DISCUSSION PAPER

Deep and Shallow Uncertainty in Messaging Climate Change

Roger M. Cooke

1616 P St. NW
Washington, DC 20036
202-328-5000   www.rff.org

RESOURCES
FOR THE FUTURE

Deep and Shallow Uncertainty in Messaging Climate Change

Roger M. Cooke

Abstract

Deep and shallow uncertainty are defined and contrasted with regard to messaging the uncertainty about climate change. Deep uncertainty is often traced back to the writings of Frank Knight, where in fact it simply meant subjective probability. Although Knight envisioned a scientifically grounded quantification of subjective uncertainty, deep uncertainty is frequently invoked to disable uncertainty quantification, with attendant problems in communicating and propagating uncertainty through chains of reasoning. These issues, together with science-based uncertainty quantification, are illustrated with recent applications to ice sheet dynamics. The issues of performance assessment and validation are addressed.

Key Words: deep uncertainty, Knightian uncertainty, risk, expert judgment, climate change

Discussion papers are research materials circulated by their authors for purposes of information and discussion. They have not necessarily undergone formal peer review.

Contents

Deep and Shallow Uncertainty in Messaging Climate Change

Roger M. Cooke*

1. The Inevitability of Uncertainty Quantification

The US National Research Council's 2010 report, *Advancing the Science of Climate Change* illustrates reasoning and communicating under uncertainty. Using the calibrated uncertainty language of the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (2005), the NRC report bases its first summary conclusion on "high confidence" (at least 8 out of 10) or "very high confidence" (at least 9 out of 10) in six (paraphrased) statements:[1]

1. Earth is warming.

2. Most of the warming over the last several decades can be attributed to human activities.

3. Natural climate variability ... cannot explain or offset the long-term warming trend.

4. Global warming is closely associated with a broad spectrum of other changes.

5. Human-induced climate change and its impacts will continue for many decades.

6. The ultimate magnitude of climate change and the severity of its impacts depend strongly on the actions that human societies take to respond to these risks.

What is the confidence that all six of these statements hold? You can have high confidence that a randomly drawn integer between 1 and 10 is strictly less than 9 and also high confidence that it is strictly greater than 2. Do you have high confidence that it is one of the six integers 3, 4, … 8? If six statements each have a 0.8 chance of being true, the chance that all are true could be anything from 0 to 0.8. In the non-formalized natural language it is not even clear whether "all statements have a 0.8 chance of being true" means "each statement has a 0.8 chance

---

* Resources for the Future, Washington DC; Delft University of Technology, Department of Mathematics; and Strathclyde University

[1] See National Research Council (2010, 4–5) http://www.nap.edu/catalog.php?record_id=12782): "As discussed in Appendix D, high confidence indicates an estimated 8 out of 10 or better chance of a statement being correct, while very high confidence (or a statement than an outcome is 'very likely') indicates a 9 out of 10 or better chance."

of being true" or "there is a 0.8 chance that all statements are true". Suppose the Nuclear Regulatory Commission licensed a nuclear reactor based on the finding that its safety each year was "virtually certain" (99 percent to 100 percent probability; Mach et al, 2010, Mastrandrea et al 2011). With 104 commercial nuclear reactors, each with a probability of 0.01 per year of a meltdown, ….

It gets worse. Consider the second statement. Are the authors highly confident that 'the earth is warming AND humans are responsible', or are they highly confident that 'GIVEN that the earth is warming, humans are responsible'? These are very different statements. Since the Earth's warming is asserted in the first statement, perhaps the second statement is meant. In that case, the likelihood of both statements holding is the product of their individual likelihoods. If the first two statements enjoy "high confidence", then both can hold with only "medium confidence". Scouring the text for the authors' meaning is futile; were there any determinate meaning, no scouring would be needed. Instead of reasoning under uncertainty, the authors do what most of us do most of the time, they just throw words at it.

The greatest barrier to communicating uncertainty is not some deficiency of the target audience; it is a deficient understanding of uncertainty on all sides. The logic of partial belief is subjective or Bayesian probability. Specialists have known how to "do" uncertainty for a long time, and it involves specialist training. You can't do it by the seat of your pants, as the National Research Council has amply shown. We are now facing decisions that could eventually impact the habitability of the planet as we know it, and we will take those decisions without knowing exactly how current actions affect the future climate. It's all about uncertainty and, not surprisingly, uncertainty has become a key part of the climate messaging: deniers and contrarians use uncertainty to shift the proof burden; alarmists use uncertainty to frighten us into action; science messengers, crafty or clumsy, magnify their certainty. The way forward starts with getting the uncertainty narrative right, and this involves more than hurling words.

In 1977, under prodding from E.A. Feigenbaum (1977), the artificial intelligence community shifted attention from computer chess to "expert systems." Studying the strategies and heuristics of "grand masters" of science, they concluded that the grand masters did not reason probabilistically, and through the 1980s explored "alternative representations of uncertainty" (Cooke 1991), including certainty factors, degrees of possibility, fuzzy sets, belief functions, random sets, imprecise probabilities, and nonmonotonic logic, among many others. The proceedings of the premier conference "Uncertainty in Artificial Intelligence" have been digitized since 1985 and provide a unique record of the development of alternative representations of uncertainty. Figure 1 shows the relative word fragment count of various

approaches in 1985. The largest component is "belief function," followed by "Bayes," "fuzzy," and "certainty factor." Bayes accounts for 26 percent of the total.

Figure 1. Word Fragment Counts in Uncertainty in Artificial Intelligence, 1985
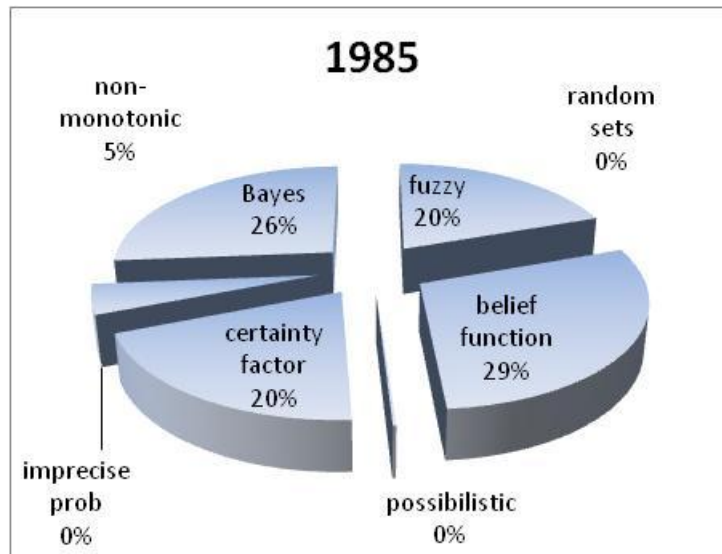
In 2000 the balance has shifted; "Bayes" now accounts for 79 percent of the count. In 2012 the count is 97 percent "Bayes".

