Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Interpreting the Probabilistic Language in IPCC Reports

COREY DETHIER
Leibniz Universität Hannover

The Intergovernmental Panel on Climate Change (IPCC) often qualifies its statements by use of probabilistic "likelihood" language. In this paper, I show that this language is not properly interpreted in either frequentist or Bayesian terms—simply put, the IPCC uses both kinds of statistics to calculate these likelihoods. I then offer a deflationist interpretation: the probabilistic language expresses nothing more than how compatible the evidence is with the given hypothesis according to some method that generates normalized scores. I end by drawing some tentative normative conclusions.

1. Introduction

The Intergovernmental Panel on Climate Change (IPCC) systematically uses probabilistic language in two different ways: in qualitative "confidence" judgments and in quantitative, but usually imprecise, "likelihood" assignments. This two-fold use of probabilistic language has attracted substantial discussion from scientists (including the authors of various IPCC reports), who worry that the two categories are not clear enough for either the scientists who are asked to make such judgments or the policy-makers and public who are asked to interpret them.[1] To date, philosophical attention to this two-fold use of probabilistic language has (understandably) focused on the relationship between the two: since the confidence language is often applied not just to hypotheses but also to the likelihood assignments, it's an interesting question

1. For a sampling, see Mach, Mastrandrea, Freeman, and Field (2017), Herrando-Pérez, Bradshaw, Lewandowsky, and Vieites (2019), and Janzwood (2020) as well as the citations therein. Note also that the IPCC's approach has been influential throughout climate science; Crimmins (2020) documents that many of the same problems can be found in U.S. governmental reports on climate change.

Contact: Corey Dethier <corey.dethier@gmail.com>

https://doi.org/10.3998/ergo.4637

how this confidence language is to be interpreted and used.[2] The likelihood language has received less attention.

In this paper, I argue that the likelihoods found in IPCC reports are neither uniformly the frequencies of classical statistics nor the subjective credences of Bayesian statistics. The IPCC uses both classical and Bayesian methods in calculating likelihoods, meaning that either interpretation will be misleading if employed generally. My proposed alternative is a "thin" or "deflated" interpretation: what the IPCC is reporting when it reports likelihoods is the normalized "score" that is given by the best available method for measuring compatibility with the evidence. Employing probabilities in this "thin" sense allows the IPCC to compare results across studies that employ a variety of methodologies, but at the cost of making it more difficult to evaluate the practical implications of the relevant results.

Section 2 briefly describes the IPCC's use of probabilistic language. Sections 3 and 4 demonstrate that the IPCC's likelihood measures are sometimes based on classical confidence levels and sometimes on Bayesian posterior probabilities. Sections 5 and 6 lay out a pair of further constraints on any interpretation of the IPCC's use of "likelihood," while Section 7 gives my deflationist reading. Section 8 ends by drawing some tentative normative conclusions regarding how the IPCC might improve their use of probabilistic language.

2. The IPCC's Use of Probabilistic Language

The IPCC frequently uses probabilistic language in their reports. In this paper, I'm (mostly) going to focus on two examples from Working Group 1's contribution to the Fifth Assessment Report.[3] The first of these concerns the contribution of greenhouse gases to the change in temperatures observed to date:

> GHGs [greenhouse gases] contributed a global mean surface warming *likely* to be between 0.5°C and 1.3°C over the period 1951–2010. (IPCC 2013: 869)

The second example concerns *equilibrium climate sensitivity* or ECS, which measures how much the temperature would increase given a doubling of atmospheric CO_2 concentration:

> ECS is positive, *likely* in the range 1.5°C to 4.5°C with *high confidence*, *extremely unlikely* less than 1°C (*high confidence*) and *very unlikely* greater than 6°C (*medium confidence*). (IPCC 2013: 81)

2. See Bradley, Helgeson, and Hill (2017), Helgeson, Bradley, and Hill (2018), and Winsberg (2018a).

3. The most recent report, IPCC (in press), adopts the same approach as the Fifth.

These two examples illustrate a few of the interesting features of the IPCC's use of probabilistic language. First, and most obviously, the IPCC distinguishes between "likelihoods" and "confidence." (Note that both of these are terms of art in the context of IPCC reports and that neither should necessarily be understood as corresponding to the concepts of "likelihood" and "confidence" found in confirmation theory and statistics.) Second, confidence judgments often modify likelihood assignments; at least at face value, the IPCC is expressing high confidence in the probabilistic claim that "ECS is *likely* in the range 1.5°C to 4.5°C" rather than simply in the hypothesis that "ECS is in the range 1.5°C to 4.5°C." Third, the IPCC doesn't always use the language of confidence. In cases where it doesn't, however, it should be understood as implicit that the result is assigned "high" or "very high" confidence (Mastrandrea et al. 2010: 3).

This use of probabilistic language is relatively unique to climate science and invites investigation by philosophers. Unfortunately, the IPCC doesn't have a lot to say that clarifies how we are to understand the relevant terms. Here's what the "guidance note" released by the IPCC says with respect to confidence:

> Confidence in the validity of a finding, based on the type, amount, quality, and consistency of evidence (e.g., mechanistic understanding, theory, data, models, expert judgment) and the degree of agreement. Confidence is expressed qualitatively. (Mastrandrea et al. 2010: 1)

> [Confidence] synthesizes the author teams' judgments about the validity of findings as determined through evaluation of evidence and agreement. ... Confidence should not be interpreted probabilistically. (Mastrandrea et al. 2010: 3)

Elsewhere (e.g., Mach et al. 2017), (many of) the same authors more explicitly state that confidence should be interpreted as a measure of how well the relevant domain is understood.

The IPCC's comments on the likelihood measure are similar:

> [Likelihoods are] Quantified measures of uncertainty in a finding expressed probabilistically (based on statistical analysis of observations or model results, or expert judgment). (Mastrandrea et al. 2010: 1)

> Likelihood ... provides calibrated language for describing quantified uncertainty. It can be used to express a probabilistic estimate of the occurrence of a single event or of an outcome (e.g., a climate parameter, observed trend, or projected change lying in a given range). (Mastrandrea et al. 2010: 3)

This description tells us that that likelihoods can be determined in a number of ways, are designed to capture uncertainty, and can be applied to individual hypotheses (of various sorts). It doesn't answer any number of philosophical questions, however. It doesn't tell us, for example, whether likelihoods are long-run frequencies or subjective credences.

