Making Confident Decisions with Model Ensembles

Joe Roussos, Richard Bradley, and Roman Frigg*†

Many policy decisions take input from collections of scientific models. Such decisions face significant and often poorly understood uncertainty. We rework the so-called confidence approach to tackle decision-making under severe uncertainty with multiple models, and we illustrate the approach with a case study: insurance pricing using hurricane models. The confidence approach has important consequences for this case and offers a powerful framework for a wide class of problems. We end by discussing different ways in which model ensembles can feed information into the approach, appropriate to different collections of models.

1. Introduction. In sciences dealing with complex systems, it is common to encounter a range of different models representing the same system. Such models might disagree deeply over the structural relations in the system or in shallower ways over the values of parameters or initial conditions. Since it is often impossible to decide between these models using available evidence, scientists work with whole collections—or "ensembles"—of models. Prominent examples are the CMIP5 ensemble of global climate models and ensembles of hurricane models for the North Atlantic. In some cases, model ensembles indicate disagreements among scientists; in other cases, they reflect agreed latitude in model construction. In either case the ensemble represents (at least partially) scientific uncertainty about the target system.

How should policy makers use model ensembles in making decisions, and how should these decisions reflect the scientific uncertainty associated with

Received March 2019; revised March 2020.

*To contact the authors, please write to: Joe Roussos, Institute for Futures Studies, Holländargatan 13, 101 31 Stockholm, Sweden; e-mail: joe.roussos@iffs.se.

†We thank Tom Philp for numerous discussions about hurricane modeling and for his helpful advice on navigating the hurricane science literature. Thanks also to Jan-Willem Romeijn, Sean Gryb, Simon Dietz, and Jonathan Livengood for their comments on earlier drafts.

Philosophy of Science, 88 (July 2021) pp. 439–460. 0031-8248/2021/8803-0004$10.00

them? Mainstream decision methods such as expected utility theory assume that all relevant uncertainty is captured by a single probability measure and so do not (without supplementation at least) provide an adequate answer to this question. Recent decades have seen the development of numerous decision rules for situations in which decision makers face what is known as "ambiguity," when precise probabilistic estimates of all decision-relevant quantities are unavailable (for a survey, see Kriegler 2007; Gilboa and Marinacci 2013; Heal and Milner 2014). There is also a nascent decision-theoretic literature on model uncertainty (see esp. Marinacci 2015).

This article aims to advance this discussion by reworking a recently developed decision theory called the *confidence approach* (Hill 2013; Bradley 2017) to tackle inputs from model ensembles. The aim is to demonstrate its fruitfulness; in particular, the benefits of its structured approach to managing ensemble uncertainty. We illustrate the approach by applying it to a scenario involving hurricane models used for insurance pricing. The lessons from this case are applicable to many policy-making scenarios. The framing of our article is crucial: we work from the decision maker's perspective. They are typically *non-experts*. The challenge is to design a normatively appealing decision procedure for their use, which is nevertheless sensitive to the state of the relevant scientific knowledge.

In section 2 we introduce hurricane modeling and our insurance case study. Section 3 argues that current practice—using a weighted average of the hurricane models' outputs—is problematic, and it would be desirable to have an alternative approach. We introduce such an approach in section 4 and apply it in a simple form in section 5. Section 6 considers various ways of constructing the main ingredient in our approach: a nested family of sets of probabilities. Section 7 concludes with a program for developing the approach.

2. Hurricane Insurance Decision-Making. Our case study is drawn from a research collaboration with scientists working for a large insurance company. The insurer is based (and regulated) in the United Kingdom but operates worldwide. As part of its US property insurance, the company offers cover for damage resulting from hurricanes. This practice relies on estimating the probability of the insured-against events (i.e., destructive hurricanes) and the damage they cause. It is often not economically efficient for insurers to invest in the expertise and capabilities required to do this, and so they buy predictive models from commercial modeling companies. These companies employ teams of environmental scientists, statisticians, and programmers to construct simulation models to determine the probability of hurricane "landfalls" along the US Atlantic coast.[1]

1. In 2015, the Florida Commission on Hurricane Loss Projection Methodology (FCHLPM) received submissions for approval to sell models to insurers from four private firms: AIR,

The modeling firms face a problem: there is significant uncertainty in hurricane modeling, derived in part from disagreements about the underlying science. The result is that there are multiple models representing the same system. The FCHLPM (2007) carried out an assessment of the modeling industry using an ensemble of 972 models (Guin 2010). Risk Management Solutions (RMS), a leading modeling firm, uses an ensemble of 13 models to generate the medium-term rate (MTR), their preferred prediction of hurricane landfall frequency (Sabbatelli and Waters 2015).

Any company selling models to insurers must decide how to navigate this landscape. Which model(s) should they build as part of their offering? Offering more than one model better represents the landscape, but presenting insurers with a collection of models creates a further problem for them: How does one decide when faced with not one model probability but 13 or 972? The most common solution when working with ensembles is to average the outputs from each model, and one task of this article is to lay bare the limitations of this process. To add specificity to the problem, and show how it arises in an important real-world application, we now give a brief overview of the RMS model ensemble. We chose RMS because they are a leading hurricane modeling firm and because they are open about their use of an ensemble of models.

A catastrophe model for insurance works in four stages, covering (1) the hazard, in this case a hurricane; (2) the physical damage it creates, which requires modeling the vulnerability of buildings and infrastructure to wind, water, and so on; (3) calculating insurer exposure, by looking at insurance policy terms; and (4) financial modeling of the insured losses that result. We will consider only the first component. Shome et al. (2018, 37) provide a classification of RMS hurricane models for the North Atlantic, with model names reflecting the sometimes competing choices made in the modeling process. Briefly reviewing these types effectively conveys the diversity of the models in this ensemble.[2]

- "Direct" models use historic hurricane landfall counts as input and make a landfall prediction.
- "Indirect" models use storm formation data from the Atlantic basin to make a prediction of hurricane activity in the basin and then convert that

Applied Research Associates, CoreLogic, and Risk Management Solutions (FCHLPM 2015). Insurers are really interested in losses from destructive hurricanes, and so the mathematical object of interest is (a function of) the probability of losses above a certain value. We focus on probabilities of the underlying events for simplicity.

2. As the RMS ensemble is proprietary, some detective work is required here. We compared Jewson et al. (2009), Sabbatelli and Waters (2015), Sabbatelli (2017), and Shome et al. (2018).

prediction into a landfall prediction using the estimated proportion of basin storms that finally make landfall along the US coastline (Jewson et al. 2009, 12).

