, though unpredictable, effects on results. For example, the annual mean temperature graph could have been presented with confidence bands illustrating the variability in yearly estimates of mean temperature. The extreme probabilities could have been set at '1 chance in 20' or '1 chance in 100'. The words 'even' and 'only' could have been used in both the first and third parts of the question, or in neither part. *Even such seemingly trivial variations in format can have an effect on results.*

5.6. *Cognitive Limitations*

Judgment and decision research has identified limitations in human information processing capacity that can affect the validity of complex judgments (e.g., Newell and Simon, 1972; Hammond *et al.*, 1975, 1980; Slovic *et al.*, 1977; Einhorn and Hogarth, 1981). The questions in the NDU study demand difficult judgments, so difficult that some experts declined to participate and others who did participate expressed reservations about doing so (NDU, 1978, p. 2). Each question required the expert to integrate many items of information in order to arrive at a judgment. Much of the information is of uncertain validity, different items are interdependent and sometimes indicate conflicting conclusions, and the relations between individual items of information and the judgment to be made are complex and generally non-linear and non-additive. It is in just such situations that humans have been found to perform most poorly. The same research on judgment that has identified some of the limitations has also produced useful aids to better judgment (Raiffa, 1968; Hammond *et al.*, 1977, 1982; see also Hogarth, 1980 for a summary). However, such aids were not used in the NDU study. *The final, and perhaps most important, concern to be raised regarding the questionnaire is that the experts received no assistance in coping with cognitive limitations when answering the questions.*

5.7. Summary

Three interrelated problems have been identified in research on subjective probability assessment: (a) assessments require judgments that exceed the limits of human cognitive ability, (b) people typically give biased responses, and (c) the responses are sensitive to seemingly inconsequential changes in question format and wording. Unless safeguards are adopted, these problems can lead to seriously misleading results.

Since no such safeguards were used in the NDU study, we conclude that the responses of the expert panel were susceptible to cognitive limitations, psychological biases, and the influences of question format and wording. Because the study used a single method and included no checks for consistency or bias, it is impossible to determine the magnitude of these effects. *We can conclude, however, that the combined effects of calibration, anchoring, and limiting the response range possibility resulted in a narrowing of the resulting distributions of temperature changes; that is, the results could indicate more certainty about future temperature change than would be found had more sophisticated procedures been used.*

6. Problems of Aggregation: Averaging the Responses

It is not surprising that the experts in this study disagree. The same lack of knowledge about climate change that produced the need for a study that relied on expert judgment virtually assures that a group of 'diverse' experts will disagree. There are several methods for coping with disagreement among experts (see Hogarth, 1977). One method is to simply report the disagreement, aided perhaps by a 'clustering' of experts into groups with similar opinions. A second method is to attempt to resolve the disagreement through feedback and structured discussion. A third method is analytic aggregation of opinion. The NDU study used an analytic aggregation procedure based on a weighted average.

Regardless of the aggregation method used in an expert judgment study, the results of individual panelists should always be reported so that the reader can form his or her own opinion regarding the diversity or homogeneity of the panel (see National Academy of Sciences, 1975, for an example). Since the NDU reports do not include results for individual panelists, the reader is forced to rely on the aggregated results. Different aggregation methods can produce different results, however, and the NDU reports do not present the data that would be needed to determine precisely how the aggregation method chosen affected the results. In the following sections, therefore, we rely on the description of the method and the relevant literature in order to evaluate the aggregation procedure and estimate its effect on the results.

The procedure for aggregating responses to question 1 for the year 2000 consisted of four steps:

(1) The responses of each panelist were used to estimate a cumulative probability distribution.

(2) The cumulative probability functions were converted to probability density functions.

(3) The probability density functions were weighted by multiplying them by expertise weights. The expertise weights were based on an average of self- and peer-ratings of expertise. The weights were either 4 (expert), 2 (quite familiar), 1 (familiar), or 0 (casually acquainted or unfamiliar).

(4) The weighted density functions were added together and the sum was normalized by dividing by the sum of the weights.

The second and fourth steps are routine computations. The first and third steps involve methodological choices and will be discussed below.

6.1. Approximating the Cumulative Probability Function

Each expert's responses to question 1 were, for each year between 1967 and 2000, direct estimates of the 0.10, 0.50, and 0.90 fractiles of his cumulative subjective probability function. It was necessary to estimate the complete cumulative probability function by fitting a curve. A piecewise linear function was used for this purpose (see NDU, 1978, pp. 7–8). Although the straight line is the simplest curve, it is not necessarily the best for this purpose. For example, a logistic function might have been used. This curve is often used in probability work because of its mathematical properties and because its S-shape provides a good approximation to many types of data. Other functional forms, such as the normal or cubic also could have been used to fit the data. The appropriate function depends on both the fit to the data and on assumptions about the underlying process to be represented by the function. *In general, the linear approximation produces narrower (more certain) probability distributions than do other functions that might reasonably have been used.* (See Appendix A.)