Philosophers aren't the only readers who are liable to find these guidelines insufficient. As has been thoroughly documented in the scientific literature, there are substantial differences in how the various terms are used throughout the IPCC reports and confusion from both users and the scientists themselves about how to interpret this probabilistic language (Janzwood 2020; Mach et al. 2017). These complaints have led to a number of different suggestions regarding how to improve both the guidelines and the IPCC's presentation of results themselves, though no one of these suggestions seems to have yet garnered sufficient support to replace the current practice.

There are many ways that philosophy might contribute to this project. So, for instance, it might be helpful for philosophers to offer an analysis of the problem situation faced by the IPCC and the best means of resolving that problem given the various competing desiderata. Such an analysis could indicate better ways to demarcate and use the relevant probabilistic concepts. I take it that this is essentially the project of Helgeson et al. (2018) and Parker and Risbey (2015), for example.

In this paper, by contrast, I'm not going to take up a philosophical analysis of the relevant concepts or situation. Instead, I'm going to offer an interpretation of the IPCC's current practice—that is, I'm going to characterize what "likelihood" means in the context of extant IPCC reports. As we'll see, I understand this as a largely descriptive project. After all, the IPCC doesn't pull the judgment that a particular hypothesis is "likely" out of thin air; these judgments are based on the results of statistical analyses. What "likely" means therefore depends on the kind of statistical analysis employed; the p-values and confidence levels generated by classical statistics simply are not posterior probabilities, and it would be a mistake to approach them as such. Though the project is mostly descriptive, the descriptive facts here have clear normative implications. For one thing, we can't properly determine whether some alternative method for presenting uncertainty would be better than the current method without a solid understanding of the latter. For another, how the relevant probabilities are in fact determined constrains how climate scientists should communicate the results, and as such we can't judge whether a given proposal is preferable without looking at the details of the statistical analyses.

One final comment. I shouldn't be seen as attempting to rule out principled *re*-interpretations of the IPCC's language. So, for instance, in many situations the tools of Bayesian and classical statistics can be justified for use by partisans

of the other framework as "good enough" approximations. It's therefore open to a Bayesian (/ frequentist) to argue that this or that classical (/ Bayesian) measure can be understood in terms of subjective probabilities (/ long-run frequencies). In what follows, I'm going to assume that this kind of re-interpretative project can only proceed once we have already determined what the relevant measures are, because (for example) while there are some contexts where we're justified in treating a p-value as telling us something about the subjective credence we should assign to a hypothesis, those contexts are not the same as the ones where we're justified in treating a posterior probability generated using "objective" Bayesian methods as telling us something about the subjective credence we should assign to a hypothesis. The goal here is to carry out this first, descriptive, step.

On its face, this is a simple task. As I detail over the next four sections, however, what we find when we carry out this analysis is that the IPCC is extremely pluralistic when it comes to statistical methodology, and thus that the language of likelihood cannot be straightforwardly interpreted in terms of either the subjective credences or the (long-run) frequencies familiar to philosophers. We need a different tool.

3. Likelihoods and Classical Statistics

Recall the first of the two probabilistic claims that opened the last section: "GHGs [greenhouse gases] contributed a global mean surface warming *likely* to be between 0.5°C and 1.3°C over the period 1951–2010" (IPCC 2013: 869). According to the IPCC, the "likely" modifier here should understood as a quantified (but imprecise) "likelihood," where the range covered by the imprecise "likely" modifier is 0.66–1.00.[4]

In interpreting what the IPCC means in this case, it's helpful to begin with what the IPCC themselves have to say about how the probabilistic claim is generated (see IPCC 2013: 883). What we find is that the 0.5–1.3°C estimate is determined by taking the smallest range compatible with the intervals given by Gillett, Arora, Matthews, and Allen (2013) and Jones, Stott, and Christidis (2013). When we follow up on these two studies, we find that these intervals are the 5–95% *confidence intervals*. Confidence intervals are a tool of classical statistics; a 5–95% confidence interval tells us that there is a probability of .05 that we would observe data as extreme as the actual data given the assumption the true contribution falls below / above the specified range. The *confidence level* associated with this confidence interval is

4. For a list of the different likelihood terms and their associated intervals, see Mastrandrea et al. (2010).

thus $1 - .05 \times 2 = .9$.[5] As a tool of classical statistics, confidence intervals and levels are usually properly interpreted in terms of frequencies: in the long run, 90% of .9 confidence intervals will include the true value.

The IPCC is explicit in endorsing this frequentist interpretation of the relevant likelihoods. As the IPCC puts it, using a hypothesis-testing example rather than an interval one:

> Attribution results are typically expressed in terms of conventional 'frequentist' confidence intervals or results of hypothesis tests: when it is reported that the response to anthropogenic GHG increase is *very likely* greater than half the total observed warming, it means that the null hypothesis that the GHG-induced warming is less than half the total can be rejected with the data available at the 10% significance level. (IPCC 2013: 878)

The IPCC's claim that the use of classical statistics is "typical" is not quite as accurate now as it was in 2013. Though there have been Bayesian attribution studies since the late 90s (e.g. Hasselmann 1998), such methods have become much more common over the last decade. Today, attribution is an area in which both Bayesian and classical approaches flourish (a fact I'll discuss more in §5). Nevertheless, classical statistics remains common in attribution studies, including those that are relied on by the IPCC (see, e.g., Gillett et al. 2021).

The upshot of the foregoing is that the IPCC often uses its likelihood terminology in cases where the relevant probabilities are measures of classical statistics such as confidence levels. As such, it cannot be interpreted as reporting posterior probabilities in these cases: confidence levels just aren't the same thing as posterior probabilities, and it's well known that treating confidence levels (or the associated p-values) as though they were posterior probabilities is fallacious. As a consequence, the IPCC's likelihood language should not—at least in this instance—be interpreted in terms of subjective credences. Or, more precisely, we cannot take it that when the IPCC says that a hypothesis is "likely" ("very likely" etc.), that means that either the organization or the authors of the particular report assign the associated subjective credence to it. Instead, these uses of likelihood terminology are much more naturally interpreted in terms of the frequencies of classical statistics.

Of course, to reiterate a point from the prior section, the fact that the IPCC employs classical measures in attribution studies doesn't mean that we can't or shouldn't reinterpret them in a Bayesian setting. For now, at least, my point is simply that the IPCC often uses likelihood language to communicate the results

5. The careful reader will note that this confidence doesn't align with the IPCC's use of "likely" (as opposed to "very likely") in this scenario. See § 6 for discussion.

from classical statistics, explicitly recognizes that it is doing so, and offers no indication that it is engaging in some sort of reinterpretative project beyond the simple pooling and comparison of (classical) results from different studies. At face value, therefore, in these cases the likelihood language should be interpreted in terms of (classical) confidence levels.

