

The use of probability elicitation in the high-level nuclear waste regulation program

Aaron R. DeWispelare^{a,*}, L. Tandy Herren^b, Robert T. Clemen^c

^aCenter for Nuclear Waste Regulatory Analyses, 6220 Culebra Road, San Antonio, TX 78238, USA

^bSouthwest Research Institute, 6220 Culebra Road, San Antonio, TX 78238, USA

^cLundquist College of Business, University of Oregon, Eugene, OR 97405, USA

Abstract

Expert judgement elicitation is expected to be used in the performance assessments (PA) of the long-term behavior of high-level waste (HLW) geologic repositories. As a preparation for an effective review of the U.S. Department of Energy (DOE) PA, the Nuclear Regulatory Commission (NRC) is evaluating the mechanics of eliciting expert judgements. One of the objectives of this evaluation is to explore techniques for generating and aggregating probabilistic judgements of future conditions at the proposed HLW repository at Yucca Mountain, Nevada. An actual elicitation was conducted as an aid to these evaluations. This paper documents this probabilistically centered elicitation and subsequent activities to explore aggregation of opinion techniques. Future climate in the Yucca Mountain, Nevada vicinity was selected as the topic for elicitation. Personnel from the NRC and Center for Nuclear Waste Regulatory Analyses (CNWRA) defined the climatic parameters of interest in conjunction with a panel of five expert climatologists. Individual elicitations were performed with each climatologist to produce probabilistic estimates of each parameter at seven points of time in the future. The elicitations employed the fractile technique to generate cumulative probability distributions representing the uncertainty in the predictions. After the individual elicitations, a group session was conducted to explore aggregation and consensus methods.

Keywords: Expert elicitation; Expert judgement; Subjective probability assessment; Climate; Uncertainty; Opinion aggregation

1. Introduction

1.1. Background

The Nuclear Regulatory Commission (NRC) is developing methods to determine compliance with its regulation for the disposal of high-level

nuclear waste (10 CFR Part 60) by the U.S. Department of Energy (DOE). Performance of high-level nuclear waste (HLW) geologic repositories has to be assessed for a regulatory period of 10 000 years as stipulated in standards developed by the Environmental Protection Agency. Because of this extremely long time period, a combination of data from site characterization, experimental methods, studies of natural analogs and mathematical models will be

* Corresponding author. Tel. +1 210 522 6072; Fax +1 210 522 6081.

used in performance assessments (PA) to demonstrate compliance with the requirements in these standards. Mathematical models are expected to be the primary tools for estimating the long-term future performance of the repository. Expert judgement elicitation is a potential source of data for PA and, like other data such as that from scientific site exploration, will require interpretation and supplementation before it can be incorporated into PA models.

This study applies a state-of-the-art expert judgement elicitation procedure to an area of considerable interest related to predicting the performance of a HLW repository. Identification of external conditions, to which the repository will be subjected, during the regulatory period is an essential requirement for applying the mathematical models. These external conditions pertain to evolving tectonism, volcanism, seismicity, climate and other factors that may affect repository performance. In the HLW program in the United States (and several other countries) these external conditions are included in PA by defining 'disruptive scenarios' (NRC, 1992). Noting that the future is highly uncertain, a full definition of a scenario consists of not only its characteristics but also its probability of occurrence. One method of defining scenarios and their probability of occurrence is the use of formal expert elicitation.

1.2. Formal expert elicitation

Expert judgement is involved in an informal manner in all scientific research. A scientific judgement is defined to be informal when there is no formal training provided to debias the expert and the documentation is lacking. Informal judgement is applied every time a researcher makes decisions about what phenomenon to study, how to initialize a model, what variables are important to the process, what data are appropriate, and how to process and use those data. Informal expert judgements are a normal part of the scientific process and, under well-controlled and understood situations, such judgements are validated, or invalidated, by the process itself. This validation is not possible for

the long-term assessment of HLW repositories.

Formal expert elicitation (commonly referred to as expert judgement elicitation) refers to a structured procedure designed to gather knowledge about a discipline or area of endeavor from individuals considered human experts in that domain. Topics of elicitation can be probability encoding, scenario development, or model selection. The elicitation procedure requires the participation of three parties: normative experts, generalists, and experts. Normative experts have a suitable background in decision theory, probability theory, and psychology. The generalists are usually project staff that (i) have a thorough understanding of the project and the manner in which the judgements will be used, and (ii) play a role in defining the issues that the experts are to address, and (iii) participate in assembling and presenting the fundamental information about the issue(s) to the experts. The experts from which the judgements are elicited have extensive expertise in the domain of interest. The elicitation procedure includes methods to select and train the experts as well as a defined approach to gathering the appropriate knowledge. The normative experts elicit judgements from the experts according to the formal process and carefully document the experts' rationale. The generalists and the normative experts together are often referred to as the elicitation team.

Compared to an informal expert judgement process, formal expert elicitation increases the credibility and defensibility of the judgements because of the careful and thorough documentation of each expert's rationale and enhances the communication of the results. The normative expert carefully documents the procedure used to obtain the judgements and the rationale for each judgement, making the process readily available for external evaluation. The training session incorporated in the formal process ensures that the participants are aware of the cognitive and motivational biases that can influence expert judgements and provides techniques to recognize and avoid them. Accuracy is improved also because the problem definition is clear and the issues addressed are explicitly stated. Finally, formal expert elicitation in-

creases the consistency of procedures across studies, facilitating comparisons among them.

1.3. Interpretations of probability

Probability elicitation is a special case of expert elicitation that focuses on collecting subjective probabilities for uncertain events. In this view, a probability is interpreted as an individual's degree of belief that an event will occur. This differs from classical or frequentist probabilities in which the relative likelihood of outcomes (e.g., the roll of a die) can be determined or in which there is empirical evidence describing the relative frequency of events (e.g., the frequency of accidents at an installation). In our case, the probability is the individual's degree of belief based on any relevant information that is available. For uncertain quantities that are continuous, an expert's estimate is often interpreted as the median of the experts' distribution and the spread of the distribution is based on the confidence or degree of belief in that value. Formal elicitation techniques have been devised to collect and interpret subjective probability assessments. A workshop sponsored by the Nuclear Energy Agency in 1987 (NEA, 1987) concluded PA of HLW repositories will be plagued by 'lack of knowledge' uncertainties in areas such as future seismicity and climate that can be quantified as probabilities representing degrees of beliefs.

The elicitation and use of expert judgement encoded as probabilities has long been a vital part of decision analysis (Raiffa, 1968; Spetzler and Stael von Holstein, 1975). Many of the techniques and procedures developed and used in decision analysis are also used in risk assessment and performance assessment. For the HLW repository PA program, obtaining probabilities through elicitation is motivated largely by basic and compelling facts: in the earth and atmospheric sciences, current models and underlying principles are still elementary in many respects, the time and space scales of the problem at hand are quite large, and the interrelationships are complex. In this situation, obtaining sufficient quantities of data to compute a reliable probability estimate will usually not be feasible. The

elicitation approach permits the formal incorporation of subjective expert judgement into the PA process. This information supplements other sources, including the results of scientific observations, characterizations, experiments, and modeling of physical and geochemical processes. The role of the experts is not to create knowledge, but to summarize the available information and to help express what is known and what is not known (level of uncertainty). That is, the experts present a state-of-the-art assessment of the scientific knowledge of a specific discipline or scientific topic based on the information available at a point in time.

2. The expert elicitation procedure

Our formal expert elicitation procedure consists of 11 steps:

- (i) Determine the objectives and goals of the study.
- (ii) Recruit the experts.
- (iii) Identify issues and information needs.
- (iv) Provide initial data to the experts.
- (v) Conduct the elicitation training session.
- (vi) Discuss and refine the issues.
- (vii) Provide a multi-week study period.
- (viii) Conduct the elicitations.
- (ix) Provide post-elicitation feedback to the experts.
- (x) Aggregate the experts' judgements (if required).
- (xi) Document the process.

The eleven steps can be broken into roughly five sequential phases: pre-training activities ((i) through (iv)), training ((v)), study period ((vi)–(vii)), elicitation ((viii)), and follow-up ((ix) through (xi)). These phases are strictly sequential while the individual steps within the phases can proceed largely concurrently. These steps should be present in high-quality formal elicitations.