6.2. Weighting the Experts

The weighting of experts was based on the intuitively appealing notion that panelists who are more expert with regard to a specific question should have more influence on the results of the aggregation (see Winkler, 1968). In the NDU study, however, the expertise ratings had very little influence on the results. *The averaging procedure implicitly weighted different points of view by the number of panelists representing that point of view not by their expertise.* (See Appendix B.)

This raises an important and difficult question: It is desirable to weight expert opinion by the number of experts holding that opinion? To accept such a weighting scheme in general implies that scientific disputes should be settled democratically, that is, science by popular vote or by consensus. The consequences of such a view are alarming. The number of experts holding an opinion depends not only on the strength of the evidence supporting that opinion but also on psychological biases and sociological and cultural factors that are irrelevant to scientific validity. Furthermore, the subset of experts who participate in a panel may not be representative of the 'population' of relevant experts. In particular, the NDU panelists were not necessarily chosen to be representative of climatologists in general; they were chosen for their expertise and diversity. Availability of time and willingness to cooperate were undoubtedly also factors in the makeup of the panel.

6.3. Justification for Aggregation

To question the weights and the methods used to aggregate expert judgment begs a much more important question: Should conflicting expert opinion be statistically aggregated? The NDU study does not discuss the justification for averaging expert opinion, nor does it provide guidance for the reader in interpreting or using the average probability distribution.

What is the meaning of an average probability distribution? It clearly does not represent a consensus of the panelists, because no procedure for reaching a consensus was included in the study. Nor does it represent the consensus that would result if the panel met and formed a consensus because the outcome of such a meeting is unpredictable. The average distribution gives the impression of consensus where none exists.

Furthermore, the average distribution does not represent an estimate of the average probability distribution of the population of experts because the sample of experts was not selected to be representative of the population. For the same reason, the distribution does not represent the diversity of experts in the field. Although the sample was chosen to be diverse, it was not selected randomly nor stratified so that the proportion of scientists holding a particular point of view in the sample corresponded to the proportion of all scientists holding the same view.

The average probability distribution is merely a statistic that describes a property of the responses of a particular group of experts to a particular questionnaire at a particular time. What is the justification for basing conclusions about climate change and its effects on this statistic? This question is not addressed in the NDU report which states only that the method is "considered appropriate when respondents base their replies on a common data base" (p. xix). No justification for this statement is given, nor is evidence presented that the respondents in fact used a common data base. Indeed, Clemen and Winkler (1983) show that the dependence among experts produced by a common data base "can have a seriously detrimental effect on the precision and on the value of information [relative to information obtained from independent experts] " (p. 18).

Despite the objections to averaging the diverse judgments of experts, the practice might be tolerated as a practical approach for pooling expert judgment if it could be shown that such averaging improves the validity of judgment. The averaging tradition has drawn support from psychometric theory (e.g., Guilford, 1954; Nunnally, 1978) which shows that, when a judgment can be viewed as the sum of a valid component and an error component, then the mean of several judgments is more reliable and valid than the individual judgments. The errors in the individual judgments, because they are assumed to be random, will 'average out'. Therefore, averaging emphasizes the similarities among individuals, which are assumed to be due to the valid component of their judgment, and it minimizes the differences between individuals, which are assumed to be due to error. Such a theory assumes, however, a lack of systematic shared biases in the individual judgments (Hogarth, 1977).

It is likely that some of the similarities in the judgments of individual NDU panelists are due to biases that are shared with other panelists. Such biases might have been produced by the questionnaire, for example, or by interactions among the panelists who generally knew one another and read each other's work. Einhorn *et al.* (1977) showed that any advantage of group average over individual judgments will decline rapidly as bias increases. *Although random errors will average out, shared biases will not* (Seaver, 1976, p. 19; Hammond *et al.*, 1980).

Research on group decision-making often indicates that the average of the judgments of group members is better than the judgments of most individuals in the group, though rarely better than the judgment of the best group member. Research also shows, however, that interaction among group members can improve the quality of judgment. This research has been reviewed by Seaver (1976) and by Rohrbaugh (1979), who conclude that the literature on group judgment and problem solving shows an advantage for interaction among group members over simple averaging of results. Armstrong (1978), however, concludes that there is little evidence that interaction improves predictions. Fischer (1981) reviewed the literature on aggregation when forecasts are expressed as subjective probability distributions and found no clear trend with regard to differences between simple averaging and interaction in the few studies available. In fact, the differences in accuracy among the judgments of individual group members, simple averages of members' judgments, and judgments produced by interacting groups vary from study to study. The overwhelming impression left by this research is that the relative advantages of different methods of aggregating individual judgments depend upon the nature of the task (see Hogarth, 1977).

Based on his review of the literature, Rohrbaugh (1979) concluded that the advantages of different methods of aggregation depend on the 'intentional depth' of the task; that is, the amount of information relevant to the judgment and the strength of the relationship between the information available and the unknown quantity to be inferred. When judgment tasks had little depth (such as judging the weights of objects), simple group-averaged judgments were significantly superior to the judgments of individuals. With tasks of greater depth (such as soldiers judging the date of an approaching armistice), the averaged judgments were not better than the judgments of individual members. The predictions of global temperature change in the year 2000 require judgments of great intentional depth, that is, a large amount of information is relevant to the judgment and the relation between the available information and future global mean temperature is weak. As a result, the averaging process may not have improved the validity of judgment.