4. Likelihoods and Bayesian Statistics

How representative is the above example? There is little discussion of the difference between classical and Bayesian methods in the extant IPCC reports, but so far as I can tell the majority of the IPCC's likelihood judgments are based on the probabilistic measures of classical statistics and are thus not straightforwardly read as posterior probabilities.[6]

There is at least one notable exception in Working Group 1's 2013 report, however: the likelihood values assigned to ECS (and the related concept of *transient climate response* (TCR))—our second example from Section 2—are determined by posterior probabilities. As the IPCC's discussion stresses, there are a number of lines of evidence that are relevant to the estimation of ECS: it can be estimated directly from paleoclimate data, or from contemporary trends; similarly, volcanic eruptions provide nice natural experiments for evaluating ECS; and theoretical knowledge embodied in global climate models also has implications for ECS (for discussion, see Winsberg 2018b). The estimates found in the IPCC report are based on evaluations that combine these different sources of evidence, such as Aldrin, Holden, Guttorp, Skeie, Myhre, and Berntsen (2012), and Olson, Sriver, Goes, Urban, Matthews, Haran, and Keller (2012).

These combined evaluations are explicitly Bayesian, as the IPCC notes:

> the probabilistic estimates available in the literature for climate system parameters, such as ECS and TCR have all been based, implicitly or explicitly, on adopting a Bayesian approach and therefore, even if it is not explicitly stated, involve using some kind of prior information. The shape of the prior has been derived from expert judgement in some studies, observational or experimental evidence in others or from the distribution of the sample of models available. (IPCC 2013: 922)

The upshot is that the ranges that the IPCC uses in estimating ECS are not confidence intervals strictly speaking. Instead, they're what are sometimes called

6. Note that that's not to say that priors have *no* role in determining the values reported. Often times, the reported values are the final results of extremely complex chains of reasoning; my claim is simply that the final outcomes of these processes are classical in character.

"credibility intervals": the 5–95% credibility interval is the narrowest range such that the true value has a .05 posterior probability of falling below / above that range.[7]

It's worth stressing—as indicated by the quote above—that the priors employed in these studies are not (necessarily) the IPCC's own.[8] For one thing, the studies aren't carried out by the IPCC itself. For another, the standard approach in this area is to use priors that are considered more "objective": either those elicited from groups of experts or mathematical constructions designed to be as uninformative as possible.[9] Furthermore, it's routine to test how sensitive hypotheses are to the choice of priors. It is thus not particularly plausible to view the IPCC's reported likelihoods here as its group credence in the hypothesis; since the relevant posterior probabilities are not derived by updating on the subjective priors held by the group, they ought to be interpreted as measuring something other than the author's confidence.[10]

Here is where we stand at present. Sometimes, the IPCC uses likelihood language to communicate how well a hypothesis or event scores according to the methods of classical statistics; but also, in other cases, it sometimes uses the same language to communicate how well a hypothesis or event scores according to the methods of Bayesian statistics. Though I've pulled out two places where the IPCC is explicit in identifying which tools are being employed, it would be a stretch to say that there's anything like a careful demarcation between the cases where likelihood language is based on classical measures and those where it is based on Bayesian ones. Indeed, it's much more accurate to say that this distinction isn't emphasized—at least not in the reports of Working Group 1.

Whatever the practical benefits of this practice, it does pose a problem for the philosopher interested in interpreting the IPCC's claims. The best that it seems like we might be able to offer is a disjunctive reading: sometimes likelihoods are frequencies and sometimes they're subjective credences, but they're never both or anything else. As we'll see, there are reasons to reject this disjunctive reading too: the IPCC's practice places additional constraints on any interpretation of its talk of "likelihoods," and these constraints further complicate the interpretative endeavor in a way that rules out the simple disjunctive view just sketched.

---

7. The confidence interval-credibility interval distinction is not as uniformly observed as it ought to be. It's not uncommon, either in climate science or more generally, to see "confidence interval" used in the context of Bayesian statistics. I'll be avoiding that practice here for obvious reasons.

8. Compare Rougier and Crucifix (2018: 372): "The IPCC reports are valuable sources of information, but no one owns the judgements in them."

9. For discussion of the sense of "objectivity" that's relevant to the IPCC reports, see Jebeile (2020).

10. There's quite a bit more to be said about the use of objective priors in situations like the IPCC's, but as that discussion is (mostly) orthogonal the present one, I'll forgo it at present.

5. Neither Classical nor Bayesian

The next two sections lay out two constraints—or, better, two sources of constraint—on any interpretation of the IPCC's use of "likelihood." The first of these are cases in which the IPCC's approach is neither straightforwardly classical nor straightforwardly Bayesian.

Consider Table 1, which is adapted from IPCC (in press: chapter 3). Table 1 compares the estimates given by different studies for the °C temperature change attributable to humans over different periods. The important piece of information here is that these studies are neither all Bayesian nor all classical. On the contrary, Ribes et al. (2021) is Bayesian while Gillett et al. (2021) is classical. So the intervals on the first line of the table are credibility intervals: the interval given is the one such that .9 of the posterior probability falls within it. And the intervals on the second line of the table are confidence intervals, meaning that the null hypothesis that the truth falls outside the interval can be rejected at a .9 significance level.

The upshot is that the IPCC doesn't just rely on classical and Bayesian statistics in estimating the likelihoods of entirely different quantities; it also sometimes relies on classical and Bayesian statistics in estimating one-and-the-same quantity. From a certain point of view—one that I'll argue we should resist in Section 7—this kind of comparison is fallacious or methodologically suspect. After all, even though the two studies both report 5–95% ranges, those numbers don't *mean* the same thing in the two different contexts. One of them says something about posterior probabilities; the other about the probability of observing data at least as extreme. It's a well-known fallacy to conflate the two. So (from this perspective) presenting these estimates together as though they measure the same thing is misleading at best.

Note that the use of these different studies gets even more complicated, however. Table 1 includes the IPCC's own estimate, which is determined by taking the smallest tenth of a °C interval compatible with all three studies. These intervals are then said by the IPCC to be "likely." How should we interpret *this* use of "likely"? Effectively, the IPCC's intervals are determined by aggregating the intervals given using both Bayesian and classical methods. As such, the probabilities that go into determining the likelihood score are (depending on one's point of view) either both posterior probabilities and confidence levels or neither of the two, meaning that neither subjectivist nor frequentist interpretation of probability really applies to this particular use of the likelihood language.