Figure 2. Word Fragment Counts in Proceedings of Uncertainty in Artificial Intelligence, 2000 (Left) and 2012 (Right)
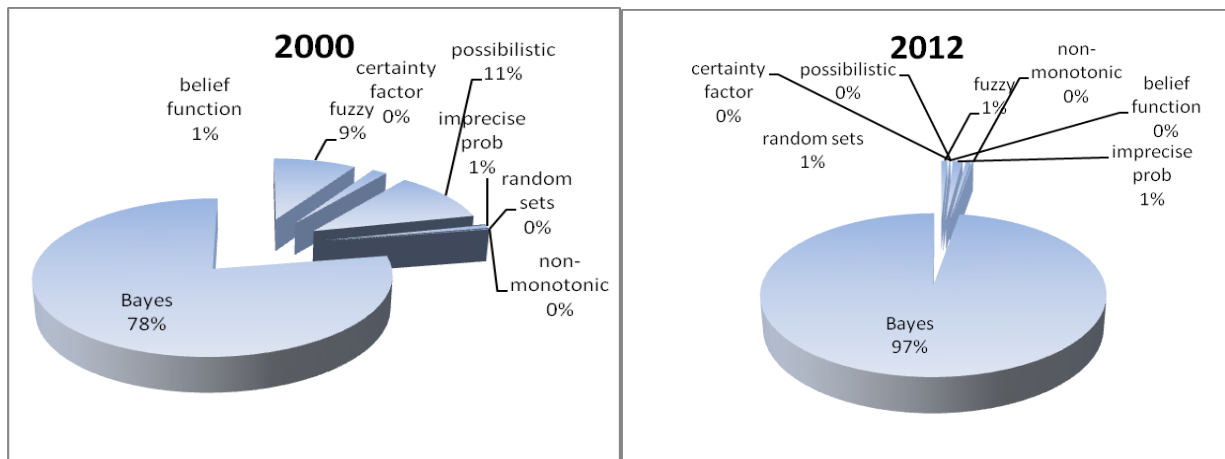
Climate change is the current theater of alternative uncertainties. This paper is a high pass over the representation of uncertainty in climate science, reducing technical jargon to the minimum. The logic of partial belief is rehearsed at the 30,000-foot level. Imprecise

probabilities, deep, and Knightian uncertainties are overflown. For some uncertainties, the apparent depth may result from the lack of operational meaning; this is called shallow uncertainty. The alternative to hurling words at uncertainty is to quantify it through a process of rational consensus based on empirical measures of performance; the last two sections introduce this approach.

2. The Logic of Partial Belief

The subjectivist interpretation of probability goes back to Thomas Bayes in 1763, but the operational measurement of partial belief as subjective probability dates from Ramsey (1931). This brief section rehearses the best modern rendering; namely, that of Savage (1954). Suppose that receiving $10,000 is better than receiving $100 in all situations. The event *France wins the next soccer World Cup* is "qualitatively more probable than" *USA wins the next soccer World Cup* for a subject if and only if (s)he prefers the lottery,

$10,000 if France…; $100 otherwise

to

$10,000 if USA…; $100 otherwise,

and if the relationship is additive:

$10,000 if France OR Belgium…; $100 otherwise

is preferred over

$10,000 if USA OR Belgium…; $100 otherwise.

Other axioms require that the qualitative probability relation is transitive, that multiplying rewards by a positive constant does not change preferences, and that there is a sufficient number of disjunct events. If your preference complies with these axioms then there is a unique probability measure P such that event A is qualitatively more probable than event B if and only if $P(A) \geq P(B)$. Ramsey's initial theory used preference equalities between "certainty equivalents" and gambles, and has led to the persistent misconception that subjective probability is restricted to betting. This is not true, and the step from a qualitative to quantitative probability is quite small. It can be shown that any event can be split into two subsets of equal qualitative probability. It follows that we can make arbitrarily large uniform partitions. Let $A_1 … A_n$ be a partition such that for every i,j, A_i is qualitatively as probable as A_j. Suppose that event X is qualitatively more probable than the union of the first k of the A's and qualitatively less probable

than the union of the first *k+1;* then the probability of *X* is approximately *k/n*, and as *n* gets large, the numerical probability is uniquely determined.

Probability in this sense is the logic of partial belief. Logic does not tell us what to believe, it tells us, for example, that if we believe both *A* and *If A then B* then we should also believe *B*. Logic is normative; it says how we *should* transfer belief. It does not describe how people actually reason. Indeed, most people think that the following is a valid argument:

Only the fittest species will survive.

Cockroaches are the fittest species.

Therefore,

Cockroaches will survive.[2]

The logic of partial belief is similar; it does not tell us the degree to which we should believe that "Bill, who is dull and good in math, is an accountant," but it says that the degree to which we believe that "Bill is a jazz musician" cannot be smaller than the degree to which we believe that "Bill is a jazz musician and an accountant" (Tversky and Kahneman 1982). Yet people commit such errors in droves. The paradoxes of Allais (1953) and Ellsberg (1961) describe choice situations in which many people exhibit behavior violating the axioms guaranteeing the representation of partial belief as subjective probability. McCrimmon (1968) found that business executives willingly corrected violations of the axioms when made aware of them. Other authors (Kahneman and Tversky 1979; Schmeidler 1989; Quiggin 1993; Wakker 2010) account paradoxical choice behavior by transforming the probabilities of outcomes into "decision weight probabilities," which play the role of likelihood in computing optimal choices but do not obey the laws of probability. Wakker (2010, p. 350) notes that decision weighting also fails to describe some empirically observed behavior patterns.

To be sure, the representation of partial belief as subjective probability has limitations.

a) First and foremost, partial belief is *personal*, that is, it pertains to an individual. If a group of individuals' partial beliefs satisfy certain mathematical constraints, then by jointly observing various phenomena and updating their beliefs on the results of

---

[2] Among third-year mathematics students at the Delft Technical University, 80% judged this a valid syllogism. To see that it is not valid, compare: "Only women get pregnant. Maria is a woman. Therefore, Maria gets pregnant."

observations, members of the group will converge to a common probability distribution about those phenomena. Absent this "natural" convergence mechanism, differing partial beliefs are unavoidable.

b) Just as propositional logic does not capture every valid passage of thought, subjective probability does not capture all valid reasoning about partial belief. Events which I myself can readily cause to occur violate Ramsey's condition of "ethical neutrality" and illustrate this limitation. My partial belief in the event that *I will clean my cellar next week* cannot be assigned a probability via lotteries: *$1 if I clean my cellar next week, $0 else* would not be preferred to *$1 if Heads with a fair coin, $0 else*; but multiplying the rewards in both lotteries by 1,000,000 would change my preferences.

c) The theory describes convergence via observation, but says nothing about other ways of reaching consensus in the face of uncertainty. One can obey the rules of logic and still be "irrational" in a wider yet less well-defined sense of the term. The same holds for partial belief.

3. Imprecision, Probability Intervals, and Fuzziness

Imprecise probabilities were introduced by Walley (1991) as differences in certainty equivalents when buying and selling lotteries. Together with fuzzy sets, they have appeared in the climate literature (see, for example, Fu et al. 2005; Hall et al. 2007; Ghosh and Mujumdar 2009; Kriegler et al. 2009). Without disputing the substantive contribution these authors have made, I briefly discuss the use of imprecise probabilities in uncertainty accounting.

The idea is that we, or experts, cannot assess a precise degree of belief, or precise subjective probability P that:

Contribution to sea level rise from ice sheets will exceed 1 meter in 2100.

