# The IPCC Uncertainty Framework: What *Some* Decision Makers Want (and Why They Shouldn't)

**Abstract**

In "Combining probability with qualitative degree-of-certainty metrics in assessment," Helgeson et al. present a mathematical model of the confidence-likelihood relationship in the IPCC uncertainty framework. Their goal is to resolve ambiguities in the framework and clarify the roles of "confidence" and "likelihood" in decision-making. In this paper, I provide a conceptual evaluation of their proposal. I argue that the IPCC cannot implement the model coherently and that adopting it could result in unclear and potentially misleading communication of uncertainty.

## 1 Introduction

Studies of climate change are afflicted by deep uncertainty, the communication of which is made fraughter still by the studies' immediate policy implications. The world of policy-making has its demands: uncertain information must be communicated in a simple, consistent, and relevant manner. To balance these demands with the need for scientific rigour, it is essential to convey uncertainty in a way that is both decision-relevant and as faithful to the science as possible.

To address this, the IPCC uncertainty guide (IPCC 2010) provides both a confidence and a likelihood metric for experts to characterize uncertainty in their findings. The confidence metric is defined on a qualitative scale with five levels ("very low", "low", "medium", "high" and "very high"). The appropriate level of confidence depends on the evaluation of two independent dimensions: evidence and agreement (Figure 1). The

likelihood metric, on the other hand, is defined on a quantitative scale with seven levels, where each level corresponds to a probability interval (Figure 2). According to the guide, this metric is meant 'to express a probabilistic estimate of the occurrence of a single event or of an outcome' [...and it] may be based on statistical or modelling analyses, elicitation of expert views, or other quantitative analyses' (ibid., 3).

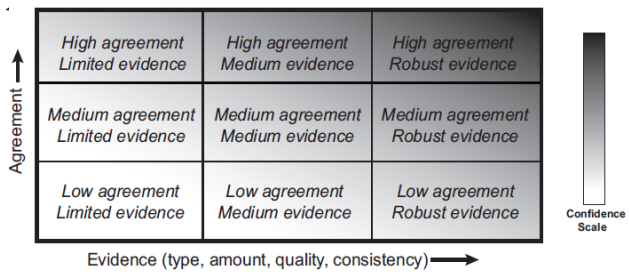


Figure 1: The confidence metric

**Table 1. Likelihood Scale**

| Term* | Likelihood of the Outcome |
|---|---|
| *Virtually certain* | 99-100% probability |
| *Very likely* | 90-100% probability |
| *Likely* | 66-100% probability |
| *About as likely as not* | 33 to 66% probability |
| *Unlikely* | 0-33% probability |
| *Very unlikely* | 0-10% probability |
| *Exceptionally unlikely* | 0-1% probability |

Figure 2: The likelihood metric

While many note that the relationship between these two metrics remains unclear,[1] Richard Bradley, Casey Helgeson, and Brian Hill (BHH) propose an account of "confidence" and "likelihood" that they believe could simultaneously resolve existing ambiguities and clarify the roles of these metrics in decision-making (Bradley et al. 2017; Helgeson et al. 2018). However, in this paper, I argue that the IPCC would face significant challenges in implementing BHH's proposal in a conceptually coherent way. Hence I contend that, despite their well-intentioned efforts, their proposal does not provide a viable solution for clarifying the relationship between "confidence" and "likelihood."

## 2 What Some Decision Makers Want ...

Decision makers are (relatively) comfortable with making decisions when faced with precise probabilities. In this case they can, for instance, rely on the orthodox normative decision theory, expected utility theory. When it comes to imprecise probabilities, they are slightly less comfortable since there is no longer an orthodox normative decision theory on which they can rely. There are, nonetheless, a host of decision rules that have been offered that can help them in these cases too.

However, as BHH observe, the IPCC's uncertainty framework differs considerably from the models typically developed by decision theorists (Bradley et al. 2017, 503).