Maybe the argument just presented is too fast, however. One way that we could motivate the IPCC's methodology here is to adopt a Bayesian perspective in which we assume that the classical probabilities reported by Gillett et al. (2021) are close enough approximations of the relevant authors' subjective views; the approach the IPCC actually adopts could then be seen as a way of pooling the

	1986–2005	1995–2014	2006–2015	2010–2019
Ribes et al. (2021)	0.52–0.77	0.69–0.94	0.81–1.08	0.89–1.17
Gillett et al. (2021)	0.32–0.94	0.63–1.06	0.74–1.22	0.92–1.30
Haustein et al. (2017)	0.58–0.82	0.75–0.98	0.87–1.10	0.94–1.22
IPCC (in press)	0.3–1.0	0.6–1.1	0.7–1.3	0.8–1.3

Table 1: The °C temperature change attributable to humans by period since 1986. Each interval given by a study is 5–95% interval; the IPCC estimate is the "likely" interval. Modified from IPCC (in press: chapter 3).

views of different experts.[11] Qua (Bayesian) justification of the IPCC's practice, I don't know whether this line of thinking can be fleshed out successfully. Whatever its prospects, it isn't what the IPCC is doing, at least not in any explicit way. Instead, the IPCC is by all appearances simply adopting a much simpler view towards the results of these different studies—namely, that all of the studies are measuring the same thing. Whether this simpler view can be justified retrospectively on Bayesian grounds is an interesting question, but (to my mind) an importantly different one from the question of what the IPCC is saying when it claims that the resulting ranges are "likely."

This example motivates two important conclusions. First, and more obviously, a simple form of disjunctivism won't allow us to give a general and consistent interpretation of the IPCC's practice. Not only does the IPCC treat the probabilistic measures yielded by classical and Bayesian statistics as comparable, it (at least sometimes) employs measures that are neither purely Bayesian nor purely classical. Nevertheless, all of these different measures are talked about using the same language. Regardless of the methods employed, the IPCC expresses the resulting judgments using the same likelihood terminology.

Second, the IPCC's approach to statistical methodology is pluralist in a way that isn't well captured by the contrast between classical and Bayesian methods. The superficial problem is that the IPCC relies on methods that don't fit neatly into this contrast—such as (e.g.) methods that aggregate the two, which neither framework consistently licenses. But the deeper problem is that the IPCC is pluralistic about methods at a finer level of grain than this contrast allows: as we'll see in the example of the next section, the IPCC often employs the same kind of aggregation method seen above across studies that differ in their methodologies but not in the interpretation of their outputs. That is, the IPCC's pluralism is first-and-foremost a pluralism about statistical methods generally speaking, and only in virtue of this general pluralism is it a pluralism about the distinction between classical and Bayesian methods.

11. Roussos, Bradley, and Frigg (2021) outline how such an approach might work, but do not suggest that the IPCC is carrying out a similar project.

This shouldn't be surprising. Climate scientists, and I suspect scientists generally speaking, tend to be pragmatic about the difference between Bayesian and classical methods, treating it as little more substantial than any other difference in statistical approach. Consider again Ribes et al. (2021). Even a brief perusal of Ribes et al. (2021) reveals that the motivation for employing Bayesian methods rather than classical ones has next to nothing to do with philosophical debates over the proper interpretation of probability. Instead, Ribes et al. (2021) stress that introducing a prior probability distribution makes it easier to account for different potential sources of error in a non ad-hoc manner. This example is indicative. In examining other uses of Bayesian statistics in climate science, from Hasselmann (1998) through Annan and Hargreaves (2017) and Katzfuss, Hammerling, and Smith (2017), one consistently finds that the reason for moving to Bayesian methods is their utility for a particular problem, not (or at least not primarily) any philosophical or theoretical advantages that they might have.

The IPCC is not itself a body that conducts research (or at least it is not primarily that). Instead, its reports serve to summarize the state of the field. *Given* that the field is pluralistic in the sense that different scientists approach problems with different tools, the IPCC has little choice but to adopt a similar pluralism insofar as it aims to simply report the results of the research that is actually conducted. I'll be doing more work with this point below; for now the important upshot is that when evaluating the IPCC's use of probabilistic language, we should recognize the IPCC's pluralistic stance towards statistical methodology. As exemplified by all of the examples employed in this paper, the IPCC often chooses to base its conclusions on a set of studies that employ different methods but that are treated as equally trustworthy. In some cases, *one* of the differences between the methods is that some of them are classical and some Bayesian, but in our analysis we should be careful to recognize the more general phenomenon.

6. Modifying Likelihood Claims

The second of the two constraints on any interpretation of the IPCC's use of likelihood language comes from the IPCC's practice of *modifying* likelihood claims. There are at least two senses in which the IPCC modifies its likelihood claims, and both constrain how we can interpret the probabilistic language.

First, the IPCC regularly "downgrades" the results given by the individual studies that it surveys to "account for residual sources of uncertainty." Here's the relevant discussion of the attribution example from Section 3:

> We derive assessed ranges for the attributable contributions of GHGs, other anthropogenic forcings and natural forcings by taking the smallest

> ranges with a precision of one decimal place that span the 5 to 95% ranges of attributable trends over the 1951–2010 period from the Jones et al. (2013) weighted multi-model analysis and the Gillett et al. (2013) multi-model analysis. We moderate our likelihood assessment and report *likely* ranges rather than the *very likely* ranges directly implied by these studies in order to account for residual sources of uncertainty. (IPCC 2013: 883)

While this passage is fairly technical, the basic picture is the following. The mechanical statistical evaluations carried out in the two studies referenced—Gillett et al. (2013) and Jones et al. (2013)—assign a confidence level of .9 to the interval 0.5–1.3°C. This would imply a "very likely" judgment on the IPCC's scale. But these studies are (by their own acknowledgement) imperfect; they rely on various idealizations and assumptions that are either risky or known to be false.

Working through an example will be helpful. Both of the attribution studies appealed to in the above quote employ a form of regression analysis known as "errors in variables" or EIV analysis.[12] So a standard regression analysis involves estimating the β_i terms in the equation $\mathbf{Y} = \sum_i \beta_i \mathbf{X}_i + v_{\mathbf{Y}}$. Here $\mathbf{Y}$ is the observed data (e.g., temperature changes), $\mathbf{X}_i$ is the effect of a single causal factor i (think greenhouse gases) on the observed system, β_i is the strength of that effect, and $v_{\mathbf{Y}}$ captures the system's internal variability.

