Climatic Change (2015) 132:47–60
DOI 10.1007/s10584-013-0813-5

On judging the credibility of climate predictions

Friederike E. L. Otto · Christopher A. T. Ferro · Thomas E. Fricker · Emma B. Suckling

Received: 12 January 2013 / Accepted: 30 May 2013 / Published online: 22 June 2013

Abstract Incorporating a prediction into future planning and decision making is advisable only if we have judged the prediction's credibility. This is notoriously difficult and controversial in the case of predictions of future climate. By reviewing epistemic arguments about climate model performance, we discuss how to make and justify judgments about the credibility of climate predictions. We propose a new bounding argument that justifies basing such judgments on the past performance of possibly dissimilar prediction problems. This encourages a more explicit use of data in making quantitative judgments about the credibility of future climate predictions, and in training users of climate predictions to become better judges of credibility. We illustrate the approach using decadal predictions of annual mean, global mean surface air temperature.

This article is part of a Special Issue on "Managing Uncertainty in Predictions of Climate and Its Impacts" edited by Andrew Challinor and Christopher Ferro.

Part of the EQUIP special issue of Climatic Change.

F. E. L. Otto (✉)
Environmental Change Institute, University of Oxford,
Oxford University Centre for the Environment,
South Parks Road, Oxford, OX1 3QY, UK
e-mail: friederike.otto@ouce.ox.ac.uk

C. A. T. Ferro · T. E. Fricker
College of Engineering, Mathematics and Physical Sciences,
University of Exeter, Harrison Building,
North Park Road, Exeter, EX4 4QF, UK

E. B. Suckling
Centre for the Analysis of Time Series, Tower One,
London School of Economics and Political Science,
Houghton Street, London, WC2A 2AE, UK

1 Introduction

Climate prediction centres produce a large variety and number of climate predictions and update them fairly frequently for coordinated efforts such as phase 5 of the Coupled Model Intercomparison Project (CMIP5, Taylor et al. 2012). There is also an increasing demand to incorporate information about future climate into planning decisions (e.g. HM Government 2012). While there is a strong consensus concerning recent climate change and the associated influence of human activities, worrying disagreements persist about the credibility of climate predictions. Many climate scientists have 'considerable confidence that climate models provide credible quantitative estimates of future climate change' (Randall et al. 2007, p. 600), while other authors argue that there is little justification for such claims in most cases (e.g. Parker 2010). In this paper, we encourage attempting to resolve this impasse by making more explicit use of available data to form quantitative predictions for the performance of climate predictions and thereby better informing decisions that depend on future climate.

Before allowing a prediction to influence our behaviour, we should judge the meaning and relevance of the prediction. We should also judge the credibility of the prediction, which is the focus of our discussion. In Section 2, we define what we mean by credibility and discuss how to judge credibility by predicting performance. In Section 3, we review two arguments for justifying predictions of performance before outlining a new argument that justifies quantitative bounds on expected performance. In Section 4, we discuss the role of different data sources in predicting performance of climate predictions. We illustrate our approach with predictions of annual mean global mean surface air temperatures in Section 5.

We close this introduction by defining some of our terminology. We consider a *prediction* to be a statement about the world. We refer to that which is predicted as the *predictand* and the creator of the prediction as the *predictor*. We distinguish forecasts and hindcasts: a *forecast* is a prediction that is issued before the predictand could be determined, while a *hindcast* is a prediction that is issued after the predictand could be determined, regardless of whether or not the predictand was in fact determined before the prediction was issued. Many predictions about future climate are referred to as *projections* (e.g. Collins 2007). One reason for this term is to emphasize that predictions of climate are often conditioned on specific future boundary conditions—such as the representative concentration pathways used for some CMIP5 experiments (Moss et al. 2010)—and that these boundary conditions are not intended as predictions but as plausible future scenarios. More recently, 'prediction' has been used to signify near-term (seasonal to decadal) simulations that are initialized with observations, while 'projection' has been used to signify uninitialized, multi-decadal simulations. The ideas presented below are relevant to all types of prediction and so we refer generically to 'predictions' throughout Sections 2 and 3, before discussing differences between near- and long-term climate predictions in Section 4.

2 Judging credibility

There are many factors that might influence our judgment of the credibility of a prediction. The track record and the theoretical or physical basis of the pre-

dictor are two common examples. We suggest, however, that such factors should affect our judgments if and only if they affect our expectations about how well the prediction will perform. We take this to be self-evident but feel that it helps to clarify what we must do to judge the credibility of a prediction. In line with this reasoning, we identify the credibility ascribed to a prediction with a prediction of its performance, and the act of judging credibility becomes the act of predicting performance. With this definition, credibility is not a binary characteristic, as in statements such as 'these predictions are credible'; rather, predictions may be more or less credible depending on how we expect them to perform.

