Assessing the quality of regional climate information

Marina Baldissera Pacchetti
University of Leeds, School of Earth and Environment and ESRC Centre for Climate Change Economics and Policy, UK
m.baldisserapacchetti@leeds.ac.uk

Suraje Dessai
University of Leeds, School of Earth and Environment and ESRC Centre for Climate Change Economics and Policy, UK
s.dessai@leeds.ac.uk

Seamus Bradley
University of Leeds, School of Philosophy, Religion and History of Science and ESRC Centre for Climate Change Economics and Policy, UK

David A. Stainforth
London School of Economics, Grantham Research Institute on Climate Change and the Environment, Centre for the Analysis of Timeseries and ESRC Centre for Climate Change Economics and Policy, UK
Department of Physics, University of Warwick, UK

Early Online Release: This preliminary version has been accepted for publication in *Bulletin of the American Meteorological Society*, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-20-0008.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

Abstract

There are now a plethora of data, models and approaches available to produce regional and local climate information intended to inform adaptation to a changing climate. There is, however, no framework to assess the quality of these data, models and approaches that takes into account the issues that arise when this information is produced. An evaluation of the quality of regional climate information is a fundamental requirement for its appropriate application in societal decision-making. An analytical framework for "science-based statements and estimates about future climate" that allows for an assessment of their quality is constructed. This framework targets statements that project local and regional climate at decadal and longer time scales. After identifying the main issues with evaluating and presenting regional climate information, it is argued that it is helpful to consider the quality of statements about future climate in terms of (1) the type of evidence and (2) the relationship between the evidence and the statement. This distinction not only provides a more targeted framework for quality, but also shows how certain evidential standards can change as a function of the purpose of a statement. The key dimensions to assess regional climate information quality are: diversity, completeness, theory, adequacy for purpose, and transparency. This framework is exemplified using two research papers that provide regional climate information and the implications of the framework are explored.

Capsule

A framework for the assessment of quality in regional climate information needs to include dimensions such as: Diversity, Completeness, Theory, Adequacy for purpose, and Transparency.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI10.1175/BAMS-D-20-0008.1.

1. Introduction

Informing the large number of actions needed to manage climate risks, reduce damages, and maximise potential opportunities is a grand challenge for climate change science (Moss et al. 2013). Adapting human and natural systems across sectors of society is necessary to improve preparedness and enhance resilience to a changing climate. Decision-relevant climate information can support society's efforts to adapt to climate change.

Different strategies for making climate adaptation decisions utilise scientific information differently in the decision-making process (Dessai and van der Sluijs 2007). Nevertheless, whenever scientific climate information is used in adaptation, "quality" is considered to be an essential characteristic that this information should have (e.g. Lu 2011, Wilby et al. 2009, for a general overview on quality of science for policy, see Funtowicz and Ravetz 1990).

The kind of long term regional climate information that is increasingly important for decision makers (see, e.g., Knutti 2019) ranges in temporal scales - from hours to multiple decades – and spatial scales – from meters to hundreds of kilometers. Global Climate / Earth System Models (GCMs and ESMs) are the dominant source of regional climate information, but increasingly downscaling, both statistical and dynamical, has been used to achieve higher resolution information (see, e.g., Pielke and Wilby 2012, Giorgi 2020). However, it is difficult to evaluate long term information at the regional scale because of the presence of deep uncertainty and because the usual empirical tests are not applicable due to the non-stationarity of the climate system (e.g. see Stainforth et al. 2007b).

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

General guidance on the dimensions of quality (e.g. Nissan et al. 2019, Benestad et al. 2017) are emerging in the literature, but this literature is fragmented. The way quality is characterized can be very different depending on whether it is discussed in physical climate science (e.g. Zeng et al. 2019; Krysanova et al. 2018), environmental social science (e.g. Lemos and Moorhouse 2005; Cash et al. 2003) or philosophy of science (e.g. Parker and Risbey 2015, Parker 2009), and on whether it is aimed at knowledge deriving from climate science research or information deriving from climate services (Barsugli et al. 2013). Characterization of the quality of climate information should also include a discussion of the roles of scientific knowledge in their social, political and economic contexts (Maxim and van der Sluijs 2011).

In order to address the quality issues that are specific to climate change adaptation decision making, we focus on what quality means for *regional climate information derived from knowledge produced by scientific research*. In particular, we target the quality of the evidence and the methods used to produce knowledge about regional climate change and its related uncertainty. Due to the knowledge evaluation issues discussed below, this knowledge cannot only be derived from the output of Earth System Models: other lines of evidence, such as expert judgment, theory and observations can and should be taken into account to produce information that goes beyond or replaces model output (see Fig. 1). In the rest of this paper, regional climate information refers to scientific knowledge about regional climate that intends to inform adaptation to a changing climate. We recognize that social values do influence methodological choices in the sciences (see e.g. Douglas 2009, Intemann 2015, Elliott and Steel 2017), but we do not directly engage with this debate here.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

The aim of this paper is to propose a framework that enables the assessment of the quality of regional climate information. The framework draws on insights from physical climate science, environmental social science and philosophy of science to identify relevant quality dimensions and is intended as a general guide to assess quality in this context. We characterize quality along five dimensions that can support users of climate information, including “non-specialist scientists” and decision-makers. Scientists might use this framework because knowledge about future regional climate is produced by experts across many different disciplines, and an expert in one discipline may not have the expertise to evaluate the knowledge produced from another discipline. Decision-makers can use this framework because they may not be trained in the science that is used to generate regional climate information. The framework raises the questions that specialists and non-specialists alike need to ask regarding the provenance of the information they provide and use.

We start this paper by further clarifying what is meant by regional climate information (Sect. 2.1). Next, we highlight the main issues that arise in evaluating and representing this information: knowledge evaluation issues (Sect. 2.2) and quantification issues (Sect. 2.3). These issues are to be understood in the context of the purpose of regional climate information discussed in Section 2.1. Next, we characterize how regional climate information is constructed (Sect. 3) and we propose a framework to evaluate its quality (Sect. 4). The framework identifies five key dimensions of quality: *diversity*, *completeness*, *theory*, *adequacy for purpose*, and *transparency* and we illustrate the application of these dimensions with two examples. We conclude with some general remarks about the framework (Sect. 5).

2. Why we need a quality assessment framework

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

Climate science serves many different purposes, ranging from improving our understanding of earth-system processes and their interaction with human activity, to informing decision making. We focus on regional climate information that has the purpose of informing climate change adaptation decisions. Regional climate information is to be understood as *statements or estimates about future regional climate that have the intention of informing adaptation to a changing climate.* We highlight the *purpose* of regional climate information because different purposes for generating information motivate the use of different methods of generating this information (Laudan 1984, pp.62-63, Shackley and Wynne 1995), and different standards of evaluation. However, there is currently no quality framework that explicitly claims that quality can change depending on the purpose that the information serves. In this section we clarify the purpose of the regional climate knowledge addressed by our framework and specify some of the main challenges that can affect quality in this context.