- "Indo-Pacific" models include the impact of sea surface temperatures (SSTs) in the Indian and Pacific oceans on hurricane formation through their effect on wind shear in the Atlantic basin.
- "Shift" models identify periods of higher or lower than average hurricane activity or SSTs in the historic data. This is due to the Atlantic multidecadal oscillation (AMO), and probabilities of transitions from high- to low-activity periods are estimated using historic data on tree ring sizes, a method due to Enfield and Cid-Serrano (2006) (Jewson et al. 2009, 14).
- "Active baseline" models, a category mutually exclusive with "shift," reflect an alternate hypothesis on the AMO: the low-activity period in the 1970s and 1980s was due to SST cooling induced by high atmospheric aerosol content, primarily volcanic aerosol (Booth et al. 2012). If correct, SSTs will not revert to a cool phase in the future, and one should not apply a probability of shifting back to a low-activity hurricane generation phase. These "active baseline" models therefore do not include the Enfield and Cid-Serrano probabilities in their forecasts and subsequently forecast higher landfall rates than the shift models (Sabbatelli 2017).

The ensemble is built up by taking combinations of the above methods. It starts with two models: direct and indirect. By adding models with Indo-Pacific SSTs, we get to four. We then add shift and active baseline variants of all four—leading to 12 models. The thirteenth is a long-term rate (LTR) model, included for comparison. RMS's LTR is a statistical model based on historical landfall and basin storm data, and it models hurricane frequency as constant in time (Shome et al. 2018, 33). RMS's certification as a modeler for the American market (by the FCHLPM) is granted on the basis of their LTR model, and so, although RMS advertises the MTR ensemble average as providing their state-of-the-art view of hurricane risk, the LTR is often used as a reference view.

This list shows that the models included in the ensemble are not merely variants of the same model (obtained, possibly, by varying parameter values). The models fall into groups that are genuinely different and in some cases mutually incompatible. How should an insurance company use this ensemble to determine the price of its policies?

3. Averaging and Its Limitations. The problem that scientists face is how to extract the information contained in the ensemble and make it available to users. This is a thorny issue because it is far from clear how to interpret

the (often conflicting) outputs from different models. A popular method is to calculate a weighted average of all model outputs and use this average for decision-making (Clemen [1989] and Armstrong [2001] provide reviews from economics and management).[3] This has the advantage of allowing standard decision methods, which require a single probability as an input, to be applied.

Averaging works in some cases, and where it works one should use it. However, it does not work in all cases, and hurricane insurance is such a case (Philp et al. 2019). The most significant problem from our perspective is that such decision methods discard useful information about the state of scientific uncertainty. As Morgan (2014) points out, averaging focuses attention on the mean projection only. However, the spread of results is itself important information. First, when we are interested in extreme events like major hurricanes, then we are explicitly concerned with the shape of the distribution and not just the mean. Second, the spread tells us something about the state of our knowledge about a question. To the degree that there is spread, it reflects scientific uncertainty about the system and our lack of precision in modeling its relevant features. This by itself is valuable information that the decision maker might want to use.

Although averaging does not preclude communication of the spread, in the expected utility paradigm it is unclear how this information is to be used, since the expected utility of an action or policy depends only on the decision maker's probability for relevant contingencies (in addition of course to the utility of outcomes). Evidently the same probability can be obtained by averaging over very different sets of candidate probability functions. Consequently, expected utility maximization precludes making decisions in a way that is sensitive to the state of scientific understanding as expressed by the spread in the ensemble projections. But in decisions with high stakes it is reasonable to seek to calibrate one's choices to the level of uncertainty contained in the scientific projections that one is drawing on (see Bradley 2017; Hill 2019). The procedure we develop in sections 4 and 5 overcomes this limitation by making structured use of ensemble spread.

Additional problems undermine decision-maker confidence in averaging procedures. Weights for averaging are typically constructed by scoring models on skill. This often involves hindcasting: reproducing a piece of the historical record (which the model has not "seen" before). This method faces the problem that the historical data set used to score these models is small,

3. RMS aggregates the outputs from the 13 models in its ensemble; Sabbatelli (2017) confirms this is a weighted average. In a recent interview RMS stated that the predictive test used for scoring involves predicting hurricane activity in every sequential 5-year period over the past 50 years (Insurance ERM 2018). The scoring rule (SR) used is not discussed, however, and this information does not appear to be in the public domain.

as large hurricanes are infrequent. HURDAT2, the standard database for hurricanes hitting the Atlantic coast of the United States, is moderate in size, with ~300 storms to date and only one-third of those counting as "major hurricanes." If we split the data set by region the numbers drop precipitously.[4] As Shome et al. (2018) point out, actuaries judge that there are insufficient data to form a reliable statistical model to predict future events. The data are also insufficient to meet regulatory requirements. The national regulator in the United Kingdom requires insurance companies to design their portfolio in a way that they go bust at most once in 200 years. Even on a generous reading there are at most 120 years of usable hurricane frequency data. But 120 years of data do not provide a reliable understanding of the tail events and the shape of the distribution at those longer return periods. Shome et al. (2018) cite this paucity of data as a reason for using quasi-physical simulation models, whereby modelers create "statistical storms" to expand and "fill in" the data set. This process, however, relies on the (scant) historical evidence, and so it cannot remove the problem of restricted evidence.

A further problem concerns the choice of a scoring rule (SR).[5] The problem is that the weights are set by the rule, and so the average value is sensitive to this choice. But the range of SRs on offer is so diverse that almost any reasonable answer could be selected by one of them (Stainforth, Allen, et al. 2007, 2155). The Australian Bureau of Meteorology maintains a website on probabilistic forecast verification techniques, which covers more than 50 SRs, visualization techniques, and analytical approaches to measure the success of probabilistic forecasts. The categories are nonexclusive, and for a given problem there may be multiple appropriate rules with different features (Australian Bureau of Meteorology 2017).[6] So, experts often disagree over which rules to use when, and individual experts may endorse more than one rule as appropriate for a given situation. The debate over these rules covers both technical matters (e.g., which is better suited to rare event predictions) and values (about what counts as a "good" prediction). Decision makers ought ideally to select a rule embodying their values (i.e., that corresponds to what they regard as important), and yet the technical complexity of the subject means that decision

4. The full database is at http://www.aoml.noaa.gov/hrd/hurdat/All_U.S._Hurricanes.html.

5. There is also a debate over the suitability of linear averaging as opposed to, e.g., geometric averaging (Dietrich and List 2016). As we advocate for a different method entirely, we do not discuss this debate.