[1]See, for example, Aven (2018), Janzhood (2020), and Harris (2021, 20-32).

Rather than offering precise or even imprecise probabilities alone, the IPCC supplements these with qualitative confidence judgments. This creates a challenge, as BHH rightly note: how should these confidence judgments be integrated into decision-making processes? To address this issue, BHH draw on a model proposed by Hill (2013, 2017), which accommodates imprecise probabilities qualified by ordinal confidence judgments. Building on Hill's approach, BHH propose what they describe as a "simple mathematical model of the confidence-likelihood relationship," aiming to resolve ambiguities in the IPCC's framework while preserving the qualitative aspects of its confidence scale (Helgeson et al. 2018, 518).

According to BHH's proposal the assignment of an imprecise probability interval to a hypothesis should always be determined by a well-defined set of probability distribution functions (pdfs); so for instance, 'assigning probability 0–.1 to outcome x means that within the set of pdfs collectively representing authors' uncertainty, the smallest probability given to outcome x by any pdf is 0 and the largest probability given to x by any pdf is .1' (ibid., 520). However, there is a caveat. Each level of confidence should be associated 'with its own set of pdfs, where higher confidence sets encompass lower-confidence sets' (ibid., 520).

They give the following toy example to illustrate how this is supposed to work. Suppose the author team starts with a set of possible pdfs concerning the value of the equilibrium climate sensitivity (ECS).[2] The author team must then sort those pdfs into what BHH call a confidence partition, which in this example is assumed to have four elements $\pi = \{M_0, M_1, M_2, M_3\}$. The pdfs in $M_0$ are supposed to be those considered to be most plausible according to the author team and the pdfs in $M_1$ 'collectively represent a second tier of plausibility. The element $M_2$ is another step down from there, and $M_3$ is the bottom of the barrel: all of the pdfs more or less ruled out by the body of research that the experts evaluated' (Bradley et al. 2017, 515). This partition of pdfs can then be used to generate a nested family of subsets of pdfs $\{L_0, L_1, L_2, L_3\}$ where $L_i$ is the union of $M_0$ through $M_i$ and each $L_i$ is associated with a level of confidence. In this example, there are two pdfs in $M_0$, hence two pdfs in $L_0$; and three pdfs in $M_1$, hence five pdfs in $L_1$ —as illustrated in Figure 3, provided by BHH (ibid., 516).

[2]The latest IPCC assessment report defines ECS 'as the equilibrium (steady state) change in the surface temperature following a doubling of the atmospheric carbon dioxide (CO2) concentration from pre-industrial conditions' (IPCC 2021, 46).

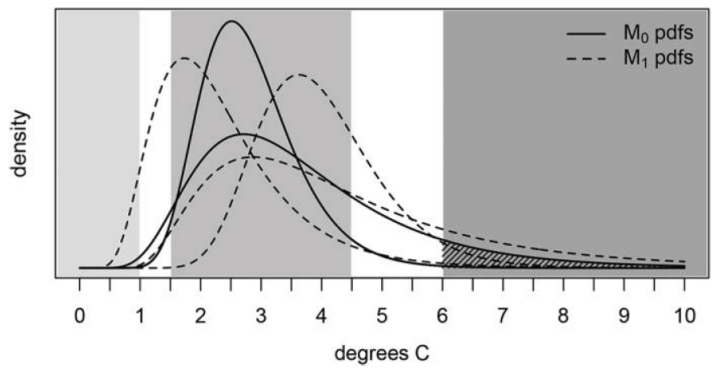


Figure 3: Illustration of a confidence partition in this toy example (the pdfs in $M_2$ and $M_3$ are not represented in the figure).