Crucially, standard regression analyses require treating each $\mathbf{X}_i$ as *known*. In fact, however, each $\mathbf{X}_i$ is estimated—usually using climate models but sometimes from some subset of the data—meaning that there is non-negligible uncertainty about the accuracy of the estimate. EIV analysis introduces an additional variation term ($v_{\mathbf{X}_i}$) to account for this uncertainty (yielding a modified regression equation $\mathbf{Y} = \sum_i \beta_i(\mathbf{X}_i + v_{\mathbf{X}_i}) + v_{\mathbf{Y}}$). But this variation term must itself be estimated, and there's no perfect method for doing so—indeed, Gillett et al. (2013) and Jones et al. (2013) use different methods. So here is one "residual source of uncertainty" that the mechanical application of statistics doesn't account for. Since the studies in question don't perfectly account for this and other sources of uncertainty, the IPCC "downgrades" the resulting likelihood judgment from "very likely" to simply "likely."

The other way that the IPCC qualifies likelihood judgments is with the use of confidence judgments, as in the claim that "ECS is positive, *likely* in the range 1.5°C to 4.5°C with *high confidence*" (IPCC 2013: 81). As in the first of the two examples, the explicit statement of confidence in a result reflects—or at

12. See Carroll, Ruppert, Stefanski, and Crainiceanu (2006) for an introduction and Dethier (2022a; 2022b) for a more indepth discussion of this example. Confusingly, "EIV" is often used in a more specific way in the climate science setting, but the details that separate these more specific methods don't matter for the present purposes.

least appears to reflect—a judgment about the trustworthiness of the methods employed in estimating the relevant likelihood or likelihoods.

In principle, these two different means of modifying a likelihood judgment can be distinguished along the following lines. Suppose that some hypothesis *H* is "very likely" according to method *m* but that *m* doesn't account for some source of uncertainty—it relies on a risky assumption or an idealization that we haven't shown to be harmless. If we're reasonably certain that accounting for that uncertainty would lead to less support for *H*, then we should downgrade the likelihood judgment itself. If, by contrast, we don't know how accounting for the uncertainty would affect the result, the qualification should come in the form of a confidence judgment. As is stressed in the scientific literature (Janzwood 2020; Mach et al. 2017), however, there are substantial differences in how the authors of the different parts of the IPCC report approach the confidence-likelihood relationship, and so we should expect that not every example will fit this pattern.

Stepping back now from the details to look at the bigger picture. As just stressed, climate scientists are often in situations where it is not possible to carry out a statistical evaluation of the evidence without relying heavily on idealizations. I take it that it is an open question what scientists should do in these cases. Should they refuse to apply the statistical tools and thus give up on presenting anything but qualitative information, or should they forge ahead with the quantitative analysis while explicitly acknowledging the limitations of their results? Whatever the merits of these two approaches in the abstract, the IPCC's strategy is clearly closer to the latter extreme.[13] That is: though the IPCC relies heavily on the application of statistical methods, it doesn't apply these methods in a mechanical way: whatever likelihoods are, a mechanical application of statistical methods does not determine them. The likelihoods that IPCC reports have often been modified (in one or both of the two senses discussed) to reflect the IPCC's judgment about how reliable the methods in question are; the possibility of this kind of modification is an important constraint on our interpretation of the likelihood language.

It's worth stressing that the takeaways from the last two sections dovetail nicely. As the present section stresses, any actual study is likely to employ idealizations and approximations. As noted in the prior section, IPCC reports are strongly pluralist with respect to statistical methodology and frequently base their official statements on sets of studies that employ different methods. From the present perspective, the IPCC's pluralism can thus be seen as reflecting a view on which the different methods employed by these different studies all have equally good claim to being the best method available.

13. Most of the extant literature on the subject pushes in the opposite direction (without necessarily advocating for the opposite extreme); see, e.g., Stainforth, Allen, Tredger, and Smith (2007), Parker (2010), and Parker and Risbey (2015). Though see Dethier (2022b) for a recent defense.

7. Likelihoods as Thin Compatibility Scores

It may be tempting, given the prior discussion, to conclude that there is no general interpretation of the IPCC's use of likelihood language that is consistent or coherent. There is perhaps some sense in which this conclusion is accurate: if by "interpretation" we simply mean a disjunction between frequencies and credences (and perhaps also propensities), then there is no general and consistent interpretation of the IPCC's likelihoods. My view, however, is that this reading of "interpretation" is overly restrictive. In this section, therefore, I'll offer a "thin" or "deflationist" interpretation according to which the IPCC's likelihood language is simply communicating the normalized scores given by the best method available in the present context. The use of language that communicates this kind of "thin" probability allows the IPCC to present and compare pluralistic research in a relatively simple and consistent way; the cost is that it is more difficult to evaluate the practical implications of the relevant results.

It's helpful to begin at a very high level. At their most basic, various methods of statistical inference are "just" different means of using evidence to quantitatively differentiate between hypotheses. In general, different statistical methods aim to measure or score what we might call the "compatibility" of the evidence with different hypotheses, though they evaluate compatibility in different ways. So, for instance, a classical hypothesis test scores hypotheses on the basis of how well it predicts evidence that is at least as extreme, whereas a Bayesian hypothesis test scores hypotheses on the basis of how plausible they are after updating on the evidence. Some methods of statistical inference yield compatibility scores that operate on the same normalized [0,1] scale—in particular, both confidence levels and posterior probabilities are normalized in this technical sense. We can thus say that the methods that yield confidence levels and posterior probabilities both give *probabilistic* measures of the compatibility of a hypothesis with the evidence.

On a "thick" interpretation, of course, these two different measures say different things about the hypothesis, because they evaluate "compatibility" in different ways. But this thick perspective only matters insofar as we're concerned with the differences between different notions of "compatibility with the evidence." We're not always concerned with those differences, however, and so in at least some cases we can treat these different scores as (attempting to) measure the same thing—that is, compatibility with the evidence—in different ways. On this "thin" interpretation, the difference between quantities like confidence levels and posteriors probabilities disappears. From this perspective, a 5–95% confidence interval and a 5–95% credibility interval have *the same* compatibility with the evidence even though these scores are generated by different methods and yield different thick interpretations.

A comparison with measures of confirmation may be helpful here. So, for example, the confirmation measure $\log[p(h \mid e) / p(h)]$, originally suggested by Keynes (1921) and normally called r, operates on the same logarithmic scale as the measure $\log[p(e \mid h) / p(e \mid \neg h)]$ suggested by Good (1984) and normally called l. As a consequence, these two measures share some of their mathematical properties—namely, the ones constitutive of being logarithms. So, for instance, there's a sense in which a score of 1 means the same thing regardless of which of the two logarithmic confirmation measures one is using. (The same isn't true of all confirmation measures, of course, a score of 1 on the difference measure $p(h \mid e) - p(h)$ means something entirely different.) Nevertheless, the two measures are importantly different in that they represent different and arguably incompatible views about what properties a measure of confirmation ought to have (Crupi 2020: §3). My claim is that (some) means of measuring the compatibility between data and hypothesis are akin to the different logarithmic measures of confirmation: though philosophically distinct, they share important mathematical properties that allow us to group them together for some purposes.