6. The problem equally arises in the context of hurricane modeling. Skill scores are among the fiercely protected trade secrets of modeling companies and insurers, and they are therefore not in the public domain. However, we know in fact that different actors in the market use different skill scores and that these can support different results (nondisclosure agreements prohibit us from saying more about this).

makers are often not in a position to participate meaningfully in a choice of rule. Crucially, the problem of choosing a SR is very similar to the original problem of choosing/formulating an answer from the range present in the model ensemble. Any method of choosing a SR requires deciding between disagreeing experts, and so we may well ask why we do not simply apply these same considerations to the "first-order" problem of model disagreement.[7]

We conclude that averages are a reliable guide to action only when uncertainty is small (and known to be so), enough data are available for meaningful scoring, and different SRs produce similar results. There may be situations that satisfy these requirements, but hurricane modeling is not one of them. Averaging is therefore not an optimal procedure to make decisions on hurricane insurance. Practitioners feel similarly: insurers have expressed some of these concerns to us, and in practice they will "factor in" their dissatisfaction by, for example, multiplying the average event probabilities by some $\alpha > 1$. They have, however, no principled way of determining the value of α, which is typically set in an ad hoc manner by managers removed from the detail of the modeling. The nature of these problems is such that they are unlikely to be resolved by tweaks to the aggregation methodology; a completely new approach is needed.

4. The Confidence Approach. In this section we outline the theoretical basis for our alternate approach to using results from the model ensemble, and in the next we apply this approach to a simple insurance problem. This work applies a relatively new theory of decision-making under ambiguity, which we call the *confidence approach*, developed by Brian Hill (2013, 2016).

The confidence approach makes use of "imprecise probabilities," sets of probability functions that generate sets of values (typically, intervals) for each event, as a way of capturing features of agents' uncertainty—specifically the empirical ambiguity (Bradley [2019] presents an overview). Imprecise probabilities support decision rules that take such sets as inputs, as opposed to the single probability function required by the classical decision theory of Savage. There are many such rules on offer, but they all face similar challenges.

First, how does one determine the relevant set of probability functions or, equivalently, the extent of the uncertainty the decision maker faces? The question has significant implications for applications of these decision rules. When the set is large, many of them lead to levels of caution that could be regarded as excessive (e.g., such that much of today's hurricane insurance

7. Some might protest: there is a best SR—the Brier score—and all should use it (e.g., Leitgeb and Pettigrew 2010). But the Brier score's purpose is to select the best prediction; it is nearly useless for relative comparisons of low-probability predictions. Brier compares predictions to the "truth," e.g., 1 if it occurs. The differences between predictions of very improbable events will be lost when they are subtracted from 1.

business would not be written). But when it is small, it may not capture all uncertainty relevant to the decision maker. Despite this, imprecise decision theories often fail to separate the uncertainty agents face from their attitude to it, for instance, by taking the set of probabilities representing the agent's uncertainty to be the one with respect to which they maximize the quantity that the decision theory in question takes to be significant, for example, minimum expected utility (Gilboa and Marinacci 2013; Hill 2019).

Second, these theories, like expected utility theory, do not allow for the differing importance of various decisions to affect how a decision maker responds to uncertainty (Hill 2013), in part because of the aforementioned failure to separate the uncertainty from the response to it. But intuitively, how much it matters that we lack scientific certainty on some question, and how cautious we want to be in making choices as a result, should depend on how much is at stake for us in these choices. The confidence approach mitigates against these concerns, while drawing on the benefits of using imprecise probabilities.

We start with a high-level description of the approach, using a trivial but intuitive example; along the way, we note how policy decisions will differ. Here is the trivial case: suppose you are deciding whether to place a bet on your favorite contestant, Baga Chipz, winning a drag competition. To bet you pay £50 up front. If she wins you are paid back your £50 and receive another £50; if she does not win you lose your £50. So, this bet has a positive expected monetary return whenever the probability of her winning is strictly greater than 0.5. We will show how the confidence approach determines whether this is a fair or advantageous bet.

First, we represent your beliefs with a family of nested sets of probabilities. Each set represents a claim that you accept about the relevant probability, while the nesting captures the logical relationship between these claims. In our example, these claims could range from the very imprecise (indeed trivial) claim that "the probability of Baga Chipz winning is between 0 and 1" to the very precise "the probability of Baga Chipz winning is 0.42." Figure 1*b* shows such a nested family schematically.

Next, we consider your confidence in each of these claims. Confidence should reflect the "weight of evidence" supporting a claim—a term, coined by Peirce (1878) and popularized by Keynes (1921, 78), describing the property of evidence that makes us more sure of our probabilistic judgment, even when the judgment itself may remain constant. "Confidence" is thus a (second-order) attitude toward a (first-order) claim, reflecting an evaluation of the state of knowledge underpinning it. It has the following logical structure: one cannot be more confident in more precise claims. So, you cannot be more confident that the probability of Baga's win is 0.42 than you are that it is in the interval [0.3, 0.5]. This is a simple consequence of the former's inclusion in the latter.

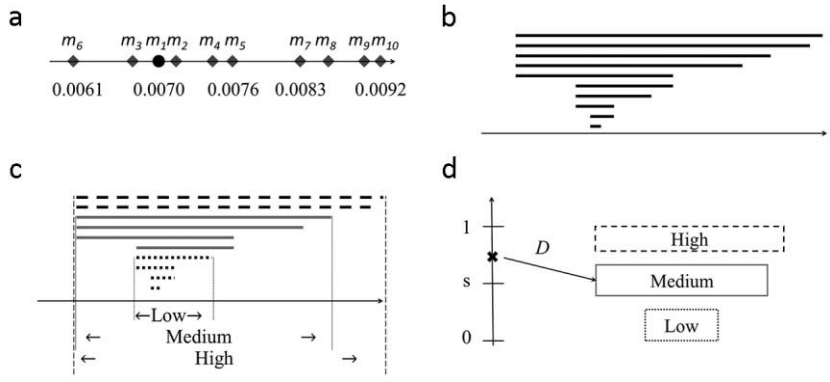

Figure 1. Confidence approach: *a*, set of point estimates of probability; *b*, nested family of sets; *c*, confidence levels; *d*, stakes and cautiousness select a level. Color version available as an online enhancement.

In principle we could have any number of sets in the nested family. For practical purposes, it is useful to coarse grain down to "levels of confidence" (e.g., low, medium, high). This involves making a judgment about which members of the family of nested sets are equivalent in terms of coarse-grained confidence and grouping these together as in figure 1*c*. The idea is that there is no decision-relevant difference between the grouped sets. As narrower sets exclude more possibilities, we therefore have a pragmatic motivation to work with only the most precise interval in each level—in figure 1*c* this corresponds to the bottom interval in each level. In our example, this means restricting your consideration to three probability intervals, say 0.42, [0.3, 0.5], and [0.2, 0.6] reflecting claims you endorse with low, medium, and high confidence.[8] Every claim wider than 0.42 but narrower than [0.3, 0.5] is considered confidence equivalent to 0.42 (i.e., low confidence), and so forth. Put another way: if you saw decision-relevant differences between the intermediate intervals, you would further fine grain.