Assuming that $L_0$ corresponds to "medium confidence" and $L_1$ to "high confidence", the author team is now able to generate various probability statements at those two levels of confidence. For instance, if the experts want to determine the probability interval to assign to the hypothesis that ECS is greater than 6°C at say "medium confidence", all they have to do is check what probabilities the various pdfs in $L_0$ assign to that hypothesis and then report the probability interval bounded by the smallest and largest of those probabilities (the hatched area in Figure 3). In this case, $L_0$ contains two pdfs, where one assigns (nearly) zero probability to the hypothesis that ECS is greater than 6°C and the other assigns just under 0.1 probability. Hence, in this case, the author team should assign a probability range of roughly $[0 - 0.1]$ which, according to the IPCC terminology, corresponds to "very unlikely" and report that "ECS is *very unlikely* greater than 6°C (*medium confidence*)". The author team can then repeat this mechanical procedure to determine the likelihood level that should be assigned to various hypotheses concerning the possible values of ECS at any confidence level.

BHH argue that their proposal aligns with the current IPCC uncertainty framework and provides a structured and transparent method for clarifying the interaction between "confidence" and "likelihood." They further suggest that their proposal demonstrates how it 'can make sense, conceptually, to answer the same question at multiple confidence levels,' thereby giving "confidence" a more defined role in decision-making. As mentioned in the introduction, their approach builds on a confidence-based decision model proposed by Hill (2013, 2017), which is designed to manage imprecise probabilities qualified by ordinal confidence judgments.

However, in the next section, I will argue that, despite the good intentions behind

BHH's proposal, it faces significant conceptual challenges that hinder its practical implementation by the IPCC.

# 3 And Why They Shouldn't

The main challenge that BHH acknowledge with their proposal relates to the calibration of confidence levels across different author teams. For this proposal to be feasible in practice, the IPCC would need to develop 'a proper calibration scale [that] would enable clear and unambiguous formulation and communication of confidence judgments across authors and actors' (Bradley et al. 2017, 518).

While I agree that developing such a calibration scale is necessary if this proposal is to be pursued, in this section I will argue that the task of developing such a scale, under the most plausible interpretation of likelihood in BHH's proposal, is significantly more challenging than they suggest (section 3.1). I will then explore how BHH's implicit recommendations for implementing their proposal in practice would lead to difficulties in maintaining a coherent interpretation of likelihood (section 3.2). These considerations raise important questions about the overall feasibility of the proposal.

## 3.1 Two Conceptually Coherent but Inapplicable Interpretations of Likelihood and Confidence

Surprisingly, BHH do not clarify what interpretation of probability they have in mind in their proposal, leaving the kind of uncertainty that the likelihood and confidence metrics are intended to represent somewhat ambiguous. I will argue that there are, at most, two conceptually coherent interpretations of likelihood and confidence in BHH's proposal, but both present distinct challenges for applicability to the IPCC.

Under the first interpretation, "likelihood" is understood as the IPCC authors' imprecise estimate of the objective probability (i.e., the chance) of an event based on the available evidence (at a given confidence level). Under this view, the set of pdfs with which BHH's procedure would begin represents possible objective pdfs for the variable in question, with the IPCC authors uncertain about which one is correct. Under this interpretation, "confidence" could be interpreted straightforwardly in Bayesian terms, as expressing the credences that the IPCC authors assign to the hypothesis that the correct objective pdf is within a certain set of possible pdfs.[3] Based on this, one could develop a

[3]In Bayesian statistics, assigning credences to a set of possible objective pdfs is a common approach. For example, Bayesian parameter estimation—a widely used technique for estimating pdfs of random variables with unknown parameters—begins with the form of the pdf known, but the parameter values unknown. A prior distribution is placed over the parameter space, and it is updated using Bayes' theorem

confidence scale that aligns with different ranges of credence values.