This study was concerned with acquiring probabilistic judgement of future climate parameters in the Yucca Mountain Nevada vicinity (YMNV). One of the areas that can affect the performance of a high-level nuclear waste repository, over a 10 000-year horizon is hydrologic

transport of the radionuclides contained in the entombed wastes, which may be significantly affected by future climates. Because the state of science and modeling in the area of climate projections has not been validated for long-term, sub-regional estimates, the uncertainty in this area can best be characterized by expert probability judgements. Expressing uncertainty in the form of probabilities allows the judgements to be directly utilized in the PA models examining the behavior of the proposed HLW repository because these models are probabilistic. Five of the steps in the formal elicitation procedure are particularly relevant to probability elicitation in general and to the climate elicitation in particular: identifying the issues (iii), conducting the elicitation training session (v), refining the issues (vi), conducting the individual elicitations (viii), and aggregating the expert judgements (x). The following sections describe each of these steps. Description of the recruiting and selection of the experts, as well as their identification and resumes, is provided in DeWispelare et al., 1993.

2.1. Identify the issues and information needs

The focus in this step was on identifying the issues to be considered. Early in this step, a broad scope was chosen to ensure that the significant issues were identified and that other issues were excluded for sound reasons. Once a list of issues was developed the most important issues were identified to enhance the likelihood of achieving the project objectives and goals. For example, the climatology elicitation included precipitation variables related to the maximum amount of precipitation likely to be experienced in the YMNV. This issue was chosen because extreme precipitation is thought to have a potentially adverse impact on repository performance. But the issue of the future state of society's impact on climate was rejected because of the inability to characterize the extreme uncertainty. Criteria used to select the issues were explicitly stated and reviewed by the experts at the initial meeting of the entire group.

The definition of the issues should be precise because their clarity is critical for the elicitation design. The issues can range from general to

specific and from complex to simple. For example, a general question might be, "What are the primary climatological controls operating at Yucca Mountain?" A more specific question might be, "What is the likely average annual precipitation amount at 3000 years after the present (AP) in the vicinity of Yucca Mountain?" The generalists and normative experts should propose initial conceptual models and scenarios for the purposes of interest to be reviewed by the experts at the first meeting.

This first step also involves formalizing the quantities to be elicited and stating all assumptions that are to influence the judgements. Winkler et al. (1978) state that all questions asked of the experts should be about observable or at least theoretically observable quantities. The questions in probability elicitation should be framed to reduce the possible impact of motivational and cognitive bias on the judgements. The normative expert should avoid ways of framing questions that may bias the response; for example, focusing on the extremes of a distribution rather than the central part can help reduce the tendency to anchor on a central value and thereby reduce the overconfidence bias (Winkler et al., 1992).

The issue statement for the climatology elicitation charged the experts with predicting the future climate conditions in the YMNV. The primary variables of interest were changes in annual precipitation and temperature as well as changes in the seasonal variability of precipitation at the site over the course of 10000 years. These variables were selected by the generalists to include those that were thought most likely to have an impact on repository performance. The issue statement also indicated the intent to evaluate these changes at a series of future time epochs: 100, 1000, 3000, 5000, 7500, and 10 000 AP (300 was added later). At each of these time steps, the experts were charged with providing distributions for the main variables and discussing the climatic controls that caused any expected changes.

2.2. Conduct the elicitation training session

The literature on expert judgement stresses

the importance of training experts on the task facing them (Merkhofer, 1987; von Winterfeldt and Edwards, 1986; Mosleh et al., 1988). A formal meeting involving the elicitation team and the experts is appropriate to accomplish this training. Training consists of familiarizing the experts with the judgement process and motivating them to provide formal judgements, giving them practice in expressing their judgements formally and educating them about possible judgement biases. There are two general classes of bias: motivational and cognitive. Motivational biases occur because the expert has a vested interest in an issue which, consciously or unconsciously, distorts his judgement. Cognitive biases occur because of a failure to process, aggregate, or integrate the available data and information (Kahneman et al., 1982). Motivational biases can generally be avoided, or at least reduced, by a careful expert selection process. The primary cognitive biases in probability judgements are (i) overconfidence, (ii) anchoring, (iii) availability, (iv) ignoring base rates, and (v) nonregressive predictions. To counter cognitive biases, experts should learn about the bias through personal experience, detailed discussion (Bonano et al., 1990) and debiasing exercises as part of the training.

The initial meeting of the elicitation team and climatology panel occurred at the CNWRA in San Antonio, Texas. The goals of this meeting were threefold: (i) to orient the experts, (ii) to refine the initial issue statement, and (iii) to conduct elicitation training. Initially, CNWRA personnel provided an overview of the proposed repository and the PA program together with a short overview of the goals of the elicitation. These overviews also described the current and past climate in the YMNV and the structure of the proposed HLW repository. The introductory session attempted to motivate the experts by showing the importance of the elicitation study and discussing how their judgements would be used in the PA program. In addition, the elicitation team fielded any concerns the experts had about making subjective judgements. Another part of the initial meeting was devoted to extensive training in probability elicitation and debiasing. The training session included an overview of

subjective judgements of uncertainty. Examples were given of overconfidence, and confirmatory bias anchoring, availability, base rate dependency, and nonregressive predictions to help the experts learn first hand about possible cognitive biases. Probability assessment was demonstrated through the use of graphical techniques. The normative expert helped elicitation team members practice responding to probability questions.

2.3. *Discuss and refine the issues*

This step can take place at the first meeting of the experts, which may coincide with the elicitation training. The experts should be familiar with the issue statement and should comment on the issues to be covered in the elicitation. Discussion is used by the experts to refine the initial conceptual models and scenarios in order to arrive at an unambiguous definition of the events or quantities to be elicited. This definition needs to include a list of all assumptions that will be shared by the group. Finally, the experts should discuss possible decompositions of the problem, noting any areas in which conditional probabilities are necessary to capture the relationship among variables. When done properly, this step results in a convenient model with which to focus discussion and proceed into an individual research effort by each expert.

At the first meeting of the climatology elicitation panel, the experts had the opportunity to refine the issues statement and generate a list of the climate-forcing factors and assumptions that would be shared by the group. Fig. 1 shows the results of this session. The primary forcing factors of the climate in the YMNV were identified as the physiography of the area, Milankovitch cycles (i.e., cyclical perturbations in the earth's orbit), solar output, shocks to the climate system, natural variability and the human influence on climate. Through their impact on large-scale, regional-scale and local-scale weather phenomena, such as atmospheric circulation, these factors determine the temperature, precipitation, seasonal variability, precipitation persistence, short-term precipitation intensity, and solar radiation at the site. The group reached consen-