How we coarse grain will partly be a matter of convention, but it is motivated by an important consideration: connecting the relative ranking of a particular decision's family of sets to a background standard of confidence. An ordinal ranking cannot say anything about how much confidence we have in any claim; it can only tell us how that claim is related to other claims. If the outcome of my bet is that I will be shot if I lose, I want to be very confident about my probability estimate; "very" reflecting the absolute importance of the decision and not just that I want more confidence in it than other estimates of drag queen odds. A decision maker can do this by developing a sense of

8. As a matter of logic, the full [0, 1] interval is always in the highest confidence level. This could be the high level above or an implicit "highest" level.

what counts as "enough evidence to warrant high confidence" and applying that standard across decision problems through the labels applied in this coarse-graining step. If there is poor evidence supporting all claims in the family, perhaps the top coarse-grained level should only be considered "medium confidence."

Coarse graining to levels pegged to such a background standard of evidence allows our notion of "confidence" to decouple from the situation-specific information in front of the agent. This allows us to make decisions in a way that reflects their importance, relative to other decisions we make. To accomplish this systematically, the confidence approach adds two more features to standard decision theory.

The first is the "stakes" of the decision: the agent's assessment of how important it is. The key feature of stakes is that it partially orders decisions (i.e., that we can compare two decisions and rank them in terms of importance). How this is done can vary, but for formal simplicity, we think of stakes as a number on a 0–1 scale.[9] The term "stakes" is chosen to imply that it should be a feature of the potential outcomes, as in our betting example where you stand to lose (or win) £50. There is a wide range of potential functions of these outcomes that could measure your stakes; Hill (2016) discusses their differences. For simplicity let us take the stakes to be a function purely of the worst possible outcome—losing £50. Assessing the relative importance of a decision in which you stand to lose £50 involves reflecting on other decisions you make, the value of £50 to you, and so forth. For the moment let us simply assume you regard this as a moderately important decision and assign $s = 0.5$.

Second, agents have as a feature of their psychology a function called "cautiousness," which determines how much confidence they require to decide, given the stakes. Cautiousness thus takes the stakes as input and outputs the coarse-grained level of confidence required to make the decision; figure 1*d* illustrates, with cautiousness denoted *D*. A different degree of caution, colloquially speaking, is represented by a different cautiousness function. "More cautiousness" means that more of the 0–1 stakes range is mapped to a high confidence level. In our simple example, the question to ask is, "How much confidence do you need in order to make moderately important decisions, with stakes around 0.5?"

Cautiousness represents an attitude to ambiguity on the part of the decision maker. It is therefore subjective and will need to be elicited.[10] Let us suppose

9. If one is happy to think that agents have outcomes they consider least (most) important, then 0 (1) represents this. If one is concerned that stakes should be unbounded, then suppose that this infinite scale has been transformed to 0–1, with (0) 1 representing (negative) positive infinity.

10. Theorem 2 in Hill (2013) proves that the cautiousness function is equivalent to measures of ambiguity aversion in decision theories that strictly separate beliefs and desires,

that after such an elicitation we determine that you require medium confidence for decisions of moderate importance—so that, for you, $D(0.5)$ = medium. You can now select a probability interval: the narrowest interval in the level of confidence that your cautiousness demands, given the stakes of the decision. In the case of your drag queen bet, that interval is [0.3, 0.5].

In a policy-making context, we may wish to do more than elicit the attitude of the individual decision maker and look to public and political debate to settle what level is appropriate. In our insurance setting, it will be natural for this to represent the firm's ambiguity attitude, a policy on the conditions under which it is allowable to sell insurance.

We now reconnect with imprecise decision theory. The probability interval selected by the confidence procedure can be used with any imprecise decision rule. For the sake of specificity, we will illustrate with one popular rule: maxmin expected utility (MMEU). This rule says that act f is preferred to act g if and only if the minimum expected utility of f, with respect to the set of probability functions, is greater than the minimum expected utility of g with respect to the same set (Gilboa and Schmeidler 1989). That is, it recommends choosing cautiously, by acting to guarantee the best outcome if things turn out to be the worst way they could be, from your perspective.

One benefit of the confidence approach is that we have two "levers" of ambiguity attitude: the cautiousness function and the decision rule. Although MMEU is highly ambiguity averse, this aversion can be attenuated by the choice of cautiousness function—specifically, by choosing a function that recommends moderate levels of confidence for a wide range of stakes. (The opposite choice could boost MMEU's ambiguity aversion.) Decision makers who are not completely ambiguity averse can thus still use MMEU, for instance, because it is a very simple rule to implement. We adopt it for precisely this reason.

Applying MMEU leads to bad news for your bet on Baga Chipz. Things turn out worst if the probability is at the low end of your medium-confidence interval, where you expect to lose money on this bet because the probability is below the priced value of 0.5. You therefore should not place the bet. In a single simple decision like this, the confidence approach may seem at once complex and permissive. In the next section, this heavy machinery will show its value.

5. Confidence, Models, and Insurance. We now apply the confidence approach to our case: making an insurance pricing decision, using input from an ensemble of scientific models. Our example simplifies some details of actual insurance pricing by considering a very simple portfolio with only one

such as the "smooth ambiguity" model of Klibanoff, Marinacci, and Mukerji (2005). Cautiousness can therefore be elicited using methods apt to such theories.

contract. This does not influence the philosophical points we wish to make about the treatment of outputs from a model ensemble, and the approach can be applied to more complex portfolios.

5.1. A Simplified Insurance Problem. Assume that an insurer wants to sell a single insurance contract on house damage due to hurricanes. It has no current contracts and plans to sell just this one, which insures against event E: "a hurricane strikes Fort Lauderdale in 2021." The contract is for a total value $v = 100{,}000$, and to simplify we will assume it is a simple binary contract, paying out either \$0 if the event does not occur or \$100,000 if it does.

The insurer plans to price this contract in the tradition of Stone's (1973) constraint equation: $\pi > yH - d$. The equation determines the minimum premium (annual price) π required to make a profit. Premium π must be larger than the difference between the annual cost of held capital, yH, and the (negative) expected damages, $\langle d \rangle$, where "damages" refers to the amount the insurer pays to its customers, which is represented by a damage function d. The insurer's probability for E is $p(E)$, which determines $\langle d \rangle$ (and, in a complicated process we will not discuss here, H). We will assume that the insurer's capital holdings must be equal to the contract value ($H = v$). If it then turns out that, say, $p(E) = 0.01$ and the cost of capital is 5% ($y = 0.05$), Stone's equation says that the minimum premium is \$6,000. The value of y is dictated by capital markets; the insurer's problem is to determine $p(E)$, given the results of the model ensemble.