The performance of a prediction is multi-faceted and our descriptions of performance should always be tailored to the decision problem at hand. In the case of a deterministic (point) prediction, performance might be described by some measure of the magnitude of its error. In the case of a probabilistic prediction, various performance measures are commonly used, such as proper scoring rules (e.g. Bröcker 2012). We may also wish to describe the performance of a set of predictions in terms of an aggregated measure such as a correlation coefficient or reliability statistic (e.g. Ferro and Fricker 2012; Fricker et al. 2013). Some decision makers may prefer predictions of performance to be expressed qualitatively, as in 'the error of this climate prediction will probably be small', while others may prefer quantitative predictions, as in 'the error of this climate prediction will be less than 1°C with probability 70 %'. However assessments are presented to decision makers, we believe that they should always be based on quantitative predictions of performance in order to be able to clarify ambiguous terms such as 'probable' and 'small', and thereby justify the guidance given.

One might judge credibility subconsciously, based on experience with the predictor and predictand, but we contend that there are situations in which making explicit, structured judgments is beneficial. Sometimes, predicting performance is equivalent to predicting the predictand. For example, a probabilistic prediction for the error of a deterministic prediction is equivalent to a probabilistic prediction for the predictand. (Suppose that the predictand is the outcome (heads or tails) of a coin toss, Ros issues the deterministic prediction 'heads', and Guy predicts the performance of Ros's prediction by specifying a probability for the event that Ros is correct. Ros is correct if and only if the coin shows heads, so Guy's prediction is equivalent to a prediction for the outcome of the coin toss.) This equivalence does not always hold, however, in which case predicting the values of a selection of performance measures can help to synthesize and summarize a large amount of information that is relevant to the decision problem.

Nevertheless, we could predict performance by forming our own prediction of the predictand, that is a prediction that reflects our personal beliefs about the predictand. If p is a prediction for a predictand x and $s(p, x)$ is a measure of performance then a prediction for x defines a prediction for $s(p, x)$. (For example, a prediction by Guy for the outcome of the coin toss defines his prediction for Ros being correct.) Indeed, this is the only possible approach to predicting performance if we have no information whatsoever about the provenance of the prediction under consideration. This approach is also appropriate, of course, if predicting performance is equivalent to predicting the predictand. In all other situations, however, predicting performance by forming our own prediction of the predictand is inappropriate for the following reasons.

First, we contend that allowing our own prediction of the predictand to be influenced by the prediction under consideration is inappropriate unless we have already judged its credibility. (If Guy wishes to take into account Ros's prediction of heads when he forms his own prediction for the outcome of the coin toss then he should first judge the credibility of Ros's prediction.) To avoid this circularity, we would have to disregard the prediction under consideration when forming our own prediction of the predictand. A consequence of this, however, is that we may fail to incorporate relevant information. To see this, suppose that Ros is always correct when she predicts heads, but is less often correct when she predicts tails. Suppose also that, before he knows Ros's prediction, Guy's prediction for the outcome of the next coin toss is that heads and tails are equally likely. On its own, Guy's prediction would lead him to predict that Ros has an even chance of being correct, whether she predicts heads or tails. (Guy predicts that heads will occur with probability 50 %, so that if Ros predicts heads then Guy predicts that she will be correct with probability 50 %, and similarly for tails.) This is unreasonable because Guy knows that Ros will be correct if she predicts heads. In most situations, therefore, we should seek to predict performance directly rather than solely via our own prediction of the predictand.

If the prediction under consideration, p, is our own prediction of the predictand, x, then we can and should still predict its performance. If p is deterministic then predicting its performance could amount to making our prediction probabilistic. If p is already probabilistic, however, then our prediction of its performance is already determined by p: our prediction for the performance measure $s(p, x)$ is defined by our prediction, p, of x. This is also the reason why predicting performance does not lead to a regress: we do not need to predict the performance of our prediction of the performance of p because this latter can already account for all of our relevant uncertainties.

3 Justifying predictions of performance

3.1 Construction-based arguments

We argued in the previous section that in order to judge the credibility of a prediction we should predict its performance. As with any non-trivial prediction, we should not expect our predictions of performance to be perfect, but we should aim to minimize errors. To this end, we scrutinize now the arguments that we might employ to justify predictions of performance.

One way to justify a prediction of performance might be to argue that performance can be deduced logically from facts or assumptions about the predictor and predictand without reference to the past performance of the predictor. This is called a design- or construction-based argument by Parker (2010, 2011).