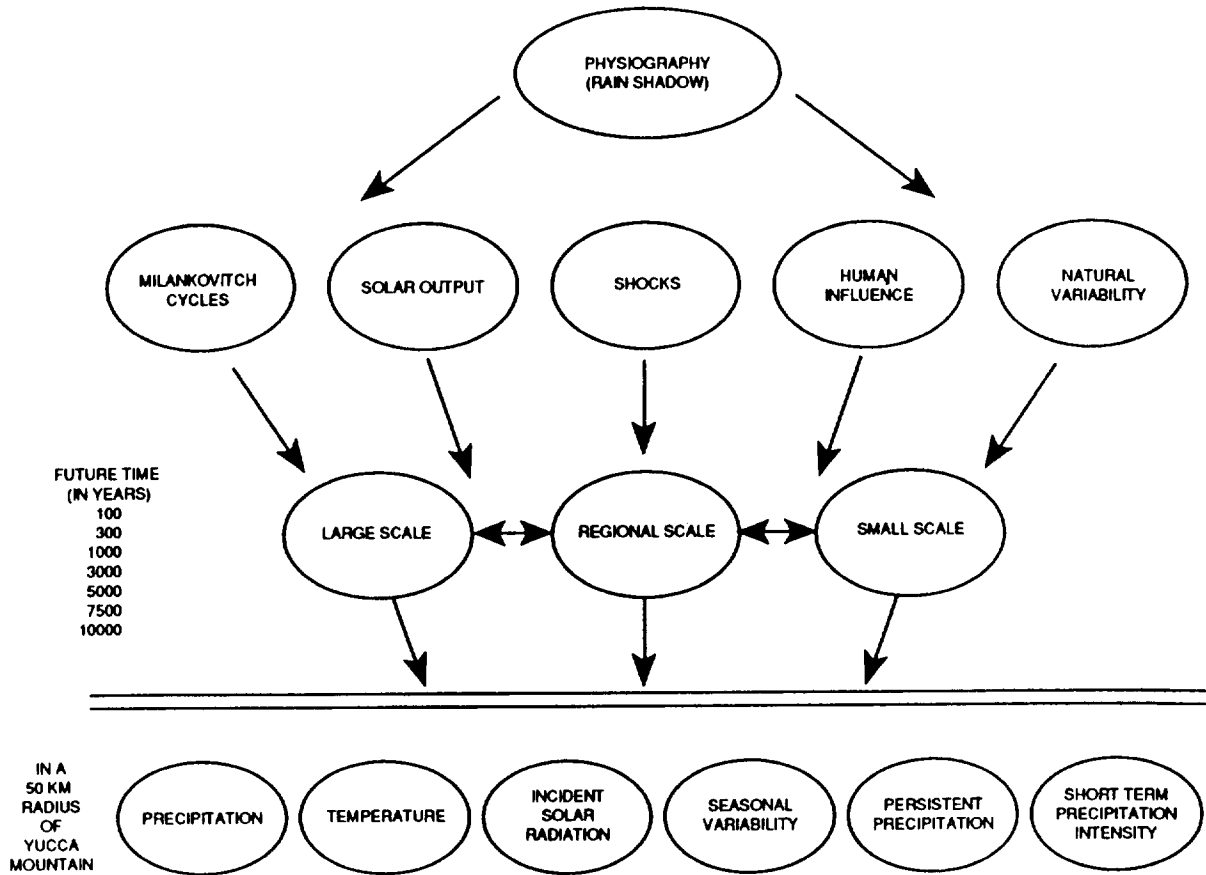


Fig. 1. Conceptual climate model.

sus on several assumptions. They agreed that the area of evaluation would be a 50-km radius around YMN and that they would estimate the various parameters for 100-year climatological normals (moving average) centered on the point in time of interest.

The following is a definition of each of the variables within the context of this study.

- (i) Precipitation:
The average amount of annual rainfall.
- (ii) Temperature:
The average annual temperature.
- (iii) Seasonal Variability:
The average summer versus winter precipitation in monthly units or in proportions.
- (iv) Short-Term Precipitation Intensity:
The largest amount of rainfall that will occur in a 10-day period.

- (v) Persistent Precipitation:

The wettest concurrent 10-year period in the 10 000-year period of investigation.

- (vi) Solar Radiation:

Changes in either cloud cover or incident solar radiation relative to current levels.

The group added the 300-year time slice because they felt it was an important step between 100 and 1000 AP. Precipitation estimates provided the absolute incident amount rather than an effective amount available for infiltration resulting from evapotranspiration. The group agreed to assume that the topography of the region would remain essentially invariant across the 10 000-year period. Finally, for purposes of tractability, the group decided to consider only those scenarios of plausible human impact on the environment that are possible at present. Unlikely but possible changes, such as the potential to

control the weather at some point in the future, were ignored.

2.4. *Conduct the individual elicitation*

The experts in the climatology elicitation had one month between training and the individual elicitation to review any literature, data, or model results they felt were relevant to their judgements and to prepare position papers relevant to their judgements. Each expert prepared his position paper independently; they had access to each other only for brief consultation to exchange data or to clarify information of common interest. The goal was to obtain their independent stances in preparation for the individual elicitation. At the end of the research period, the experts each prepared a short paper that documented the basis of their judgements. At the start of the individual elicitation, the experts were assembled as a group and asked to provide short presentations of the main conclusions contained in their position papers. The aim of this meeting was to expose the experts to each others' reasoning processes in a non-confrontational environment without attempting to alter their individual stances. Questions at the end of each presentation were limited to clarification of the expert's reasoning.

After the group meeting, the experts individually provided judgements and probability estimates for the variables identified in the issue statement in a formal elicitation. The individual elicitation was held in a quiet conference room and was attended by the elicitation team and an individual expert. At the start of the session, the normative expert mapped out the general tasks and summarized the issues to be covered. The normative expert explicitly reviewed all definitions and assumptions agreed to by the group during the initial meeting. The expert then described in his own words the current climatic controls in effect at the site. The first judgements provided by an expert were trends in precipitation and temperature across the 10 000-year time period. This provided an initial look at the expert's fundamental approach to predicting the climate in the area and allowed for subsequent

consistency checks at each point in time, (called a time slice). The expert also provided a trend for solar radiation (or its surrogate, cloud cover) across the time slices. The normative expert then asked the expert to provide specific answers to questions about the variables considered, encouraged the expert to discuss the reasoning behind the estimates, ensured that the required information was obtained and checked the consistency of the information. The information was documented for subsequent analysis.

After this, the elicitation progressed time slice by time slice. At each time slice, the expert provided probability distributions for temperature, precipitation and short term intensity as well as the most likely precipitation distribution by season. At the end of the session, the expert estimated the persistent precipitation. Cumulative distribution functions (CDFs) and/or probability density functions (PDFs) were produced for each variable of interest at each time slice. The fractile method was used as the probability encoding technique to elicit information to construct the probability distributions. In the fractile method, the z -fractile is the magnitude x_z of the uncertain quantity x such that there is a probability of z that the true magnitude falls below x_z or is equal to x_z and a $1-z$ probability that x falls above x_z . The lower bound is the 0.0 fractile and the upper bound is the 1.0 fractile. The tails of a distribution, for example, 0.01 or 0.05 fractiles and 0.95 or 0.99 fractiles, were explored first to reduce the effects of overconfidence and anchoring to a central value. Questions like the following were posed: "Is this the point at which you see only a 5% chance of getting a lower value of the variable and a 95% chance of getting a higher value?" The extreme events needed to have credible explanations that the expert deemed probable at the given level. Each distribution was systematically constructed from five to nine points representing the various fractiles (depending on the degree to which the expert could propose the distribution shape). Because of the significant uncertainty involved in identifying the 0.0 and 1.0 fractiles, lower tail values of either 0.01 or 0.05 and upper tail values of 0.95 or 0.99 were agreed upon as satisfactory

and practical distribution end points by both the generalists (representing the modelers or users of the data) and the experts. The elicitations lasted approximately 4 h each and were documented by careful notes and videotape.

2.4.1. Results of the individual elicitation sessions

The experts reviewed the results to ensure that they accurately reflected the information provided in the elicitation. They also provided comprehensive rationales with each of their distributions (DeWispelare, et al. 1993). Appendix A contains a sample from one of the transcripts to show the way the elicitation team collected the probability distributions and justifications. The following two sections discuss the results for the future temperature and precipitation in the YMNV as a sample of the data obtained in the elicitations.

2.4.2. Likely future temperature in the YMNV

Distributions associated with the temperature estimates for each of the five expert climatologists (identified in the legend) are shown in Fig. 2. The 1993 annual temperature was assumed to be 16°C (DOE, 1988). Generally all the experts agreed that the next 100 years will be getting warmer by as much as 3°C, on average, due to the greenhouse warming associated with anthropogenic activities. Following 100 years AP, the temperature trends across the 10 000-year period represented two distinct beliefs in the relative impact of the forcing factors on the YMNV climate. Both groups cited the influence of anthropogenic global warming and Milankovitch (solar radiation variability due to orbital fluctuations) forcing in their trends, but two of the experts believed that anthropogenic warming would have a more persistent impact across the 10 000-year period. These two experts predicted that temperatures would rise to a maximum of a 3.5°C increase between 300 and 1000 years AP and then would decrease somewhat across the remainder of the period. However, at no point would temperatures decrease below current levels.