5.2. The Scientific Input and Aggregation. Consider a simplified scenario that captures the salient features of the RMS ensemble from section 3. Our scientific modelers construct 10 models, $m_1, \ldots, m_{10}$, which encode different views about, for example, how hurricanes move across the Atlantic and how the factors influencing their generation will turn out in 2021. As the details will not matter here, we will not describe how these models work except that they generate $p(E)$ and that one of them—m_8—was built for a different region but works for Florida. Table 1 shows 10 numbers that we will use as our model outputs. The "standard" approach would be to score these models on their predictive skill, as described in section 3. Suppose that we have done this, using a popular SR R. The normalized scores and outputs for $p(E)$ are shown in table 1. Using these normalized scores, we can calculate the ensemble average: $p^A(E) = 0.0072$. This is what is standardly used for

TABLE 1. Model Outputs for Toy Example

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
$p(E)$	.0070	.0071	.0068	.0074	.0076	.0061	.0083	.0086	.0091	.0092
Weight (%)	23.7	20.7	15.8	11.6	11.5	7.3	3.2	3.0	1.7	1.6

decision-making, especially in cases in which it has higher "skill," according to rule R, than even the best model.

Let us consider how the standard approach would price the contract. Using $p^A(E)$, the expected damage is $\langle d \rangle^A = p^A(E)d(E) = 0.0072 \times -100{,}000 = -720$. The required holdings are $H = 100{,}000$. If we take the cost of capital to be $y = 0.05$, then we have minimum price $\pi^A > yH - \langle d \rangle^A = 5{,}720$.

As we noted above, skeptical underwriters often introduce an inflationary factor for safety. This may take the form of crudely doubling the aggregate probability $p^A(E)$. Going through the calculations with that probability, we get the "safety" price $\pi^S > 6{,}440$, which is much higher than the "technical" price π^A. This is clearly ad hoc, but decision makers reported making such moves to us, in the absence of a structured method for dealing with ambiguity.

5.3. Pricing with Confidence. To apply the confidence approach we must formulate claims about $p(E)$. Here is a simple method of doing so; we discuss more elaborate alternatives in the next section. We will assume that SR R has reliably identified the best model, m_1, and build intervals around it. Our "lowest," most specific claim is that $p(E) = 0.007$. We form wider intervals by including predictions in order: $I_1 = 0.007$, $I_2 = [0.007, 0.0071]$, ... , $I_{10} = [0.0061, 0.0092]$.

We now coarse grain and form levels. As our insurer has made no decisions of this sort before, it does not have an ongoing assessment of evidential quality and so constructs the confidence levels using only the model outputs. For simplicity we can suppose that these 10 models represent all the major relevant scientific views. As a first pass, the insurer could decide to coarse grain by model support: using up to 4 models will yield low confidence, 5–7 medium, and 8–10 high. However, in consultation with the modelers the insurer notes that this would result in the narrowest interval in the high level being I_8. But the modelers doubt m_8, as it was built for a different region. This is a scientific reason to avoid having a decision depend on it, and so the insurer revises the coarse graining so that low involves models 1–4; medium, 5–8; and high, 9 and 10:[11]

$$\mathcal{L} = \{0.007, [0.007, 0.0071], [0.0068, 0.0071], [0.0068, 0.0074]\}$$

$$\mathcal{M} = \{[0.0068, 0.0076], [0.0061, 0.0076], [0.0061, 0.0083], [0.0061, 0.0086]\}$$

$$\mathcal{H} = \{[0.0061, 0.0091], [0.0061, 0.0092]\}.$$

11. This is a simplified example of how scientific facts about the models can inform confidence level formation. A further move of this sort might be to look at the evidence used in the construction of each model: different evidential bases (perhaps because of the scientific disagreements generating the ensemble) can generate different incremental gains in confidence. For the moment we will neglect such considerations.

The agent now regards the intervals within each level as being of equivalent confidence and so will make decisions using the narrowest interval available at a level. Figures 1*a*–1*c* illustrate the process just described.

How can we characterize the stakes facing this insurer? This contract will constitute its whole business, and so the risk of ruin is high. Still, no one's life is at stake, and there is no impact on anything outside of the realm of this decision (no other business that might be taken down). So, the insurer concludes that its stakes are moderately high, $s = 0.75$. Next, we describe the insurer's ambiguity attitude (which corresponds to cautiousness). As insurance of natural catastrophes involves significant ambiguity, it seems reasonable to assume that this insurer is not overly ambiguity averse. Here is one cautiousness function that exhibits only moderate ambiguity aversion: stakes below 0.6 require low confidence, 0.6–0.9 require medium, and above 0.9, high. Applying this to the example outlined above, we see that the decision maker resolves to use level $\mathcal{M}$ and its narrowest interval $I_5 = [0.0068, 0.0076]$.

We can now apply our chosen decision rule. In insurance, higher probabilities represent worse payoffs for the insurer, and so MMEU selects the highest probability in the range: $p^C(E) = 0.0076$. We therefore have the following expected damages $\langle d \rangle^C = 0.0076 \times -100{,}000 = -760$. The holdings are exactly as before. We therefore get $\pi^C > yH - \langle d \rangle^C = 5{,}760$.

So, the confidence model recommends pricing the same contract (at least) 0.7% higher than the averaging approach. Crucially, π^C is 10.5% below the "safety" price π^S. This is a very large difference when pricing large insurance contracts—representing money lost to undue, and unjustified, caution. The difference between π^C and π^A, although smaller, is still significant, but note also that its size is an artifact of our toy example: selling only one contract imposes exceedingly high capital costs. If we sold 20, and spread the capital cost evenly among them so that $h = H/20$, the prices would be

$$\pi^A_{20} > yh - \langle d \rangle^A = \$970, \quad \pi^S_{20} > \$1{,}690, \quad \pi^C_{20} > \$1{,}010,$$

in which case the confidence price is 4.1% higher than the aggregate price and 40% below the "safety" price. Our structured approach to uncertainty classifies a number of sales as imprudent that would go ahead under the aggregate price but far fewer than are excluded by the ad hoc safety price. This shows that "rule of thumb" uncertainty management (i.e., making ad hoc adjustments to the average probability) is not only baseless; it is not cost effective. Note also that the stakes and cautiousness functions, while subjective, are stable attitudes of the decision maker that persist across decisions. The confidence approach therefore ensures consistency across sets of decisions in a manner that ad hoc uncertainty management cannot.

Let us pause here to consider a potential objection: while the ad hoc safety price is obviously unjustified, in cases in which the average has greater "skill"

than the best model, the decision maker should simply rely on the average. In reply we say: decision-making using the average is overly reliant on the scoring process, which we noted in section 3 suffers from a paucity of data and a degree of arbitrariness in the choice of a SR. The confidence approach mitigates these worries by introducing a flexible degree of robustness. As we discuss below, there are different ways one might construct the nested sets, and it may well be reasonable to center them on an average. But the key to the confidence approach is that, as the stakes increase, one uses wider intervals around the center, thereby guarding against concerns about how that center was identified. So even when the average has greater "skill" than even the best individual model, one should prefer the confidence approach to using simply the average.

6. Methods for Constructing Nested Sets. The simple implementation of the confidence approach to model outputs described above is by no means the only option. Our view is that there is no "one size fits all" method for the construction of nested sets, given the diversity of target systems and modeling endeavors. Instead the set-construction method will depend on the specifics of the ensemble. In this section we make a start on a "toolbox" for model-based decisions: outlining several potential set-construction methods and identifying what each requires of the ensemble and when it is likely to perform well.