Instead, experts should give an interval, say [0.1, 0.5] in which P is certain to lie. The bounds of this interval are quite precise, but second, third, or higher order imprecisions can deal with that. Why not take a distribution over this interval and use its expected value as an estimate of P? That would confuse imprecision with probability, goes the response: you can't put a probability on imprecise numbers. Denying the applicability of probability distributions within probability intervals has been explained as follows by Ferson et al. (2007, 19):

> The interval lacks any concentration of probability or likelihood within the interval, so the actual value is not more likely to be at any one place or another

within the interval. But neither is there necessarily a uniform probability distribution over the interval. Instead the actual value has complete latitude to be anywhere within the interval with probability one.

Bounding or "simple interval measures" of course are not new; the question is the extent to which they can aid complex uncertainty accounting. The *Probabilistic Risk Assessment Procedures Guide* (US Nuclear Regulatory Commission 1983, 12) gets it right:

> The simplest quantitative measure of variability in a parameter or a measurable quantity is given by an assessed range of the values the parameter or quantity can take. This measure may be adequate for certain purposes (e.g., as input to a sensitivity analysis), but in general it is not a complete representation of the analyst's knowledge or state of confidence and generally will lead to an unrealistic range of results if such measures are propagated through an analysis.

If different probability intervals are generated by different experts, what do we do with them? Hall et al. (2007) take weighted combinations of the lower bounds to be "conservative." If one has no way to distinguish between good and poor probability interval assessments, then one can do anything else with equal justice. A very large measure of arbitrariness is introduced in this way.

Fuzziness as a representation of uncertainty has also appeared in the context of emissions scenarios (Fu et al. 2005). Regarding the question of whether fuzziness represents uncertainty, the discussion can be very brief. Suppose you get an email from an unknown person named Quincy, and you are equally uncertain that Quincy is a man or a woman (Cooke 2003). The uncertainty that Quincy is a man would be represented by a fuzzy membership function, $\mu_M(Q)$, taking a value in the interval [0, 1] that reflects the degree to which Quincy is believed to be in the set of men (M). A similar function, $\mu_W(Q)$, describes the degree to which Quincy is believed to be in the set of women (W). The uncertainty that Quincy is a man AND a woman would be represented as the membership $\mu_{M \cap W}(Q)$ in the intersection of the sets of M and W. In the original theory, this would be the minimum of $\mu_M(Q)$ and $\mu_W(Q)$. Since you are equally uncertain whether Quincy is a man or a woman, $\mu_M(Q) = \mu_W(Q) = \frac{1}{2}$; and Quincy is a man AND a woman with value ½. It may well be that this sort of heuristic underlies National Research Council's reasoning when concatenating high confidence statements about climate change. Fuzzy combination rules have proliferated, but they all share this feature: the uncertainty of belonging to an intersection of two sets (M and W) is some function of the uncertainties $\mu_M(Q)$ and $\mu_W(Q)$ and does not depend on M and W themselves. Partial belief simply does not behave that way.

4. Deep and Knightian Uncertainty

Denying the application of probability within intervals of imprecision is related to the notion that there are "deep" uncertainties which defy quantification. One often hears that climate change is rife with deep uncertainty. The first documented use of the term appears to be the 2003 Senate testimony of the late Stephen Schneider:

> In fact, the climate change debate is characterized by deep uncertainty, which results from factors such as lack of information, disagreement about what is known or even knowable, linguistic imprecision, statistical variation, measurement error, approximation, subjective judgment, and disagreement about structural models, among others (see Moss and Schneider 2000).

A search for a precise definition came no farther than: "By deep uncertainty we mean uncertainty that results from myriad factors both scientific and social, and consequently is difficult to accurately define and quantify" (Kandlikar et al 2005). For an uncertainty analyst, that is standard fare. Although Moss and Schneider (2000, p. 36) advocate a Bayesian or subjectivist approach in which experts assess their subjective probability distributions, "deep uncertainty" seems to have morphed into apology for not quantifying uncertainties. Modelers assigning values of dubious pedigree to variables need not concern themselves with quantifying uncertainty because it is deep. The sobriquet is that deep uncertainty is "Knightian."

Economist Frank Knight's book, *Risk, Uncertainty, and Profit* appeared in 1921, the same year as John Maynard Keynes' *Treatise on Probability*, well before F.P. Ramsey's (1931) operational definition of partial belief in terms of observed preference behavior and before R. von Mises' (1928) frequentist interpretation of probability. The latter works have framed most of the subsequent discussions on the foundations of probability. Keynes believed that probabilities of various events were incommensurable and that probabilities should be organized as partial orderings. For Knight, "risk proper" is measurable by resolving outcomes into equi-probable alternatives (Knight 1921, III.VII.34). By aggregating risks, the law of large numbers converts losses into fixed costs, and these would not give rise to profits. "Uncertainty," in contrast, concerns "partial knowledge" for which "the conception of an objectively measurable probability or chance is simply inapplicable" (Knight 1921, III.VII.47). Many authors have seized on such statements to argue that uncertainty in climate change is unquantifiable. A typical example is Claude Henry (2006), who also performs a post-mortem baptism of Heisenberg into his faith:

> Keynes and Knight make a clear distinction between two kinds of uncertainty: the first one, called risk, may be characterized by probabilities, while this is not possible for the second one. Here we deal with decision-making under

genuine uncertainty, no probability distributions being available ….Uncertainty in quantum mechanics is strictly probabilistic, and Werner Heisenberg, had he been an economist, would have called "principle of risk" his famous Uncertainty Principle.

Many economists, not the least of whom is Sir Nicholas Stern (2008), have averred that where "we don't know the probability distribution, then "Knightian uncertainty" kicks in, which cannot be characterized by probabilities. Regrettably, these authors did not read further in Knight (1921, III.VIII.1): "We can also employ the terms 'objective' and 'subjective' probability to designate the risk and uncertainty respectively, as these expressions are already in general use with a signification akin to that proposed." Knight, writing in 1921, did not know how to measure subjective probabilities. Neither did he know how to measure "risk" or objective probabilities. The Laplace (1814) interpretation of probability to which Knight's definition of risk appeals, has been dormant, if not dead, since the end of the 19th century. It was, after all, von Mises (1928) who emphasized that objective probabilities can be measured as limiting relative frequencies of outcomes *in a random sequence*. Like many authors of this period, Knight appears to have been unaware of the role of (in)dependence assumptions and believed that objective probabilities are much more objective than modern probability warrants. It is indeed significant that economists claiming that climate uncertainty cannot be described with probability harken back to a period when probability, both objective and subjective, was not well understood.

This is not to gainsay that Knight was ahead of his time. The idea of calibrating the probability judgments of the individual "business man" prefigures the modern use of structured expert judgment (Knight 1921, III.VII.43):

> A still more interesting complication, and one of much greater practical significance, is the possibility of forming a class of similar instances on entirely different grounds. That is, instead of taking the decisions of other men in situations more or less similar objectively, we may take decisions of the same man in all sorts of situations. It is indisputable that this procedure is followed in fact to a very large extent and that an astounding number of decisions actually rest upon such a probability judgment….

5. Shallow Uncertainty

Shallow uncertainty is uncertainty resulting from undefined terms, careless formulation, lack of operational definitions and overall intellectual sloth. The good news is that shallow

uncertainty, originating from our own insouciance regarding the meaning of words, is much easier to remove than uncertainty about our future climate.