Construction-based arguments are valid in some circumstances. If the performance measure is independent of the predictand then performance can be deduced from the prediction alone. Examples include criteria that check the prediction for self-contradictions (e.g. daily minimum temperatures exceeding daily maximum temperatures), or that assess the confidence (e.g. sharpness) of the prediction.

If the performance measure does depend on the predictand then additional information must be used in order to justify a prediction of performance. One such construction-based argument that has been discussed in the context of climate predictions concerns the 'perfect model scenario' (Smith 2002, 2006). The argument here is that an ensemble of predictions will be reliable if the ensemble-generating model is structurally similar to the predicted system in all relevant aspects and if its inputs are sampled to represent the uncertainty (e.g. the known distribution of measurement error) in the initial conditions of the predicted system. See also Betz (2006, ch. 9) and Stainforth et al. (2007). Which aspects of the predicted system are relevant depends on the prediction problem, so a model may be adequate for one type of prediction but not for others. Some authors, however, consider the structure of today's climate models to be insufficiently similar to that of the climate system for such an argument to provide a strong justification of reliability for most prediction problems (e.g. Betz 2006; Frame et al. 2007; Parker 2010). Smith (2006) doubts that weather and climate models sufficiently isomorphic to the climate system will ever be realized. See Parker (2010), Allen et al. (2006) and Otto (2012) for similarly negative conclusions regarding arguments based on 'imperfect model scenarios'.

Unless rather trivial measures of performance are of interest, it seems that construction-based arguments are unable to justify predictions for the performance of most climate model predictions because climate models are imperfect and it is difficult to trace the effects of imperfections through complex systems. This applies equally to predictions of good and bad performance—construction-based arguments do not provide strong justification for predictions of poor performance—and so we must look to other arguments.

3.2 Performance-based arguments

Another way to justify a prediction of performance might be to argue that performance can be extrapolated from information about the past performance of the predictor. This is called a performance-based argument by Parker (2010).

Performance-based arguments proceed by identifying a class of predictions that contains the prediction, p, whose performance is under consideration, such that one has no reason to believe in advance that any particular prediction in the class will perform better than any other prediction in the class. The performances of a sample of predictions in the class are then measured and the performance of p is subsequently inferred following standard statistical procedures.

The key stage in performance-based arguments is to justify the choice of reference class. Membership of the class should be determined by characteristics of the prediction problems, derived from knowledge about both the predictor and the predictand. For example, if future conditions are expected to differ from past conditions only in ways that are unlikely to affect the performance of predictions then past predictions and future predictions may be judged to belong to the same class.

Performance-based arguments are familiar in weather forecasting, where past performance is often taken to be a face-value indicator of future performance. Parker (2010) contends that such arguments fail to justify predictions for the performance of climate predictions, however, because past cases whose performances can be measured differ significantly from predictions into the future. In particular, the relevance

of hindcasts is compromised by the possibility that climate models have been tuned to perform well in the hindcast period, and the future forcings (such as concentrations of greenhouse gases) prescribed in climate predictions differ significantly from the forcings prescribed in the hindcasts. The studies by Reifen and Toumi (2009) and Weigel et al. (2010) attest to changes in performance over time. See also Frame et al. (2007).

While performance-based arguments may not provide strong justifications for predictions of the performance of climate predictions, there does seem to be scope for them to provide some justification. Even a small number of past predictions similar to the future predictions under consideration can help to give a rough indication of likely performance, and this may be very informative if the indication differs markedly from any prior expectations of performance.

3.3 Bounding arguments

In this subsection we propose a third type of argument for justifying quantitative predictions of the performance of climate predictions. While there may be few similar prediction problems from which we can infer the performance of climate predictions via the performance-based arguments of the previous subsection, there are many other climate-model experiments providing data that are commonly used to justify judgments about credibility. We might judge that some of these experiments are sufficiently similar to one another to form a reference class, and the statistical properties of the performance of members of the class can then be inferred from a sample of cases, as before. Even if the prediction whose performance is in question is not judged to be a member of this class, we might be willing to judge that it is either a harder or easier prediction problem than those in the reference class. In other words, we expect the performance of the prediction in question to be either worse or better than the performance of randomly selected members of the reference class. In this case, the inferred performance for the members of the reference class provides an upper or lower bound on the performance of the prediction in question.