In section 5 we constructed our intervals by starting with the best model, m_1, and including the next best (according to rule R) each time. But we could also describe what we did as starting with m_1 and including the next *closest* model each time, with respect to the Euclidean distance between outputs. In the toy example these procedures generate the same result, and we did not specify which we were following at the time. But in general, we may not have a reliable rule (sec. 3), and these two orderings may diverge. We now outline a decision tree for how to construct a nested hierarchy in the general case.

The first question is: Can you identify a model output or outputs to act as a center for the nesting? (We consider different ways you might successfully accomplish this in sec. 6.1) If yes, then: Do you also have a reliable ordering of model outputs? If yes, then we recommend forming the nesting in line with this ordering (sec. 6.1.1). But note that in section 3 we outlined various problems with scoring models in the hurricane case—at this point in the decision tree, our case likely yields a no. In this case, we recommend forming the nesting by including models in distance order (sec. 6.1.2).

If you cannot identify a center model (sec. 6.2), then we ask: Can you defend the use of one of a suite of statistical methods that construct a center (and a nesting to go with it)? In section 6.2.1 we consider centering around the average using either a partial ordering or distance, and in section 6.2.2 we discuss a method that uses central intervals of a Gaussian distribution to define

confidence levels. If you cannot do either of these, then you are in our worst-case scenario and must use only the widest envelope of your model outputs (sec. 6.2.3).

6.1. Cases with an Identified Center Model. Let us consider cases in which we can identify one model as best, and so we use it as the center. First, this identification might use a skill score or multiple scores. Recall that in section 3 we presented a number of limitations of using the weighted average of the hurricane ensemble outputs, two of which also speak against the use of a SR to rank models: (1) there are many such rules and choosing between them is a complex matter over which experts disagree, and (2) there may be limited data for testing, in which case the scores may be unreliable.

In the simplest case when neither of these problems is salient, we will have a single SR that makes use of sufficient data to identify one model as best. If so, we use it to form the center. Note that we need not always center on a point output. In situations in which we are uncertain, it may be natural to have the most precise claim we are willing to accept be interval valued: an uncertainty range around the best output, reflecting the uncertainty in even our best model. More complex cases will involve multiple plausible SRs. If they agree on the best model, we not only have a starting point for the hierarchy but can regard it as having a degree of robustness. If the rules disagree, we are in a difficult situation in which there are multiple best models (Betz 2009). In such a situation we can still follow the robustness thought and form a central interval from the best model identified by each SR. Finally, we may have a method of identifying a center that does not rely on a skill score, for example, if experts tell us one model is best without providing a performance-based rationale. (The same considerations discussed for SRs apply.)

Given that we have a center, we now need to form the nesting. Here the natural question is: Can we form a reliable partial ordering of models, reflecting their strength? We consider first the positive answer case, then the negative.

6.1.1. Nesting Using a Partial Ordering. As one of the main ways of identifying a center is using a skill score, we will first consider the case in which we trust a rule (or rules) to partially order the models. As with the center, good cases using a SR order are those in which there is a natural rule and plenty of data. Here the rule's ordering gives evidence of model strength, and we can follow it as in the toy example. If there is more than one SR on the table, we can attempt to form a SR order by consulting each of them. In the best case, they agree, and we use the resulting order.[12] This would confer some

12. We only ever use the ordering provided by a rule. When there are multiple rules, agreement means ordinal rather than cardinal agreement. We are therefore always in a better position than averaging with respect to sec. 3.

robustness on the ordering and the resulting hierarchy. If they disagree, we are back in a difficult case. Following the thinking above, we might try to form the interval about the center by including all the second-best model outputs, and so on.[13] This is a rather cautious approach, and relatively small differences between the rules could lead to a very coarse-grained hierarchy.

A less cautious approach is to break the tie between, for example, two models each ranked second by some SR, using the distance of each output from the last interval in the nesting. This produces a finer-grained hierarchy, which may be helpful when the SR order is too coarse to allow for the desired number of confidence levels.

6.1.2. Nesting Using Distance. If we have no reliable ordering information about models, other than the identified center, then we can use the distance of models from the center to form a naive ordering. A hierarchy built on this ordering will respect the logic of confidence and will produce relatively fine-grained hierarchies (unless many model outputs happen to be equally spaced), which can then be coarse grained to form confidence levels. This method is conservative, in that it uses only model outputs to form the hierarchy, unlike methods discussed below.

The problem with it is that distance ordering need not track any facts about model strength. When we use a SR order, we know something about the confidence gains resulting from moving to a wider interval: each step up in the hierarchy delivers weakly less incremental confidence than the previous step. Using a distance ordering does not ensure this, and so the resulting hierarchy is less informative. This makes sense in our more uncertain case, but it is why we do not endorse distance ordering when there is a defensible SR order available.

6.2. Cases without an Identified Center Model. We now consider cases in which we cannot identify any center. Here the only facts available to a decision maker are the model outputs themselves; we are in a case of more severe uncertainty and can use only distributional properties of the ensemble to generate our hierarchy.

6.2.1. Nesting around the Average. It is conceivable that there are cases in which the model average has more skill than any individual model according to the SR being used. In such a scenario one can use the methods discussed in section 6.1 and nest the models around the mean using either a partial ordering or a distance.

13. Nothing in our method requires us to consider a sequence of points when constructing intervals. If the models generate interval-valued outputs, we can conduct the confidence procedure using any of the options outlined in this decision tree.

6.2.2. Nesting Using Statistical Methods. In cases in which there is neither a center model nor a meaningful partial ordering or distance, one may nevertheless construct a nesting using statistical methods. The thought here is that the ensemble contains useful information about the phenomenon of interest, at the level of individual model outputs, but that we are unable to extract it through model comparisons like performance testing. Treating the models statistically, we can attempt to structure this information at the level of the ensemble and use it to guide our decision-making.

There are many statistical methods, and comprehensive discussion of their uses in the context of the confidence approach is a project for future research. We here briefly outline a simple method from a natural science setting closer to our case study: the Coupled Model Intercomparison Project (CMIP5) for general circulation models (GCMs). The method uses point estimates and centers a nesting around the average. The foundation of this approach is "one model, one vote" (each model is treated equally), with results generated by simple statistical analysis. To begin, we calculate a straightforward arithmetic mean of model outputs, $\bar{m}$, and use this as the center of the nesting (Stocker et al. 2014, 754). We then calculate the variance of the output set, defined simply as $s^2 = \Sigma_i(m_i - \bar{m})^2/n$. Assuming error is Gaussian, one can then input these into a Gaussian $G(x) = c\exp[(x - \bar{m})^2/2s^2]$, where c is a normalization constant. With this in place we can calculate nested intervals directly from the distribution. We can center on the mean and then consider various centered intervals of the distribution: the central 50%, central 80%, and so on. These form the sets of the nested hierarchy.