Of course, there's a substantial step from grouping together confidence levels and posterior probabilities (for some purpose) and saying that the IPCC's talk of likelihoods refers to probabilistic scores in a thin sense. This is where the examples from §5 are crucial: as we saw, the IPCC doesn't treat confidence levels and posterior probabilities in a simple disjunctive way—they not only compare them but sometimes pool together or combine them. In the analogy, it's as though we're analyzing the practice of someone who treats various logarithmic measures of confirmation as interchangeable and sometimes even averages l and r together to create a score that is logarithmic but that doesn't share all of the unique properties of either measure. In this imagined scenario, it makes sense to say that the person in question isn't referring to r or l when talking about the degree of confirmation, but instead to a general or thin sense of logarithmic measures of confirmation. I'm claiming that we should interpret the IPCC in just this way.

The cost of adopting this kind of thin interpretation is that there are many conclusions that we can draw when employing a thick interpretation that we can't draw when employing a thin one. So, for instance, the .9 confidence level found in a 5–95% confidence interval tells us something about how frequently we're likely to make certain sorts of errors. The thin .9 compatibility with the evidence doesn't license that sort of conclusion. Similarly, the .9 posterior probability found in the 5–95% credibility interval can be straightforwardly plugged into a decision matrix. The same is not generally true of a thin .9 compatibility with the evidence. Of course, these thicker concepts might be viewed as special cases of the thin concept and so there will some cases in which it's legitimate to use the thin probabilities in this way, but the legitimacy of these moves depends

wholly on the details of the specific case. From the thin perspective, we're simply talking about scores that have certain mathematical properties—nothing about frequencies or expected utilities follows generally.

That's not to say that these scores are useless. Recall that the statistical methods used in climate science often rely on idealizations and risky assumptions (hence the IPCC's practice of modifying likelihood claims). This has two important implications for the present discussion. First, it means that we're often not justified in adopting a thick interpretation; even if the method in principle yields long-run frequencies or the subjective credences that a rational agent would adopt, the actual method should be understood as (at best) providing estimates of those quantities. That is, borrowing a point from Thompson and Smith (2019), even when purely Bayesian (/ classical) methods are employed in a study, it's a substantive step to practical conclusions about real error rates or the subjective credence that one should assign. Second, the idealized character of the statistical methods employed means that there's often reasonable disagreement between practitioners as to the best method—as to which idealizations or assumptions will generate the most reliable and accurate results. Sometimes, as in the examples discussed in the Section 5, different author groups even "disagree" as to whether to use Bayesian or classical approaches, but (as we've seen) the differences between approaches run much deeper than just this contrast.

Adopting language that communicates the thin notion of compatibility with the evidence allows the IPCC to present the findings of various studies without committing to those studies accurately representing the "true" probabilities or getting bogged down in trying to adjudicate which of the different methodologies employed in different studies is to be preferred. The true value of the thin notion is thus that adopting it allows the IPCC to communicate the results of the science while retaining a relatively pluralistic view towards the different methods that different climate scientists deem appropriate for their given projects. In other words, when trying to summarize a literature that uses a wide variety of statistical methods—some of which are Bayesian, some classical, and all of which are idealized—it's helpful to adopt general concepts that can be used to summarize and compare all of the different results in as simple a way as possible. From this perspective, the likelihood judgments of the IPCC are neither frequencies nor credences but simply the compatibility scores delivered by the method that the authors deemed best in the circumstances.

Still, one might push back here by stressing that confidence levels and posterior probabilities don't license the same inferences and so cannot be grouped together in this way. I think this objection puts things too starkly, however. For one thing, it's often the case that we can draw qualitative conclusions from probabilistic scores without knowing whether they're confidence levels or posterior probabilities. Again, it's helpful to compare measures of confirmation. Though *r*

and *l* differ in ways that philosophers deem important, they share many qualitative implications. If you tell me that the degree of confirmation of *h* by *e* according to *some* logarithmic scale is close to 0, for example, I don't need to know which scale you're using to draw general qualitative conclusions. The same is true of probabilistic measures of compatibility with the evidence; that a particular hypothesis scores a .95 can provide important qualitative information about the relationship between the hypothesis and the evidence even if we don't know whether the .95 score is a posterior probability, a confidence level, or neither.

Moreover, the imagined objection—that the thin probabilities offered by the IPCC are useless in comparison to posterior probabilities or confidence levels—relies on an idealized picture of how statistics works. To be sure, the ideal output of a statistical study would be an expert function: a posterior probability distribution that one could reliably treat as giving the true chances. But, as stressed above, this ideal output is not achievable: there's simply too much uncertainty to think that the probabilities found in climate science are anything more than imperfect evidence with respect to the uncertainty that we should have.[14] Regardless of whether the idealized methods in question generate posterior probabilities, confidence intervals, or something else, it would be a mistake to think that one can simply "read off" the true frequencies or the warranted subjective probabilities from these studies (compare Sprenger 2019; Thompson & Smith 2019)—certainly, one shouldn't plug the results into a decision matrix except with extreme care. Insofar as the thin probabilities delivered by the IPCC are harder to interpret and use than posterior probabilities or confidence intervals are thought to be, that has less to do with the use of thin probabilities per se and more to do with the idealized and imperfect nature of the analyses that they're based on.

At least in their present form, the value of the probabilistic judgments presented by the IPCC is that they communicate how well our present evidence supports a given hypothesis. While we philosophers often think about evidential support in terms of perfectly rational Bayesian agents, things are much messier in the context of climate science: to compare a hypothesis with the evidence, climate scientists must make a large number of substantive assumptions. The result is that any score intended to measure compatibility of evidence and hypothesis must be understood as being generated by a particular method. The thin notion of "compatibility with the evidence" that I've outlined here is well-suited for precisely this purpose: it allows the IPCC to make comparisons between how well the evidence supports different hypotheses—or between how well supported a given hypothesis is according to different studies—without either assuming

14. See Katzav, Thompson, Risbey, Stainforth, Bradley, and Frisch (2021) for a recent, but to my mind overly pessimistic, argument for this conclusion.

that the methods employed in these different contexts are the same or imparting thick implications on the resulting scores that are not be warranted due the idealizations involved.