The other three experts viewed anthropogenic

warming differently. These three experts agreed that beginning 300 years AP temperatures will drop and will continue dropping through the remainder of the 10 000-year time frame of interest. All five of the experts agreed that anthropogenic warming would extend to between 1000 and 3000 years AP, but the remaining three experts forecast some cooling (relative to current temperature) following that time, due to a decreased impact of anthropogenic forcing and increased influence of Milankovitch forcing. One expert predicted maximum warming after 300 years but before 1000 years AP, while another expert predicted that the anthropogenic warming effects will continue through the 10 000 years, but superimposed on the warming will be cooling associated with Milankovitch forcing with a slight cooling overall in the last 5000 years. Final temperatures at 10 000 years AP ranged from -2°C to +2°C with one expert saying that temperatures in the YMNV 10 000 years from now will be about the same as present temperatures. All experts agreed that the overall effect from 3000 years AP to 10 000 years AP will be a net cooling from the 3000 years AP atmospheric temperatures.

2.4.3. Likely future precipitation in the YMNV

The expert's distributions associated with the precipitation estimates are shown in Fig. 3. Each climatologist expressed the opinion that temperature and precipitation were not consistently correlated in the YMNV. Depending on the conditions and controls in place at a specific future point in time, it was equally likely that temperature could rise with increasing precipitation, or that temperature could decrease with increasing precipitation or vice versa. 1993 precipitation was assumed to be 150 mm annually (DOE, 1988). For the first 100 years, four of the five experts agreed that precipitation might increase by as much as 10% on an annual basis while one researcher suggested that precipitation might decrease by a comparable quantity. One expert predicted a significant lessening of precipitation from 100 to 300 years AP while four experts predicted similar precipitation quantities with relatively little change over this time period.

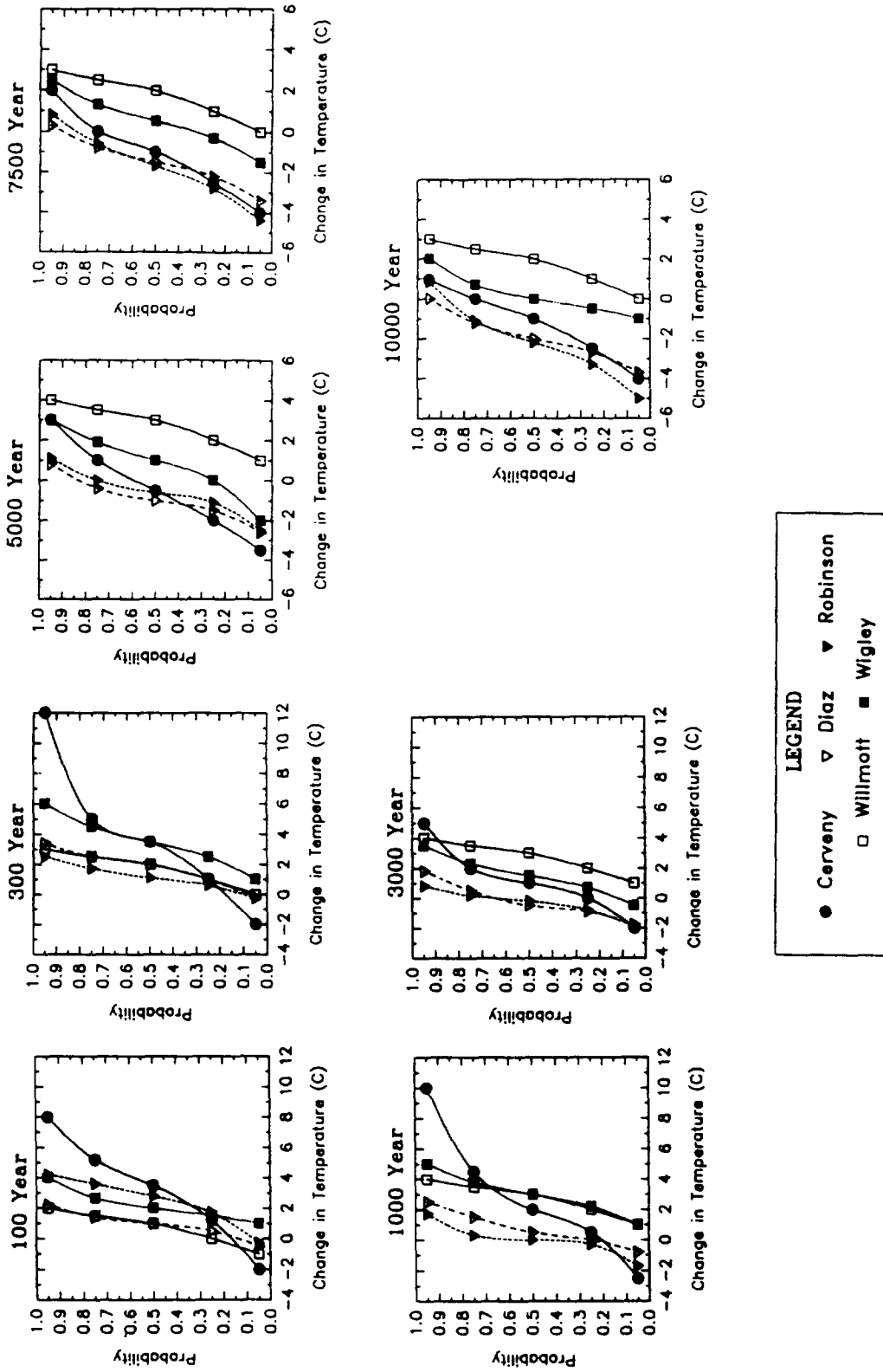


Fig. 2. Cumulative probability distributions for temperature.

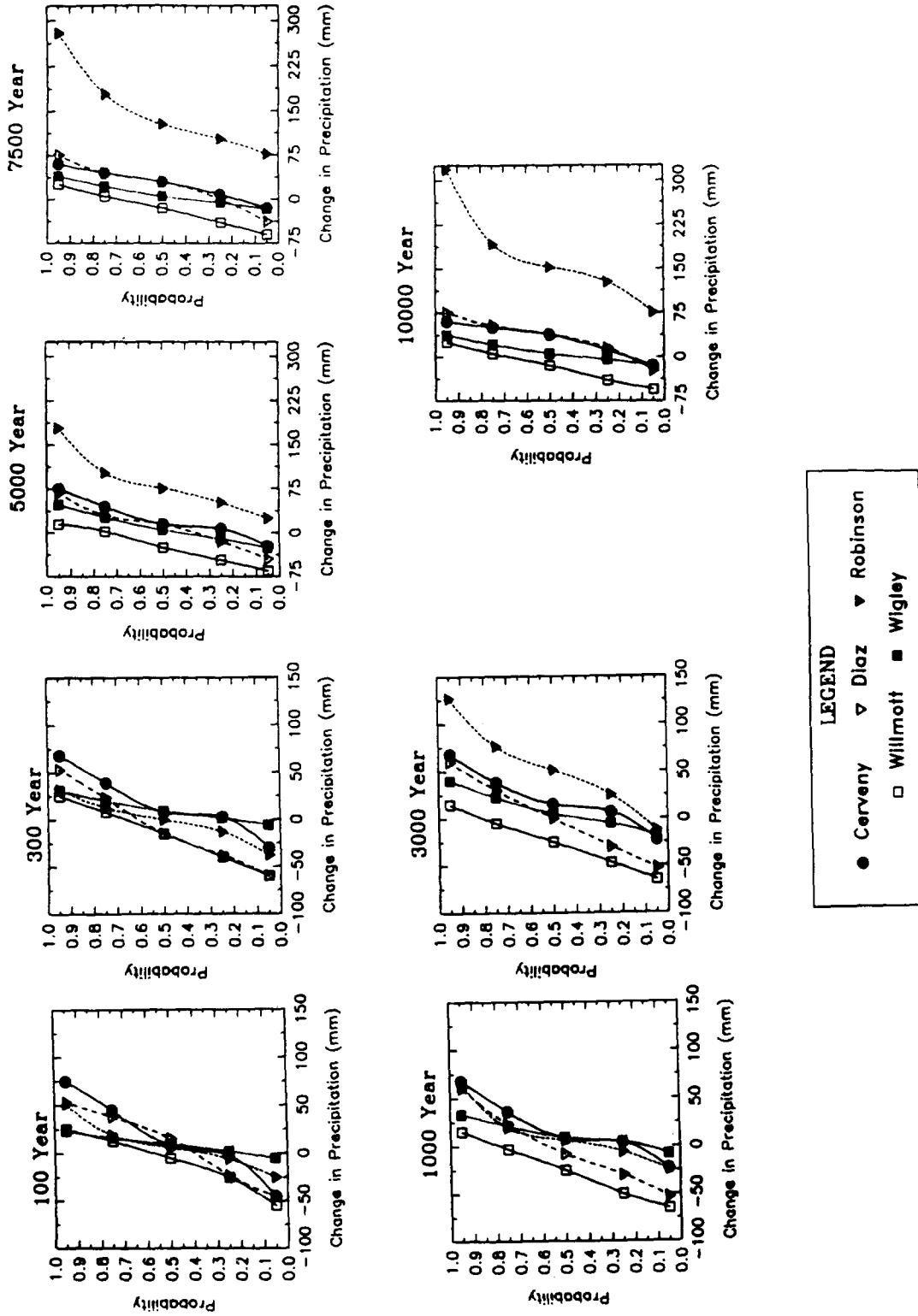


Fig. 3. Cumulative probability distributions for changes in precipitation.