2.1 The purpose and epistemic reliability of regional climate information

The issue of model purpose has been addressed by researchers interested in how to justify model-based inferences that are relevant for policy. Recently, Thompson and Smith (2019) have highlighted the distinction between “model land”, the land of statements about models derived from models, and statements derived from models about the real world. They argue that when making decision relevant model-based statements, scientists need to be aware of and explicit about the consequences of the assumptions introduced in a model, since decision-relevant statements will be taken to be about the real world, rather than the model. Risbey et al. (2005), Parker (2009, 2020), Baumberger et al. (2017), and Nissan et al. (2020) argue that when we assess a model, we are not assessing the whole model, but only the aspect of the model that addresses the particular question asked of the model, or, in other words, the

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

purpose that the model serves. So, how does the purpose of informing climate change adaptation feature in this quality assessment framework?

If the purpose of regional climate information is to inform adaptation to a changing climate then it should be epistemically reliable. Epistemic reliability of a statement or an estimate, in this context, refers to whether the statement or estimate (and its associated confidence/uncertainty) about the future is likely to capture or estimate a state of the climate that will actually realise. The epistemic reliability of regional climate information is difficult to assess due to the non-stationarity of the climate system and the timescales of change under consideration which remove the option of applying the usual empirical tests to these statements/estimates. Epistemic reliability of multi-decadal projections is different from the meaning of reliability as it is used by the weather and seasonal forecasting community (e.g. Weisheimer and Palmer 2014), which relies on past model performance in forecasting the system. Winsberg (2006, p. 17) suggests that in some cases one can define a reliable process in terms of how well it fits with the methods, physical intuition, and data of a given field, rather than in terms of the relative frequency with which a model produces an accurate statement. Baldissera Pacchetti (2020), however, adds that when assessing uncertainty from the structural differences between climate models it is important to be able to explain *why* we might believe that such climate knowledge is epistemically reliable.

Providing these explanations is especially important when an evaluation of the accuracy of model output cannot be established over a time frame of meaning to the assessment. We know whether a weather forecasts is reliable when there are collections of past forecasts and out-of-sample verifications thereof. This is the sense in which reliability is understood in the weather and seasonal forecasting community. But out-of-sample verifications are not always

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

available. Consider the following: we know that a damped harmonic oscillator provides reliable predictions of the behavior of an oscillating weight in a viscous medium if the oscillation is slow enough–and we know this because it is derived by means of Hooke's law, Newton's second law, etc. This theoretical knowledge, together with some empirical results about viscosity, will allow one to have an epistemically reliable expectation of how a weight will behave in a new medium for which one knows the viscosity but for which there are no experimental results of weights oscillating in that medium. While the case of generating regional climate information is considerably more complex, explanations of a similar nature can help us assess to what extent regional climate information captures the potential future state of affairs, in a way that can inform climate change adaptation decisions. When the purpose of regional climate information is to inform adaptation to a changing climate, it should be epistemically reliable in the sense that one should be able to explain why a statement about future regional climate is likely to happen, or, in other words, why it is credible.

Epistemic reliability of regional climate information is particularly important when consequential decisions are made on the basis of this information. This concept is therefore central to our quality evaluation framework: the higher the quality of a statement or estimate about future climate, the more reasonable it is to believe that we are making a credible statement or estimate for its societal purpose. This is how "quality" is most usefully interpreted in the context of regional climate information for adaptation planning. In order to be clear how knowledge about future regional climate should be evaluated, we first review some of the major shortcomings of current climate knowledge evaluations and of how current knowledge about regional climate is presented.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

2.2 Knowledge evaluation issues

While it is recognized that uncertainty in climate projections is in some sense irreducible (see, e.g., McWilliams 2007), gaps in our knowledge about future regional climate can also be due to theoretical, observational and computational constraints. For example, there are limitations to the empirical tests of the climate models from which this knowledge is derived. Truly forward looking tests about the accuracy of climate projections can only be made with earlier generations of models and even then the available observations are limited to those between the date the projection was produced and the present day; a short period in the context of climate change and climate variability (see, e.g., Dessai and Hulme 2008, Grose et al. 2017, Housfather et al. 2020 for exceptions).

Climate models are also evaluated based on how they simulate past observational or reanalysis data. But shared assumptions in Global Climate / Earth System Models (GCMs and ESMs) and reanalysis data generated with such models open the possibility of shared biases that are difficult to detect and isolate. Lenhard and Winsberg (2010, p. 257) have argued that GCMs are subject to a kind of "confirmation holism": the complexity of both the interactions between the modules of the GCM and the model development process, often makes it virtually impossible to assign improvements in model performance to improvements of the representation of the physical processes in the code. Further, assessments of model performance with observational and reanalysis data do not directly test the prognostic accuracy of the model, as successful reproduction of past data does not imply that models will accurately predict a changing climate on long temporal scales (e.g., Reifen and Toumi 2009). Climate projections are extrapolatory (see e.g. Stainforth et al. 2007b), because the

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

conditions to which the model is applied are different to those for which we have observations, making past successes less relevant. Last but not least, the nature of this kind of evaluation runs the risk of evaluating the models against features of datasets which have been knowingly or unknowingly used in model development or tuning (see Shackley et al. 1998 for an early analysis of the use of GCMs for policy making).

Model evaluation is often sought in a context of demonstrating a degree of *robustness.* One common understanding of robustness is the one used when performing sensitivity analysis, which philosophers have called "inferential robustness" (Woodward 2006). Inferential robustness refers to a robust inferential process. Its core idea is that the statement is robust if it is insensitive to various competing assumptions, models, or, for the case of regression analysis, choice of competing explanatory variables. Following Woodward (2006), suppose that E_i are the different assumptions, models, etc., and R is the statement derived from the inferential process. Then, if the same statement R is obtained for any choice of E_i, then R is robust, and likely to be true. This reasoning underlies some of the interpretations of multi-model ensembles, and has been criticized on the basis that model genealogy, shared assumptions between GCMs, and the use of GCMs in producing reanalysis data, undermine the strength of this inference and the associated uncertainty (Parker 2011, but see Lloyd 2015 for an alternative interpretation of robustness and ensembles). Woodward notes that this notion of robustness relies on the assumption that *all possible* competing assumptions, models or explanatory variables are considered.

Another relevant notion of robustness is what Woodward (2006) calls "measurement robustness" (also discussed in Wimsatt 1981), which refers to the confidence one has in an empirical value that is measured with different instruments. So, for example, a measurement

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

of temperature at location x and time t is robust if one gets the same value with a mercury thermometer, a thermocouple or an infrared thermometer. While this notion of robustness has not been formalized, philosophers have argued that the *independence* of these measurements is what is valuable for inferring that the reading is correct (Woodward, 2006, p. 234). The reasoning behind this argument is similar to the reasoning behind the importance of independence in statistical sampling, i.e. it is valuable because it removes possible biases.

Independence is a term discussed in physical climate science in the context of multi-model ensembles (MME) (see Knutti et al. 2017 and references therein) and in the philosophy of science (Parker 2011; Lloyd 2009, 2010, 2015). This is particularly problematic for MMEs, where models are often analysed as if they were independent (Pirtle et al. 2010; see also Parker 2011) but such an assumption is not warranted (Parker 2011, Knutti et al. 2010, Masson and Knutti 2011).