As an example (Cooke 2013), consider the social discount rate (SDR), often written as:

$$SDR = \rho + \eta G(t),$$

where ρ is the rate of pure time preference, η is the coefficient of constant relative risk aversion, and $G(t)$ is the time average growth rate of per capita consumption out to time t. It is generally recognized that the discount rate is an important driver, if not *the* most important driver, in the economic models for climate change. Some (e.g., Stern 2008) see a strong normative component. Others infer values for ρ and η from data (Evans and Sezer 2005). Nordhaus (2008) equates the SDR to the observed real rate of return on capital with a constant value for $G(t)$ and sees ρ and η as "unobserved normative parameters" (Nordhaus 2008, p. 60) or "taste variables" (p. 215) that are excluded from uncertainty quantification. Pizer (1999) assigns distributions to ρ and η. Nordhaus and Popp (1996) put a distribution on ρ. Weitzman (2001) fits a gamma distribution to SDR based on an expert survey. Frederick et al. (2002, 352) note that "virtually every assumption underlying the DU [discounted utility] model has been tested and found to be descriptively invalid in at least some situations." They also cite the founder of DU, Paul Samuelson: "It is completely arbitrary to assume that the individual behaves so as to maximize an integral of the form envisaged in [the DU model]" (Frederick et al. 2002, 355). Weitzman (2001) and Newell and Pizer (2003) show that uncertainty in the discount rate drives long-term rates down.

We may distinguish variables according to whether their values represent

a) Policy choices

b) Social preferences

c) Unknown states of the physical world.

Uncertainty quantification is appropriate for (b) and (c); it is not appropriate for (a) if we are the choosers of policies. Various authors assign time preference and risk aversion to both (a) and (b).

Questions arise as to the operational meaning of society's rate of pure time preference. How, with sufficient means and license, would this be measured? Similar questions pose themselves regarding society's coefficient of risk aversion, the utility of consumption, the preferences of a representative consumer, etc. If the modeling community has not agreed on

operational definitions for such terms, then the models are not ready for uncertainty analysis. If we don't know what it means, we can't be uncertain about it.
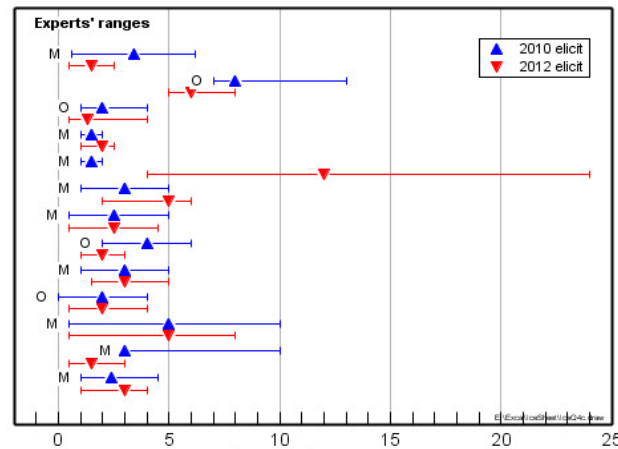
6. Experts

Saying "we don't know the probability distribution" implies that probability is an objective property of the world that we happen not to know. Partial belief is not a property of the world in this sense. Introspection suffices for quantifying partial belief; there is nothing to not know. The rub is that each individual is entitled to his/her own subjective probability distributions. Those most knowledgeable about a scientific field, experts, will also have differing degrees of partial belief about variables of interest. Expert disagreement is inherent in choice under uncertainty. Indeed, if the science 'isn't there yet', experts are supposed to disagree; without disagreement there could be no progress. This section broaches the issue of using experts to quantify uncertainty. If shallow uncertainty regarding the meaning of words is under control, there is no impediment to experts quantifying their degrees of belief as subjective probability distributions.

Data is helpful at this point. Figure 3, from Bamber and Aspinall (2013), shows the results of eliciting subjective probability distributions from 12 experts regarding the contribution to sea level rise from the Greenland ice sheet in the year 2100 in millimeters per year. The same experts were elicited in 2010 and in 2012. "M" refers to an expert who is primarily a modeler, and "O" denotes an observationalist. Experts assessed the 5th, 50th, and 95th percentiles of their subjective probability distributions. A value of 3 millimeters per year in 2100 means that by 2100 the Greenland ice sheet will have contributed a total of roughly 15 centimeters to sea level rise.[3]

---

[3] This assumes effectively zero contribution in 2000 and linear growth rate of the contribution to its 2100 value.

Figure 3. Contribution to Sea Level Rise from the Greenland Ice Sheet in 2100 (in millimeters per year) from Bamber and Aspinall (2013, Supplementary Figure 8)

Notes: M = modeler; O = observationalist.

The experts' 95th percentiles for this total contribution range from 15 to 120 centimeters for the 2012 elicitation. What are we to do with this? Clearly, we could cherry pick different experts and convey very different messages. Some authors argue that the experts' distributions should not be combined; the diversity of expert views is the proper message to the decision makers (Morgan et al. 2009; Morgan and Henrion 1990). Indeed, if the raw, uncombined expert data provides sufficient information to make a decision, then combination is unnecessary. In many cases, like the Greenland ice sheet example, the raw data is not sufficient; if we are 95 percent sure that this contribution will be less than 15 centimeters in 2100, our decisions will be different from those we would make if the chance of a contribution exceeding 1.2 meters were 1 in 20.

A strong argument against combination is that any combination procedure requires justification to deflect allegations of cherry picking. Combining the experts' distributions with equal weight (EW) entails justifying that experts perform these probabilistic assessment tasks equally well. What does it mean to perform probabilistic assessment tasks well or badly? A discussion of performance must precede a discussion of combination. Expert performance data was collected for the ice sheet study, but was not published;[4] we thus turn to other sources.

---

[4] Publications with performance data are in preparation at this writing.

6.1 Expert Performance

An expert assessing an unknown quantity can be regarded as a new sort of measuring instrument. The scientific employment of this instrument requires trying it out on things we know before applying it to things we don't know. With sufficient time and patience, we could ask experts to give point predictions of things we later observe and assess each expert's error distribution. It doesn't sound easy, but it sounds easier than it is. How do we add the error distance for an item expressed in nanometers with another expressed in kilograms per square kilometer? If we change the first to meters, should we change the second? Even if the variables are both in years, an error of three years in assessing the mean age of an aircraft fleet is not the same as an error of three years in assessing the age of your youngest daughter. To assess performance, the variables must be converted to a common scale, and the common scale is probability. Instead of an error distance between an estimated and observed quantity, performance will focus on "how surprised" the expert should be by the observed quantity. That requires that experts assess their subjective probability distributions for the unknown quantities. Experts then become statistical hypotheses, amenable to empirical control via standard statistical methods.

A recent study of information security (Ryan et al. 2012) used 10 calibration questions to score performance. Figure 4 shows the 5th, 50th, and 95th percentiles of the 2 best and worst performing experts from a panel of 13 as well as the realizations (denoted by #). The calibration questions concerned attack rates on different information systems. Expert 1 catches 1 of the 10 realizations in his 90 percent central confidence band, whereas expert 2 catches 9 out of 10. Should we use such information, and if so, how?