Although this interpretation is internally consistent, it does not align well with how uncertainty is typically handled in climate science. Specifically, it misrepresents how one should interpret the many pdfs for variables like ECS and other key climate metrics found in the literature.[4] To clarify why this is, it's useful to briefly overview how the many pdfs for ECS in the literature are produced.[5]

In AR6, the assessment of ECS represents a significant shift by incorporating a broader range of evidence compared to earlier reports. While previous assessments primarily relied on climate models, AR6 integrates multiple lines of evidence, including process-based understanding, the instrumental temperature record, paleoclimate data, and emergent constraints (IPCC 2021, sec. 7.5). The assessment of ECS across these lines of evidence relies on a multi-stage process. Key steps include estimating Effective Radiative Forcing (ERF), evaluating top-of-atmosphere (TOA) energy imbalances, and assessing the climate feedback parameter ($\alpha$). Each of these stages involves various approximations, such as assuming a constant $\alpha$ in some cases, though AR6 acknowledges that $\alpha$ can vary with spatial patterns of warming, adding complexity to ECS estimates (IPCC 2021, sec. 7.4.3). Each line of evidence provides specific information about the climate's response to rising CO2 levels—for example, paleoclimate data show past climate conditions, while the instrumental record captures recent temperature trends. However, each also carries its own uncertainties: paleoclimate data are limited by uncertainties in radiative forcing and reconstructions, while the instrumental record, though more direct, is constrained by its relatively short timescale and uncertainties related to aerosol forcing.

The various pdfs for ECS reviewed by the IPCC are derived using various methods, each reflecting different lines of evidence and sources of uncertainty. Monte Carlo simulations are commonly used to handle uncertainties in key climate parameters. For example, in studies estimating ECS based on historical changes in Earth's energy

as new data becomes available.

[4]See Winsberg (2018) for an argument suggesting that this interpretation of "likelihood" stems from a misunderstanding of the role objective chances play in climate science, where their influence is often minimal or absent (ibid., 104). Whether or not one finds Winsberg's argument persuasive, as discussed below, the spread in pdfs for ECS and other key climate variables is largely driven by uncertainties in how different lines of evidence constrain physical relationships, rather than by natural variability or objective chances.

[5]As a reviewer noted, ECS is somewhat atypical compared to other climate variables. Unlike variables such as monsoon rainfall or hurricane-related precipitation—arguably more relevant for decision-making—ECS has a long history of probabilistic assessments and is more easily constrained by physical limits and data. While ECS is more amenable to analysis, I aim to show that BHH's proposal is unfeasible even for well-understood variables like ECS, highlighting its conceptual issues for managing uncertainty, even in the best-case scenarios.

balance, Monte Carlo methods sample from probability distributions of uncertain factors, such as aerosol forcing or temperature response. By running numerous simulations with these sampled inputs, researchers generate a wide range of potential ECS outcomes. This process results in a pdf for ECS, incorporating uncertainties from both the data and models. The input probability distributions used in Monte Carlo methods are often explicitly recognized as subjective (see, e.g., Snyder 2019) and as Wagner and Weitzman (2018) emphasize, these input distributions frequently rely on assumptions that can significantly influence the final pdf for ECS, making the choice of assumptions a critical aspect of the analysis. Bayesian methods are also often employed to derive a posterior pdf for ECS by integrating data from different lines of evidence, such as paleoclimate records from various time periods. Other studies may combine data from multiple sources, such as historical warming records and paleoclimate data (see e.g., Annan and Hargreaves, 2006). However, it is important to note that most studies rely on only a subset of the available evidence, which can lead to overly broad or less precise ECS estimates.[6]

The pdfs derived from these methods are epistemic in nature, representing uncertainty about ECS based on specific lines of evidence rather than objective probabilities tied to physical randomness. In other words, they capture uncertainties stemming from incomplete knowledge and limitations in the data and methods used, rather than offering objective probabilities of different outcomes. As such, interpreting the likelihood metric as reflecting objective probabilities is not an appropriate fit for the IPCC framework since it would be at odds with how climate science typically addresses uncertainty. Therefore, while this interpretation of likelihood and confidence is conceptually coherent, it is not applicable to the IPCC.