To be precise, let S denote the performance of a randomly selected prediction from a reference class C and suppose that the value of S must lie on the positive real line with smaller values indicating better performance. Suppose also that we have used a sample of cases from C to estimate the probability distribution of performances in C. Let this distribution be denoted by the cumulative distribution function $F(s) = \Pr(S \leq s)$ for all $s > 0$. Now let S' denote the performance of a randomly selected prediction from a class C' that contains the prediction under consideration, and let $F'(s) = \Pr(S' \leq s)$ denote the unknown distribution of performances in C'. If we judge that the prediction problems in C' are harder than the prediction problems in C then we obtain the bound $F'(s) < F(s)$ for all $s > 0$, i.e. the chance of the prediction in question achieving a performance as good as s is at most $F(s)$. Similarly, if we judge that the prediction problems in C' are easier than the prediction problems in C then we obtain the bound $F'(s) > F(s)$ for all $s > 0$, i.e. the chance of the prediction in question achieving a performance as good as s is at least $F(s)$. We shall give numerical examples of such bounds in Section 5. Simultaneous upper and lower bounds may be obtained if both a harder reference class and an easier reference class can be identified.

The key stage in the performance-based argument in the previous subsection is to justify the choice of reference class. Such a justification is also needed here to define the bounding reference class, but now the class need not include the prediction under consideration, and so there is more scope to identify such classes. On the other hand, the bounding argument proposed here also requires us to justify a judgment that the prediction problem of interest is either harder or easier than the prediction problems in the bounding reference class. This is a similar type of judgment to that used to define reference classes as it should be based on similar information, namely characteristics of the prediction problems. However, the judgment that a prediction problem is harder or easier than other problems is a stronger judgment than that required to define a reference class because the direction of departure from the reference class must be specified. Furthermore, some of the characteristics of the prediction problem under consideration may suggest that the problem is easier than those in the reference class, while other characteristics may suggest that the problem is harder. A further complication is the possibility that some levels of performance may be judged to be harder to achieve for the prediction in question than for the predictions in the reference class, while other levels of performance may be judged to be easier to achieve. In such circumstances, the ordering of the probabilities $F(s)$ and $F'(s)$ would be judged to vary with s. Justifying such detailed judgments is likely to be difficult, at least for climate predictions. There is no straightforward solution to these complications but we reiterate that we should not expect perfect predictions of performance; we merely seek to improve current predictions, and believe that the simple harder/easier judgment proposed above has the potential to meet this goal in the case of climate predictions at least.

4 Empirical evidence

In this section, we discuss some of the data that may be used to define bounding reference classes, and how different classes might relate to climate prediction problems in terms of their difficulty. Predictors such as climate models that are used to forecast the future climate are also used to produce a variety of other predictions. For example, climate models are used to generate out-of-sample forecasts and in-sample hindcasts, both of the real world and, in perfect and imperfect model experiments, of other simulations. These predictions are made for assorted predictands, at a range of lead times, and under an array of initial and boundary conditions. The performance of such predictions is commonly cited as evidence for the expected performance of climate forecasts (e.g. Randall et al. 2007). If we can group these predictions into reference classes and judge how they relate to the climate prediction problems of interest then they can inform our expectations about climate forecasts and help us to make more quantitative predictions for their performance.

The more similar a reference class is to the prediction of interest, the easier it is likely to be to justify the relationship between the two, and the tighter the bound on performance is likely to be. Such reference classes can act as strong guides for our expectations. Indeed, the performance-based argument of Section 3.2 may be viewed as a limiting case of our bounding argument as the reference class becomes more and more similar to the prediction of interest. This reasoning suggests forming reference classes of predictions with the same predictand as the prediction

of interest. For example, if we wish to bound the performance of predictions of global mean, annual mean surface air temperature, then we might do well to form reference classes from other experiments in which this quantity is simulated. For the same reason, bounding arguments are likely to be more effective if the reference classes comprise predictions issued by the predictor whose predictions are under consideration, rather than predictions issued by another predictor. Predictions from other predictors might be useful, nevertheless, particularly if the predictors form a hierarchy with the predictor in question. For example, it may be possible to argue that a higher resolution version of a climate model is expected to perform better than a lower resolution version of the same model. This reasoning also suggests that we should design hindcast experiments to be as similar as possible to climate forecasts by mirroring the extent of tuning possible for out-of-sample forecasts. This might include running perturbed-physics ensembles to detune climate models, and cross-validating performance measures in case model output is bias corrected or otherwise post-processed.

Although more similar reference classes will tend to yield tighter bounds, we should also collect data on as wide a range of prediction problems as possible. Very dissimilar reference classes may have little impact on our expectations for performance, but the information that they provide can do no harm. More data can only sharpen our bounds, and if we encounter a reference class whose performance is unexpectedly good or bad then this may radically alter our expectations about the performance of the prediction of interest. Ideally, therefore, we should conduct experiments with predictors such as climate models that explore the full range of physical processes that may influence forecasts of future climate, and a wide range of initial conditions and boundary conditions. The range of conditions that can be explored is limited when predictions of the real world are made, but perfect and imperfect model experiments have a useful role in a fuller testing of climate models.