We're now nearly in a position to state my proposed interpretation of the IPCC. The final piece of the puzzle is the point made in the last section relating to the frequent modification of likelihood scores. As we saw, these modifications were motivated by unaccounted-for sources of uncertainty; in particular, I argued the "downgrading" option could be motivated by sufficiently high confidence that accounting for the additional sources of uncertainty would lower the score given to the hypothesis. In these cases, the best method available for estimating the compatibility of the evidence with the hypothesis is not the mechanical statistical method alone, but rather the mechanical method plus the additional correction embodied in the downgrade. When the IPCC is not sufficiently confident to warrant correcting the results in one way or another, however, the best method (at least the one judged best by the authors) is the purely mechanical one. So in either case, the likelihood reports rely on what the authors judge to be the best methods available.

What the likelihood language expresses, then, is simply a coarse-grained version of the thin compatibility on the evidence score that is delivered by what the IPCC judges to be the best method available in that context. So, for example, when the IPCC says that "ECS is positive, *likely* in the range 1.5°C to 4.5°C with *high confidence*" that means that (a) the best method or methods available assign a compatibility score between .66 and 1 and (b) the IPCC authors have high confidence that these methods are reliable, where the relevant sense of reliability reflects how worrying the IPCC considers the remaining potential sources of error or uncertainty.[15]

It's worth reiterating the caveat that my claim here is primarily intended to be descriptive. My arguments so far have motivated the view that there's no "thicker" way of interpreting the IPCC's likelihood judgments that is consistent with the actual statistics that goes into determining the character of those judgments, and in this section I've sketched a thinner interpretation that is consistent with this practice, but at the cost of much of the practical informativeness of thicker interpretations. It's a separate question as to whether the IPCC *should* adopt practices that allow for a thicker interpretation (see the next section). It's also a separate question to whether we can and should proactively re-interpret

15. There's more to be said about the second clause here given that the IPCC's likelihood categories are overlapping: "likely" expresses a probability range of .66 to 1 while "virtually certain" expresses a probability range of .99 to 1. It is not obvious to me how to interpret the IPCC's confidence judgments given this fact—e.g., whether the IPCC's expression of high confidence in "likely" should be read straightforwardly or instead be taken as expressing high confidence in "likely but not virtually certain."

the IPCC's practice—for example, whether arguments can be given to justify treating the IPCC's likelihoods as subjective probabilities in some or all contexts. I've argued that they aren't subjective probabilities, but it's an open question whether they're "close enough" for this or that purpose.

8. Tentative Normative Lessons

In this paper, I've argued that the likelihood language found in IPCC reports cannot be straightforwardly interpreted either in terms of credences or in terms of (long-run) frequencies. The problem is the IPCC's methodological pluralism: so long as the IPCC treats different statistical methods as equally worthwhile, it won't be possible to give what I've termed a "thick" interpretation of the likelihood language. All that we can say is that likelihoods are normalized ways of measuring the compatibility of some hypothesis with the evidence and that they are based on the method (or methods) that the IPCC judges to be best. This "thin" interpretation renders the practice consistent but at the cost that it's less clear how readers should use the relevant judgments.

The arguments in this paper should be understood as aiming towards a bigger and more important normative question: how *should* the IPCC (and other scientists) use probabilistic language to represent its uncertainty? I won't endeavor to answer the normative question here. Nevertheless, a couple of tentative—largely conditional—conclusions can be extracted from the foregoing.

First, while *in principle* it would be desirable for the IPCC to express uncertainty in a way that allowed for a "thicker" interpretation—that is, to consistently use posterior probabilities based on a particular choice of "objective" priors—such a shift would require the IPCC to dramatically alter its current pluralistic approach. As noted above, climate science is itself pluralistic in the sense that different climate scientists use different methods. Insofar as the IPCC's goal is to merely summarize the pluralistic research found in climate science, therefore, it has no choice but to adopt methodological pluralism in *some form*. That's not to say that the IPCC has to adopt the approach that it has. One alternative would be for the IPCC to take a more proactive role in interpreting the results of various studies. So, for example rather than simply presenting the different studies or pooling their results together, the IPCC could present what it takes to be the proper credential response to the evidence available. Perhaps this is what the IPCC should do. My point here is simply that we cannot suggest that the IPCC alter its approach to probability without addressing much larger questions about the proper role of the organization: should it merely summarize results or should it interpret and analyze them in a more robust sense?

Second, and similarly, the IPCC's practice of using both confidence and likelihoods to communicate uncertainty, while potentially more confusing than a single measure, is in a large part determined by the goal of communicating the results of imperfect statistical methods. As stressed above, actual applications of statistics rely heavily on idealizations, meaning that there remains uncertainty about the accuracy of the results that they generate. As such, the only accurate way of presenting the results of such studies involves qualification like that found in the IPCC reports (though of course the exact form might differ). One could argue that these applications are so flawed that we should prefer either (a) to give up on quantitative evaluations of the evidence entirely or (b) simply rely on expert judgment for the quantitative evaluation of the evidence. Again, my point here is simply to identify the cost: we can't motivate a move away from the two-tier approach without addressing these larger questions.

Third, and more concretely, the IPCC should be more consistent in how it expresses uncertainty about the results of statistical methods. As noted above, the present system allows authors to express uncertainty in the actual application of statistical methods either by modifying the likelihood judgments or by qualifying the likelihood judgment with a confidence judgments. I've sketched a principled way of distinguishing between these two options, but the IPCC would be well-served by further clarification regarding the difference between these two options and when it is appropriate for authors to adopt one rather than the other.

Finally, let me note a general meta-normative point. One of the lessons of the paper is that there's often little to be gained in satisfying the theoretical desiderata of philosophers. The IPCC's practice can be consistent and coherent without allowing for any "thick" interpretation. *Much* more important is that the IPCC's practice is valuable to its target audience, and whether the concerns raised in this paper are relevant to that question depends on how readers are using the IPCC's results. We have a problem, for instance, if decision makers are trying to plug the IPCC's likelihoods into an actual decision matrix. I therefore suspect that the most useful information about how to improve the IPCC's presentation of uncertainty will come from empirical investigations into what information decision makers want and need rather than from high-level philosophical arguments.

Acknowledgments

My thanks to the audience at EPSA2021 as well as to four anonymous reviewers at *Ergo* for their comments on earlier versions of this paper.

References

Aldrin, Magne, Marit Holden, Peter Guttorp, Ragnhild Bieltvedt Skeie, Gunnar Myhre, and Terje Koren Berntsen (2012). Bayesian Estimation of Climate Sensitivity Based on a Simple Climate Model Fitted to Observations of Hemispheric Temperatures and Global Ocean Heat Content. *Environmetrics, 23*(3), 253–71.

Annan, James D., and Julia C. Hargreaves (2017). On the Meaning of Independence in Climate Science. *Earth Systems Dynamics, 8*, 211–24.