Under a second more plausible interpretation, likelihood is meant to express the range of credences that the IPCC authors assign to a hypothesis given the available evidence (at a given confidence level). Hence, in line with the literature on imprecise probabilities (see e.g. Bradley (2019)), the set of pdfs with which BHH's recommended procedure would begin is the family of posterior credence functions that the IPCC authors assign to a hypothesis given the available evidence.

What about confidence? Here is where things get complicated. Under this interpretation, there are a variety of reasons why there might be multiple pdfs. It could be due to, for instance, uncertainty in the priors or in the likelihood functions for each line of evidence. Indeed, this is often the case in scientific applications of the Bayesian approach. Consider, for instance, Sherwood et al. (2020), one of the first ambitious effort ever made to combine various different lines of evidence with a Bayesian approach to

[6]This is why a key role of the IPCC is to somehow combine the various pdfs derived from different lines of evidence and methods to arrive at a comprehensive estimate for ECS as I will discuss further in section 3.2.

arrive at a unique posterior pdf for the ECS. Sherwood et al. (2020, 67) acknowledge that their Bayesian calculation has several limitations since it 'is based on one particular prior, assumes independence between lines of evidence, and makes no allowances for "unknown unknowns"'. And while they, themselves, perform a number of sensitivity tests to explore these limitations (e.g. they consider how a different prior distribution affects the posterior pdf), they admit that substantially more work will be needed to further explore them (e.g. 'In future work it could be possible to develop a range of plausible alternative likelihoods for each line of evidence and apply these as sensitivity tests in the synthesis' (ibid., 72)). However, there is no mention of the feasibility nor of the desire to sort all the pdfs consistent with the available evidence in a confidence partition of any kind (most of which have not yet been derived to begin with!). Instead, the general hope seems simply to be that were one to fully explore these limitations, the resulting posterior pdf would only be modestly affected.

The issue with this interpretation, therefore, appears to be twofold. First, there is a practical challenge: the current state of practice makes it unrealistic to assume that IPCC authors could begin with such a set of posterior credence functions. Very few attempts have been made to derive a posterior credence function for any given climate variable that comprehensively synthesizes various lines of evidence. As mentioned above, Sherwood et al. (2020), published only a few years ago, provides one of the few and earliest attempts for ECS, a variable that—unlike others such as monsoon rainfall or hurricane-related precipitation—has a long history of probabilistic assessments and is better constrained by physical limits and data. The atypicality of ECS compared to other climate variables that lack such probabilistic assessments, combined with the fact that even for ECS there have been very few attempts, makes it unreasonable to assume that IPCC authors are currently in a position to start with a comprehensive set of posterior credence functions for most (if not all) variables, or that they will be in such a position anytime soon. If the IPCC authors do not have a set of posterior credence functions for a variable given the available evidence, as I contend is the case for most variables, BHH's procedure cannot begin in the first place.

Second, there is a principled challenge. As mentioned above, there are multiple reasons why posterior credence functions might exist, such as uncertainty in priors or likelihood functions. To the best of my knowledge, nowhere in the literature on imprecise probabilities nor in any scientific application of Bayesian statistics has there ever been any attempt to distinguish credence functions that are consistent with the available evidence in any principled way. This, I contend, is not merely a gap in the literature but a reflection of the task's inherent complexity. Any such ranking procedure would require well-defined criteria for evaluating and comparing pdfs derived from different assumptions, such as varying priors, differing likelihood formulations, or distinct treatments of independence between lines of evidence. However, it is far from clear what

these criteria should be or how they could be justified. Without explicit guidelines or an established theoretical foundation, developing a principled ranking system poses an unresolved and profoundly challenging problem, both conceptually and practically.