Between 300 and 1000 years AP four experts predicted relatively little change in precipitation. A fifth expert, however, predicted a 25-mm precipitation decrease which would remain for 4000 years before increasing to 15 mm per year in the last 5000 years. This would still represent a 15 mm per year decrease (relative to current) at 10 000 AP. One of the four remaining experts forecast precipitation remaining very similar to 1993 precipitation for the last 7000 years of the 10 000-year time frame, while two researchers showed increases by as much as 40 mm (27%) above present values. The final expert showed a substantial increase in precipitation from 3000 years AP to a high at 10 000 years AP. Final values for this expert at 10 000 years AP are 150 mm more than the current annual average. Other individuals suggested precipitation maximums that were no more than about 30% wetter than the current average. In summary, all but one of the experts predicted increased precipitation during the 10 000-year time period. The median precipitation amounts at 10 000 years AP ranged from a 15-mm decrease (10% decrease) to a 150-mm increase (100 percent increase) over current values.

3. Aggregation of the experts judgements

One of the objectives of this effort was to examine techniques and approaches that would satisfy the occasional requirement for combining or aggregating the judgements of experts. Specifically, we wanted to know whether behavioral aggregation is convenient and appropriate for obtaining a consensus (following individual expert judgements) or if mechanical mathematical combination by the elicitation team (normative experts and generalists) is more appropriate to achieve aggregation.

There were a number of motivations for this experiment. First, an objective was to explore whether the experts could arrive at a consensus opinion following an elicitation which focused on their individual judgements. That is, could the experts agree to a common probability distribu-

tion? If this were possible, it would at least leave open the possibility of generating behavioral consensus distributions for use in risk analysis and PA. However, if it were not possible, this would suggest that either a mechanical aggregation approach or a sensitivity-analysis approach would be preferred. A second motivation was to attempt a comparison of behavioral consensus distributions and mechanically aggregated distributions. The hope was to be able to show that the behavioral consensus mimicked one of the mechanical methods, either in terms of the experts' reasoning about how to generate the distribution or in terms of the shape of the combined distributions. Such a demonstration would in a sense "validate" that particular mechanical aggregation scheme and support the latter's use in lieu of the more expensive behavioral consensus approach.

3.1. Aggregation approaches

The first nine steps of the expert elicitation were used to acquire individual probability distributions from each expert. The advantages of acquiring individual probability distributions include attribution and traceability of judgements to an individual and the voicing of disparate views and perspectives to ensure issue coverage. If only a single combined judgement is required from the group of experts, a panel format or other group opinion technique is generally used (Winkler et al., 1992). These bypass the individual elicitation step previously described.

Although a representative panel of experts may be assembled, it is no surprise that their individual opinions may differ (perhaps substantially) due to different backgrounds, assumptions, and interpretations. Nevertheless, the application of models of a HLW repository in PA may require that individual judgements from multiple experts on any given issue or variable be combined into a single coherent representation, possibly a single probability distribution; thus, obtaining an aggregated representation of the individual judgements may be a critical step in the process. Two ways of aggregating the judgements of experts are behavioral and me-

chanical which are discussed in the next two sections.

3.2. Behavioral aggregation of probability distributions

In behavioral aggregation, the experts themselves produce a combined or consensus view. One method for achieving behavioral consensus is a structured discussion. A wide variety of such techniques are available. Most require a facilitator to guide the discussion. The essence of these techniques is that all participants have an opportunity to express their opinions and that the environment is as non-threatening as possible. In addition to structured discussion, there are non-interactive consensus-building techniques such as the Delphi Method (Linstone and Turoff, 1975). The Delphi Method does not involve face-to-face interaction but relies on a cycle of judgemental assessment and controlled communication of the assessments and reassessments to all participants via the project manager.

While behavioral aggregation is seemingly a natural possibility for the individual experts to use to generate a consensus probability distribution, it brings with it some difficult issues. Having created their individual distributions, they must develop a consensus distribution. To whom, though, does this consensus distribution belong? It is not clear whether an individual expert will actually change his or her opinion in arriving at a consensus. Instead, finding a consensus may be more a matter of negotiation; an expert may agree to concede a point in order that the group appear "of one mind." Another possibility is that an expert's opinion is somewhat loosely held, and thus it is relatively easy to agree with the consensus distribution. Given the lack of commitment to the original assessment, the consensus distribution may not appear to be so different from the expert's assessed distribution.

3.3. Mechanical aggregation of the probability distributions

In contrast to the behavioral approach in which the experts themselves generate consen-

sus, there is a large literature on methods for aggregating individual opinions by means of mathematical formulas. Such formulas are referred to as 'mechanical' aggregation methods because generally the only inputs required from the experts are their individual distributions. A central decision maker or analyst is responsible for the aggregation which is accomplished by applying an aggregation formula. The term 'mechanical' is somewhat misleading in that it suggests that the process is entirely impersonal and always straightforward. This is not strictly true, as application of any given aggregation approach can require subjective judgements on the part of the decision maker or elicitation team and, in some cases, the experts themselves.

The distributions of each expert can be combined using a number of algorithms to produce aggregations (Clemen, 1989). However, in prior forecasting studies the simple average of point forecasts has tended to do as well, (and often better) than more complex methods. Although the simple average has some conceptual drawbacks (e.g., it is not sensitive to differential expert information, quality or dependence) it is computationally efficient. Other mechanical aggregation formulae attempt to account for differential expert information, quality, and dependence, although their application in practice may be difficult (Cooke, 1991; Chhibber and Apostolakis, 1993). The results of these simple weighted averages, with each expert's distributions being equally weighted, are shown in Fig. 4.

3.4. Aggregation issues

Research has demonstrated that behaviorally combined forecasts often perform well relative to the individual forecasts (Hastie, 1986). The studies that have supported this, however, have been conducted largely with students as subjects (Einhorn et al., 1977; Uecker, 1982). It is unclear whether this result would emerge with groups of expert subjects. Studies comparing the accuracy of mechanical aggregation and behavioral consensus techniques are equivocal (e.g., Rohrbaugh, 1979; Lawrence et al., 1986; Flores and White, 1989).

