Combining probability with qualitative degree-of-certainty metrics in assessment*

Casey Helgeson[†], Richard Bradley[‡], and Brian Hill[§]

***Working paper,* 01/17/2018**

Abstract: Reports of the Intergovernmental Panel on Climate Change (IPCC) employ an evolving framework of calibrated language for assessing and communicating degrees of certainty in findings. A persistent challenge for this framework has been ambiguity in the relationship between multiple degree-of-certainty metrics. We aim to clarify the relationship between the *likelihood* and *confidence* metrics used in the Fifth Assessment Report (2013), with benefits for mathematical consistency among multiple findings and for usability in downstream modeling and decision analysis. We discuss how our proposal meshes with current and proposed practice in IPCC uncertainty assessment.

1 Introduction

Beginning with its Third Assessment Report (2001), characterization and communication of uncertainties by the Intergovernmental Panel on Climate Change (IPCC) has been informed by a series of *guidance notes* (Moss and Schneider, 2000; Manning, 2005; Mastrandrea et al., 2010) that share best practices and promote consistency across chapters and working groups. The guidance note for authors of the fifth and most recent assessment report (AR5) provides two "metrics for communicating the degree of certainty in key findings." The first is *confidence*, a qualitative metric ranging from *very low* to *very high* and based on assessments of the underlying *evidence* (type, amount, quality, and consistency) and degree of *agreement* (Figure 1, left). The second metric is probability, also called *likelihood*, and is conveyed through a menu of pre-defined terms,

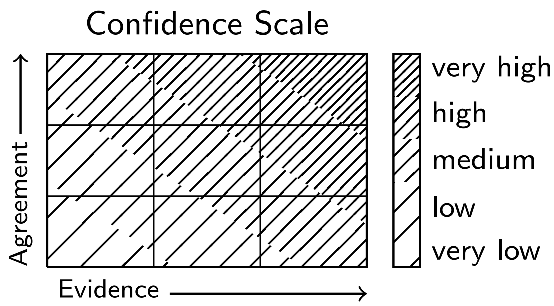
*This work was supported by the National Science Foundation (GEO-1240507), UK National Environmental Research Council (NE/P016367/1), UK Arts and Humanities Research Council (AH/J006033/1), and French National Research Agency, L'Agence Nationale de la Recherche, (ANR-14-CE29-0003- 01)

[†]Earth and Environmental Systems Institute, Pennsylvania State University, cmh455@psu.edu

[‡]Department of Philosophy, Logic and Scientific Method, London School of Economics

[§]GREGHEC, CNRS–HEC Paris

where, e.g., *unlikely* means 0–33% probability and *very unlikely* means 0–10% probability (Figure 1, right).



Likelihood Scale

virtually certain	99–100%
very likely	90–100%
likely	66–100%
about as likely as not	33–66%
unlikely	0–33%
very unlikely	0–10%
exceptionally unlikely	0–1%

Figure 1: Confidence and likelihood scales for communicating degree of certainty in key findings of the IPCC AR5 (Mastrandrea et al., 2010).

Across assessment cycles and iterative refinements to the guidance notes, a persistent challenge for both authors and consumers of the reports has been confusion over *the relationship* between multiple degree-of-certainty scales (Kandlikar et al., 2005; Risbey and Kandlikar, 2007; Swart et al., 2009; Shapiro et al., 2010; Jonassen and Pielke, 2011; Mastrandrea and Mach, 2011; Aven and Renn, 2015; Mach et al., 2017).

In the AR5, this issue is most visible where authors use *both* confidence *and* likelihood terms together in a single statement, as seen, for example, in the findings on equilibrium climate sensitivity (ECS): "ECS is *likely* in the range 1.5°C to 4.5°C with *high confidence*. ECS is positive, *extremely unlikely* less than 1°C (*high confidence*), and *very unlikely* greater than 6°C (*medium confidence*)" (Stocker et al., 2013, 84). But the issue is really much wider, since even where confidence terms do not appear in the text, *all likelihood statements* should be read as confidence-qualified, where the implicit level of confidence is *high* or *very high* if otherwise unstated (Mastrandrea et al., 2010, criteria E, F). Deep down, all AR5 likelihood statements share the form of the ECS findings quoted above.