As noted, BHH argue that for their proposal to be feasible, the IPCC would need to develop 'a proper calibration scale [that] would enable clear and unambiguous formulation and communication of confidence judgments across authors and actors' (Bradley et al. 2017, 518). However, the practical and principled issues associated with the most plausible interpretation of likelihood and confidence make it unrealistic to assume that IPCC authors are currently—or will be anytime soon—in a position to construct a family of credence functions consistent with the available evidence (for various climate variables) and sort them into a confidence partition in a principled manner. This makes the prospect of developing such a calibration scale highly unlikely. Hence, although BHH's proposal might initially seem aimed at aligning IPCC practice with minimal changes to connect it to a compatible decision model from the literature, a closer examination reveals otherwise. Under the most plausible interpretation of likelihood, the proposal is significantly disconnected from current IPCC practices. No straightforward solution to bridge this disconnect appears feasible. As a result, the proposal is ultimately unworkable for the IPCC.

Before concluding this subsection, I should mention that although BHH do not clarify how one should interpret likelihood in their proposal, they do offer an interpretation of confidence, one that seems closely related to the current IPCC confidence scale (see Figure 1), in contrast to the two interpretations that I just offered:

> When used in conjunction with likelihood, we understand confidence to express something like Keynes' (1921/1973) "weight of evidence"[7] behind a probability statement, where the weight he refers to includes the quantity, quality and diversity of evidence underpinning a claim. (Helgeson et al. 2018, 522)

However, this interpretation of confidence is not consistent with either of the interpretations of likelihood discussed above. Consider, for instance, the second interpretation under which the set of pdfs with which BHH's recommended procedure would begin is the family of posterior credence functions that the IPCC authors assign to a hypothesis given the available evidence. This means that any subset of pdfs whatsoever is based on the available evidence, and hence the very same evidence. So if

[7] A critical analysis of this notion is beyond the scope of this article. However, it is worth mentioning that Runde (1990) distinguishes at least two distinct conceptions of the weight of evidence defended by Keynes (1921) (neither of which includes the quality and diversity of the evidence underpinning a probabilistic statement). See Harris (2021, 205-219) for a critical discussion of these conceptions.

we interpret confidence as expressing something like Keynes' "weight of evidence", then any subset of pdfs would have to be assigned the same confidence level. But this is incompatible with BHH's proposal.

BHH themselves do not attempt to justify this interpretation of confidence in their proposal.[8] I will now outline the only way I see to relate confidence to something like Keynes' "weight of evidence" in a way that makes likelihood an intelligible concept. As we will see, this approach is incompatible with BHH's proposal, and thus this interpretation of confidence is not justified.[9]

Under this interpretation of confidence, likelihood would represent the range of posterior credences that the IPCC authors assign to a hypothesis based on a *subset* of the available evidence. This would require gradually incorporating more evidence at each confidence level. At the lowest confidence level, a small subset of the evidence would be used to generate a set of posterior credence functions, determining the likelihood assigned to a hypothesis. At the next confidence level, additional evidence would be considered, and a new set of credence functions recalculated, which would adjust the likelihood assigned to the hypothesis at that higher level, continuing in this fashion.

It is difficult to see the epistemic value in requiring the IPCC authors to follow this process, and it seems highly impractical as well. However, setting aside the question of epistemic value and feasibility, my reconstruction of what would be required for assigning likelihood levels at various confidence levels is incompatible with BHH's own proposal. Rather than starting with a full set of pdfs to be sorted into confidence partitions, under this interpretation of "likelihood" and "confidence" the authors would need to recalculate new sets of pdfs as more evidence is introduced. As such, there's no reason to assume that the pdfs associated with a particular confidence level would include those from lower confidence levels, as BHH seem to recommend. Hence while BHH suggest a relationship between confidence and likelihood akin to Keynes' "weight of evidence," the practical requirements for this interpretation appear to conflict with their proposal and hence can't be a valid interpretation.