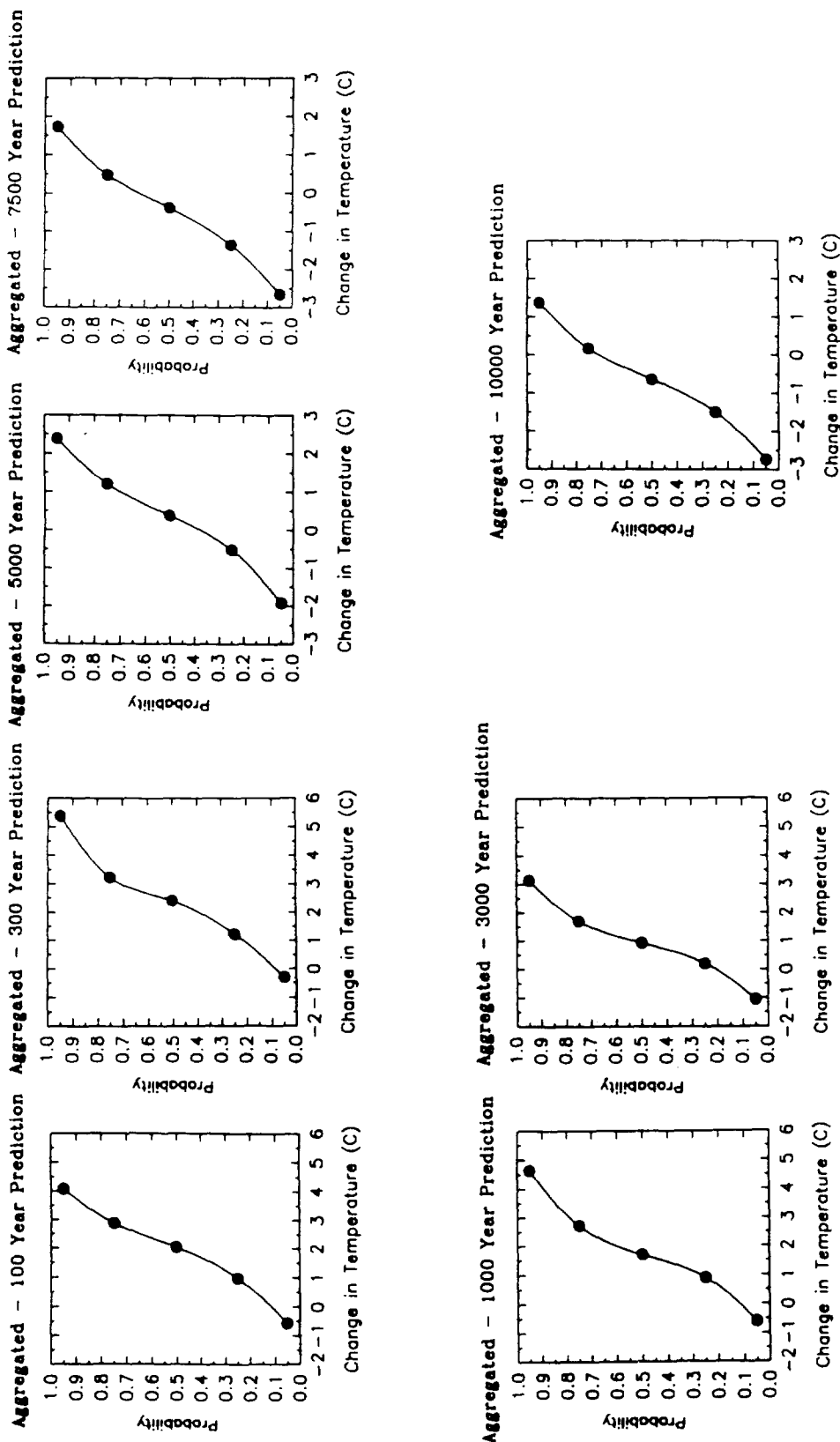


Fig. 4. Mechanically combined (equal weighted average) cumulative probability distributions for change in temperature.

It is not clear that aggregating opinions is always the best thing to do. In fact, it is important to preserve the individual opinions if only because these opinions contain more information than the aggregated distribution (Clemen, 1987). For example, by looking at a combined or consensus distribution, a user of the information cannot reconstruct the individual distributions. Such information may have an important impact on how the decision maker uses the information.

In an intriguing discussion, Bunn (1988) suggests that decision makers may be better served by having the individual distributions and using them as a basis for a sensitivity analysis. For example, the decision maker may simply take the range of opinions and factor each one into his or her decision model. If the chosen decision remains the same, regardless of the distribution used, this indicates that the decision is not sensitive to the uncertainty in that parameter. If the decision is sensitive, it is possible to return to the experts and delve deeper into their reasoning in an effort to refine the information and reduce the uncertainty.

In summary, aggregating judgements across a group of experts provides a distribution that attempts to summarize the state of knowledge across the group. This often needs to be done to meet the needs of the users of the judgements. All aggregation alternatives need to be considered and evaluated in the context of the overall program and selection of a method may depend on the way the users intend to manipulate and present the data. Whichever aggregation technique is used, it is the responsibility of the assessment team to provide the aggregation of the data, following the individual elicitation, rather than the experts themselves.

From a regulatory standpoint, it may be advantageous to retain the individual expert judgements so that the regulator can determine the potential impact of different aggregation rules on the regulatory decision. Preservation of the individual judgements also permits the regulator to examine the diversity of opinion and, hence, obtain a sense of the uncertainty regarding a given variable, quantity, and/or phenomenon

(Fehring and Coplan, 1992). However, this is not necessarily the best approach from a modeling standpoint. Therefore, mechanical aggregation and behavioral consensus techniques were examined in the climatological risk assessment of YMNV.

3.5. *Consensus session results*

To examine these issues, an informal experiment was conducted in the context of the climatology study. In this experiment, structured discussions were held in which the experts were asked to arrive at a consensus probability distribution. In other discussions, they were asked to try to arrive at a consensus only in terms of their reasoning about underlying issues. In an effort to explore, first hand, the various behavior aggregation techniques for combining expert judgements (the climate forecasts in the form of probability distributions were the judgements of interest here), three exercises were conducted which attempted to achieve consensus among the experts on selected distributions. Different supporting information was presented in each of the exercises. The first exercise utilized the 100-year temperature distribution. The experts saw a graph of all five individual temperature distributions and attempted to generate a consensus distribution, either by selecting one of the five temperature distributions or by creating a new distribution. The process of attempting to achieve consensus was fairly structured. The group saw the five distributions, each member commented on his stance and then the group discussed the issues involved. After approximately 1 h, the experts arrived at a consensus distribution (see Fig. 5(a)). This distribution differed from a simple average of the individual distributions because one member of the group conceded part of his stance before the group generated a consensus distribution. New information had become available between the individual elicitation sessions and the group meeting that allowed one member to modify his individual probability distribution. This change allowed the group to agree quickly on a consensus distribution.