Post-AR5 commentary indicates continuing challenges for the interpretation and consistent usage of such findings (Aven and Renn, 2015; Mach et al., 2017), with the most comprehensive analysis of uncertainty terms used across the AR5 concluding that "In practice and perhaps out of necessity, author teams adopted a spectrum of approaches in interpreting the relationship between confidence and likelihood. Redundancies and interactions across layers of characterized uncertainties were often handled differently and not transparently" (Mach et al., 2017, 9).

Heading into the sixth assessment report, we contribute a proposal to solve this particular challenge. We propose a simple mathematical model of the confidence–likelihood relationship that resolves outstanding ambiguities while respecting the qualitative nature of the confidence scale. Our proposal also preserves the conceptual distinction between the scientific estimate of chances (the likelihoods) and the evaluation of the evidential basis underpinning these estimates (the confidence assessments), which is an important motivation for reporting confidence as well as likelihood. In what follows, we briefly examine the interpretive problem a bit further, then present the solution, then discuss.

2 Reasoning with likelihood and confidence

To further motivate the problem we mean to solve, we illustrate how ambiguity in the confidence–likelihood relationship can challenge attempts to reason rigorously from AR5 likelihood statements. Continuing with the ECS findings quoted above, note that the statements addressing the middle and left tail of the distribution are made with *high confidence* while the statement on the right tail is made with *medium confidence* (Figure 2). A natural reading is that these confidence terms flag differences in the evidence base underpinning likelihood statements about one *value range* versus another: the evidence on small and middling values for ECS warrants *high confidence* whereas the evidence on more extreme values is weaker and allows for only *medium confidence*.

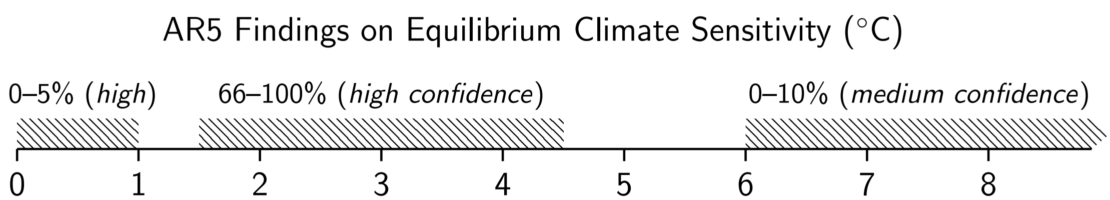


Figure 2: AR5 findings on equilibrium climate sensitivity (ECS). Probability ranges are the numerical translations of (from left to right) *extremely unlikely*, *likely*, and *very unlikely*.

But this reading gets muddled when we try to draw out some of the mathematical consequences of the individual likelihood statements. For example, starting from the *medium confidence* statement on the right tail, assigning probability 0–.1 (*very unlikely*) to the range $ECS > 6$ is mathematically equivalent to assigning .9–1 (*very likely*) to $ECS \leq 6$. Supposing that mathematically equivalent statements should enjoy the same level of confidence, we can equally regard the authors as reporting with *medium confidence* that ECS is *very likely* less than 6°. But this new *medium confidence* statement covers the same low and midrange ECS values where the other findings say *high confidence* is appropriate.

Or if we start from the *high confidence* finding on $1.5 < ECS < 4.5$, the likelihood assignment .66–1 (*likely*) significantly constrains the probabilities of ECS values *outside* that range, as there is at most .33 probability left to go around. How much is leftover for $ECS > 6$, for example? Supposing that deductive mathematical consequences retain the confidence level of the statement from which they are derived, we can infer that ECS is (at most) *unlikely* (0–.33) greater than 6° (*high confidence*). Again this contradicts the reading with which we began, according to which only *medium confidence* could be achieved for likelihood assignments on ECS values beyond 6.

How should we understand, and resolve the tension between, these overlapping and potentially competing *high* and *medium confidence* likelihood assignments?