Figure 4. Best and Worst Performing Experts from Ryan et al. (2012)
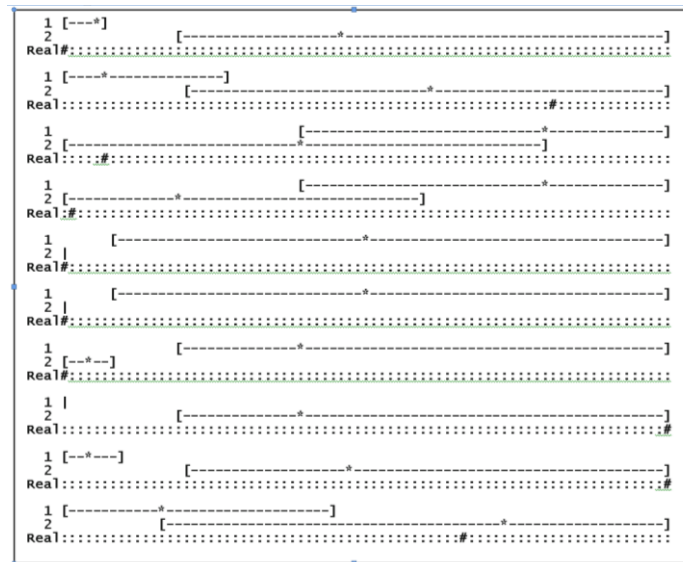
Notes: 1 = expert 1 (worst performing); 2 = expert 2 (best performing); [--*--] = 90% central confidance band, with median ' *; # is the true value (unknown at the time of the elicitation).

6.2 Bayesian Approaches

Most statisticians, confronted with this sort of data for the first time, would seek to combine the experts' performance information through some form of Bayesian updating. Researchers have made many proposals in this direction, starting with Morris (1977), yet they have come to naught. With a bit of formalism, it is easy to understand why (Cooke 1991).

Suppose we have two unknown continuous quantities, X_1 and X_2. One expert gives his cumulative distribution functions F_1, F_2 for X_1 and X_2. The decision maker also has prior distributions G_1, G_2 for X_1 and X_2. The decision maker wants to use the observed value of X_1, say x_1, to update his distribution about X_2. The decision maker has a joint distribution for X_1 and X_2, and he should simply update that, whereby the information from the expert plays no role. However, suppose that somehow the decision maker can't do that and instead wants to update his opinion about this expert based on the observation of x_1. For any *r* in *(0, 100)*, *{$F_1(X_1) \leq r/100$}* denotes the event that X_1 takes a value less than or equal to the r-th percentile of the expert's distribution. The expert evidently believes that this event has probability *r/100*, but the decision maker may assign it a different probability. If the decision maker thinks this expert is likely to underestimate X_1, then his probability of *{$F_1(X_1) \leq r/100$}* will be less than *r/100*. The decision maker will update his probability of event *{$F_2(X_2) \leq r/100$}* based on observing *{$F_1(X_1) \leq r/100$}*.

This requires specifying the decision maker's joint probability of these events. However, if he does that, problems will arise. The simplest case is that these two events are *exchangeable*, for any *r*. This entails that, for the decision maker, these two events are equally likely for any *r*. Now we do a simple calculation, where P denotes the decision maker's probability and, assuming the distribution functions are all invertible,

$P(\{F_1(X_1) \leq r/100\}) = P\{(X_1 \leq F_1^{-1}(r/100)\}$

$= G_1F_1^{-1}(r/100) = P(\{F_2(X_2) \leq r/100\})$

$= G_2F_2^{-1}(r/100).$

From this, we infer that, for any *r*,

$F_2^{-1}(r/100) = G_2^{-1}G_1F_1^{-1}(r/100).$

The expert's distribution for X_2 is completely determined by his distribution for X_1 and the decision maker's priors. The decision maker needn't consult experts at all!

6.3 Rational Consensus

The Bayesian paradigm is a paradigm for *individual* decision making. An expert judgment study can have any of several possible goals: an expert survey, a political consensus seeking an equilibrium of stakeholders, or a rational consensus. Rational consensus is not formally defined, but the idea is this: for combining their opinions, experts pre-commit to a method of combination that satisfies necessary conditions of the scientific method. Then the method is executed. Experts needn't adopt the results as their personal probability distributions, but withdrawing from the rational consensus incurs a proof burden. They must argue that other necessary conditions for the scientific method have not been met. Withdrawing for other reasons—for example, that the results are hostile to one's interests—does not threaten rational consensus.

For details on combining experts see Cooke (1991), Cooke and Goossens (2008), Aspinall (2010) and Aspinall and Cooke (2013). A few points can be recalled here. The most important requirement of the scientific method is empirical control. Each expert–hypothesis must be empirically tested on the basis of calibration variables from the experts' field whose values are known post hoc. Figure 4 is an example. For each item, each expert's percentiles (5th, 50th, and 95th) define four intervals to which (s)he assigns realization probabilities 0.05, 0.45, 0.45, and 0.05. The statistical hypothesis being tested is that the realizations can be viewed as

independent samples from this distribution.[5] The *p*-value of each hypothesis, given the realizations, is the probability of falsely rejecting the hypothesis that the expert's probability statements are true. For expert 1 in Figure 4, the *p*-value of the corresponding statistical hypothesis is 5.4×10^{-8}. For expert 2, it is 0.314. I hasten to add that we are not rejecting expert–hypotheses; instead we are using the language of hypothesis testing to measure how well an expert's probability statements accord with the realizations. Note the eight-orders-of-magnitude difference in these *p*-values. Statistical likelihood is a fast function. If you believe that each of 10 independent events has a 90 percent probability of occurring, then the likelihood of seeing only 1 of the 10 occur is $10 \times 0.9 \times 0.1^9 \sim$

10^{-8} (the computation for the data in Figure 4 is a bit more complicated, owing to the presence of the 50th percentile and a chi square approximation). Each expert's statistical accuracy is measured by the corresponding *p*-value. The maximal value is one, the minimal value is zero. Less important than statistical accuracy, but still important, is the expert's informativeness. Intuitively, this is measured as the expert's ability to concentrate high probability mass in small intervals. The measure of informativeness must not depend on the scale of the underlying variables. Details are found in the above references. These performance measures are combined so as to satisfy a strictly proper scoring rule constraint: in an appropriate long-run sense, an expert maximizes his/her expected score by, and only by, stating percentiles corresponding to his/her true beliefs. It must be emphasized that measuring expert performance and combining experts satisfies constraints devolving from the scientific method but are in no way unique. In the same way, a bicycle satisfies Newton's laws but does not follow from these laws; it is designed to serve a purpose. The performance-based combinations of expert judgments are designed to serve the purpose of finding rational consensus on the quantification of uncertainty. Better bicycles are always welcome, but not having the perfect bicycle is a weak argument for walking.

---

[5] In this context, independence is not an assumption about the experts' joint distribution; it is a desideratum of the decision maker. If the experts' joint distributions are dependent, then the expert will learn from observing the calibration variables. Expert learning is not a goal of the elicitation, and the decision maker desires experts who have already learned. If the calibration variables are multiple realizations of the same random variable, or are otherwise confounded, then standard techniques for reducing the power of the statistical test apply.

