Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics
Discussions
Open Access
EGU

On the meaning of independence in climate science

James Annan and Julia Hargreaves

BlueSkiesResearch.org.uk
Settle, UK

Correspondence to: James Annan (jdannan@blueskiesresearch.org.uk)

Abstract. The concept of independence has been frequently raised in climate science, but has rarely been defined and discussed in a theoretically robust and quantifiable manner. Improved understanding of this topic is critical to better understanding of climate change. In this paper, we introduce a unifying approach based on the statistical definition of independence, and illustrate with simple examples how it can be applied to practical questions.

1 Introduction

Approximately 30 climate models contributed to the CMIP database, and they all agree, at least on broad statements: the world is warming, anthropogenic emissions of CO_2 is the major cause of this, and if we continue to emit it in large quantities then the world will continue to warm at a substantial rate for the foreseeable future (Stocker et al., 2013). The consensus across models is also strong for more detailed statements regarding for example the warming rates of land versus ocean, high versus low latitudes, and the likely changes in precipitation over many areas. Even where models disagree qualitatively amongst themselves (for example, concerning changes in ocean circulation and some regional details of precipitation patterns), their range of results is still quantitatively limited. Climate models are probably the most widely-used tool for predicting future climate changes, and their spread of results is commonly used as an indication of what future changes might occur in our climate.

But should the consensus between models really lead to confidence in these results? If we were to re-run the same scenario with the same model 30 times, we would get the same answer 30 times, whether it be a good or bad model. This repetition of results would not tell us how good the model is, and the behaviour of the real climate system would inevitably lie outside this (empty) range of results. Different model development teams share code, and even when code is rewritten from scratch, algorithms and methods are often copied, and many fundamental theories are common across all models (Knutti et al., 2013). So how much confidence can we draw from the fact that multiple models provide consistent answers? How likely is it that all models are biased in the same ways, and thus too similar for their spread of results to be useful? These questions have proved difficult to answer, and indeed there appears little consensus as to how we can even address them. Further questions arise from the increasingly common situation where a single modelling centre contributes multiple model simulations to the CMIP archive, some of which only differ in terms of the settings of uncertain parameters, or even just the initial state of the atmosphere/ocean system. A common heuristic when performing multi-model analyses based on a generation of the CMIP ensemble is to use a single simulation from each modelling centre, but it is not clear where to draw the line when different

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics Discussions
Open Access
EGU

centres may share a common core or sub-models. Is there a better way to select models, and should we use a weighted ensemble — in which case, further questions arise as to how the weights should be defined, in terms of either model performance relative to observations of the real climate, or else in terms of their relationship to other models?

Another important question that has been posed in recent years, is whether the scientific community could design or select ensemble members in a more rational and scientifically defensible way than the current ad-hoc 'ensemble of opportunity'. It may be possible to address this issue in terms of statistical sampling and experimental design, but appropriate methods and even language do not yet appear to be well developed in this area.

A second motivation (which, as we shall see, is distinct but related) for considering the question of independence is in synthesising various constraints on the equilibrium climate sensitivity. The equilibrium climate sensitivity S represents the equilibrium change in global mean surface temperature following a doubling of atmospheric CO_2, and while this is far from a complete description of our future climate, it is commonly used as an indication of the magnitude of changes. Different approaches have been proposed for constraining S, for example using data drawn from the modern instrumental period (which we refer to for convenience as the 20th century, although the data available does extend a little into the 19th and 21st centuries), or looking to the paleoclimate record and particularly the Last Glacial Maximum, or searching for constraints that emerge when process studies consider how well different models simulate various aspects of the climate system. Are these analyses 'independent', and how do we synthesise these different approaches into an overall estimate (Annan and Hargreaves, 2006)?

In this paper, we consider these questions and discuss how they may be addressed. We present an approach which links the usage in climate science to the statistical definition of independence, We do not claim to provide a full and final solution, but hope that the ideas and theory presented here may form the basis for future progress. We start by reviewing in Section 2 how the concept of independence has been discussed in the recent literature. In Section 3 we present a theoretical and statistical viewpoint of independence within the Bayesian paradigm, which we argue has direct relevance to this question. We consider how this statistical viewpoint relates both to the question of model independence, and also to the independence of constraints on climate system properties. In this section, we also sketch out some plans for how to make practical use of these ideas.