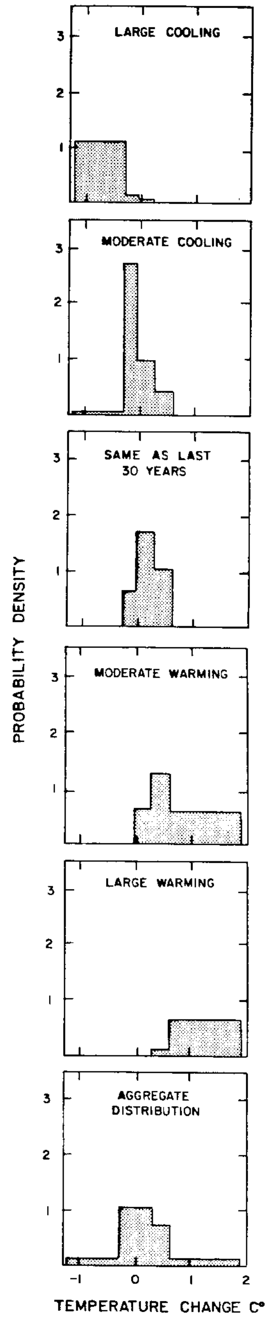


Fig. 2. Probability density functions for each temperature category and aggregate of all panelists (based on Tables I–3 and I–4 in NDU, 1978). Note: The vertical axes of probability density functions are chosen so that the total area under the curve is 1.0. Therefore, the area above any interval on the horizontal axis is equal to the probability of that interval.

Since it has not been demonstrated, either empirically or theoretically, that averaging always improves validity, the effect of averaging on the validity of the NDU results is unpredictable.

Choosing a diverse panel, and then averaging their responses does tend to cancel out the extremes and produce a moderate result. As illustrated in Figure 2, the diversity of the probability distributions of panelists classified into 5 temperature-change groups is not reflected in the aggregate distribution. This contradicts the assertion in the 1978 NDU report that the method of weighted averages "has a tendency to preserve and possibly to overstate uncertainty" (p. xix). As stated above, the aggregate is largely determined by the numbers of panelists holding various opinions. The fact that most panelists cluster in the middle of the range may be due to the selection procedure or to psychological and sociological influences that have no bearing on the validity of that position. Differences among experts resulting from, for example, access to different information or use of different conceptual schemes or metaphors (Einhorn and Hogarth, 1982) should be made explicit and used to improve the validity of group judgment. *Instead, the aggregation procedure used in the NDU study discards the advantage of having a diverse panel by averaging out differences among experts without attempting to understand them.*

The major problem with the aggregation procedure is not the weights used or the particular method of averaging. It is that the responses of a diverse group of experts were subjected to an arbitrary averaging process. Averaging may be of value under special conditions, but those conditions were not demonstrated to exist in the NDU study. *Averaging has the advantage of providing a quick answer, but that advantage is far outweighed in this context by the lack of theoretical or empirical justification for averaging.*

7. The Expert Halo Effect

The expert judgment method used in the NDU studies was similar to the Delphi method which was developed by Olaf Helmer and Norm Dalkey at the Rand Corporation (for a description of the Delphi method, see, e.g., Linstone and Turoff, 1975). The major difference between the NDU method and the Delphi method is that the NDU method lacked the feedback and revision of opinion that is considered critical to the Delphi method.

In his Rand-supported evaluation of the Delphi technique, Sackman (1975) lists 16 specific problems of the Delphi method and its applications, most of which apply to the NDU method as well. One of the most important problems that he discusses is the 'expert halo effect':

> Delphi is enmeshed in a pervasive expert halo effect. The director, the panelists, and the users of Delphi results tend to place excessive credence on the output of 'experts'. Panelists bask under the warm glow of a kind of mutual admiration society. The director has the prestige of pooled authority behind his study, and the uncritical user is more likely to feel snug and secure under the protective wing of an impressive phalanx of experts.
>
> The result of the expert halo effect for Delphi is to make no one accountable. The director merely

> reports expert opinion objectively according to prescribed procedure; he is not responsible for what the experts say. Everyone has an out, no one needs to take serious risks, and no one is ultimately accountable (p. 36).

Although the NDU panelists were not anonymous (their names are listed in the report), their individual responses were confidential. Therefore, as Sackman argues, no one is really accountable for the results. *Even though the validity of the study results is highly uncertain, the study may be given credence by methodologically unsophisticated users because of the "halo effect".*

8. Concluding Comments

It is important to note at the outset of this section that we favor the use of expert judgment methodology in order to determine the beliefs held by experts, especially on issues related to public policy. However, it is imperative that such research techniques be properly applied. The implication for public policy of a poorly designed study may prove to be worse than having undertaken no study at all.