Physical climate scientists are developing strategies to approximate independence in MMEs by designing schemes to weight the models (Sanderson et al., 2015, Knutti et al., 2017) but there are still open questions about whether this strategy is effective. The main point of Parker (2011) is that we cannot think of an ensemble of models to be a *random sample* from the space of possible models but since it is unclear how to define a space of all possible models, model weighing is inherently problematic. Most recently, Jebeile and Crucifix (2020) have discussed the difficulties of MME optimization.

The complexity of GCMs and the difficulties of model evaluation implies that regional climate knowledge needs to rely on more than just the models (or ensemble of models) to be able to evaluate its epistemic reliability and hence its quality. Further, the above limitations

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

should always be clearly stated when producing decision relevant climate knowledge based on GCMs.

2.3 Quantification issues

The issues outlined above can be exacerbated by an excessive focus on quantification. Policy makers want, or are thought to want, quantified climate information (Heaphy, 2015) but scholars interested in the science-policy interface have increasingly called attention to the pitfalls of such a focus and the overconfidence that it produces (Porter 1995, Supiot 2017, Kovacic 2018).

Adopting a “one size fits all” approach for quantifying knowledge and uncertainty can lead to several issues. Parker and Risbey (2015) argue that this kind of approach can lead to a false sense of precision regarding the uncertainty associated with a particular distribution of future states of the climate. Such false precision, they continue, may influence the choice of decision making strategy adopted by the policy maker (e.g. a top down instead of a bottom up approach, see Dessai and van der Sluijs 2007).

The focus on quantified information may also suggest that such information is somehow better than other ways of representing knowledge claims and associated uncertainties (e.g. ranges with low precision, or direction of change), but this is not the case. See, for example, the discussion of quantified information provided by the IPCC found in Risbey and Kandlikar (2007). In that paper, the authors argue that the distinction made by the IPCC between likelihood and confidence (Mastrandrea et al. 2011) is not a useful one, because likelihood and confidence are supposed to separate the frequentist and subjective interpretations of

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI10.1175/BAMS-D-20-0008.1.

quantified model output, but these cannot be clearly separated (Risbey and Kandlikar 2007, p. 24). Relatedly, philosophers have argued that even in a Bayesian framework, societal and ethical values can influence the evaluation of probabilistic model output (Parker and Winsberg 2018). So, quantification may lead to a false perception of lack of subjectivity.

In sum, the current focus on GCM/ESM evaluation and output quantification is not generally adequate to achieve the kind of epistemic reliability that is required for informing decision making. In the rest of this paper, we focus on how regional climate information that intends to inform decision making is constructed and indicate the quality dimensions that directly address the issues that have been presented so far.

3. Towards a quality assessment framework

We can now ask how to approach quality assessment for regional climate information that intends to inform decision making. Decision makers may want to know how likely it is that a particular statement about future climate will realize, so the epistemic reliability of a statement is an important component of statements that aim to inform decision making. The relation between epistemic reliability and how these statements or estimates are presented, e.g. how precise a particular statement is and what form it takes (probability distribution or qualitative estimate), depend on how the information is produced.

Risbey and Kandlikar (2007) suggest that scientists should formulate statements with different levels of precision based on the available evidence and the strength of the justifications for the statements. Precision, in this case, refers to whether the information appears in the form of a probability distribution function, bounds on estimates, and so on. The

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

quality of an estimate depends on the quality of the evidence in the sense that we can make better estimates with better evidence. But we can *choose* what precision to report and that choice also depends on the quality of the evidence and on how the accuracy is assessed.

The relation between quality of model output and quality of evidence is most clear in short term forecasting (e.g. weather forecasting): in this case, instances of past successes of the models, and well-established methodological choices provide support for the accuracy of future probabilistic forecasts and, as a consequence, their quality. It should be noted however, that probabilistic weather forecasting may incur similar problems as climate models when evaluating forecasts of extreme events for which there are few examples in the observations.

For the case of regional climate information however, instances of past success of a model do not directly imply that the model will be accurate in the future, since the conditions to which the model is applied are different (see section 2). We therefore fine-grain the analysis of the relation between evidence and scientific statements to better articulate how quality can be evaluated for regional climate information. We consider two aspects of this information relevant for quality:

(1) the evidence which underlies this information (e.g. observational or model time-series data, proxy data, expert judgment, theoretical understanding, etc.), and

(2) the relationship between the evidence and the information (e.g. validity of the methodological details regarding how the information is extracted from the evidence, or how different lines of evidence are aggregated, etc.).

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

Considering this distinction is helpful for evaluating the quality of statements or estimates about future regional climate, in so far as it allows for a systematic representation of the way this information is produced. We exemplify the utility of this distinction in Table 1, where we introduce two papers that we will use to illustrate the quality dimensions discussed in the next section: Risbey et al. (2002) (R02 hereafter) and Tebaldi et al. (2004) (T04 hereafter). Both papers produce scientific regional climate information that intends to inform adaptation. Both papers target changes in precipitation under climate change but they present the information in qualitative and quantitative terms respectively.

Table 1 shows R02's qualitative statement about future regional climate, the evidence used and how the evidence is aggregated to produce the statement. Table 1 also shows T04's statement, and their use of a Bayesian method to estimate probability distributions of present and future precipitation using observations and model output. Their conclusion is similar to R02 in that there is large inter-regional variability, but they claim more precision about what areas are affected, in what way, and provide quantified estimates.

4. The quality assessment framework

Our framework utilises five dimensions that are indicative of the quality of statements about future climate that are relevant for adaptation. These are *diversity*, *completeness*, *adequacy for purpose*, *theory* and *transparency.* This section describes these dimensions and how they apply to our framework. The dimensions embody an in practice unattainable standard for quality but should nevertheless be used as a standard toward which regional climate information should strive.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Diversity. This dimension of quality indicates that *different types of evidence* should be taken into account when producing knowledge about future regional climate of high quality. It is motivated by the knowledge evaluation issues of section 2.2 and the importance of variety-of-evidence discussed by Vezer (2016) (see also Lloyd 2009). Recall the discussion of robustness in that section: it is clear that MME are not robust in the senses that have been spelled out by the philosophers discussed in section 2.2, and hence evaluation of regional climate information in terms of MME is insufficient for high quality. Recent discussions in the philosophy of climate science (e.g. Winsberg 2018, Lloyd 2015) suggest the focus of robustness should be not only the *independence* of the lines of evidence but also the *types* of evidence. In Figure 1, we show a possible typology of evidence that can contribute to climate knowledge, such as theoretical understanding, model output, paleoclimate data, etc. Note that this typology may not be exhaustive and does not have strict boundaries. For example, reanalysis data is a hybrid between model output and observations, and shares characteristics with both types of evidence.