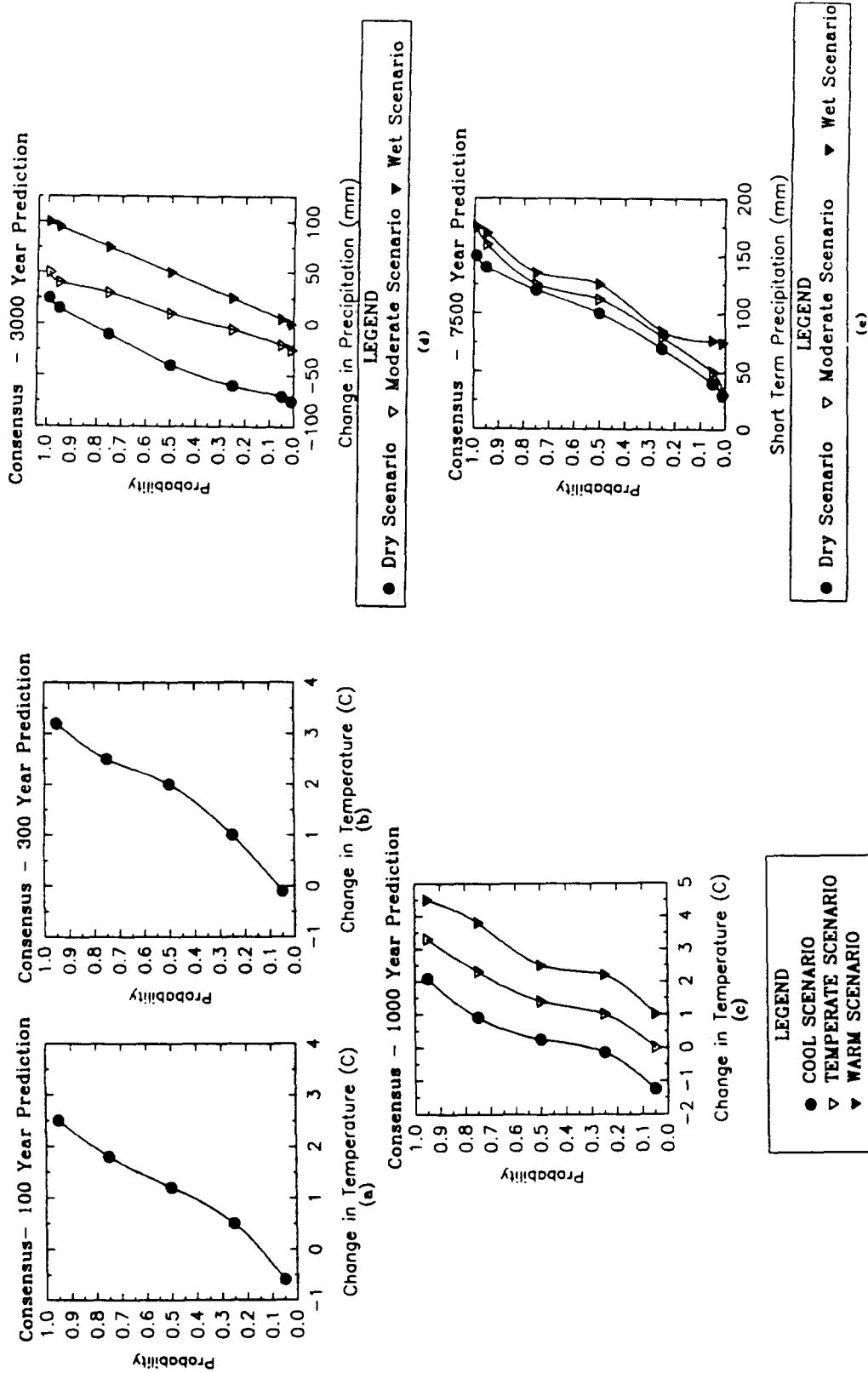


Fig. 5. Cumulative probability distributions from consensus exercises.

The group saw different supporting material in the second exercise. They were charged with generating a consensus distribution after seeing four different mechanically aggregated distributions for precipitation at 3000 years AP. The process was the same as in the first exercise. However, this exercise proved considerably more difficult. Because the elicitation team had agreed to allow the experts to use analyses supplemented by their own data, they were not forced to reach the same conclusion about the current state of the YMNV climate. One of the experts elected to use a slightly lower average annual precipitation level (125 mm), based on his analysis of precipitation records throughout Nevada, than the other experts who used an average of 150 mm. This inconsistent base for annual precipitation confounded the process considerably, making it impossible for the experts to achieve consensus. In the end, the experts agreed to support three separate precipitation curves for use by PA analysts to conduct a sensitivity analysis to determine the impact of the uncertainty in the change in precipitation on the PA results. The curves represent both of the extreme bounds on precipitation with a moderate distribution in between (see Fig. 5(d)).

The difficulty in resolving the precipitation distribution also stemmed from differences in the forecasts of how changes in climate-forcing mechanisms would affect the region. Two of the experts indicated that the YMNV would be somewhat wetter throughout the period. One expert argued that, toward the end of the period, it would be substantially wetter. Two experts thought it would be a somewhat drier through a large part of the 10 000-year time frame. These three possible scenarios defined the boundaries of the consensus discussions and polarized the group. The group reached consensus only by defining three separate curves that preserved their original stance with only minor compromise.

In the third exercise the group saw both the individual and mechanically aggregated distributions. They were charged with generating a consensus distribution for short-term intensity at

7500 years AP. Again, the process was the same: initial presentation followed by open discussion. The group had difficulty generating a consensus distribution. Instead, they generated three separate curves as they had for the 3000-year precipitation distributions [See Fig. 5(e)]. They did eventually agree that, in lieu of a sensitivity analysis, modelers could take the lowest point on the three curves, the median of the middle curve, and the high point on the top curve to generate an approximate consensus distribution for this variable. Although the discussion began with a substantive analysis of the climate forcing mechanisms, it eventually centered primarily around the procedural aspects of devising a curve that would aid the modeling endeavor.

A fourth session, also consisting of a structured discussion, was not centered around a single probability distribution. Rather, the discussion concentrated on substantive issues identified by the generalists in which there was substantial disagreement by the expert panel. The intent of this session was to help the generalists appreciate the diversity of opinion represented by the experts. The first issue was the duration and magnitude of anthropogenic effects on the climate. Some of the experts argued that the maximum warming, due to increased greenhouse gases, would occur relatively soon, for example, within the next 300 years, and taper off out to 3000 years AP. After that, other forcing mechanisms, such as Milankovitch orbital variation, would dominate. Other experts asserted that the effect of anthropogenic warming would be more pronounced and would increase temperature relative to today, throughout the 10 000-year period. Cooling due to Milankovitch orbital forcing would not return temperatures to their present-day values.

To explore this discrepancy, the experts engaged in an open discussion of the issues. They decided to evaluate the 300 and 1000 years AP temperature distributions. Consensus in a single distribution was achieved for 300 years AP (Fig. 5(b)), but the experts could only produce warmer, cooler, and moderate temperature cumulative distributions for temperature for 1000

years AP (Fig. 5(c)). After this, they examined the 10 000-year trends. The experts concluded that, although they agreed on the primary forcing mechanisms (and even, to some extent, when those forcing mechanisms would affect the climate) they could not agree on the magnitude of the effects. This resulted again in agreement on three separate curves for temperature change.

4. Summary and conclusions

4.1. *Application of expert elicitation to produce probability distributions in support of licensing activity*

An extensive amount of data remains to be collected in support of a license application for a HLW repository. Most of this data will be gathered through activities such as site characterization and experimental studies. However, certain data may not be readily obtainable through such investigations, particularly in regard to the prediction of future events which may affect the performance. In these cases, expert judgements may be used to supplement the other data sources. Expert judgements, elicited through a formal expert elicitation process, will be scrutable and the rationales used to reach the judgements will be documented. Probabilities are a natural medium for expressing uncertainty and can be manipulated mathematically. This makes them particularly attractive for the PA program for the proposed HLW repository at YMN. Traceability and rationality of judgements are important considerations during review of a license application.

It is often desirable and expedient to aggregate the results of individual expert judgement elicitation. Conversely, from a regulatory standpoint, it is advantageous to retain the individual expert judgements. Preservation of the individual judgements permits the regulator to examine the diversity of opinion leading to a basis for the uncertainty regarding a given variable, quantity, or phenomenon and to determine the potential

impact of different aggregation rules on the outcome.

4.2. *Elicitation procedure summary and conclusions*

One of the goals of this study was to utilize a state-of-the-art expert elicitation exercise to generate future climate forecasts using probability distributions to represent the uncertainty in the climatological predictions. Probability distributions of four variables (annual temperature, annual precipitation, precipitation intensity, and persistent precipitation) at seven time slices were obtained successfully from five climate experts.

The five experts who participated in the study reported little difficulty in representing their judgements as probability distributions. They felt that the training session was essential to acquaint them with this process. They were also comfortable with generating and evaluating CDFs and did not need to rely on PDFs. The fractile technique proved a very effective elicitation technique. The group can be sensitized to the need for the data they are providing and make reasonable recommendations for its use.

4.3. *Expert judgement aggregation summary and conclusions*

A second objective of this paper was to explore the aggregation of individual expert judgements to determine the adequacy and feasibility of behavioral consensus as compared to mechanical aggregation. Once experts have publicly provided their individual analyses, they rarely admit a change. Our experts agreed to modifications and consensus distributions only on the pretext of establishing a group opinion. They indicated that they did not actually change their minds about their own stance. Finally, the consensus distributions generated behaviorally are not necessarily similar to those generated from mechanical combination techniques. Ultimately, the decision of whether and how to combine the individual distributions is the responsibility of the decisionmaker and the elicitator.

tion team. Efficiency would dictate that a well documented mechanical aggregation of the individual experts judgements (probability distributions) provided by the elicitation team is both a feasible and appropriate way to satisfy the needs of the PA modelers for aggregation and the regulatory concerns for documented individual expert judgements.

The findings of the consensus study are informative but not definitive. This area is appropriate for future research.

Acknowledgements

This paper was prepared to document work performed by the Center of Nuclear Waste Regulatory Analyses (CNWRA) for the Nuclear Regulatory Commission (NRC) under Contract No. NRC-02-93-005. The activities reported here were performed on behalf of the NRC Office of Nuclear Material Safety and Safeguards, Division of Waste Management. This paper is an independent product of the CNWRA and does not necessarily reflect the views or regulatory position of the NRC. This paper is based on a report (DeWispelare et al., 1993) done in support of the NRC.