3 Proposal

We start from the idea that the underlying metrics *evidence* and *agreement* are best judged with respect to a concrete statement: evidence must be evidence *for something*, and agreement must be agreement *on something*. The head-scratching elicited above ultimately stems from an ambiguity in what that *something* is. Here we explore the consequences of judging evidence and agreement, and hence confidence, with respect to a concrete *likelihood assignment* (such as $ECS > 6$ is *unlikely*).

One consequence of this move is that confidence now participates in a trade-off with *the precision* of the likelihood assignment: confidence can be raised (without going out and doing more research) by widening the probability interval; conversely, a likelihood assignment can be made sharper (more narrow) by accepting lower confidence. In terms of the evaluation of *evidence* and *agreement* that underwrites confidence assessments in the AR5 framework, wider, less informative probability intervals may enjoy greater confidence because they are supported by additional studies or lines of evidence from which sharper probabilistic conclusions cannot be drawn (*evidence*) and/or because broader statements generally make agreement between different lines of evidence more likely (*agreement*).

And because confidence terms attach to the likelihood rather than the outcome directly, two findings can address the same outcome despite using different confidence levels. There is no *logical* inconsistency in reporting, for example, that the probability of ECS exceeding 6° is 0–.1 (*very unlikely*) with *medium confidence*, and 0–.33 (*unlikely*) with *high confidence*. The two statements *complement* one another, together giving an indication of the prevailing trade-off between confidence and precision. Informally, these findings say: “We have good evidence that the probability is less than one tenth, and very strong evidence it is no more than one third.” On our approach, all of the ECS findings discussed above—both original and derived—turn out to be mutually consistent and complementary.

Usage in the AR5 conforms to this understanding of the confidence–likelihood relationship insofar as authors have sometimes traded off precision in the likelihood assignment against the level of confidence—a practice noted by Mach et al. (2017), calling it *adjustment of likelihood as a function of confidence*. We take our proposal to offer a clear rationale for such "adjustments," and a route to making them more principled and more transparent.

An important step towards greater rigor can be taken by incorporating both confidence and likelihood into a formal *consistency check* among multiple likelihood-plus-confidence findings that address the same uncertain quantity. What is needed to check consistency is an overarching mathematical structure that can relate one finding to another and systematize the logical constraints that each finding puts on the others.

When working with *exact probabilities*, the numbers assigned to different outcome ranges ought to be consistent with a single probability density function (pdf). But here we want to address *probability intervals*. The analogous construct for probability intervals is a *set* of pdfs. 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. (Equivalently, we can say that 0 and .1 are the narrowest bounds such that every pdf in the set agrees on the probability of x falling within the interval.)

So a set of pdfs systematizes, and enforces consistency among, the probability intervals assigned to different ranges of a single quantity such as ECS. The question is how *confidence levels* fit into the picture. We associate each level of confidence *with its own set of pdfs*, where higher-confidence sets encompass lower-confidence sets (Figure 3). This nesting of sets naturally encodes the trade-off between confidence and precision, since more inclusive sets of pdfs translate to wider probability intervals for any given outcome. Multiple findings addressing the same uncertain quantity are mutually consistent if and only if a single such mathematical object can be constructed so as to underwrite those findings (in practice, this would typically require consideration of at most two sets).

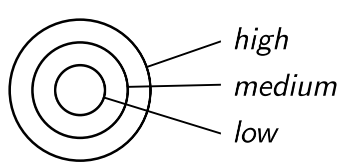


Figure 3: Each confidence level is associated with its own set of pdfs. The nested structure reflects the trade-off between confidence and the precision of likelihood assignments.

Clarifying what constitutes mathematical consistency among several findings not only provides a quality control mechanism in assessment (Parker and Risbey, 2015), but is also an essential prerequisite for applying more formal expert elicitation protocols, as recommended by the TAR, AR4, and AR5 guidance notes as well as a chorus of commentators (Reilly et al., 2001; Oppenheimer et al., 2007; Shapiro et al., 2010; Moss, 2011; Yohe and Oppenheimer, 2011; Morgan, 2014; Thompson et al., 2016; Oppenheimer et al., 2016; Mach et al., 2017).