Incorporating different types of evidence is important to address some of the issues around shared biases between climate models–and, to a lesser extent, with reanalysis data. Doing so somewhat approximates independent lines of evidence and the features of measurement robustness that have been discussed by philosophers such as Woodward (2006) and Wimsatt (1981). Diversity of evidence is mostly a dimension that applies to the evidence that underlies regional climate information, but it can also inform the relation between the evidence and the information. While it would be most convenient if diverse sources of evidence supported the same narrow range of values, they are still useful even when that is not the case. Having different sources of evidence that disagree is still better than relying on just one of them, since this allows the scientist to have an appropriate level of uncertainty.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

To further illustrate this dimension of quality, consider the two example papers of Table 1. R02 mainly use three different types of evidence. The first is “dynamical thinking”, which is an expert evaluation of possible future climate based on theoretical insights. This type of evidence combines “expert judgment” and “theory” (see Fig. 1). The second type of evidence is climate model output. Dynamical thinking and climate model output are then supplemented by a third type of evidence, reanalysis data, which it is used to illustrate large scale synoptic features identified by the experts. T04 rely on model output and observational data.

To evaluate diversity, we need to ask about the relation between these types of evidence. It is clearly stated in R02 that dynamical thinking is used to interpret model output, but the possible shared assumptions between model output, reanalysis data and dynamical thinking are not specified (e.g., are the experts the same individuals that have built the models the output of which is used?). On the other hand, the observational data used by T04 shares fewer assumptions with model output data. So, while T04 use fewer types of evidence, the types of evidence used are more *diverse* than in R02. However, the lack of detailed information about the relation between sources of evidence in both R02 and T04 makes this a difficult dimension to assess.

Completeness. Completeness refers to how many of the potential sources of evidence are taken into consideration. This characterization of completeness draws from the discussion of completeness of uncertainty assessments found in Parker and Risbey (2015). Completeness is also discussed as a necessary assumption in the context of robust inferential processes, and Woodward (2006) criticizes it in so far as it is an unattainable standard for robustness. Nevertheless, good statements about future climate draw from all possible and relevant

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

sources of evidence (all the elements that contribute to climate knowledge in Fig. 1), and completeness is a dimension that, together with diversity, captures the value of maximizing the different types of evidence for improving the quality of regional climate information. Because of the structural similarities and shared assumptions of climate models, using all possible models in an MME would not count as complete for the purposes of delivering information for adaptation.

Some reasons for which technically sophisticated model intercomparison projects may be insufficient are the following. First, as discussed in section 2.2, model ensembles do not suffice to produce a probabilistic projection which reflects our uncertainty: models cannot be considered to be elements from a random sample of all possible models (Parker 2010, 2011). Relatedly, the hawkmoth effect (Frigg et al. 2014) implies that small differences in (non-linear) model structure can lead to diverging differences in model output. Even with model weighting, therefore, multi-model ensembles may still produce a biased representation of the uncertainties in model-based projections. Second, Deser et al. (2012) have shown that different (micro) initial-condition (Stainforth et al. 2007a) ensemble sizes are needed depending on the variable of interest (such as sea level pressure, precipitation, or surface air temperature), and computational constraints limit the ensemble size to below what is required for many variables. Third, Hawkins et al. (2016) have shown that details of the distributions of the model output from the model ensembles strongly depends on the (macro) initial conditions (Stainforth et al. 2007a) used for these experiments. We therefore suggest that regional climate information needs additional lines of evidence for satisfying the dimension of completeness.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Take the R02 and T04 cases above. R02 mainly take three possible types of evidence into consideration: “dynamical thinking”, climate model output and reanalysis data. T04, on the other hand, explicitly state that they want a non-heuristic approach to produce a weighted average of model output. This suggests that they purposefully leave out evidence that cannot be formalized (e.g. “dynamical thinking”). So, only two types of evidence (models, observations) are used. T04 would therefore have lower completeness than R02.

Completeness should, of course, be evaluated in conjunction with diversity. The number of different types of evidence (aspect 1 in Section 3.2) needed to satisfy the completeness dimension may depend on the relation between the evidence and the statement (aspect 2 in Section 3.2). In R02, the expert reasoning used to augment and interpret climate model information provides a more complete assessment of the uncertainty in the statements about future climate than the model-based information alone. Dynamical reasoning is, in this case, used as a tool for evaluating model deficiencies and interpreting model output. Furthermore, the authors recognize that model output and dynamical reasoning could be compared with observations to improve the credibility of their statements still further. But in T04 the two types of evidence (model output and observations) are less “complete” than they appear, as a consequence of how they are combined to inform the statements. In the quantification of uncertainty in model-based projections the choice of observational dataset can itself be a source of bias (Singh and AchutaRao 2020). These methodological choices affect T04’s results because their model performance measures (see Table 1) are based on model performance against past observations.

Theory. Theory refers to the theoretical underpinning of statements about future climate, along with the representation of the underlying theory. Climate models are sometimes

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

thought of as theoretical tools (e.g. see Lloyd 2015) but the complexity of climate models implies that it is difficult to explain epistemic reliability without explicitly resorting to the theoretical understanding behind the interpretation of model output (Lenhard and Winsberg, 2010). So the strength of the theoretical underpinning is an important source for the quality of these statements. Ebi (2011), for example, argues for the importance of distinguishing theoretical support from other sources of evidence, as theoretical support can provide useful information to policy makers about the state of scientific understanding behind particular statements. For example, theoretical understanding may point to processes that are considered important for producing an estimate about future regional climate but are not adequately represented or assessed in models. Bony et al. (2011) and Giorgi (2020) make a similar argument and highlight the importance of "understanding" when no direct observations are available. In other situations, outside the domain of climate change, strong theoretical support is not always necessary of course. However, theory becomes increasingly important when other sources of evidence (like repeatable experiments or the appropriate data to test models) are not available. This is the case for climate information for adaptation, where estimates about never before observed states of the climate are needed (Stainforth et al. 2007b).

An example of the theoretical underpinning of statements about future climate is the understanding of the processes that are responsible for generating a particular weather pattern. How this is taken into consideration as evidence in regional climate knowledge is best exemplified by the way R02 uses "dynamical thinking" as theoretical support for evaluating model output. When understood as such, *theory* is a quality dimension that applies to the evidence that is used for statements about future climate (aspect 1).

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI10.1175/BAMS-D-20-0008.1.

T04, on the other hand, do not include any discussion of the physical theory underlying their analysis. Indeed, the multiscale nature of the mechanisms responsible for precipitation are not very well understood and not yet modelled successfully (see e.g. Risbey and O'Kane 2011, Deser et al. 2012). So, while we understand that models are based on theory, the absence of a discussion of how the lack of such theoretical understanding may influence future regional precipitation estimates implies that *theory* is a quality dimension that ranks lower in T04 than in R02.

Adequacy for purpose. Adequacy for purpose refers to the empirical adequacy that is required of a statement about future regional climate that intends to inform decision making. This dimension is similar to empirical adequacy more broadly but puts an emphasis on the fact that the level of empirical adequacy that is required for a statement depends on the purpose of the statement. For example, Risbey and Stone (1996) investigate whether GCMs are adequate for regional climate change assessments by analyzing how GCMs reproduce those large-scale atmospheric phenomena that are relevant for regional climate. Adequacy for purpose usually refers to how adequate the evidence is for the statement (aspect 2). This characterization draws from insights from Risbey et al. (2005), Parker (2009, 2020), Baumberger et al. (2017), and Nissan et al. (2020), discussed in Section 2.1. So, if one wants to assess the empirical adequacy of a model for predicting precipitation, one cannot *just* evaluate the model's performance on the basis of its empirical adequacy about temperature. Rather, one needs to be explicit about how the empirical evaluation contributes to the epistemic reliability of the information.