Reference classes may provide either upper or lower bounds on our expectations for performance. Reference classes of prediction problems that are harder than climate predictions of interest might be formed from hypothetical, worst-case predictions (e.g. predictions of unphysical values) or possibly from climate models that are known to be fundamentally inadequate. It may be difficult, though, to identify harder reference classes that provide very tight bounds on our expectations. Even if reference classes of only easier prediction problems are available, however, this can still help to prevent over-confidence (Frame et al. 2007).

One example of an easier reference class might be predictions in a perfect model experiment, if we expect a climate model to perform better when predicting its own simulations than when predicting the real world. Whether imperfect model experiments, on the other hand, produce predictions that are harder or easier than predictions of the real world is less clear. Climate models are structurally more similar to one another than they are to the real climate system (Smith 2002), but new studies (e.g. Knutti et al. 2013) indicate that simulations produced by the current generation of climate models can be less similar to one another than they are to observations of the real world. The genealogy of climate models (Masson and Knutti 2011) may provide useful information when deciding if imperfect model predictions yield harder or easier reference classes.

Out-of-sample predictions might be considered harder than in-sample predictions, with the relative difficulty being related to the extent to which the predictor is tuned to the in-sample prediction problems and how far out of sample the out-of-sample predictions are. This point is germane to long-term climate predictions: we might expect a climate model to perform worse when it is forced by boundary conditions that differ significantly from those conditions under which the model has been tested and developed. However, this judgment should also account for the possibility that the model may be known to respond accurately to certain changes in forcing, or may be forced into a state where it tends to perform well. Predictions with long lead times will often be considered harder than predictions with short lead times, although some account must be taken of the physical processes involved and how well these processes are represented by the predictor. In some applications, it may also be known that some initial states correlate with better predictive performance. These are the sort of considerations that help to define bounding relationships between reference classes and thereby help to bound expectations of performance.

Another feature of climate predictions that is often used to justify their credibility is agreement among different predictors (e.g. Knutti 2008). Smith (2006) and Parker (2011) argue that this justification is weak, for the same reasons that construction- and performance-based arguments are weak. The value of model agreement could be investigated empirically by forming reference classes of predictions for which there exist different levels of agreement. If agreement is indicative of performance then these reference classes can be used to bound the expected performance of other predictions.

The data discussed above are already used to judge the credibility of climate predictions. We have merely argued for a more explicit use of the data to obtain quantitative bounds on expected performance. This approach forces us to interrogate our assumptions and can help to identify gaps in our understanding. Bounding arguments still rely significantly on subjective judgment to form reference classes and to specify their performance relationships, and so they are still susceptible to errors of judgment. Importantly, however, we can train ourselves to become better judges of these matters using existing climate experiments. For example, we can define two reference classes of prediction problems, say one from a perfect model experiment and one from an imperfect model experiment, and decide which class we think comprises the harder prediction problems. Then we can measure the performance of predictions in the two classes and check if our judgment was accurate. We illustrate this in the next section using decadal predictions of global mean annual mean temperature as an example. The extent to which such training helps us to improve our judgments about the performance of specific prediction problems may depend on whether the factors that cause us to make errors of judgment are similar in both the training and target problems. Once again, therefore, the more similar the training problems are to the predictions of interest, the more useful they may be for improving our judgments. In the specific case of long-term climate predictions, for which we may feel that we have no similar training problems, examples of shorter term predictions or of imperfect model predictions in which the effects of unresolved processes are critical may help us to develop a better intuition for the impacts on performance of such model limitations. More generally, a wide range

of training problems may help to correct any general disposition towards over- or under-confidence in performance.

5 Illustrative numerical examples

We illustrate the bounding arguments outlined above with decadal climate predictions from the Met Office general circulation model HadCM3 (Gordon et al. 2000) that form part of CMIP5. An ensemble of ten hindcasts, formed by perturbing the initial conditions, is launched every year during the hindcast period, 1960 to 2000. We analyze annual mean, global mean surface air temperature anomalies and measure the performance of ensemble members with their absolute errors. We also use an ensemble member from the Météo France model CNRM-CM5 (Voldoire et al. 2013) and the observed temperature record HadCRUT3 (Brohan et al. 2006) as verifications in some of our examples. Anomalies for each time series are calculated with respect to its own mean temperature, which is estimated from the temperature values for the first year of each hindcast launch, from 1960 to 1990. Figure 1a displays a subset of the hindcasts.