A detailed technical assessment and critical appraisal of the methods as well as the substantive findings of the NDU study are clearly warranted because of the apparent widespread national and international attention it has received and because it has been held up as a model for other similar studies (NDU, 1983, p. 4). We have identified two general concerns regarding the study. The first is based on an important omission – the judgment process used by the experts was not made explicit. Perhaps one of the most serious criticisms that can be made of the NDU study is that it relies on covert judgment processes that involve unstated implicit assumptions. Studies that rely on expert judgment should not be conducted in this manner (see Hammond *et al.*, 1982, for an alternative approach). Little or no attempt was made to make use of methods available at the time of the study that could have aided or improved judgments specifically by making the judgment process explicit. To do so would have assisted those involved to expose and correct sources of biases and of inconsistencies. Although it is likely that the panelists could be good probability assessors and could provide some policy-relevant information about future climate, the study methods did not help them do so.

The second general objection concerns the way in which the judgments were analyzed. Despite the substantial, widespread diversity in the responses of the experts, they were averaged in order to produce a single probability distribution. The NDU report does not examine the sensitivity of the results to different aggregation methods, nor does it present results for the individual panelists so that readers can form their own opinions.

These two general concerns are, by themselves, sufficient to cast serious doubts on the results of the NDU study. Yet, there are additional specific technical problems (previously discussed) whose effects on the validity of the study's findings have yet to be determined or can be surmised from the results of past research on judgment methods. These problems and the direction of their possible effects have been summarized in Table II.

TABLE II: Methods used in the NDU Study and possible effect on results.

Method of eliciting judgment	*Direction of probable effects*
1. Use of questionnaire rather than interview	Unknown
2. Lack of feedback, communication among experts, and opportunity to revise judgments	Unknown
3. Lack of aid in making difficult and unfamiliar judgments	Unknown
Formulation of global mean temperature question	*Direction on probable effects*
1. Response mode (direct estimation, fixed-probability method)	Overconfidence Reduces variability
2. Lack of instructions, training, or methods to improve calibration	Overconfidence Reduces variability
3. Anchoring of responses	Reduces variability
4. Use of 1 °C range on temperature scale	Reduces variability
5. Question format and wording	Unknown
Method of averaging responses	*Direction of probable effects*
1. Use of linear approximation to produce cumulative probability distribution	Reduces variability
2. Use of expertise weights	Little effect
3. Aggregation of diverse opinions by averaging	Reduces variability, forced means temperature change toward zero

An important conclusion from the table is that whenever the direction of an effect could be predicted, it biased the results by either (a) increasing the level of certainty surrounding the results (i.e., reducing the variance of the subjective probability density function), or (b) both increasing the level of certainty and producing a 'middle-of-the-road' conclusion such as 'the climate will be the same as that of the last 30 yr' by forcing the mean of the distribution toward zero. *The procedure used in the NDU study had a built-in bias toward arriving at a moderate conclusion and toward minimizing the diversity of expert opinions.*

We are not able to estimate the *degree* of bias in the study. Such an estimate would be desirable because it could be used to correct or "debias" the result. Estimating the degree of bias in this case is not possible, however, because (a) the methods used in the study ignored potential biases and provided no data for estimating the degree of bias, and (b) judgment research and theory does not yet provide a basis for making quantitative estimates of biases. Because of the pervasiveness of biases and the lack of adequate theory to predict them, it is important that bias-estimating procedures be built into expert judgment studies.

The conclusions of the NDU study might have been predicted from a knowledge of the prevailing 'spirit of the times' (i.e., the prevailing mood in the science community) when the first part was conducted. This was an interesting time in recent history of climate studies. One could effectively argue that in the early 1970s the prevailing view was that the earth was moving toward a new ice age. Many articles appeared in the scientific literature as well as in the popular press speculating about the impact on agriculture of a 1–2 °C cooling.

By the late 1970s that prevailing view had seemingly shifted 180 degrees to the belief that the earth's atmosphere was being warmed as a result of an increasing CO_2 loading of the atmosphere. Interestingly, the NDU study's most crucial first part was undertaken during the transition in the scientific community of the dominance of these opposing views. One must ask what might have been the influence of the spirit of the times on the experts' judgments? What, for example, might have been their expert opinion had the NDU survey been undertaken in the mid-1980s, when there appears to be a growing scientific consensus that a CO_2-induced global warming is underway?

There is no 'cookbook' for conducting expert judgment forecasting studies. There is a body of research indicating that method affects results and that the effect depends on the nature of the judge and the judgment problem. No method is best for all judges and all judgment problems, and no single method can guarantee valid judgments in a context where it has not been tested. Conducting an expert judgment study requires a specialist in judgment and decision research, just as research in the physical and natural sciences requires specialists in relevant disciplines. A competent specialist will (a) use methods for externalizing the basis for judgment, (b) use more than one judge, (c) use more than one method, and (d) use an iterative procedure to reconcile differences among judges and