Appendix A: Sample transcript from an individual elicitation

This transcript starts after the expert had provided his probability distribution for the likely values of temperature at 100 AP. The following is his temperature distribution:

Probability	Change in °F
0.05	-0.05
0.25	3
0.50	5
0.75	6.5
0.95	7.5

At this point, the elicitation team checked his

reasoning to ensure that the spread of the distribution represented his uncertainty and to allow him to articulate the rationale behind his distribution.

Normative Expert: Let's take a look at the tips of the distribution here. When you said 95%, this point here, there's only a 1 in 20 chance that it will be greater than 7.5?

Expert: Yes, and a 1 in 20 chance that it'll be less than -0.5.

Normative Expert: So, you're very confident it's going to warm at 100 years.

Expert: Yes, (laughs) very confident.

Normative Expert: Now, about your median. You're saying it's equally likely to be cooler than 5 degrees as warmer than 5 degrees.

Generalist: These are degrees Fahrenheit?

Expert: Yes, degrees F.

Normative Expert: So, this captures your logic of fairly radical atmospheric perturbation.

Expert: And that, by the way, is consistent with the models (GCM models) that I relied on, although I modified from the models, as I understand it from the latest runs I'm still straight down the middle of the road in my answer for the next 100 years. I'm not making any radical judgement, in fact

it's a very conservative judgement.

Generalist: What's the control operating here? Please restate your position.

Expert: This is almost entirely due to greenhouse gas control leading to global warming. However, I think one would say here that by accepting the models I have accepted the controls that they include, which are driven primarily by increases in greenhouse gases but they do have the topographic controls build in a crude way and the more recent ones do have oceanographic controls as well.

Generalist: Would you say that 5 degrees is what you'd predict with a doubling of greenhouse gases.

Expert: Yes.

Generalist: And how would you explain the 7.5 extreme?

Expert: Ah, two important points to bear in mind. First, all of the models give different results, which are in a sense equally likely or equally unlikely to occur. So, I have worked toward what I believe is the middle of the road of their projections, that is their projections based on 1 to 2 times CO₂ and dealing with very large grid cells, some include San Diego. This is my best esti-

mate bearing in mind that the models vary among themselves and the models refer to a variety of areas. So, by the time you bring it down to the small area of Yucca Mountain, you have increased the uncertainty and these numbers reflect my uncertainty.

References

- 10 CFR Part 60 (U.S. Code of Federal Regulations), *Disposal of High-level Radioactive Wastes in Geologic Repositories*, Part 60, Chapter I, Title 10, "Energy."
- Bonano, E.J., S.C. Hora, R.L. Keeney, and D. von Winterfeldt, 1990, *Elicitation and Use of Expert Judgement in Performance Assessment for High-Level Radioactive Waste Repositories*, Albuquerque, NM: Sandia National Laboratories. SAND89-1821. NUREG/CR-5411.
- Bunn, D.W., 1988, Combining forecasts, *European Journal of Operational Research* 33, 223–229.
- Chhibber, S. and G. Apostolakis, 1993, Some approximations useful to the use of dependent sources of information, *Reliability Engineering and System Safety* 42, 67–86.
- Clemen, R.R., 1987, Combining overlapping information. *Management Science*, 33, 373–380.
- Clemen, R.T., 1989, Combining forecasts: a review and annotated bibliography, *International Journal of Forecasting*, 5, 559–583.
- Cooke, R.M., 1991, *Experts in Uncertainty* (Cambridge University Press, Cambridge).
- DeWispelare, A., L.T. Herren, M. Miklas and R. Clemen, 1993, *Expert Elicitation of Future Climate in the Yucca Mountain Vicinity – Iterative Performance Assessment Phase 2.5*. CNWRA 93-016 (Center for Nuclear Waste Regulatory Analyses San Antonio, TX).
- Einhorn, H.J., R.M. Hogarth and E. Klempner, 1977, Quality of group judgement, *Psychological Bulletin* 84, 158–172.
- Fehring, D. and S. Coplan, 1992, Uncertainty in regulatory decision-making, In: *Proceedings of the Third International Conference, High Level Radioactive Waste Management*, Vol. 1, Las Vegas, NV (American Nuclear Society, Inc) 106–109.
- Flores, B.E. and E.M. White, 1989, Subjective versus objective combining of forecasts: an experiment, *Journal of Forecasting*, 8, 331–341.
- Hastie, R., 1986, Experimental evidence on group accuracy *Information Pooling and Group Decision Making* (JAI Press, Greenwich, CN).

- Kahneman, D., P. Slovic and A. Tversky, 1982, *Judgement Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, MA).
- Lawrence, M.J., R.H. Edmundson and M.J. O'Connor, 1986, The accuracy of combining judgemental and statistical forecasts, *Management Science* 32, 1521–1532.
- Linstone, H.A. and M. Turoff, 1975, *The Delphi Method, Techniques and Applications* (Addison-Wesley, Reading, MA).
- Merkhofer, M.W., 1987, Quantifying judgemental uncertainty: methodology, experience, and insights, *IEEE Transactions on Systems, Man and Cybernetics*, SMC-17, 741–752.
- Mosleh, A., V.M. Bier and G. Apostolakis, 1988, A critique of current practice for the use of expert opinions in probabilistic risk assessment, *Reliability Engineering and System Safety*, 20, 63–85.
- Nuclear Energy Agency, 1987, Uncertainty analysis for performance assessments of radioactive waste disposal system, In *Proceedings: NEA Workshop*. Seattle, WA (OECD/NEA, Paris).
- Nuclear Regulatory Commission, 1992, in R. Codell, N. Eisenberg, D. Fehringer, W. Ford, T. Margulies, T. McCartin, J. Park and J. Randall, eds., *Initial Demonstration of the NRC's Capability to Conduct a Performance Assessment for a High-Level Waste Repository*. (NUREG-1327. Washington, DC: Nuclear Regulatory Commission).
- Raiffa, H., 1968, *Decision Analysis* (Addison-Wesley, Reading, MA).
- Rohrbaugh, J., 1979, Improving the quality of group judgement: Social judgement analysis and the Delphi technique, *Organizational Behavior and Human Performance*, 24, 73–92.
- Savage, L.J., 1954, *The Foundation of Statistics*, (John Wiley & Sons, New York).
- Spetzler, C.S. and C.A. Stael von Holstein, 1975, Probability encoding in decision analysis, *Management Science*, 22 340–352.
- Uecker, W.C., 1982, The quality of group performance in simplified information evaluation, *Journal of Accounting Research*, 20, 388–402.
- von Winterfeldt, D. and W. Edwards, 1986, *Decision Analysis and Behavioral Research*; (Cambridge University Press, New York).
- Winkler, R.L., S.C. Hora and R.G. Baca, 1992, The Quality of Experts' Probabilities Obtained Through Formal Elicitation Techniques. CNWRA Technical Letter Report. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.
- Winkler, R.L., W.S. Smith and R.B. Kulkarni, 1978, Adaptive forecasting models based on predictive distributions, *Management Science*, 24, 977–986.

Biographies: Aaron R. DeWISPELARE received his Ph.D. in Systems Engineering from the University of Virginia. He is a principal engineer at the Center for Nuclear Waste Regulatory Analysis (Southwest Research Institute) where he is involved in the design and implementation of computer databases including relational and text-search applications on multiple hardware platforms and networked support systems. His research interests included systems engineering techniques, multi-attribute decision theory, optimization techniques, and expert judgment elicitation.

L. Tandy HERREN is a senior research scientist in the Training Systems and Simulators Department at Southwest Research Institute. She has been involved in research and development projects in areas such as knowledge acquisition methodologies, computer-based training systems, intelligent tutoring systems, and interactive multimedia user interfaces. Herren received her Ph.D. in Psychology and a MS in Computer Science from Ohio State University.

Robert T. CLEMEN holds a Ph.D. in Business from Indiana University and is Associate Professor in the Lundquist College of Business at the University of Oregon. His research interests include decision analysis, decision theory, and the use and aggregation of expert information. His articles have appeared in a variety of scholarly journals, including *Management Science*, *International Journal of Forecasting*, and *Operations Research*. He is the author of the widely used text *Making Hard Decisions: An Introduction to Decision Analysis* (Belmont, CA: Duxbury, 1991).