7. Expert Judgment Validation

Forty-five professionally contracted expert judgment studies with domain experts from fields including nuclear safety, aerospace and aviation risk, environmental transport, finance, volcanology, and public health were reviewed by Cooke and Goossens (2008) in a special issue of *Reliability Engineering and System Safety*. The data, including calibration variables, were made available to researchers. Since 2008, the number of studies has nearly doubled, expanding into areas like ecosystem modeling, project management, and climate modeling. All studies, both pre and post 2008, compared an equal weight (EW) combination with a performance-based weighted (PW) combination. Measuring performance transforms the field of expert judgment from a confessional to a science-based and data-driven activity. Any combination of experts' distributions can be tested against the set of calibration variables and its performance scored. When scored on the calibration variables, EW usually—but not always—produces statistically acceptable performance at the expense of low informativeness. PW combinations usually—but not always—perform better on both counts. Both PW and EW may be scored for statistical accuracy and informativeness; combining scores for statistical accuracy and informativeness yields a "combined score" (Cooke and Goossens 2008).

In a few studies, variables of interest were later observed, enabling true out-of-sample validation (Cooke and Goossens 2008). In most cases, the variables of interest are not observable on timescales relevant for the decision problem. Therefore, various forms of cross-validation have been suggested.

The most extensive cross-validation study is by Eggstaff et al. (2014), who initialize the model on *all* nonempty subsets of calibration variables and in each case compute scores based on predictions of the complementary subset. Thus, each score is based on out-of-sample predictions. In total, 62 expert judgment studies were analyzed.

Studies differ in subject matter, number and training of experts, the methods of recruitment, and methods of elicitation. To study the effects of these factors, a numerical measure of out-of-sample validity at the study level would be desirable. For N calibration variables, for each $K = 1 ... N-1$, Eggstaff et al. (2014) initialize the PW combination on each K-tuple of variables in the training set and compare PW and EW scores based on prediction of $N-K$ out-of-sample variables in the test set. The same experts, the same calibration variables, and the same information background measures apply for all training set choices within one study. However the statistical power of the test set goes down as the test set size decreases (training set size increases). Moreover, there are many more K-tuples for values of K near $N/2$, and these

cases have overlapping training sets. With this in mind, the PW and EW combined scores are averaged for each size *K,* for *K = 1 … N–1*. The ratio of PW and EW scores is dimensionless and can be compared across training set sizes. In aggregating ratios of positive numbers, we must take the geometric mean, or geomean.[6] For details and references, see (Cooke 2014).

The geomeans of the ratios of combined scores of all comparisons for each of the 62 studies are plotted in Figure 5. In 45 of the 62 studies (73 percent), the geomean of combined score ratios, PW/EW, was greater than unity. When the PW combined score exceeded that of the EW, it tended to exceed by a greater amount than when the EW combined score exceeded that of the PW. The best eyeball assessment is to compare the mass of lines above and below the baseline of 1. The geomean of the geomeans for each study was 2.46. Summarizing, PW outperforms EW in out of sample cross validation on more than two thirds of the studies, and the combined score of PW is more than twice that of EW.

The accuracy of a Decision Maker (DM) in terms of the proximity of the median to the true value is not directly related to the scoring variables of statistical accuracy and informativeness. Eggstaff et al. (2014) report an accuracy advantage of PW over EW comparable to the ratios of combined scores.

---

[6] To see this suppose on two comparisons the scores were (PW=4, EW=1) and (PW=1, EW=4). The performance is identical, but the average of ratios is 1/2(4+1/4) =2.125. The Geomean is $(4 \times 1/4)^{1/2}$=1. Eggstaff et al report only the average scores for each size of the training sets, so we consider the ratios of averages. Since the average is always greater or equal to the geomean, the numerator and denominator in these comparisons would both be smaller if we took the geomeans of combined scores of each separate *K*-tuple of training variables. It's impossible to say if there is an overall effect of this choice.

Figure 5. Geomeans of Comparisons of PW/EW Combined Score Ratios per Study for 62 Studies
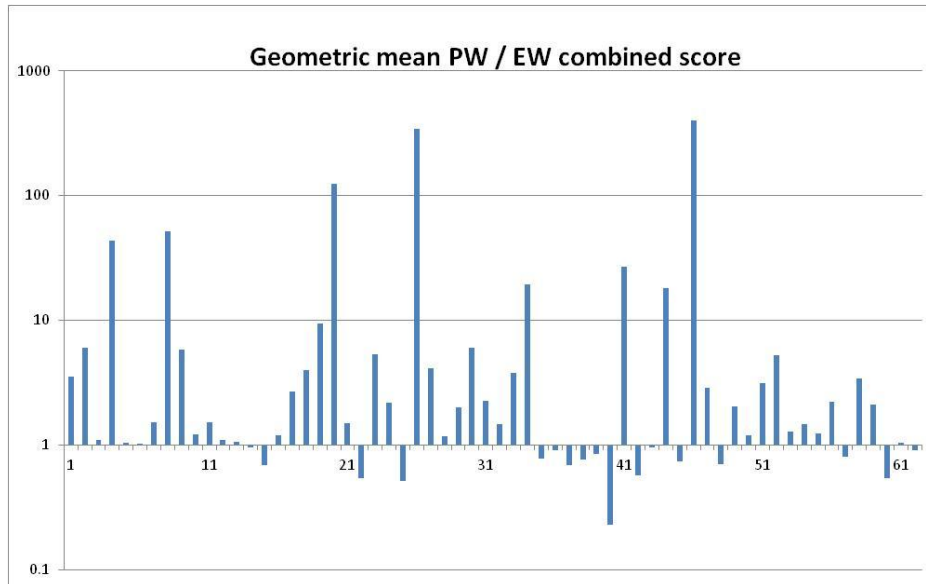
The extensive dataset of expert elicitations has afforded other insights into expert judgment. Based on a number of studies in which experts were very well published, it was possible to derive weights based on experts' citation indices (Cooke et al. 2008). Performance was comparable to EW and inferior to PW. On a few occasions, the analyst asked experts to rank each other and compared these ranks with their performance on calibration variables. The findings for elicitations on ice sheets and dam safety in Figure 6 speak for themselves.

Figure 6. Expert Mutual Self-Weights and Performance Ranking: Ice Sheets (November 2012 elicitation, unpublished), Top, and Dam Safety, Bottom (Aspinall and Cooke 2013)
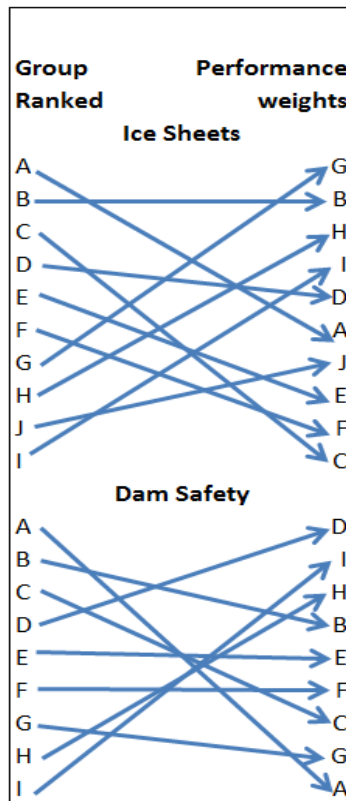
This should not come as a total surprise. Little in an expert's training prepares him or her for probabilistic assessment. The example from the National Research Council reminds us that reasoning under uncertainty is hard, and facile intuitions easily mislead. Some experts are able to give very accurate and informative assessments, but many others are less able to do this. Unfamiliarity with probabilistic assessment will be reflected not only in expert scores but in experts' ranking of colleagues for probabilistic assessment tasks. This simple observation has noteworthy consequences. Expert nominations for advisory roles, blue ribbon panels, and the like may identify experts with substantive knowledge, influence, and gravitas in their field, but the supposition that this is a good way to quantify uncertainty is not supported by the evidence.