One worry about qualitative metrics such as confidence is that the information they convey cannot be systematically propagated to "downstream" modeling and decision analysis (e.g., Morgan, 2014). By building confidence assessments into a formal belief representation (the nested sets), we facilitate this propagation. The decision theory literature has developed and defended a range of decision models that reserve a principled role for something like confidence assessments (Gilboa and Marinacci, 2013; Hill, 2016b; Bradley, 2017); our proposal enables a bridge from IPCC uncertainty assessment to this literature—the most direct link (see Bradley et al., 2017) pointing to a model of confidence-based decision from Hill (2013, 2016a). Similar constructs have been defended in econometrics (Manski and Nagin, 1998; Manski, 2013) and decision analysis (Ben-Haim, 2006).

4 Discussion

To provide context and highlight issues for further discussion, we note points of contrast between our proposal and two elements of existing or proposed IPCC practice: (1) the *sequential assessment* of confidence followed by likelihood that is implied in the AR5 guidance (Mastrandrea et al., 2010) and explicit in Mastrandrea et al. (2011), and (2) a recommendation from Mach et al. (2017)—looking forward to the AR6—to foreground *quantitative* characterization of uncertainties by making presentation of *only the likelihood assignment* the preferred reporting option for IPCC findings. We discuss these in turn.

Regarding the sequential procedure (1), systematically assessing confidence *before likelihood enters the picture* makes that assessment a function of the outcome range alone. It follows that a given outcome range (like $ECS > 6$) must be associated with a *fixed confidence level*, which in turn rules out trade-offs between confidence and precision of likelihood. But it is precisely through legalizing such trade-offs that our proposal allows multiple findings to coherently address the same outcome using different confidence terms, thus resolving the apparent contradictions that arise when trying to reason mathematically from AR5-style findings (the ECS example, §2). Such "same-outcome" findings cannot be reconciled within the constraints imposed by a strictly sequential, confidence-first procedure. Competition between stated and derived findings must instead be averted by restricting either what can be stated or what can be derived. We can see two ways to do this; both risk relinquishing any meaningful characterization of the evidence and agreement underpinning a likelihood assignment.

The first option is to prohibit the use of different confidence terms for different ranges of the same quantity (e.g. ECS). Were medium confidence used on the mode as well as the tail of ECS, then likelihood constraints derived from the mode and applied to the tail would carry medium confidence (not high confidence as in the §2 example), in line with the confidence level already assigned to the tail. But this solution imposes an artificial constraint that restricts otherwise appropriate and helpful use of the qualitative metric. It also raises questions about the meaning and validity of "evidence" and "agreement" when those assessments are forced to remain constant across outcomes ranges and likelihood judgements.

The second option is to deny that a likelihood assignment and its consequences enjoy the same level of confidence. If constraints derived from assigned likelihoods carry no particular level of confidence, then our illustration fails to generate a confidence-level conflict on the tail of the distribution. So you can preserve the option of different confidence terms for different outcome ranges only at the cost of making those terms (so to speak) *non-transferrable* to the mathematical consequences of a likelihood assignment. This solution gives up altogether on propagating qualitative assessments to downstream modeling and decision analysis.

This brings us to *likelihood only* as the preferred reporting option (2). One rationale for (2) might be the thought that likelihood—when used—displaces and supplants confidence such that the qualitative language contains no additional information beyond what is now expressed by the likelihood. This rationale conflicts with our proposal. When used in conjunction with likelihood, we understand confidence to express something like Keynes' (1921) "weight of evidence" behind a probability statement. And a key feature of this notion is that the same probability statement can convey a different overall picture of uncertainty when underpinned by different "weights." For example, the *likely* range for ECS given by the AR5 (1.5°C to 4.5°C) is the same as that given by the First (1990), Second (1995), and Third (2001) Assessment Reports (Cubasch et al., 2001, 67). But the body of research underpinning those likelihood statements grew between 1990 and 2013. This difference can be expressed through confidence assessments.