Different variables used to inform adaptation (e.g. temperature, precipitation) may have different levels of empirical adequacy, depending on the availability and consistency of past

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

data, for example. We can ask whether data is fine grained enough, whether the data has gaps, and whether model output is produced and analysed at the scales that are needed for answering a particular question. In many cases, however, the data that is needed to evaluate the models is not accessible: long term simulations of climate variables (especially at the local scale) may not suffice to test adequacy because the climate system is not a stationary system, and variability may change in unexpected ways (Smith 2002). Because of these limitations, adequacy for purpose as assessed by empirical tests is an important but not conclusive dimension to evaluate the quality of information (Oreskes 1998, Oreskes and Belitz 2001).

Consider again the statements about future regional climate change in T04 and R02. To evaluate adequacy for purpose, we need to ask: Is the evidence adequate for making a statement about future climate that can inform adaptation? R02 clearly state that "one could devise a set of diagnostics to discern whether the climate of a particular region was tending more towards one scenario or another" (p.1048), which suggests that there is more empirical evidence that should be taken into account to have a better statement about future climate. The evidence in this case is therefore not as adequate for purpose as it could be. T04 is more difficult to assess. Their methodology relies on another paper by the same authors (Tebaldi et al. 2005), which is not aimed at informing adaptation but at exploring a particular methodology. The intention of T04, by contrast, is to derive probability distribution functions for precipitation to make statements about actual future climate. However, as discussed above, precipitation is difficult to predict and T04 do not discuss the theoretical, computational and observational constraints to projecting precipitation patterns. So T04 does not address the adequacy for purpose of the information that they produce and ranks low on this dimension.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Transparency. Transparency requires that all the components of statements about future climate are accessible and traceable: a user of climate information should be able to identify the sources of evidence (aspect 1) and the methods used to derive the statements (aspect 2). This dimension is necessary for the evaluation of the dimensions described above, and for clearly defining the applicability of the approach, since there are different requirements for the quality of the evidence and methods depending on the purpose of the information. Transparency is also valuable because it allows for accountability and explicit communication of scientific and social values in the scientific process. These elements become particularly problematic in collaborative research (Winsberg et al. 2014), and hence need to be taken into consideration.

There are different ways in which transparency can be met. First of all, the data and the methods should be available. Both R02 and T04 clearly discuss their methods and their data sources. However, observational data used in T04 is not directly cited in the paper, and the consequences of only using formal methods and one particular dataset to quantify the uncertainty tied to estimates about future climate are not discussed explicitly. Explicitly discussing the limitations of using particular methods or data sets is important as these limitations may not be obvious to all possible users, since not all users share the scientists' expertise. There are different ways to facilitate access to this information in a way that promotes transparency and governments and other organizations are working on methods to achieve this.

How to best achieve transparency is still being researched (see, for example, Weil et al. (2013) for an argument in favour and John (2018) for an argument against transparency in

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

science communication). One suggestion is to use progressive disclosure of information, where information gets tailored to the expertise and needs of the target audience (van Bree and van der Sluijs 2014). A type of information disclosure, called “nontechnical summaries”, can make assumptions and limitations explicit for non-expert users. However, these summaries can only reveal relevant assumptions and limitations in so far as they are mediated by particular groups of experts and typically experts are only aware of a small subset of the assumptions they are making. Note that these kinds of disclosures are not just important in the context of communicating information to users, but also for collaborative scientific projects that involve experts from different disciplines. Another suggestion includes the “traceable accounts” approach of the Fourth US National Climate Assessment (USGCRP 2018, see Chapter 2), which is an explicit attempt at communicating the evidence and methodology that went into each key statement of the report.

5. Concluding remarks

In this paper, we have described issues associated with regional climate information and clarified that by this information we intend scientific statements about future regional climate that have the purpose of informing adaptation to a changing climate. We further described how this information is structured, and, finally, provided a framework for assessing the epistemic quality of climate information for adaptation. The current focus on regional climate information makes the need for a framework for epistemic quality clear: the perceived demand for precise quantification, the limits to evaluating statements about future climate, and the fast growth of sources of climate information for adaptation can pose serious challenges for the decision maker. Our approach to attenuating these challenges has been to clarify the purpose and construction of climate information for adaptation (sections 2 and 3)

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

and to identify a set of quality dimensions motivated by the literature in physical climate science, environmental social science and philosophy of science (section 4).

We note, however, that the framework outlined above does not provide a list of necessary and sufficient conditions for quality. Rather, the dimensions we have selected are a set of quality dimensions. These dimensions may not be comprehensive: special situations in which more indicators are needed, or some indicators become redundant can arise. For example, there are cases in which *theory* is so well established and well developed, that other indicators (such as *completeness* and *diversity*) become irrelevant. The overwhelming theoretical support for the relation between greenhouse gas concentrations and global average temperature is one such example. Of course, the theoretical support for the causal connection between greenhouse gas concentration and global average temperature is the result of a relatively long history of research (see e.g. Edwards 2010), during which the other dimensions of quality were relevant.

We also note that while the dimensions are largely independent, there are some connections among them. Furthermore, the nature of the dimensions is such that to obtain an overall assessment of quality we cannot simply average across them (see the theory example above). The extent to which overall quality is satisfied will be dependent on the specific cases for which it is assessed. Once the assessment has been completed, the user can decide whether the information is of sufficient quality to satisfy her needs.

Nevertheless, we believe that our framework is an important starting point that can have broad applicability: the framework is intended to be used as a guide for scientists and for decision makers interested in using such information. For example, a decision maker may use

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

the framework to realize that different types of evidence are needed for the information to satisfy *completeness*. When exploring a climate service portal, the decision maker can assess the extent to which this dimension is satisfied by reading a nontechnical summary that explains the methods used by the climate service provider. The nontechnical summary, however, also needs to satisfy the *transparency* dimension. It needs to reveal the assumptions and limitations of the information, and to do so to a satisfactory degree it needs to be mediated by a diverse range of experts. The framework can also therefore be a useful normative framework for scientists who produce regional climate information that is intended to inform decision-making on adapting to a changing climate.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

Acknowledgements

We are indebted to James Risbey and two anonymous reviewers for their helpful and constructive comments. Every comment has contributed to improving this manuscript. This research was supported by the UK Economic and Social Research Council (ES/R009708/1) Centre for Climate Change, Economics and Policy (CCCEP).

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

References

Baldissera Pacchetti, M., 2020. Structural uncertainty through the lens of model building. *Synthese.* 1-17.

Barsugli, J. J., and Coauthors, 2013: The practitioner's dilemma: How to assess the credibility of downscaled climate projections. *Eos, Trans. American Geophys. Union*, **94**, 424–425.

Baumberger, C., R. Knutti, and G. Hirsch Hadorn, 2017: Building confidence in climate model projections: an analysis of inferences from fit. *WIREs Climate Change*, **8**, e454.