Returning to the Greenland ice sheet, Figure 7 shows combinations based on self-weights, EW, and PW, for elicitations in 2010 and 2012, respectively.

Figure 7. Greenland Ice Sheet Contribution to Sea Level Rise in 2100, with Combined Experts (Bamber and Aspinall 2013, Supplementary Figure 8)
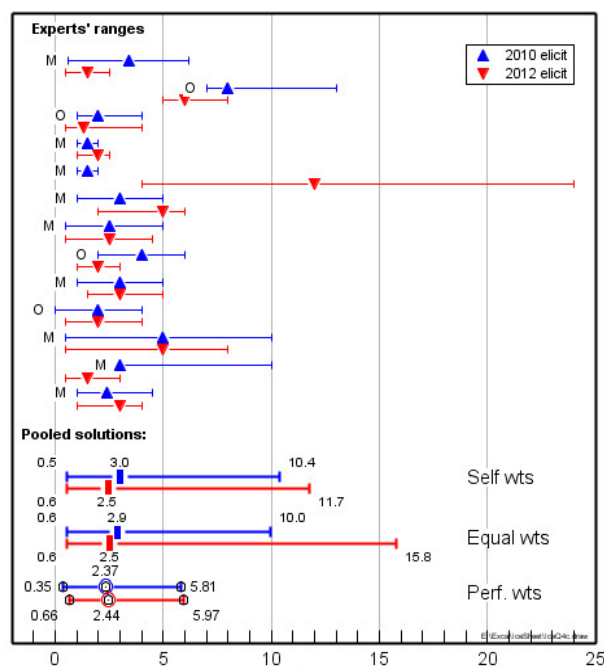
From the more recent elicitation, the equal weights decision maker's 95 percentile for total sea level rise in 2100 due to the Greenland ice sheet is about 80 cm, and for self weights about 60cm; for the performance-based decision maker the value is close to 30 cm.

8. Conclusion: Risk and Uncertainty

Quantitative uncertainty analysis emerged with risk analysis with structured expert judgment (reviewed in Cooke 2013) and has spread in areas where decision making under uncertainty is paramount. As yet it has played a small role in the climate debate. This is undoubtedly related to the fact that no government agency is charged with managing or regulating climate. The IPCC does not do research and cannot commission uncertainty studies; it can only report on what has been done by others. However, the semantics of uncertainty that the IPCC has adopted and published as guidance for lead authors is unhelpful and ultimately insufficient. Even though uncertainty qualifiers, such as "likely" and "confident," are given a precise meaning, they cannot be propagated through a chain of reasoning and, more importantly, they encourage defective reasoning under uncertainty. This is not due to an affliction of the nontechnical lay public. The experts at the National Research Council, after ripe deliberation, have dispelled that idea most convincingly.

Whether deep and Knightian uncertainty in the climate debate will go the way of fuzzy sets, imprecise probabilities, random sets, belief functions, and degrees of possibility in artificial intelligence remains to be seen. The best argument against the claim that something can't be done is to do it. Science-based uncertainty quantification in climate change can be done, has been done, and should be done much more often.

References

Allais, M. (1953) Le comportement de l'homme rationnel devant le risque. *Econometrica* 21, 503–546.

Aspinall, W. (2010) A route to more tractable expert advice. *Nature* 463, 294–295.

Aspinall, W.P., Cooke, R.M. (2013) Quantifying scientific uncertainty from expert judgement elicitation. In "Risk and Uncertainty assessment in Natural Hazards." Hill, L., Rougier, J.C., Sparks R.S.J. (eds). Cambridge University Press, Chapter 4, 64-99. ISBN: 9781107006195

Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S., communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. *Philosophical Transactions of the Royal Society of London* 53, 370–418. http://rstl.royalsocietypublishing.org/content/53/370.full.pdf+html (accessed April 16, 2014).

Bamber, J.L., and Aspinall, W.P., (2013) An expert judgement assessment of future sea 1evel rise from the ice sheets, Nature Climate Change,   PUBLISHED ONLINE: Janurary 6, 2013 | DOI: 10.1038/NCLIMATE1778.

Cooke, R.M. (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York, NY: Oxford University Press.

Cooke, R.M. (2003) Elicitation of expert opinions for uncertainty and risks. *Fuzzy Sets and Systems* 133, 267–268.

Cooke, R.M. (2008) Response to comments. *Reliability Engineering & System Safety* 93 (5), Special issue on expert judgment, 775–777.

Cooke, R.M. (2013) Uncertainty analysis comes to integrated assessment models for climate change … and conversely. *Climatic Change* 117 (3), 467–479. http://dx.doi.org/10.1007/s10584-012-0634-y (accessed April 16, 2014).

Cooke, R.M., ElSaadany, S., and Huang, X. (2008) On the performance of social network and likelihood based expert weighting schemes. *Reliability Engineering & System Safety*, 93 (5), Special issue on expert judgment, 745–756.

Cooke, R.M., Goossens, L.H.J. (2008) TU Delft Expert Judgment Data Base. Reliability *Engineering & System Safety*, 93 (5), Special issue on expert judgment, 657–674.

Cooke, Roger M. (2014) Validating Expert Judgments with the Classical Model" in Experts and Consensus in Social Science - Critical Perspectives from Economics, Sociology, Politics, and Philosophy. Editors: Carlo Martini and Marcel Boumans, Series title: Ethical Economy - Studies in Economic Ethics and Philosophy, Springer

Eggstaff, J.W., Mazzuchi, T.A., and Sarkani, S. (2014) The effect of the number of seed variables on the performance of Cooke's classical model. Reliability Engineering & System Safety 121, 72–82.

Ellsberg, D. (1961) Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics* 75, 643–669.

Evans, D.J., and Sezer, H., (2005), Social discount rates for member countries of the European Union. *Journal of Economic Studies* 32 (1), 47–59.

Feigenbaum, E.A. (1977) The art of artificial intelligence: I. Themes and case studies of knowledge engineering. In *Proceedings of the International Joint Conference on Artificial Intelligence and the National Computer Conference*. http://ijcai.org/Past%20Proceedings/IJCAI-77-VOL2/PDF/092.pdf (accessed April 16, 2014). Also published as Stanford Heuristic Programming Project Memo (112) HPP-77-25, and Computer Science Department Memo STAN-CS-77-621.

Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W., and Ginzburg, L. (2007) *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. SAND2007-0939. Albuquerque, NM: Sandia National Laboratories.

Fredrick, S., Loewenstein, G., and O'Donoghue, T. (2002) Time discounting and time preference: A critical review. *Journal of Economic Literature* 40 (2), 351–401.