This method has limitations. The key assumption here is that all models are of equal value—this underlies the simple arithmetic mean and uniform variance analysis. This may seem implausible, either because not all models are on a par or because they are not independent and so "voting" may double count (see Knutti 2010). The center is also sensitive to the number of models in a way that scoring approaches are not: the addition of duplicate models may move the center without adding additional scientific information. This approach is therefore best used in situations in which there is a fixed and small number of models and no method to rank them and all of their output values are plausible—a description many climate scientists believe holds for GCMs.

Statistical methods are also common in economics (where the term "model" is often used differently, to refer to a function of the underlying data), for instance, in the robustness method of Hansen and Sargent (1982). We will not discuss the large range of options available in this case—including maximum entropy, Bayesian model averaging, and so on. These methods typically use richer information than we have presented in this article—such as a full probability distribution rather than merely a point estimate. The confidence approach works with each of them, and at a high level of abstraction the procedure is the same: center the hierarchy on the constructed central estimate

of the relevant probability and then form confidence levels using distributional facts.

6.2.3. Working without Nesting. In the worst cases, we will not be able to rely on any of the foregoing methods. We may not believe any SR can adequately measure model skill, be unable to identify a best model or models, and have reasons to doubt the applicability of distribution fitting or other statistical techniques.

Stainforth, Allen, et al. (2007) argue that this is the case for GCMs in the CMIP5 ensemble. They argue that today's GCM ensembles provide only a "nondiscountable envelope" of outcomes (i.e., a set of possible outcomes). No individual model can provide a reliable central estimate, and therefore the ensemble should not be used to create one through aggregation. Any construction of a probability density function such as through the method described above, is therefore likely to mislead decision makers through false precision (2158). Worse, they provide only a lower bound on the range of uncertainty, because further uncertainty exploration is likely to increase it (Stainforth, Downing, et al. 2007, 2166). This is an extreme view—if it were widely accepted, the Intergovernmental Panel on Climate Change (IPCC) process would not be seen as generating anything of decision relevance—but it is a useful limit case when considering the options within our approach.

Stainforth, Allen, et al.'s (2007) arguments for these conclusions are complex, but at heart the issue is multiple uncertainties, each severe and in combination so limiting that we cannot use these models to make point predictions. The members of the ensemble are so interdependent, they argue, that we should also not believe that model agreement lends any additional confidence. All we can present is the range of results generated by our models and the range of uncertainties accompanying them. These are useful: they represent *informed assessments of possibility*, formulated by our best experts. They therefore determine a region of output space that is "nondiscountable" (i.e., that we should not expect the truth to lie outside).

In situations like this, in which the ensemble is thought to represent only a part of our uncertainty and the model results are not particularly reliable, what can the confidence approach say? We could follow the recipe of one of the statistical methods above to form a hierarchy and, therefore, provide some sense of more and less confidence-generating claims. But, when we coarse grain to confidence levels, even the widest set in the hierarchy must be regarded as having low confidence—which this is now interpreted in the sense of being nondiscountable. In order to gain more confidence, we must move to yet wider sets, and here we may have little to guide us. The confidence approach tells us that if our decision is high stakes, and our cautiousness dictates high confidence, we will have to use some wider interval than any supported by the model ensemble (in the extreme, including [0, 1]).

An additional problem is that there may be serious possibilities that are not reflected in the range of model outputs, and in such a situation it is unclear why the envelope of the model results can be seen as narrowing down the nondiscountable option (Betz 2013). The IPCC recognizes this possibility and in response has endorsed the practice of "downgrading" prediction confidence. Here, outcomes that are generated by examining the 5%–95% range of model results (e.g., for global mean temperature change in 2100, under a particular forcing scenario) are reported as merely "likely" (>66% probability) rather than "very likely" (>90% probability; Stocker et al. 2014, table SPM.2). This way of catering for the possibility that something that the models do not simulate happens uses expert judgment (Frigg, Thompson, and Werndl 2015, 973). Insofar as this reassignment reflects information that scientists hold about limitations in the prevailing modeling, it is surely more transparent to reflect it through the confidence grading of different probability ranges than by downgrading the probabilities themselves, for example, by reporting that the results are "very likely" at medium confidence but "likely" at high confidence (see Mach and Field 2017; Helgeson, Bradley, and Hill 2018).

7. Conclusion. The standard approach to working with model ensembles is beset with problems. Aggregation relies on a nonunique predictive test and SR, whose choice is difficult to motivate to decision makers. It requires significant data, which may not be available. Crucially, it misrepresents the state of scientific knowledge to decision makers by producing a single value for $p(E)$, without reflecting the underlying uncertainty. This is compounded by decision makers not knowing what to do with uncertainty information, were it to be given to them.

In the confidence approach we are as explicit as possible about uncertainty at every stage. Decision makers are presented with a variety of options: different sets of probabilities, each with an attached cost to their use in the form of the confidence it can support. One can always demand more specificity, but it is clear what is given up when doing so. There is a natural, and we think valuable, link between the importance of the decision, the confidence that importance demands, and the formulation of decision input.

In our insurance case study the benefits are marked. Current practice tries to build the "missing" uncertainty back in, in a costly and ad hoc manner. The confidence approach, however, allows decision makers to respond to uncertainty in a principled but flexible manner. In practical terms this would be done by replacing an opaque "technical" process (aggregating model outputs) with a structured process of value elicitation, in order to formulate the stakes and cautiousness functions.

The major research question facing this approach is how to construct the hierarchy of nested sets. In this article we have begun a partial taxonomy of

methods for set construction and the conditions under which they are applicable. Careful work is required to determine where specific cases lie, and there are surely additional methods not covered here. One obvious candidate, much discussed in the climate literature, is expert elicitation, which could also be used to construct the confidence levels.

The approach outlined here is not restricted to insurance or hurricane modeling. In principle, the approach can be expanded to cover any decision support using a model ensemble—including nonprobabilistic outputs. Doing so would better reflect uncertainty and strike a balance between cautious decision-making in the face of uncertainty and avoiding complete decision paralysis.

REFERENCES

Armstrong, J. Scott. 2001. "Combining Forecasts." In *Principles of Forecasting*, 417–39. Boston: Springer.

Australian Bureau of Meteorology. 2017. "Forecast Verification." https://web.archive.org/web/20171125111801/https://www.cawcr.gov.au/projects/verification/.

Betz, Gregor. 2009. "Underdetermination, Model-Ensembles and Surprises: On the Epistemology of Scenario-Analysis in Climatology." *Journal for General Philosophy of Science* 40:3–21.

———. 2013. "In Defence of the Value Free Ideal." *European Journal for Philosophy of Science* 3:207–20.