Benestad, R., K. Parding, A. Dobler, and A., Mezghani, 2017: A strategy to effectively make use of large volumes of climate data for climate change adaptation. *Climate Serv.*, **6**, 48–54.

Bony, S., M. Webb, C. Bretherton, S. Klein, P. Siebesma, G. Tselioudis, and M. Zhang, 2011: CFMIP: Towards a better evaluation and understanding of clouds and cloud feedbacks in CMIP5 models. *CLIVAR Exchanges*, No. 56, International CLIVAR Project Office, Southampton, United Kingdom, 20–24.

Cash, D., W. C. Clark, F. Alcock, N. M. Dickson, N. Eckley, D. H. Guston, J. Jaeger, and R. B. Mitchell, 2003: Knowledge systems for sustainable development. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8086–8091.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: the role of internal variability. *Climate Dyn.*, **38**, 527–546.

Dessai, S., A. Bhave, C. Birch, D. Conway, L. Garcia-Carreras, J. P. Gosling, N. Mittal, and D. A. Stainforth, 2018: Building narratives to characterize uncertainty in regional climate change through expert elicitation. *Environ. Res. Lett.*, **13**, 074005.

Dessai, S. and M. Hulme, 2008. How do UK climate scenarios compare with recent observations?. *Atmospheric Science Letters*, **9**, 189–195.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Dessai, S., and J. P. van der Sluijs, 2007: Uncertainty and climate change adaptation: a scoping study. *Copernicus Institute for Sustainable Development and Innovation,* Utrecht, NL.

Douglas, H., 2009: *Science, policy, and the value-free ideal*. University of Pittsburgh Press, 211 pp.

Ebi, K. L., 2011: Differentiating theory from evidence in determining confidence in an assessment finding. *Climatic Change*, 108(4), 693–700.

Edwards, P. N., 2010: *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press, 552 pp.

Elliott, K. C., and D. Steel, 2017: *Current controversies in values and science*. Routledge, 186 pp.

Frigg, R., S. Bradley, H. Du, and L.A. Smith, 2014. Laplace's demon and the adventures of his apprentices. *Philos. of Sci.*, **81**, 31–59.

Funtowicz, S. O. and J. R. Ravetz, 1990. *Uncertainty and quality in science for policy*. Kluwer Academic Publishers, 231 pp.

Giorgi, F. 2019. Thirty years of regional climate modeling: where are we and where are we going next?. *J. of Geophys. Res.: Atmospheres*, **124**, 5696–5723.

Giorgi, F. 2020. Producing actionable climate change information for regions: the distillation paradigm and the 3R framework. *Eur. Phys. J. Plus*, **135**, 1–15.

Giorgi, F., and L. O. Mearns, 2002. Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method. *J. of Climate*, **15**, 1141–1158.

Grose, M.R., J.S., Risbey, and P.H. Whetton, 2017. Tracking regional temperature projections from the early 1990s in light of variations in regional warming, including "warming holes". *Climatic Change* **140,** 307–322.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Hausfather, Z., H. F. Drake, T. Abbott, and G. A. Schmidt, 2020: Evaluating the performance of past climate model projections. *Geophys. Res. Lett.*, ***47***, e2019GL085378.

Hawkins, E., R. S. Smith, J. M. Gregory, and D. A. Stainforth, 2016: Irreducible uncertainty in near-term climate projections. *Climate Dynamics*, ***46***, 3807–3819.

Heaphy, L. J., 2015: The role of climate models in adaptation decision-making: the case of the UK climate projections 2009. *Eur. J. Philos. Sci.*, ***5***, 233–257.

Hewitson, B. C., J. Daron, R. G. Crane, M. F. Zermoglio, and C. Jack, 2014: Interrogating empirical-statistical downscaling. *Climatic Change*, ***122***, 539–554.

Intemann, K., 2015: Distinguishing between legitimate and illegitimate values in climate modeling. *European Journal for Philosophy of Science*, **5**, 217–232.

Jebeile, J. and M. Crucifix, 2020. Multi-model ensembles in climate science: mathematical structures and expert judgements. *Studies in History and Philos. of Sci. Part A.* https://doi.org/10.1016/j.shpsa.2020.03.001

John, S., 2018: Epistemic trust and the ethics of science communication: Against transparency, openness, sincerity and honesty. *Soc. Epistemology*, **32**, 75–87.

Kovacic, Z., 2018. Conceptualizing numbers at the science–policy interface. *Sci., Technology, & Human Values*, **43**, 1039–1065.

Knutti, R., 2008: Should we believe model predictions of future climate change? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, **366**, 4647–4664.

Knutti, R., 2019: Closing the knowledge-action gap in climate change. *One Earth*, **1**, 21–23.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. of Climate*, ***23***, 2739–2758.

Knutti, R., J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring, 2017: A climate model projection weighting scheme accounting for performance and

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

interdependence. *Geophys. Res. Lett.*, *44*(4), 1909–1918. https://doi.org/10.1002/2016GL072012

Krysanova, V., C. Donnelly, A. Gelfan, D. Gerten, B. Arheimer, F. Hattermann, and Z. W. Kundzewicz, 2018: How the performance of hydrological models relates to credibility of projections under climate change. *Hydrological Sci. J.*, **63**, 696–720.

Laudan, L., 1984. *Science and values: The aims of science and their role in scientific debate*. Univ. of California Press.

Lemos, M. C., and B. J. Morehouse, 2005: The co-production of science and policy in integrated climate assessments. *Global Environmental Change*, **15**, 57–68.

Lenhard, J. and Winsberg, E., 2010: Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philos. of Sci. Part B*, **41**, 253–262.

Lloyd, E. A., 2009: Varieties of support and confirmation of climate models. *Aristotelian Soc. Supplementary Volume*, **83**, 213–232.

Lloyd, E. A., 2010: Confirmation and robustness of climate models. *Philos. Sci.*, **77**, 971–984.

Lloyd, E. A., 2015: Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philos. of Sci. Part A*, **49**, 58–68.

Lu, X., 2011: Provision of climate information for adaptation to climate change. *Climate Research*, **47**, 83–94.

Masson, D. and R., Knutti, 2011. Climate model genealogy. *Geophys. Res. Letters,* **38**, L08703.

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**, 675–691

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Maxim, L. and J. P. van der Sluijs, 2011: Quality in environmental science for policy: Assessing uncertainty as a component of policy analysis. *Environmental Sci. & Policy*, **14**, 482–492.

McWilliams, J.C., 2007. Irreducible imprecision in atmospheric and oceanic simulations. *Proceedings of the National Academy of Sciences*, **104**, 8709–8713.

Moss, R.H., and Coauthors, 2013: Hell and high water: practice-relevant adaptation science. *Sci.*, **342**, 696–698.

New M. G., M. Hulme, P. D. Jones 1999. Representing twentieth-century space time climate variability. Part I: development of a 1961–1990 mean monthly terrestrial climatology. *J. of Climate*, **12**, 829–856.

New M. G., M. Hulme, P. D. Jones 2000. Representing twentieth-century space time climate Fields. Part II: development of a 1901–1996 mean monthly terrestrial climatology. *J. of Climate*, **13**, 2217–2238.