Fu, G., Hall, J., and Lawry, J. (2005) Beyond probability: New methods for representing uncertainty in projections of future climate. Working paper 75. Norwich, UK: Tyndall Centre for Climate Change Research.

Ghosh, S., and Mujumdar, P.P. (2009) Climate change impact assessment: Uncertainty modeling with imprecise probability. *Journal of Geophysical Research* 114, D18113, doi:10.1029/2008JD011648.

Hall, J.W., Fu, G., and Lawry, J. (2007) Imprecise probabilities of climate change: Aggregation of fuzzy scenarios and model uncertainties. *Climatic Change* 81 (3–4), 265–281.

Henry, C. 2006. *Decision-Making Under Scientific, Political and Economic Uncertainty*. Chaire Développement Durable, Cahier DDX–06–12. Paris, France: Laboratoire d'Econométrie de l'École Polytechnique.

IPCC (2012) *Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties*. IPCC Cross-Working Group Meeting on Consistent Treatment of Uncertainties, Jasper Ridge, CA, 6–7 July 2010. Geneva, Switzerland: Intergovernmental Panel on Climate Change.

Kahneman D., and Tversky, A. (1979) "Prospect theory: an analysis of decision under risk", *Econometrica*, 47, 263-291.

Kandlikar, M., Risbey, J., and Dessai, S. (2005) Representing and communicating deep uncertainty in climate-change assessments. *C.R. Geoscience* 337, 443–455.

Keynes, John Maynard (1921), Treatise on Probability, London: Macmillan & Co

Knight, F.H. (1921). *Risk, Uncertainty, and Profit*. New York, NY: Harper & Row. http://www.econlib.org/library/Knight/knRUPCover.html (accessed April 16, 2014).

Kriegler, E., Hall, J.W., Held, H., Dawson, R., and Schellnhuber, H.J. (2009) Imprecise probability assessment of tipping points in the climate system. *PNAS* 106 (13), 5041–5046.

Laplace, P.S. (1814; English edition 1951) *A Philosophical Essay on Probabilities*. New York, NY: Dover Publications.

Lin, S.-W., and Bier, V.M. (2008) A study of expert overconfidence. *Reliability Engineering & System Safety* 93 (5), 775–777.

Lin, S.-W., and Cheng, C.-H. (2008) Can Cooke's model sift out better experts and produce well-calibrated aggregated probabilities? In *Proceedings of the 2008 IEEE IEEM*. Department of Business Administration, Yuan Ze University, Chung-Li, Taiwan.

Lin, S.-W., Cheng, C.-H. (2009) The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management* 4 (2), 149–161.

Lin, Shi-Woei, Huang, Ssu-Wei (2012) "Effects of Overconfidence and Dependence on Aggregated Probability Judgments" *Journal of Modelling in Management* 7(1).

Mach, K.J., Matschoss, P.R., Plattner, G. K., Yohe, G.W., and Zwiers, F.W. (2010) *Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties*. Intergovernmental Panel on Climate Change

Mastrandrea, M., Mach, K., Plattner, G.-K., Edenhofer, O., Stocker, T., Field, C., Ebi, K., Matschoss, P. (2011) The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups. *Climatic Change* 108, 675–691. http://www.climate.unibe.ch/~stocker/papers/mastrandrea11cc.pdf (accessed April 16, 2014).

McCrimmon, K.R. (1968). Descriptive and normative implications of the decision-theory postulates. In *Risk and Uncertainty*. London, UK: MacMillan, 3–24.

Morgan, G.M., Dowlatabadi, H., Henrion, M., Keith, D., Lempert, R., McBride, S., Small, M., and Wilbanks, T. (2009) *Best Practice Approaches for Characterizing, Communicating, and Incorporating Scientific Uncertainty in Climate Decisions*. Synthesis and Assessment Product 5.2. Washington, DC: US Climate Change Science Program.

Morgan, M., and Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. New York, NY: Cambridge University Press.

Morris, P. (1977) Combining expert judgments: A Bayesian approach. *Management Science* 23 (7), 679–693.

Moss, R.H., and Schneider, S.H. (2000) Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting. In *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*. Pachauri, R., Taniguchi T., and Tanaka, K. (eds.). Geneva, Switzerland: World Meteorological Organization, 33–51. http://www.ipcc.ch/pdf/supporting-material/guidance-papers-3rd-assessment.pdf (accessed April 16, 2014).

Newell, R.G., and Pizer, W.A. (2003) Discounting the distant future: How much do uncertain rates increase valuations? *Journal of Environmental Economics and Management* 46, 53–71.

Nordhaus, W. D. (2008). *A Question of Balance: Weighing the Options on Global Warming Policies*. New Haven, CT: Yale University Press.

Nordhaus, W., and Popp, D. (1996) What is the value of scientific knowledge: An application to global warming using the PRICE model. Discussion paper no.1117. New Haven, CT: Cowles Foundation for Research in Economics, Yale University.

NRC (2010) *Advancing the Science of Climate Change*. Washington, DC: National Research Council. http://www.nap.edu/catalog.php?record_id=12782 (accessed April 15, 2014).

Pizer, W.A. (1999) The optimal choice of climate change policy in the presence of uncertainty. *Resource and Energy Economics* 21, 255–287.

Quiggin,J. (1993) Generalized Expected Utility Theory – the Rank-Dependent Model Kluwer, Dordrecht.

Ramsey, F.P. (1931; originally published 1926) Truth and probability. In *The Foundations of Mathematics and Other Logical Essays.* Braithwaite, R.B. (ed.). London, UK: Kegan, Paul, Trench, Trubner & Co.; New York, NY: Harcourt, Brace and Company, 156–198.

Ryan, J.C.H. Mazzuchi,T.A. Ryan,D.J., Lopez de la Cruz, J. and Cooke, R. M. (2012) "Quantifying information security risks using expert judgment elicitation", Computers & Operations Research Journal 39, 774–784. , doi:10.1016/j.cor.2010.11.013

Savage, L. J., (1972) The Foundations of Statistics, Dover, New York, first published in 1954, John Wiley and Sons.

Schmeidler, D. (1989) "Subjective probability and expected utility without additivity" Econometrica 57, 571-587.

Stern, N. (2008). The economics of climate change. *American Economic Review* 98 (2), 1–37.

Tversky, A., and Kahneman, D. (1982) Judgments of and by representativeness. In *Judgment under Uncertainty: Heuristics and Biases*. Kahneman, D., Slovic, P., and Tversky, A. (eds.). Cambridge, UK: Cambridge University Press, 84–98.

US Nuclear Regulatory Commission (1983). *PRA Procedures Guide. A Guide to the Performance of Probabilistic Risk Assessments for Nuclear Power Plants. Final Report*. NUREG/CR-2300. Washington, DC: US Nuclear Regulatory Commission. http://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr2300/vol2/cr2300v2-a.pdf (accessed April 16, 2014).

von Mises, R. (1928) *Wahrscheinlichkeit, Statistik und Wahrheit*. 2nd edition, 1936. Vienna, Austria: Springer-Verlag.

Wakker, P.P. 2010 "Prospect theory for risk and ambiguity" Cambridge U. Press, Cambridge, 2010.)

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London, UK: Chapman and Hall.

Weitzman, M.L. (2001) Gamma discounting. *American Economic Review* 91 (1), 260–271.