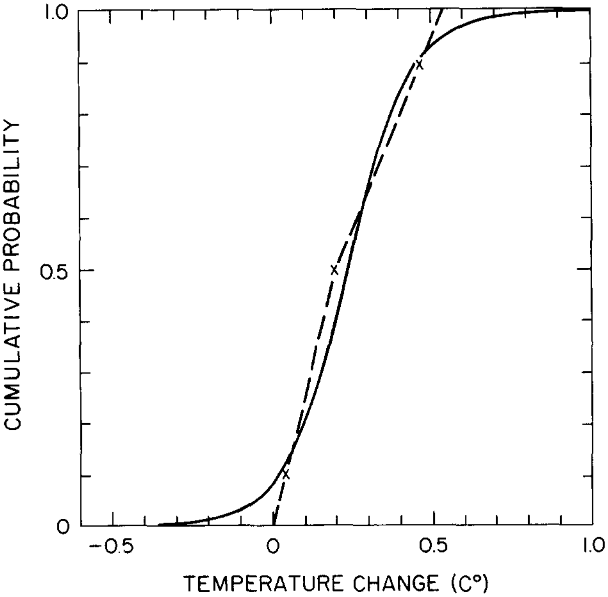


Fig. A-1. Example of cumulative probability functions derived by piecewise linear approximation (dotted line) and by fitting a logit function (solid line). Linear approximation based on Figure I–4, NDU (1978).

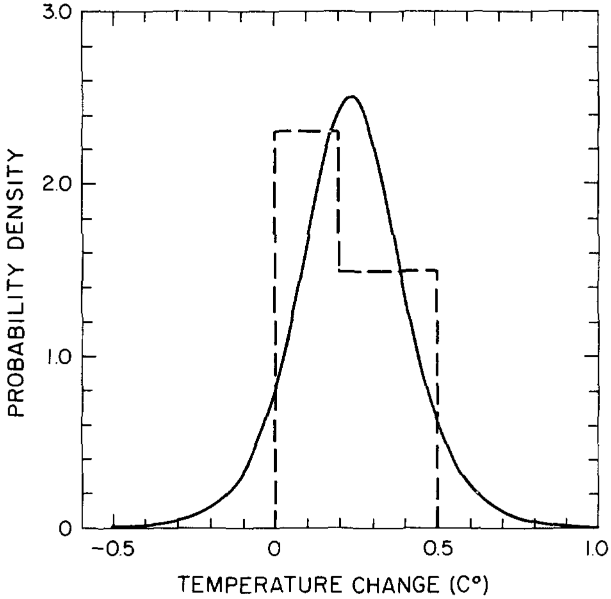


Fig. A-2. Probability density functions derived from the cumulative probability functions of Figure A-1.

methods (Stewart, 1984).

With proper aids and safeguards expert judgment can help to close the gap between scientific knowledge and the pressing needs of policymakers for the best possible information. It is unfortunate that a study as potentially important as the National Defense University's study on climate change to the year 2000 was apparently conducted without knowledge of a large body of literature on judgment and subjective probability. While a great deal remains to be learned about putting expert judgment to best use in the public policy process, much more is known about the methods of expert judgment than has been reflected in that study.

Acknowledgment

We are indebted to Maria Krenz for her suggestions and careful editing and to Bev Chavez and Jan Stewart for their expert preparation of many drafts of this paper. Special thanks are due to the scientists at the National Center for Atmospheric Research, the Center for Rsearch on Judgment and Policy, and elsewhere who took the time to read and comment on an earlier draft of the paper.

Appendix A: Approximating the Cumulative Probability Function

Figure A-1 compares the piecewise linear approximation used in the NDU study with a logit function.

Note that the logit function fits the three data points well but extends beyond the linear approximation on the tails. Figure A-2 illustrates the probability density functions derived from the cumulative density functions of Figure A-1. The logit-based function has longer tails, indicating more dispersion and less certainty about the estimates than does the linear function. In this example, the linear approximation implies zero probability that the temperature change will fall outside the range 0–0.5 °C. It is unlikely that the panelist would endorse this representation of his judgment. The logit function implies, more realistically, that the probabilities approach zero asymptotically.

Appendix B: Weighting the Experts' Responses

Both self ratings and peer ratings of expertise were obtained. The only evidence given for the validity of the ratings is a correlation of 0.52 between self and peer ratings. Although this is a statistically significant correlation, it is not a high one. It indicates that 73% of the variation in the two sets of ratings is *not* shared. In the absence of other evidence, a much higher correlation would be required to show that the ratings are measuring the same things. The correlation of 0.52 suggests that at least one of the two ratings is not valid or is unreliable.

Despite the low correlation between the self and peer ratings, they were averaged to obtain expertise weights. It is possible to show that unless two variables are perfectly correlated, the standard deviation of their mean is less than the mean of their standard deviations, and that a lower correlation between the two variables produces a lower standard deviation in the average. Therefore, averaging of the two weakly correlated ratings produces expertise ratings with reduced standard deviations, generally caused by a reduction in extreme values. The result is a bias toward equal weighting of experts rather than an increase in the validity of the expertise weights.

If the judgments of all the panelists were similar, then the relative weighting of their opinions would have little effect on the aggregate results. In this study, however, there were substantial differences among experts with regard to the global temperature question. Although the report unfortunately does not include the responses of individual panelists, nor does it discuss the differences among them, it does provide a table (NDU, 1978, p. 13) describing grouped probability densities for respondents grouped into five categories ranging from 'large cooling' to 'large warming'. The table indicates substantial differences among these groups. Figure 2 in the body of the paper is based on that table and shows probability densities (estimated by the same linear approximation method used in the study) for the five groups. Note that there is no overlap between the large cooling group and the large warming group, that is, the upper limit of the large cooling distribution is below the lower limit of the large warming distribution. Furthermore, there is little overlap between the large cooling group and moderate warming group. This spread in the probability densities means that the aggregate density function will be sensitive to the weights that are applied to the groups.