Booth, Ben B. B., Nick J. Dunstone, Paul R. Halloran, Timothy Andrews, and Nicolas Bellouin. 2012. "Aerosols Implicated as a Prime Driver of Twentieth-Century North Atlantic Climate Variability." *Nature* 484 (7393): 228–32.

Bradley, Richard. 2017. *Decision Theory with a Human Face*. Cambridge: Cambridge University Press.

Bradley, Seamus. 2019. "Imprecise Probabilities." In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University. https://plato.stanford.edu/entries/imprecise-probabilities/.

Clemen, Robert T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5 (4): 559–83.

Dietrich, Franz, and Christian List. 2016. "Probabilistic Opinion Pooling." In *Oxford Handbook of Probability and Philosophy*, ed. Alan Hajek and Christopher Hitchcock. Oxford: Oxford University Press.

Enfield, David B., and Luis Cid-Serrano. 2006. "Projecting the Risk of Future Climate Shifts." *International Journal of Climatology* 26 (7): 885–95.

FCHLPM (Florida Commission on Hurricane Loss Projection Methodology). 2007. "Report to the Florida House of Representatives Comparison of Hurricane Loss Projection Models." https://www.sbafla.com/method/Portals/Methodology/Meetings/2007/20071105_RubioReport.pdf.

———. 2015. "Current Year 2015 Modeler Submissions." https://www.sbafla.com/method/ModelerSubmissions/CurrentYear2015ModelerSubmissions.aspx.

Frigg, Roman, Erica Thompson, and Charlotte Werndl. 2015. "Philosophy of Climate Science." Pt. 2, "Modelling Climate Change." *Philosophy Compass* 10 (12): 965–77.

Gilboa, Itzhak, and Massimo Marinacci. 2013. "Ambiguity and the Bayesian Paradigm." In *Advances in Economics and Econometrics: Theory and Applications*, ed. D. Acemoglu, M. Arellano, and E. Dekel. Cambridge: Cambridge University Press.

Gilboa, Itzhak, and David Schmeidler. 1989. "Maxmin Expected Utility with Non-unique Prior." *Journal of Mathematical Economics* 18 (2): 141–53.

Guin, Jayanta. 2010. "Understanding Uncertainty." AIR Worldwide, March 16. http://www.air-worldwide.com/Publications/AIR-Currents/2010/Understanding-Uncertainty/.

Hansen, Lars Peter, and Thomas J. Sargent. 1982. *Robustness*. Princeton, NJ: Princeton University Press.

Heal, G., and A. Milner. 2014. "Uncertainty and Decision Making in Climate Change Economics." *Review of Environmental Economics and Policy* 8:120–37.

Helgeson, Casey, Richard Bradley, and Brian Hill. 2018. "Combining Probability with Qualitative Degree-of-Certainty Assessment." *Climatic Change* 149 (3–4): 517–25.

Hill, Brian. 2013. "Confidence and Decision." *Games and Economic Behavior* 82:675–92.

———. 2016. "Incomplete Preferences and Confidence." *Journal of Mathematical Economics* 65 (August): 83–103.

———. 2019. "Confidence in Beliefs and Rational Decision Making." *Economics and Philosophy* 35 (2): 223–58.

Insurance ERM. 2018. "RMS Responds to AIR's Attack on Hurricane Risk Modelling." Insurance ERM, May 29. https://www.insuranceerm.com/news-comment/rms-responds-to-airs-attack-on-hurricane-risk-modelling.html.

Jewson, Stephen, Enrica Bellone, Thomas Laepple, Kechi Nzerem, Khare Shree, Manuel Lonfat, Adam O. Shay, Jeremy Penzer, and Katie Coughlin. 2009. "Five Year Prediction of the Number of Hurricanes That Make United States Landfall." In *Hurricanes and Climate Change*, ed. James B. Elsner and Thomas H. Jagger. New York: Springer.

Keynes, John Maynard. 1921. *A Treatise on Probability*. London: Macmillan.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. "A Smooth Model of Decision Making under Ambiguity." *Econometrica* 73 (6): 1849–92.

Knutti, Reto. 2010. "The End of Model Democracy?" *Climate Change* 102:395–404.

Kriegler, Elmar. 2007. "Updating and Testing Beliefs: An Open Version of Bayes' Rule." In *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, ed. Gert de Cooman, Jiřina Vejnarová, and Marco Zaffalon, 271–80. Prague: Action M.

Leitgeb, Hannes, and Richard Pettigrew. 2010. "An Objective Justification of Bayesianism I: Measuring Inaccuracy." *Philosophy of Science* 77:201–35.

Mach, Katharine J., and Christopher B. Field. 2017. "Toward the Next Generation of Assessment." *Annual Review of Environment and Resources* 42:569–97.

Marinacci, Massimo. 2015. "Model Uncertainty." *Journal of the European Economic Association* 13:1022–100.

Morgan, M. Granger. 2014. "Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy." *Proceedings of the National Academy of Sciences of the USA* 111 (20): 7176–84.

Peirce, Charles S. 1878. "The Probability of Induction." *Popular Science Monthly* 12.

Philp, Tom, Tom Sabbatelli, C. Roberston, and Paul Wilson. 2019. "Issues of Importance to the (Re)Insurance Industry: A Timescale Perspective." In *Hurricane Risk*, vol. 1, ed. Jennifer Collins and Kevin Walsh. Cham: Springer.

Sabbatelli, Tom. 2017. "Catastrophe Modeling." Pt. 2. *RMS Blog*, September 2. http://www.rms.com/blog/tag/catastrophe-modeling/page/2/.

Sabbatelli, Tom, and Jeff Waters. 2015. "We're Still All Wondering: Where Have All the Hurricanes Gone?" *RMS Blog*, October 27. http://www.rms.com/blog/2015/10/27/were-still-all-wondering-where-have-all-the-hurricanes-gone/.

Shome, Nilesh, Mohsen Rahnama, Steve Jewson, and Paul Wilson. 2018. "Quantifying Model Uncertainty and Risk." In *Risk Modeling for Hazards and Disasters*, ed. Gero Michel, 3–46. Amsterdam: Elsevier.

Stainforth, David A., M. R. Allen, E. R. Tredger, and Leonard Smith. 2007. "Confidence, Uncertainty and Decision-Support Relevance in Climate Predictions." *Philosophical Transactions of the Royal Society* A 365 (June): 2145–61.

Stainforth, David A., T. E. Downing, R. Washington, A. Lopez, and M. New. 2007. "Issues in the Interpretation of Climate Model Ensembles to Inform Decisions." *Philosophical Transactions of the Royal Society* A 365 (1857): 2163–77.

Stocker, Thomas F., et al. 2014. *Climate Change 2013: The Physical Science Basis*. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. New York: Cambridge University Press.

Stone, James. 1973. "A Theory of Capacity and the Insurance of Catastrophe Risks." Pt. 1. *Journal of Risk and Insurance* 40 (2): 231–43.