Confidence assessments not only add information about the state of scientific understanding in some domain, they may also be relevant for policy making that draws on this science. Consider two potential policies, each of which is justified by scientific findings reported with the same likelihood level, but underpinned by different amounts of evidence and agreement, and hence enjoying differing degrees of confidence. It seems reasonable in such cases to prioritize the policy backed up by evidence-rich, high-confidence findings. Likelihood-only reporting cannot support such a practice.

Yet even if confidence terms do contain substantive and decision-relevant information beyond what is conveyed by the likelihood, one might still worry that this information is too difficult for downstream users to interpret and make use of. One response to these worries—and the one embodied by our proposal—is to clarify and enrich the structure of the qualitative assessments such that the information they are meant to convey might be transmitted more cleanly and used more systematically downstream. Indeed, our proposal connects neatly into decision-making approaches that incorporate confidence into the evaluation of options (Hill, 2016, Bradley, 2017). (Of course, this theoretical point about the possibility and relevance of consistent confidence reporting does not preclude *de-emphasizing* confidence terms in the text of the report—perhaps pushing them into the traceable account—for more practical reasons, such as simplicity and readability for the reports' more casual readers.)

5 Conclusion

A persistent shortcoming of the IPCC's evolving expert judgement framework has been the ambiguous relationship between multiple uncertainty scales. The issue arises in the AR5 between the confidence and likelihood scales, contributing to difficulties both in applying the framework and interpreting the findings. We have proposed a rigorous, principled solution that resolves outstanding ambiguities while also enabling mathematical reasoning from individual likelihood assignments and providing a consistency check on sets of likelihood assignments.

Designing an uncertainty framework for use in assessment involves striking a balance between a host of important considerations, many of which have not been addressed here. We offer our proposal with the aims of stimulating further discussion and highlighting the importance of maintaining mathematical consistency among findings, communicating the weight of evidential support, and enabling downstream use of assessment outputs.

References

Aven, T. and O. Renn (2015). An evaluation of the treatment of risk and uncertainties in the IPCC reports on climate change. *Risk Analysis 35*(4), 701–712.

Ben-Haim, Y. (2006). *Info-gap decision theory: Decisions under severe uncertainty* (Second ed.). Academic Press.

Bradley, R. (2017). *Decision Theory With a Human Face.* Cambridge University Press.

Bradley, R., C. Helgeson, and B. Hill (2017). Climate change assessments: Confidence, probability, and decision. *Philosophy of Science 84*(3), 500–522.

Cubasch, U., X. Dai, Y. Ding, D. J. Griggs, B. Hewitson, J. T. Houghton, I. Isaksen, T. Karl, M. McFarland, V. P. Meleshko, J. F. B. Mitchell, M. Noguer, B. S. Nyenzi, M. Oppenheimer, J. E. Penner, S. Pollonais, T. Stocker, and K. E. Trenberth (2001). Technical summary. In *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Gilboa, I. and M. Marinacci (2013). Ambiguity and the Bayesian paradigm. In *Advances in Economics and Econometrics, Tenth World Congress*, Volume 1.

Hill, B. (2013). Confidence and decision. *Games and Economic Behavior 82*, 675–692.

Hill, B. (2016a). Confidence in beliefs and rational decision making. *mimeo, HEC Paris*.

Hill, B. (2016b). Incomplete preferences and confidence. *Journal of Mathematical Economics 65*, 83–103.

Jonassen, R. and R. Pielke (2011). Improving conveyance of uncertainties in the findings of the IPCC. *Climatic change 108*(4), 745–753.

Kandlikar, M., J. Risbey, and S. Dessai (2005). Representing and communicating deep uncertainty in climate-change assessments. *Comptes Rendus Geoscience 337*(4), 443–455.

Keynes, J. M. (1973/1921). A treatise on probability. In *The Collected Writings of John Maynard Keynes*, Volume 8. London: Macmillan.

Mach, K. J., M. D. Mastrandrea, P. T. Freeman, and C. B. Field (2017). Unleashing expert judgment in assessment. *Global Environmental Change 44*, 1–14.

Manning, M. R. (2005). Guidance notes for lead authors of the IPCC fourth assessment report on addressing uncertainties. Technical report, Intergovernmental Panel on Climate Change (IPCC).