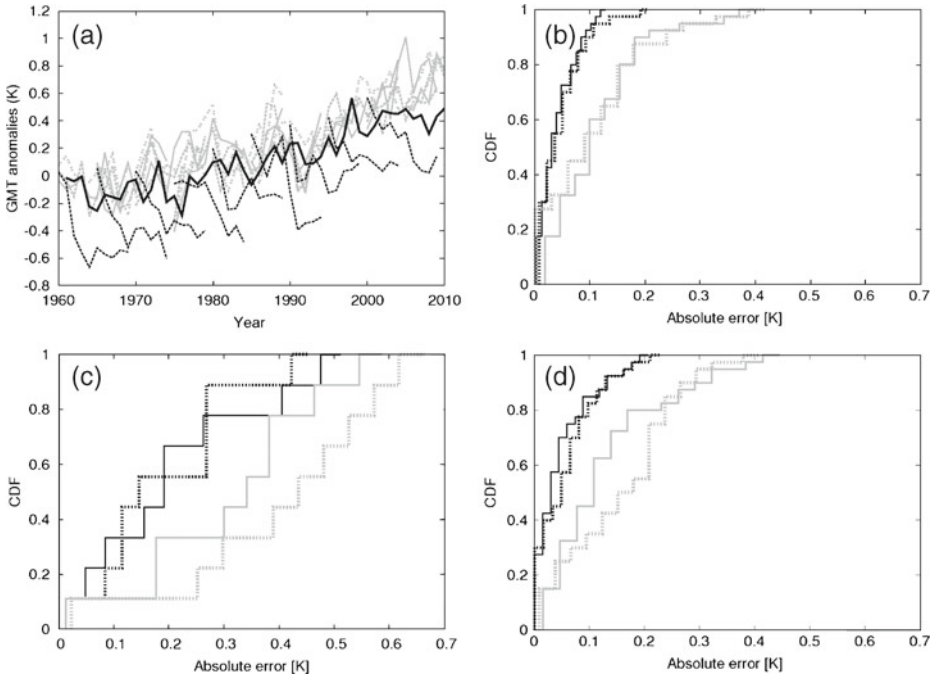

Fig. 1 **a** Annual mean, global mean surface air temperature anomalies: HadCRUT3 (*solid black*), CNRM-CM5 (*dashed black*) and HadCM3 members 1, 2 and 3 (*grey*) launched every five years; **b** c.d.f.s of the absolute errors for HadCM3 members 2 (*dashed*) and 3 (*solid*) at lead times of one year (*black*) and ten years (*grey*) when HadCM3 member 1 is the verification; **c** as **b** but CNRM-CM5 is the verification; **d** as **b** but HadCRUT3 is the verification

We begin with an illustration of a performance-based argument rather than a bounding argument. We consider a perfect model experiment in which we investigate how well HadCM3 ensemble members predict other members of the same ensemble. We take member 1 to be the verification, consider a lead time of one year, and define a reference class of predictions to comprise the 369 predictions that span the 41-year hindcast period from the nine other members. Now we measure the 41 absolute errors for member 2, whose cumulative distribution function (c.d.f.) is plotted in Fig. 1b (dashed black line). The c.d.f. depicts the proportion of errors that are no larger than the value on the horizontal axis. For example, about 90 % of the errors are less than 0.1 K for member 2 in Fig. 1b (dashed black line). If we were interested now in the performance of member 3 then we could use the c.d.f. for member 2 as a prediction for the performance of member 3 because we have judged that members 2 and 3 belong to the same reference class. In this example, moreover, we can check if this judgment is accurate by calculating the c.d.f. of the 41 errors for member 3 and comparing it to the c.d.f. for member 2. The c.d.f. for member 3 is also shown in Fig. 1b (solid black line) and is indeed similar to the c.d.f. for member 2, thus showing that our prediction of performance of these one-year ahead climate predictions is accurate.

Now we illustrate a bounding argument. Suppose that we are interested in the performance of member 3 at a lead time of ten years. How is our prediction of its performance guided by the c.d.f. of absolute errors for member 2 at a lead time of one year? Our expectation might be that forecasting ten years ahead is a harder prediction problem than forecasting one year ahead. (It is possible, for some states and forcing of the climate system, that we might expect the opposite, but in most cases we expect ensemble members to diverge and errors to grow with lead time.) In this case, we would take the c.d.f. for member 2 at one year in Fig. 1b to be an upper bound on the c.d.f. corresponding to our prediction of the errors of member 3 at ten years. Figure 1b shows that this expectation would be correct: the errors at ten years (grey lines) tend to be greater than the errors at one year (black lines). The bound has provided a useful guideline for our expectations.