In the next section, I will take a step back and consider how BHH implicitly suggest their proposal should be implemented by the IPCC. Specifically, I will examine where BHH assume the pdfs that the IPCC authors are supposed to sort into confidence

[8]It is worth mentioning, however, that Hill (2019) himself argues that his normative account of belief and decision making (which BHH's proposal is based on) allows for 'a role for something akin to the Keynesian concept of "weight of evidence" in choice' (ibid., 235), unlike Bayesianism. However, the example Hill uses to support this is unconvincing, as the agent can increase confidence by lowering the precision of their credal judgments without any change in the underlying evidence (ibid., 235-37). This suggests that the conceptual foundations of Hill's account are confused, but an extensive defense of this claim is beyond the scope of this article.

[9]This is to demonstrate that I have considered all reasonable interpretations of confidence and likelihood in BHH's proposal.

partitions would originate. I will show that what BHH have in mind is conceptually flawed, further underscoring why the IPCC should approach this proposal with caution.

## 3.2 A Potentially Feasible but Conceptually Incoherent Approach to Likelihood and Confidence

Where do all these pdfs come from in the first place? As discussed in Section 2, BHH provide a simple example to illustrate how their proposal is intended to work. In this example, they assume that the author team starts with a well-defined set of possible pdfs concerning the value of ECS but they do not explain where those pdfs would come from in the first place. However, to justify why, "for concreteness," each pdf is assumed to be lognormally distributed, Bradley et al. (2017, 515) include the following footnote:

> Most studies aiming to constrain climate sensitivity with observations do indeed indicate a similar to lognormal probability distribution of climate sensitivity (Meehl et al. 2007, sec. 10.5.2.1).[10]

Indeed, in Meehl et al. 2007, sec. 10.5.2.1, one finds a summary of the evidence on ECS (Box 10.2) and a variety of pdfs concerning the value of ECS, obtained from different studies and lines of evidence, many of which suggest a lognormal distribution. This suggests that BHH might envision their recommended procedure beginning with the set of pdfs published in the studies reviewed by the author team. The IPCC authors would then be expected to sort these pdfs into a confidence partition according to some criterion. Under this interpretation, it becomes clearer why BHH might think that confidence in their proposal could be understood as relating to Keynes' concept of "weight of evidence." In their view, higher confidence sets would encompass more pdfs, and therefore, the probability intervals they produce would be based on a broader body of evidence. This connection could lead one to think that what I have argued thus far misses the mark.

This conclusion would be mistaken because, if this is indeed how BHH envision their proposal functioning, it is conceptually flawed. As discussed in Section 3.1, the pdfs reviewed by the IPCC to assess ECS are specific to individual lines of evidence. These pdfs are often derived from particular datasets or methods, encapsulating the uncertainties inherent to those specific approaches, and do not provide a comprehensive synthesis of all available evidence. Consequently, the various pdfs for ECS reviewed by

[10] As a reviewer rightly pointed out, Meehl et al., 2007 is from the AR4 assessment report, published 17 years ago, and is therefore not representative of how the IPCC currently estimates ECS. However, the fact that BHH cite this source provides insight into where they implicitly assume the pdfs would come from and how they envision their proposal being implemented by the IPCC.

the IPCC cannot be interpreted as posterior credence functions representing the total available evidence from the perspective of the reviewing authors (or the authors of the original studies).

But one might ask, as a reviewer suggested,[11] whether it is really so far-fetched to consider each of the pdfs as an attempt to estimate ECS, albeit imperfectly. Could we not think of it as consulting various experts, each reasoning imperfectly but offering roughly correct answers? If so, could the confidence assignments then indicate which of the experts or estimates we judge to be most plausible, skillful, or reliable, based on what we know? Wouldn't "weight of evidence" be an appropriate description as we move through different confidence levels (medium, high, etc.), if each pdfs acts as a piece of evidence?