Manski, C. F. (2013). *Public policy in an uncertain world: Analysis and decisions*. Harvard University Press.

Manski, C. F. and D. S. Nagin (1998). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociological Methodology 28*(1), 99–137.

Mastrandrea, M. D., C. B. Field, T. F. Stocker, O. Edenhofer, K. L. Ebi, D. J. Frame, H. Held, E. Kriegler, K. J. Mach, P. R. Matschoss, G.-K. Plattner, G. W. Yohe, and F. W. . . Zwiers (2010). Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Technical report, Intergovernmental Panel on Climate Change (IPCC). Available at <http://www.ipcc.ch>.

Mastrandrea, M. D. and K. J. Mach (2011). Treatment of uncertainties in IPCC assessment reports: Past approaches and considerations for the Fifth Assessment Report. *Climatic Change 108*(4), 659–673.

Mastrandrea, M. D., K. J. Mach, G.-K. Plattner, O. Edenhofer, T. F. Stocker, C. B. Field, K. L. Ebi, and P. R. Matschoss (2011). The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups. *Climatic Change 108*(4), 675–691.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences 111*(20), 7176–7184.

Moss, R. H. (2011). Reducing doubt about uncertainty: Guidance for IPCC's third assessment. *Climatic change 108*(4), 641–658.

Moss, R. H. and S. H. Schneider (2000). Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistenet assessment and reporting. In R. Pachauri, T. Taniguchi, and K. Tanaka (Eds.), *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*, IPCC Supporting Material. Intergovernmental Panel on Climate Change (IPCC).

Oppenheimer, M., C. M. Little, and R. M. Cooke (2016). Expert judgement and uncertainty quantification for climate change. *Nature Climate Change 6*(5), 445–451.

Oppenheimer, M., B. C. O'Neill, M. Webster, and S. Agrawala (2007, September). The limits of consensus. *Science Magazine's State of the Planet 2008-2009: with a Special Section on Energy and Sustainability 317*(5844), 1505–1506.

Parker, W. S. and J. S. Risbey (2015). False precision, surprise and improved uncertainty assessment. *Phil. Trans. R. Soc. A 373*(20140453), 1–13.

Reilly, J., P. H. Stone, C. E. Forest, M. D. Webster, H. D. Jacoby, and R. G. Prinn (2001). Uncertainty and climate change assessments. *Science 293*(5529), 430–433.

Risbey, J. S. and M. Kandlikar (2007). Expressions of likelihood and confidence in the IPCC uncertainty assessment process. *Climatic Change 85*(1-2), 19–31.

Shapiro, H. T., R. Diab, C. de Brito Cruz, M. Cropper, J. Fang, L. Fresco, S. Manabe, G. Mehta, M. Molina, P. Williams, et al. (2010). Climate change assessments: Review of the processes and procedures of the IPCC. Technical report, InterAcademy Council, Amsterdam.

Stocker, T., D. Qin, G.-K. Plattner, L. Alexander, S. Allen, N. Bindoff, F.-M. Bréon, J. Church, U. Cubasch, S. Emori, P. Forster, P. Friedlingstein, N. Gillett, J. Gregory, D. Hartmann, E. Jansen, B. Kirtman, R. Knutti, K. K. Kumar, P. Lemke, J. Marotzke, V. Masson-Delmotte, G. Meehl, I. Mokhov, S. Piao, V. Ramaswamy, D. Randall, M. Rhein, M. Rojas, C. Sabine, D. Shindell, L. Talley, D. Vaughan, and S.-P. Xie (2013). Technical summary. In T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 33–115. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Swart, R., L. Bernstein, M. Ha-Duong, and A. Petersen (2009). Agreeing to disagree: uncertainty management in assessing climate change, impacts and responses by the IPCC. *Climatic Change 92*(1-2), 1–29.

Thompson, E., R. Frigg, and C. Helgeson (2016). Expert judgement for climate change adaptation. *Philosophy of Science 83*, 1110–1121.

Yohe, G. and M. Oppenheimer (2011). Evaluation, characterization, and communication of uncertainty by the intergovernmental panel on climate change—an introductory essay. *Climatic Change 108*(4), 629–639.