Our next example of a bounding argument considers an imperfect model experiment in which we use the CNRM-CM5 ensemble member as the verification. We might expect the performance of HadCM3 member 3 at a given lead time to be worse in the imperfect model scenario than in the perfect model scenario, assuming that model formulation is the dominant difference between the simulations. Comparing the solid lines in Fig. 1c with the corresponding solid lines in Fig. 1b shows that, for either lead time, the errors do indeed tend to be larger in the imperfect model scenario, and so our approach has again provided a correct bound for our expectations of performance.

Our final example of a bounding argument uses the HadCRUT3 observations for the verification and corresponds to bounding the expected performance in the real world, rather than in a perfect or imperfect model example. For each lead time, we might expect HadCM3 to perform worse in this situation than in both the perfect and imperfect model scenarios. Comparing the lines in Fig. 1b–d shows that this expectation is wrong for both lead times: the errors tend to be smaller when hindcasting reality compared to the imperfect model scenario but slightly larger than in the perfect model scenario. This is in line with the findings of Knutti et al. (2013) mentioned in Section 4. Even if we had formed such mistaken expectations in this

case, the bound would nonetheless have prevented us from being overconfident: we would have assumed that things would be worse than they actually are and thus applied a cautionary approach (Frame et al. 2007). Moreover, this new evidence might cause us to revise future judgments, particularly if we can determine why our original expectation was misguided.

From these examples, we can see the potential difficulties of correctly predicting performance, even if we aim only to bound our expectations. For example, we found that performance was worse in the imperfect model scenario than when hindcasting reality. If we wanted to predict the performance of *future* predictions of annual mean, global mean surface air temperatures from HadCM3, this experience might make us wary of using the results of imperfect model experiments to bound our expectations. On the other hand, we might be willing to adopt the c.d.f.s from the perfect model scenario as upper bounds on our expectations for the performance at the two lead times.

6 Discussion

We have noted the importance of being able to justify quantitative predictions for the performance of predictions. We have also agreed with other authors that construction- and performance-based arguments provide little justification in the case of future climate predictions. Instead, we have proposed bounding arguments as one possible way of using data more explicitly to justify quantitative predictions for the performance of future climate predictions. Our examples have shown the difficulty of predicting performance correctly, and that the confidence with which we make bounding arguments may depend on various features of the prediction problem, such as the predictand, predictor, post-processing method, and performance measure. We have recommended using existing experiments to train ourselves to be better judges of performance. Whenever our expectations are overthrown, we should try to understand the causes of our misjudgment. Not only will this lead us to make better judgments in the future, but this can also prompt us to investigate ways in which our climate prediction systems might be improved.

While we hope that bounding arguments demonstrate the potential to use data more effectively, the bounding approach that we have presented is, nonetheless, rather simple-minded. We envisage developing these ideas to make fully probabilistic judgments about performance, possibly along the following lines. If S denotes the performance of a prediction in a class C then we might select a probability model $F(s \mid \theta)$ to represent its distribution, where θ is a model parameter. We might also describe our uncertainty about the performance, S', of the prediction of interest with the distribution $F(s \mid \phi\theta)$, where ϕ is a second parameter that characterizes the difference between the distributions of the performances of S and S'. If we represent our prior beliefs about θ and ϕ with probability density functions g and h, and if we measure the performances, $D = (s_1, \ldots, s_n)$, from a sample of n cases from C, then the predictive distribution for S' can be expressed as

$$\Pr(S' \leq s \mid D) = \int F(s \mid \phi\theta) g(\theta \mid D) h(\phi) \, \mathrm{d}\theta \, \mathrm{d}\phi.$$

A key part of such an approach is the specification of our beliefs about ϕ. Bounding arguments may help us to formulate these beliefs and thus render explicit our implicit assumptions about a prediction's credibility.

Acknowledgements We thank Ed Hawkins, Reto Knutti, David Stainforth and an anonymous reviewer for their helpful comments on earlier versions of this paper. This work was funded by Natural Environment Research Council Directed Grant NE/H003509/1.33.

References

Allen M, Frame D, Kettleborough J, Stainforth D (2006) Model error in weather and climate forecasting. In: Palmer T, Hagedorn R (eds) Predictability of weather and climate. Cambridge University Press, pp 391–427

Betz G (2006) Prediction or prophecy? The boundaries of economic foreknowledge and their socio-political consequences. Deutscher Universitats-Verlag

Bröcker J (2012) Probability forecasts. In: Jolliffe IT, Stephenson DB (eds) Forecast verification: a practitioner's guide in atmospheric science, 2nd edn. John Wiley & Sons, Ltd, Chichester, pp 119–139

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. J Geophys Res 111:D12106. doi:10.1029/2005JD006548

Collins M (2007) Ensembles and probabilities: a new era in the prediction of climate change. Philos Trans R Soc A 365:1957–1970. doi:10.1098/rsta.2007.2068

Ferro CAT, Fricker TE (2012) A bias-corrected decomposition of the Brier score. Q J R Meteorol Soc 138:1954–1960. doi:10.1002/qj.1924

Frame DJ, Faull NE, Joshi MM, Allen MR (2007) Probabilistic climate forecasts and inductive problems. Philos Trans R Soc A 365:1971–1992. doi:10.1098/rsta.2007.2069

Fricker TE, Ferro CAT, Stephenson DB (2013) Three recommendations for evaluating climate predictions. Meteorological Applications. In press.