While this perspective might seem appealing, it is ultimately misguided because it does not align with how these pdfs can be reasonably understood. Each credence function based on a single piece of evidence provides only a partial perspective. These functions are inherently incomplete, as they fail to incorporate the full body of available evidence—each represents merely a fragment of the overall information. Therefore, it is unreasonable to treat these partial credence functions as a representation of anyone's belief about the true value of a variable, given that they do not account for the totality of evidence. From a Bayesian perspective, the appropriate approach is to combine all available evidence into a single posterior credence function that synthesizes the information comprehensively. This combined function provides a coherent and complete representation of uncertainty. On the other hand, attempting to rank these incomplete components (the individual pdfs) is conceptually problematic, as they do not represent anyone's final belief and hence cannot provide meaningful insights when considered in isolation. In short, these individual components are only meaningful within the context of a combined framework; trying to rank them would fail to yield anything conceptually coherent or practically useful. Therefore, it is implausible to interpret a pdf from a specific line of evidence as anyone's—including the original study author's—estimate of ECS based on the overall available evidence, which extends beyond individual studies. As Sherwood et al. (2020, 2) observe, 'the available evidence consists of diverse strands, none of which is conclusive by itself. This requires that the strands be combined in some way.' In other words, each pdf, considered in isolation, lacks a meaningful probabilistic interpretation in the broader evidential context.

My interpretation aligns with the IPCC's own approach, which recognizes the need to integrate different lines of evidence to arrive at an overall ECS assessment. In AR6, the IPCC presented a best estimate of ECS at 3°C, with a "likely" range of 2.5 to 4°C and a "very likely" range of 2 to 5°C (noting only "medium confidence" in the upper end of the

[11] I thank a reviewer for pressing me on this issue.

"very likely" range). Although these estimates did not result from formal Bayesian methods, they were derived using a pragmatic approach that combined multiple lines of evidence, drawing on certain Bayesian principles where applicable.[12]

Therefore, if BHH's proposal were implemented as implicitly suggested, the "probability" assignments would lack a coherent interpretation of probability. The probability intervals resulting from different confidence sets of pdfs would effectively become arbitrary numerical ranges, rather than meaningful probability estimates. In other words, this approach would lead to the IPCC communicating uncertainty by assigning various sets of arbitrary numbers to a hypothesis at different arbitrary confidence levels, instead of providing useful probability intervals with a meaningful interpretation of probability. Given this, it is clear that the IPCC should approach BHH's proposal with caution since it risks resulting in incoherent and misleading communication of uncertainty.

# 4 Conclusion

The question of how the IPCC should conceptualize and communicate uncertainty is undeniably challenging. While I agree with BHH that clarifying how decision-makers and the broader public should understand IPCC findings is an important goal, I contend that BHH's proposal is not a practical or effective way to achieve this. While BHH's account might, in theory, be applicable in cases where likelihood could be interpreted objectively, this interpretation is not well-suited to the IPCC's context. Regarding the subjective interpretation of likelihood, without a clear example or framework that addresses the challenges discussed, I remain unconvinced of the applicability of BHH's account, even in principle.

Alternative approaches have been proposed to improve the IPCC's communication of uncertainty. For instance, Mach et al. (2017) offer a radically different proposal that eliminates the confidence metric altogether. Their proposal, grounded in a detailed understanding of IPCC practices and challenges, highlights the diversity of perspectives on this issue. My aim here was not to evaluate these alternatives, nor was it to propose my own. Instead, it was simply to demonstrate that BHH's proposal is unlikely to help the IPCC improve its communication of uncertainty. Therefore, to strengthen the IPCC uncertainty framework, other options should be explored.

[12]For instance, the AR6 notes that when two independent lines of evidence each suggest a low probability for an outcome (e.g., ECS being less than 2°C), the combined probability for that outcome becomes even lower. Similarly, when one line of evidence provides more robust constraints, the final assessment tends to favor the narrower range suggested by that stronger evidence (IPCC 2021, 1006).

# 5 Statements and Declarations

## 5.1 Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

## 5.2 Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

## 5.3 Author Contributions

All authors contributed to the study conception and design and there was no Material preparation, data collection and analysis.

## 5.4 Data Availability

This research includes no new data.