Nissan, H., L. Goddard, E. C. de Perez, J. Furlow, W. Baethgen, M. C. Thomson, and S. J. Mason, 2019: On the use and misuse of climate change projections in international development. *Wiley Interdisciplinary Rev.: Climate Change*, **10**, e579.

Nissan, H., Á. G. Muñoz, and S. J. Mason, 2020: Targeted model evaluations for climate services: a case study on heat waves in Bangladesh. *Climate Risk Management*, 100213.

Oreskes, N., 1998: Evaluation (not validation) of quantitative models. *Environmental Health Perspectives*, **106**, 1453–1460.

Oreskes, N. and K. Belitz, 2001: Philosophical issues in model assessment. *Model validation: Perspectives in Hydrological Sci.*, Anderson, M. G. and Bates, P. D., Ed., Wiley and Sons, 23–41.

Parker, W. S., 2009: Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Soc. Supplementary Volume*, **83**, 233–249.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

Parker, W. S., 2010: Whose probabilities? Predicting climate change with ensembles of models. *Philos. Sci.*, **77**, 985–997.

Parker, W. S., 2011: When climate models agree: The significance of robust model predictions. *Philos. Sci.*, **78**, 579–600.

Parker, W. S., 2020: Model evaluation: An adequacy-for-purpose view. *Philosophy of Science,* **87**, 457–477.

Parker, W. S. and J. S. Risbey, 2015: False precision, surprise and improved uncertainty assessment. *Philos. Trans. Roy. Soc. A*, **373**, 20140453.

Parker, W.S. and E. Winsberg, 2018. Values and evidence: how models make a difference. *European J. for Philos. of Sci.*, **8**, 125–142.

Pielke Sr, R. A., and R. L. Wilby, 2012: Regional climate downscaling: What's the point?. *Eos, Trans. American Geophys. Union*, **93**, 52–53.

Pirtle, Z., R. Meyer, and A. Hamilton, 2010: What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Sci. & Policy*, **13**, 351–361.

Porter, T., 1995. *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton University Press.

Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill?, *Geophys. Res. Letters*, **36**, L13704, doi:10.1029/2009GL038082.

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

Risbey, J.S., P.J. Lamb, R.L. Miller, M.C. Morgan, and G.H. Roe, 2002. Exploring the structure of regional climate scenarios by combining synoptic and dynamic guidance and GCM output. *J. of Climate*, **15**,1036–1050.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Risbey, J.S. and T. J. O'Kane, 2011. Sources of knowledge and ignorance in climate research. *Climatic Change*, **108**, 755–773.

Risbey, J., J. van der Sluijs, P. Kloprogge, J. Ravetz, S. Funtowicz, and S. Corral Quintana, 2005. Application of a Checklist for Quality Assistance in Environmental Modelling to an Energy Model. *Env. Mod. & Assessment.* **10**, 63–79.

Risbey, J. S., and P. H. Stone, 1996. A Case Study of the Adequacy of GCM Simulations for Input to Regional Climate Change Assessments. *J. Climate.* **9**, 1441–1467.

Rummukainen, M., B., Rockel, L., Bärring, J. H., Christensen, & M. Reckermann, 2015. Twenty-first-century challenges in regional climate modeling. *Bulletin of the American Meteorological Society*, **96**, 135–138.

Sanderson, B. M., R. Knutti, and P. Caldwell, 2015: Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J. Climate*, **28**, 5150–5170.

Shackley, S. and B. Wynne, 1995. Integrating knowledges for climate change: pyramids, nets and uncertainties. *Global Environmental Change*, **5**, 113–126.

Shackley, S., P. Young, S. Parkinson, and B. Wynne, 1998. Uncertainty, complexity and concepts of good science in climate change modelling: are GCMs the best tools?. *Climatic change*, **38**, 159–205.

Singh, R. and K. AchutaRao, 2020. Sensitivity of future climate change and uncertainty over India to performance-based model weighting. *Climatic Change*, 1–22.

Smith, L. A., 2002: What might we learn from climate forecasts? *Proc. of the National Academy of Sci. of the United States of America*, **99**, 2487–2492.

Stainforth, D. A., M. R. Allen, E. R. Tredger, and L. A. Smith, 2007a: Confidence, uncertainty and decision-support relevance in climate predictions. *Philos. Trans. Roy. Soc. A*, **365**, 2145–2161.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

Stainforth, D. A., T. E. Downing, R. Washington, A. Lopez, and M. New, 2007b: Issues in the interpretation of climate model ensembles to inform decisions. *Philos. Trans. Roy. Soc. A*, **365**, 2163–2177.

Supiot, A., 2017. *Governance by numbers: The making of a legal model of allegiance*. Hart Publishing. 310 pp.

Tebaldi, C., L.O. Mearns, D. Nychka, and R.L. Smith, 2004. Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophys. Res. Letters*, **31**, L24213.

Tebaldi, C., R.L. Smith, D. Nychka, and L.O. Mearns, 2005. Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. of Climate*, **18**, pp.1524–1540.

Thompson, E. L., and L. A. Smith, 2019: Escape from model-land. *Economics: The Open-Access, Open-Assessment E-Journal*, **13**, 1–17.

USGCRP, 2018: *Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II*. Reidmiller, D.R., C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and B.C. Stewart, Eds., U.S. Global Change Research Program, Washington, DC, USA, 1515 pp.

Van Bree, L., and J. P. Van Der Sluijs, 2014: Background on uncertainty assessment supporting climate adaptation decision-making. *Adapting to an Uncertain Climate,* Capela Lourenço, T., Rovisco, A., Groot, M., Nilsson, C., Fuessel H. M., van Bree, L., Street, R. B., Eds., Springer. 17-40.

Vezér, M. A., 2016. Computer models and the evidence of anthropogenic climate change: An epistemology of variety-of-evidence inferences and robustness analysis. *Studies in History and Philosophy of Science Part A*, **56**, 95–102.

Weisberg, M., 2006: Robustness analysis. *Philosophy of Science,* **73**, 730–742.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Weisheimer, A. and T.N. Palmer (2014), On the reliability of seasonal climate forecasts. *J. R. Soc. Interface*, **11**, 20131162.

Weil, D., M. Graham, and A. Fung, 2013: Targeting transparency. *Science*, **340**, 1410–1411.

Winsberg, E., 2006. Models of success versus the success of models: Reliability without truth. *Synthese*, **152**, 1–19.

Winsberg, E., B. Huebner, and R. Kukla, 2014: Accountability and values in radically collaborative research. *Studies in History and Philosophy of Science Part A*, **46**, 16–23.

Winsberg, E., 2018. What does robustness teach us in climate science: a re-appraisal. *Synthese*, 1–24.

Wilby, R. L., Troni, J., Biot, Y., Tedd, L., Hewitson, B. C., Smith, D. M., and Sutton, R. T., 2009: A review of climate risk information for adaptation and development planning. *International Journal of Climatology*, **29**, 1193–1215.

Wimsatt, W. C., 1981: Robustness, reliability, and overdetermination. *Scientific Inquiry and the Social Sciences*, Brewer, M. and Collins, B., Eds., Jossey-Bass, 124–163.