Goddard L, Kumar A, Solomon A, Smith D, Boer G, Gonzalez P, Deser C, Mason SJ, Kirtman BP, Msadek R, Sutton R, Hawkins E, Fricker T, Kharin S, Merryfield W, Hegerl G, Ferro CAT, Stephenson DB, Meehl GA, Stockdale T, Burgman R, Greene AM, Kushnir Y, Newman M, Carton J, Fukumori I, Vimont D, Delworth T (2013) A verification framework for interannual-to-decadal predictions experiments. Clim Dyn 40:245–272. doi:10.1007/s00382-012-1481-2.

Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, Johns TC, Mitchell JFB, Wood RA (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16:147–168. doi:10.1007/s003820050010

HM Government (2012) UK climate change risk assessment: government report. The stationery office. http://www.defra.gov.uk/environment/climate/government/risk-assessment/. Accessed 20 Aug 2012

Knutti R (2008) Should we believe model predictions of future climate change? Philos Trans R Soc A 366:4647–4664. doi:10.1098/rsta.2008.0169

Knutti R, Masson D, Gettelman A (2013) Climate model genealogy: generation CMIP5 and how we got there. Geophys Res Lett 40:1194–1199. doi:10.1002/grl.50256

Masson D, Knutti R (2011) Climate model genealogy. Geophys Res Lett 38:L08703. doi:10.1029/2011GL046864

Meehl GA, Goddard L, Murphy J, Stouffer RJ, Boer G, Danabasoglu G, Dixon K, Giorgietta MA, Greene AM, Hawkins E, Hegerl G, Karoly D, Keenlyside N, Kimoto M, Kirtman B, Navarra A, Pulwarty R, Smith D, Stammer D, Stockdale T (2009) Decadal prediction—can it be skillful? Bull Am Meteorol Soc 90:1467–1485. doi:10.1175/2009BAMS2778.1

Moss RH et al (2010) The next generation of scenarios for climate change research and assessment. Nature 463:747–756. doi:10.1038/nature08823

Otto FEL (2012) Modelling the earth's climate—an epistemic perspective. Dissertation, Freie Universität Berlin

Parker WS (2010) Predicting weather and climate: uncertainty, ensembles and probability. Stud Hist Philos Modern Phys 41:263–272. doi:10.1016/j.shpsb.2010.07.006

Parker WS (2011) When climate models agree: the significance of robust model predictions. Philos Sci 78:579–600. doi:10.1086/661566

Randall DA et al (2007) Climate models and their evaluation. In: Solomon S et al (eds) Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 590–662

Reifen C, Toumi R (2009) Climate projections: past performance no guarantee of future skill? Geophys Res Lett 36:L13704. doi:10.1029/2009GL038082

Smith LA (2002) What might we learn from climate forecasts? Proc Natl Acad Sci 99:2487–2492. doi:10.1073/pnas.012580599

Smith LA (2006) Predictability past, predictability present. In: Palmer T, Hagedorn R (eds) Predictability of weather and climate. Cambridge University Press, pp 217–250

Stainforth DA, Allen MR, Tredger ER, Smith LA (2007) Confidence, uncertainty and decision-support relevance in climate predictions. Philos Trans R Soc A 365:2145–2161. doi:10.1098/rsta.2007.2074

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498. doi:10.1175/BAMS-D-11-00094.1

Voldoire A, Sanchez-Gomez E, Salas y Mélia D, Decharme B, Cassou C, Sénési S, Valcke S, Beau I, Alias A, Chevallier M, Déqué M, Deshayes J, Douville H, Fernandez E, Madec G, Maisonnave E, Moine M-P, Planton S, Saint-Martin D, Szopa S, Tyteca S, Alkama R, Belamari S, Braun A, Coquart L, Chauvin F (2013) The CNRM-CM5.1 global climate model: description and basic evaluation. Clim Dyn 40:2091–2121. doi:10.1007/s00382-011-1259-y

Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of model weighting in multimodel climate projections. J Climate 23:4175–4191. doi:10.1175/2010JCLI3594.1