Bradley, Richard, Casey Helgeson, and Brian Hill (2017). Climate Change Assessments: Confidence, Probability and Decision. *Philosophy of Science, 84*(3), 500–22.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.). Chapman & Hall/CRC.

Crimmins, Allison (2020). Improving the Use of Calibrated Language in U.S. Climate Assessments. *Earth's Future, 8*(11), 1–15.

Crupi, Vincenzo (2020). Confirmation. In Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/spr2021/entries/confirmation

Dethier, Corey (2022a). Calibrating Statistical Tools: Improving the Measure of Humanity's Influence on the Climate. *Studies in the History and Philosophy of Science, 94*, 158–66.

Dethier, Corey (2022b). When is an Ensemble Like a Sample? 'Model-Based' Inferences in Climate Modeling. *Synthese, 200*(52), 1–20.

Gillett, Nathan P., Vivek K. Arora, Damon Matthews, and Myles R. Allen (2013). Constraining the Ratio of Global Warming to Cumulative CO_2 Emissions Using CMIP5 Simulations. *Journal of Climate, 26*(18), 6844–58.

Gillett, Nathan P., Megan Kirchmeier-Young, Aurélien Ribes, Hideo Shiogama, Gabriele C. Hegerl, Reto Knutti, . . . , Tilo Ziehn (2021). Constraining Human Contributions to Observed Warming since the Pre-Industrial Period. *Nature Climate Change, 11*, 207–12.

Good, Irving J. (1984). The Best Explicatum for Weight of Evidence. *Journal of Statistical Computation and Simulation, 19*(4), 294–99.

Hasselmann, Klaus (1998). Conventional and Bayesian Approach to Climate-change Detection and Attribution. *Quarterly Journal of the Royal Meteorological Society, 124*(552), 2541–65.

Haustein, Karsten, Myles R. Allen, Peter M. Forster, F. E. L. Otto, D. M. Mitchell, H. D. Matthews, and D. J. Frame (2017). A real-time Global Warming Index. *Scientific Reports*, *7*, 1–6.

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

Herrando-Pérez, Salvador, Corey J. A. Bradshaw, Stephan Lewandowsky, and David R. Vieites (2019). Statistical Language Backs Conservatism in Climate-Change Assessments. *BioSciences, 69*(3), 209–19.

IPCC (2013). *Climate Change 2013: The Physical Science Basis.* Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Thomas F. Stocker, Dahe Qin, Gian-Kasper Plattner, Melinda M. B. Tignor, Simon K. Allen, Judith Boschung, . . . , Pauline M. Midgley (Eds.). Cambridge University Press.

IPCC (in press). *Climate Change 2021: The Physical Science Basis.* Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Valérie Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, . . . , B. Zhou (Eds.). Cambridge University Press.

Janzwood, Scott (2020). Confident, Likely, or Both? The Implementation of the Uncertainty Language Framework in IPCC Special Reports. *Climatic Change, 162*, 1655–75.

Jebeile, Julie (2020). Values and Objectivity in the Intergovernmental Panel on Climate Change. *Social Epistemology*, *34*(5), 453–68.

Jones, Gareth S., Peter A. Stott, and Nikolaos Christidis (2013). Attribution of Observed Historical Near-Surface Temperature Variations to Anthropogenic and Natural Causes Using $CMIP_5$ Simulations. *Journal of Geophysical Research: Atmospheres, 118*(10), 4001–24.

Katzav, Joel, Erica L. Thompson, James Risbey, David A. Stainforth, Seamus Bradley, and Mathias Frisch (2021). On the Appropriate and Inappropriate Uses of Probability Distributions in Climate Projections, and Some Alternatives. *Climatic Change, 169*(15), 1–20.

Katzfuss, Matthias, Dorit Hammerling, and Richard L. Smith (2017). A Bayesian Hierarchical Model for Climate Change Detection and Attribution. *Geophysical Research Letters, 44*(11), 5720–28.

Keynes, John Maynard (1921). *A Treatise on Probability*. Macmillan.

Mach, Katharine J., Michael D. Mastrandrea, Patrick T. Freeman, and Christopher B. Field (2017). Unleashing Expert Judgment in Assessment. *Global Environmental Change, 44*, 1–14.

Mastrandrea, Michael D., Christopher B. Field, Thomas F. Stocker, Ottmar Edenhofer, Kristie L. Ebi, David J. Frame, . . . , Francis W. Zwiers. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. Retrieved from https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf

Olson, Roman, Ryan Sriver, Marlos Goes, Nathan M. Urban, H. Damon Matthews, Murali Haran, and Klaus Keller (2012). A Climate Sensitivity Estimate Using Bayesian Fusion of Instrumental Observations and an Earth System Model. *Journal of Geophysical Research: Atmospheres, 117*(D4), 1–11.

Parker, Wendy S. (2010). Predicting Weather and Climate: Uncertainty, Ensembles and Probability. *Studies in the History and Philosophy of Modern Physics, 41*, 263–72.

Parker, Wendy S., and James S. Risbey (2015). False Precision, Surprise and Improved Uncertainty Assessment. *Philosophical Transactions of the Royal Society Part A, 373*(3055), 20140453.

Ribes, Aurélien, Saïd Qasmi, and Nathan P. Gillett (2021). Making Climate Projections Conditional on Historical Observations. *Science Advances*, *7*(4), 1–9.

Rougier, Jonathan, and Michel Crucifix (2018). Uncertainty in Climate Science and Climate Policy. In Elisabeth A. Lloyd and Eric Winsberg (Eds.), *Climate Modeling: Philosophical and Conceptual Issues* (361–80). Palgrave Macmillan.

Roussos, Joe, Richard Bradley, and Roman Frigg (2021). Making Confident Decisions with Model Ensembles. *Philosophy of Science, 88*(3), 439–60.

Sprenger, Jan (2019). Conditional Degrees of Belief and Bayesian Inference. *Philosophy of Science, 87*(2), 319–35.

Stainforth, David A., Myles R. Allen, Edward R. Tredger, and Leonard A. Smith (2007). Confidence, Uncertainty and Decision-Support Relevance in Climate Predictions. *Philosophical Transactions of the Royal Society Series A, 365*(1857), 2145–61.

Thompson, Erica L., and Leonard A. Smith (2019). Escape from Model-Land. *Economics, 13*(1), 1–15.

Winsberg, Eric (2018a). *Philosophy and Climate Science.* Cambridge University Press.

Winsberg, Eric (2018b). What does Robustness Teach us in Climate Science: A Re-Appraisal. *Synthese, 198*(S21), 5099–122.