Woodward, J., 2006. Some varieties of robustness. *Journal of Economic Methodology*, **13**, 219–240.

Zeng, Y., and Coauthors, 2019: Towards a traceable climate service: Assessment of quality and usability of essential climate variables. *Remote Sensing*, **11**, 1186. https://doi.org/10.3390/rs11101186

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

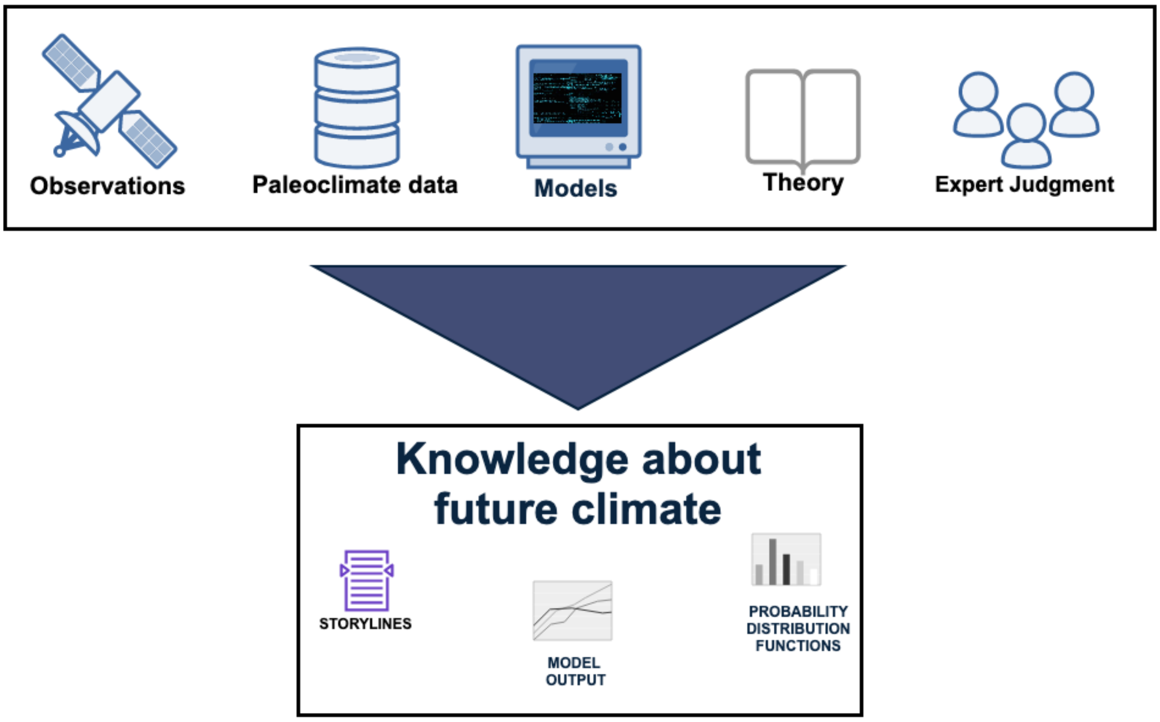


Fig. 1 Typology of evidence that can be used to support knowledge claims about future climate (top box) and selected ways in which knowledge claims about future climate can be presented (bottom box). The blue triangle indicates that relation between the evidence and the statement about future regional climate that contributes to the quality of knowledge about future regional climate.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0008.1.

Paper	Risbey et al. 2002. Exploring the structure of regional climate scenarios by combining synoptic and dynamic guidance and GCM output. *J. of Climate*	Tebaldi et al. 2004. Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophys. Res. Letters*
Statement or estimate about future climate	Qualitative statement: Small changes in large scale atmospheric dynamics can lead to large changes in regional climate in some regions and very small or no changes in other regions.	Probabilistic projection of precipitation change at seasonal and regional scales. 22 regions, under IPCC SRES A2 and B2 scenario for DJF and JJA. The information is presented in boxplots for precipitation change (as percentage) in each region
Evidence	Expert judgment of atmospheric scientists on effects of synoptic features of climate on region of expertise, both for winter and summer, to build "scenarios", where scenarios are possible futures assuming a more or less equivalent of doubling CO_2. NCEP reanalysis data. Available model output found in the literature: mostly (but not exclusively) AGCM output from NASA Goddard Institute for Space Studies and NOAA Geophysical Fluid Dynamics Laboratory.	Precipitation data from the observational dataset from the Climatic Research Unit (CRU) of the University of East Anglia (New et al. 1999, 2000) aggregated in seasonal and regional 30 year means (1961-1990). Precipitation data from multi-model ensemble output of 9 AOGCMs aggregated in seasonal and regional 30 year means (2070-2099)
Relationship between evidence and statement	Experts describe how synoptic features affect the seasonal cycle of regional precipitation and temperature. Experts then evaluate how the synoptic features and its relationship to regional precipitation and temperature may change in light of changes in GHG concentrations (“scenarios”). These scenarios for particular regions are subsequently compared with reanalysis data and with relevant AGCM output found in the literature.	Bayesian analysis: Model output and observational data is used to update priors (uniform distributions) to posteriors. The joint posterior probabilities are approximated through Markov chain Monte Carlo simulation. Posteriors are weighted by dividing by a measure of natural variability. The percentage precipitation change and the derived new values of natural variability are calculated. Model independence is calculated by estimating “model bias” and “model convergence” based on the Reliability ensemble average (REA) method of Giorgi and Mearns (2002). Assumption: ensemble average of projections is “best approximation” to truth and

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.

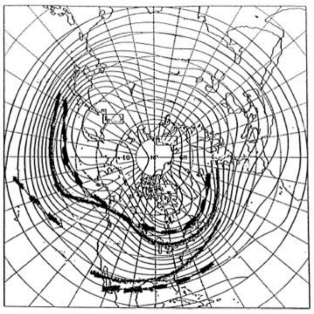
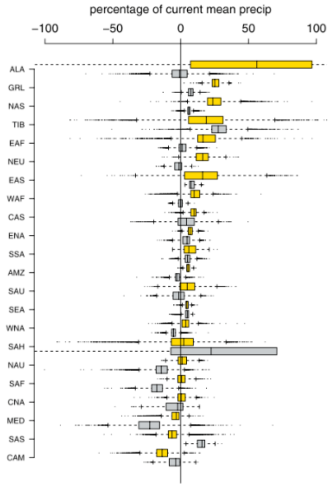
		bias is deviation of any one projection from the ensemble average.
Graphical representation of the statement or estimate	 This image shows how the current position of large scale features such as the wintertime polar and subtropical jet stream (thick solid line) can change location under a first guess climate change scenario (thick dashed line). Changes in location of large scale features influences regional climate.	 Probabilistic distribution of mean precipitation change for different regions for DJF (yellow/top) and JJA (grey/bottom), averaged over the A2 and B2 scenarios.

Table 1 Statements or estimates about future regional climate of Risbey et al. (2002) and Tebaldi et al. (2004), the evidence used for these statements, the relationship between the statements or estimates and the graphical representation of the information presented in these papers.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0008.1.