The NDU report does not include a sensitivity analysis of the weights, nor does it include the individual response data that would be needed to conduct such an analysis.[1]

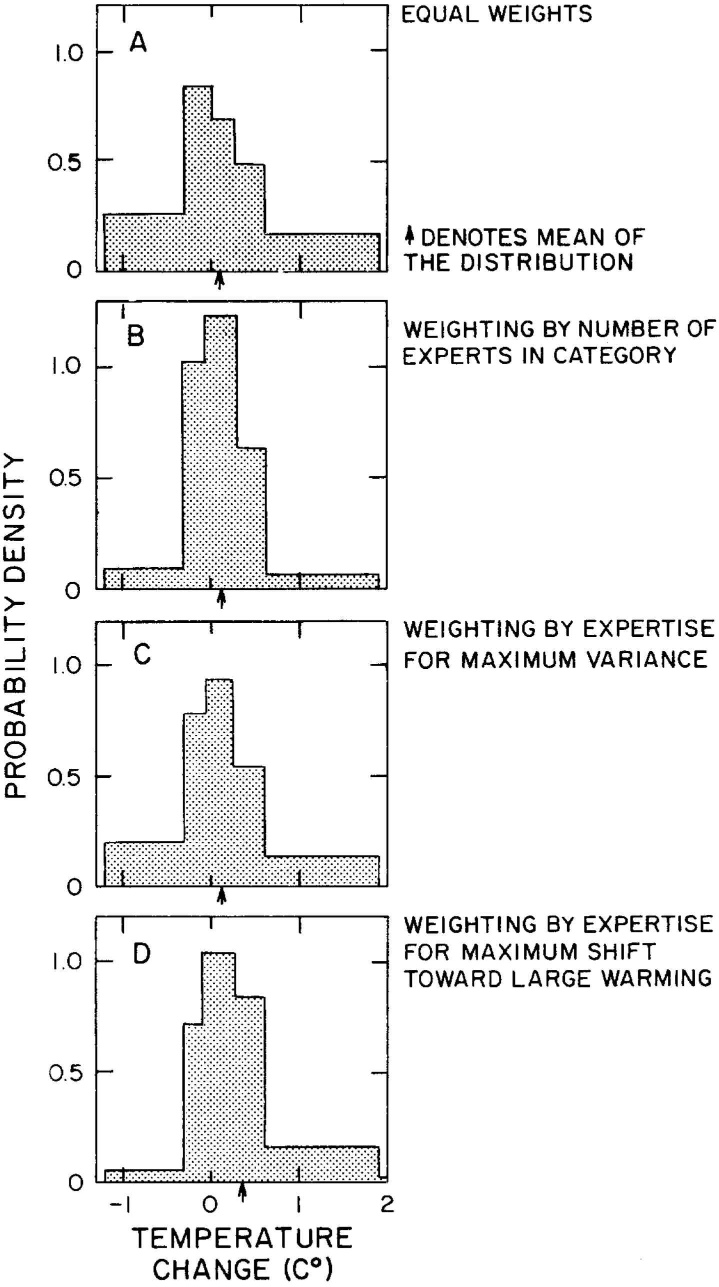


Fig. B-1. Probability density functions obtained by applying different weighting schemes to the temperature categories.

The data on grouped probability densities in the report does, however, provide a basis for making some inferences about what the sensitivity analysis might have shown had it been carried out.

Figure B-1 is the result of aggregating the probability density functions for each temperature group (Figure 2) by four different weighting schemes following the procedure used in the NDU study. Scheme A places equal weight on each group. Scheme B weights each group by the number of respondents in that group. Scheme C weights each group by the product of the number of respondents in that group and an expertise rating for the group which was chosen to maximize the variance in the distribution.[2] Weighting scheme D achieves a maximal shift in the mean of the distribution toward the 'large warming' group by assessing expertise weights of 4 to the 'moderate' and 'large warming' groups, and weights of 1 to the other groups. The major conclusion to be drawn from Figure B-1 is that the weighting of groups by number of respondents has a greater effect on the results than do the expertise ratings.

The results of weighting schemes A, B, and C differ primarily with regard to the variability of the distribution. Equal weighting (A) produces the most variable distribution. When the groups are weighted by the number of panelists in the group, a much greater proportion of the distribution is near the middle because of the large number of members in the middle groups relative to the extreme groups. Assigning the greatest expertise weights to the most extreme groups (weighting scheme C) has little effect on the variability of the distribution, and a highly asymmetrical weighting of expertise (D) shifts the mean only slightly (from 0.11 °C to 0.29 °C).

Since the NDU study was based on the aggregation of individual panelists, rather than of groups (as in Figure B-1), it implicitly weighted each point of view by the number of panelists representing that point of view. Indeed, the aggregate probability distribution for the year 2000 that was used in the study (see Figure 2) is very similar to weighting scheme B, which, in effect, assigned equal expertise weights to each panelist. Thus, for this question, the aggregate probability distribution is determined by the number of panelists representing different points of view and the expertise ratings have very little influence. The statement in the NDU report that "the aggregated curves are not acutely sensitive to the weighting system used" (p. 9) supports this conclusion.

References

Armstrong, J. Scott: 1978, *Long-Range Forecasting: From Crystal Ball to Computer*, Wiley, New York.

Brown, L. R.: 1970, *Seeds of Change*, Praeger, New York.

Clemen, R. T. and Winkler, R. L.: 1983, 'Limits for the Precision and Value of Information from Dependent Sources', Unpublished manuscript, University of Indiana, Bloomington, Indiana.

[1] One of the organizers of the study indicated that the raw data had been destroyed, and a request to the Institute for the Future for information about the sensitivity of the results to the weights went unanswered.

[2] This was accomplished by assigning the largest weight to the most extreme groups. Thus, the 'large cooling' and 'large warming' groups received weights of 4, the 'moderate cooling', 'moderate warming', and 'same as last 30 yr' groups received weights of 1 (weights of 0 were not used because it is assumed that only panelists who were familiar with the question were included in Table 1–4 of the report).

Congressional Research Service: 1976, *A Primer on Climatic Variation and Change*, U.S. Government Printing Office, Washington.

Council on Environmental Quality: 1980, 'The Global 2000 Report to the President: Entering the Twenty-First Century', Washington, D.C. U.S. Government Printing Office.

Daan, H. and Murphy, A. H.: 1982, 'Subjective Probability Forecasting in The Netherlands: Some Operational and Experimental Results', *Meteorol. Rund.* **35**, 99–112.

Einhorn, H. J. and Hogarth, R. M.: 1982, 'Prediction, Diagnosis, and Causal Thinking in Forecasting', *Journal of Forecasting* **1**, 23–36.

Einhorn, H. J. and Hogarth, R. M.: 1981, 'Behavioral Decision Theory: Processes of Judgment and Choice', *Annual Review of Psychology* **32**, 53–88.

Einhorn, H. J., Hogarth, R. M., and Klempner, E.: 1977, 'Quality of Group Judgment', *Psychological Bulletin* **84**, 158–172.

Fischer, G. W.: 1981, 'When Oracles Fail – A Comparison of Four Procedures for Aggregating Subjective Probability Forecasts', *Organizational Behavior and Human Performance* **28**, 96–110.

Fischhoff, B.: 1982, 'Debiasing', in D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York.

Fischhoff, B. and MacGregor, D.: 1982, 'Subjective Confidence in Forecasts', *Journal of Forecasting* **1**, 155–172.

Glantz, M. H., Robinson, J., and Krenz, M.: 1982, 'Improving the Science of Climate-Related Impact Studies: A Review of Past Experience', in William Clark (ed.), *Carbon Dioxide Review: 1982*, Oxford University Press, New York.

Guilford, J. P.: 1954, *Psychometric Methods*, McGraw Hill, New York, p. 25.

Hammond, K. R., Anderson, B. F., Sutherland, J., and Marvin, B.: 1982, 'Improving Scientists' Judgments of Risk, University of Colorado at Boulder, Center for Research on Judgment and Policy, Report No. 239.

Hammond, K. R., McClelland, G. H., and Mumpower, J.: 1980, *Human Judgment and Decision Making: Theories, Methods and Procedures*, Praeger, New York.

Hammond, K. R., Rohrbaugh, J., Mumpower, J., and Adelman, L.: 1977, 'Social Judgment Theory: Applications in Policy Formation', in M. F. Kaplan and S. Schwartz (eds.), *Human Judgment and Decision Processes in Applied Settings*, Academic Press, New York.

Hammond, K. R., Stewart, T. R., Brehmer, B., and Steinmann, D. O.: 1975, 'Social Judgment Theory', in M. F. Kaplan and S. Schwartz (eds.), *Human Judgment and Decision Processes*, Academic Press, New York.

Hogarth, R. M.: 1980, *Judgment and Choice: The Psychology of Decision*, Wiley, Chichester, England.

Hogarth, R. M.: 1977, 'Methods for Aggregating Opinions', in H. Jungermann and G. de Zeeuw (eds.), *Decision Making and Change in Human Affairs*, D. Reidel Publ. Co., Dordrecht, Holland, pp. 231–255.

Hogarth, R. M.: 1975, 'Cognitive Processes and the Assessment of Subjective Probability Distributions', *J. Amer. Statist. Assoc.* **70**, 271–289.

Hopkins, R. F. and Puchala, D. J. (eds.): 1978, *The Global Political Economy of Food*, U. of Wisconsin Press, Madison, Wisconsin, p. 2.

Koriat, A., Lichtenstein, S., and Fischhoff, B.: 1980, 'Reasons for Confidence', *J. Experimental Psych.: Human Learning and Memory* **6**, 107–118.

Kraemer, R. S.: 1978, 'Meeting Reviews: Session on Climatic Futures at the Annual Meeting of the AAAS, 17 February, 1978', *Bulletin of the American Meteorological Society* **59**(7), 822–23.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D.: 1982, 'Calibration of Probabilities: The State of the Art to 1980', in D. Kahneman, P. Slovic, and A. Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York.

Linstone, H. A. and Turoff, M.: 1975, *The Delphi Method: Techniques and Applications*, Addison-Wesley, Reading, Mass.

McKay, G. A. and Williams, G. D. V.: 1982, 'Crop Yields and Climate Change to the Year 2000, Volume I (Book Review)', *Bulletin of the American Meteorological Society* **63**, 427.

Murphy, A. H. and Brown, B. G.: 1984, 'A Comparative Evaluation of Objective and Subjective Weather Forecasts in the United States', *Journal of Forecasting* **3**.

Murphy, A. H. and Daan, H.: 1984, 'Impacts of Feedback and Experience on the Quality of Subjective Probability Forecasts: Comparison of Results from the First and Second Years of the Zierikzee Experiment', *Monthly Weather Rev.* **112**, 413–423.

Murphy, A. H. and Winkler, R. L.: 1974, 'Credible Interval Temperature Forecasting: Some Experimental Results', *Monthly Weather Rev.* **102**, 784–794.

National Academy of Sciences: 1975, *Environmental Impact of Stratospheric Flight.* National Academy of Sciences, Washington, D.C.

NDU: 1983, *The World Grain Economy and Climate Change to the Year 2000: Implications for Policy*. National Defense University Press, Washington, D.C.

NDU: 1980, *Crop Yields and Climate Change to the Year 2000*, National Defense University, Washington, D.C.

NDU: 1978, *Climate Change to the Year 2000*, National Defense University, Washington, D.C.

Newell, A. and Simon, H. A.: 1972, *Human Problem Solving*, Prentice-Hall, Inc., Englewood Cliffs, N.J.

Nunnally, J. C.: 1978, *Psychometric Theory*, 2nd ed., McGraw Hill, New York.

Raiffa, H.; 1968, *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Addison-Wesley Publishing Co., Reading, Massachusetts.

Rohrbaugh, J.: 1979, 'Improving the Quality of Group Judgment: Social Judgment Analysis and the Delphi Technique', *Organizational Behavior and Human Performance* **24**, 73–92.

Sackman, H.: 1975, *Delphi Critique*, D.C. Heath and Company, Lexington, Mass.

Schuman, H. and Presser, S.: 1981, *Question, and Answers in Attitude Surveys.* Academic Press, New York.

Science News: 1978, 'The 25-Year Forecast: Group Prediction', **113**(8), 116.

Seaver, D. A.: 1976, 'Assessment of Group Preferences and Group Uncertainty for Decision Making', Technical Report SSRI 76–4, Social Science Research Institute, University of Southern California, Los Angeles, California.

Seaver, D. A., von Winterfeldt, D., and Edwards, W.: 1978, 'Eliciting Subjective Probability Distributions on Continuous Varibales', *Organizational Behavior and Human Performance* **21**, 379–391.

Sellers, W. D.: 1979, 'Climate Change to the Year 2000: A Book Review', *Bulletin of the American Meteorological Society* **60**(6), 686.

Slovic, P., Fischhoff, B., and Lichtenstein, S.: 1977, 'Behavioral Decision Theory', *Annual Review of Psychology* **28**, 1–39.

Slovic, P. and Lichtenstein, S.: 1971, 'Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment', *Organizational Behavior and Human Performance* **6**, 649–744.

Spetzler, C. S. and Stael von Holstein, C.-A.S.: 1975, 'Probability Encoding in Decision Analysis', *Management Sci.* **22**, 340–358.

Stewart, T. R.: 1984, 'Judgment and Forecasting: Methodological Implications of Judgment Research'. Paper prepared for the Conference on Forecasting in the Social and Natural Sciences, Boulder, CO, June 10–13.

Tversky, A. and Kahneman, D.: 1981, 'The Framing of Decisions and the Rationality of Choice', *Science* **211**, 453–458.

Tversky, A. and Kahneman, D.: 1974, 'Judgment under Uncertainty: Heuristics and Biases', *Science* **185**, 1124–1131.

Tversky, A. and Kahneman, D.: 1971, 'The Belief in the Law of Small Numbers', *Psychological Bulletin* **76**, 105–10.

U.N. FAO: 1979, 'Scanning the Future for Climate Change', *CERES* **12**, 7.

U.S. Central Intelligence Agency: 1974, *A Study of Climatological Research as It Pertains to Intelligence Problems*, Document Expediting Project, Library of Congress, Washington, D.C.

Wallsten, T.S. and Budescu, D.V.: 1983, 'Encoding Subjective Probabilities: A Psychological and Psychometric Review', *Management Science* **29**, 151–173.

Winkler, R.L.: 1968, 'The Consensus of Subjective Probability Distributions', *Management Science* **B 15**, 61–75.

Winkler, R. L. and Murphy, A.H.: 1979, 'The Use of Probabilities in Forecasts of Maximum and Minimum Temperatures', *Meteorological Magazine* **108**, 317–329.

(Received 6 April, 1984; in revised form 22 May, 1984)