2 The literature on independence in climate research

The question of independence has featured widely in climate research, but the research community has not yet arrived at a clear and unambiguous definition. Different authors have approached the question of independence in different ways, and these approaches are often mutually inconsistent. In this section, we firstly explore the literature as it pertains to climate model independence, and then consider the literature on independence of constraints on climate sensitivity.

2.1 The literature on model independence

While the concept of model independence has been frequently discussed in recent years, a clear consensus of understanding has not yet emerged from the literature.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics Discussions
Open Access
EGU

One common approach has been to interpret model independence as meaning that the models can be considered as having errors which are independent, identically distributed (i.i.d. in common statistical parlance) samples drawn from some distribution with zero mean (Tebaldi and Knutti, 2007). This is the so-called 'truth-centred' hypothesis. An ensemble of samples drawn from such a distribution would be an incredibly powerful tool. If we could build models with these properties, we could generate arbitrarily precise statements about the climate, including future climate changes, merely by proceeding with the model-building process indefinitely and using the ensemble mean, without the need for any additional understanding of how to accurately simulate the climate system. For example, if the 19 CMIP3 models listed in Table 8.2 of Randall et al. (2007) provided independent estimates (in this sense) of the equilibrium climate sensitivity S, then we could immediately generate a 95% confidence interval for the real value for S of 3.2 ± 0.3°C (based on the first order assumption that the samples are drawn from a Gaussian distribution of unknown variance).

However, the truth-centred hypothesis is clearly refuted by numerous analyses of the ensemble. In particular, the errors of different models are observed to be strongly related, as can be shown by the positive correlations between spatial patterns of biases in climatology (Knutti et al., 2010, Figure 3). As a corollary of this, although the mean of the model ensemble generally outperforms most if not all the ensemble's constituent models (Annan and Hargreaves, 2011b), it does not actually converge to reality as the ensemble size grows. Rather, the ensemble mean itself appears to have a significant bias. There have been some attempts to compensate for this shared bias, for example by estimating the number of 'effectively independent' models contained in the full ensemble (Jun et al., 2008b, a; Pennell and Reichler, 2010). However, the theoretical basis for these calculations does not appear to be clearly justified, and the results presented have startling implications. If we accept the arguments of Pennell and Reichler (2010) that the CMIP3 ensemble contains 8 'effectively independent' models then their full range of sensitivity values, 2.1–4.4°C, would still be a valid 99% confidence interval, as the probability of 8 independent (in this sense) estimates all lying either below or above the truth is only 1 part in 2^7. The same argument would apply to any other output or derived parameter of the model climates. That is, we should be extremely confident that the model ensemble bounds the behaviour of the climate system generally. This conclusion does not seem very realistic, which must lead us to question the validity of the assumptions underlying such analyses.

Abramowitz and Gupta (2008) define independence purely in terms of inter-model differences and suggest down weighting models that are too similar in outputs. This approach has the weakness that models that agree *because they are all accurate* will be discounted, relative to much worse models, without any allowance being made for their good performance relative to reality. A challenge for this and similar approaches is that the use of a distance measure does not readily suggest a threshold at which models can be considered truly independent. All models are designed to simulate the real climate system, and are tuned towards observations of it (Hourdin et al., 2016). Therefore it should not be surprising that climate models appear broadly similar, since the maximum distance (in any relevant metric space) between a pair of models can be no more than the sum of the distances between each of these models and reality.

Some approaches to model independence have been less quantitative in nature. Masson and Knutti (2011) defines their interpretation as "independent in the sense that every model contributes additional information" but information in this context is not further defined or quantified. In fact the cluster analysis presented by Masson and Knutti (2011) may be more precisely

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

described by the phrasing in the related paper by Knutti et al. (2013), which states that independence is used "loosely to express that the similarity between models sharing code is far greater than between those that do not". While these papers certainly establish that point convincingly, there is again no indication of how much similarity should be expected or tolerated between truly 'independent' models, or whether complete independence is a meaningful concept in their terms. The interesting philosophical discussions of Parker (2011) and Lloyd (2015) both consider the interpretation and implications of consensus across an ensemble of models that are not independent, but the premise of non-independence is adopted from the literature and these two authors do not themselves attempt to further define this term in a quantifiable manner.

Perhaps the most constructive and complete approach to date is that of Sanderson et al. (2015). In this work, dependence is again defined in terms of inter-model differences in output, and this distance measure is used to remove or downweight the models which are most similar to other models in output. By comparing the inter-model distances both to model-data differences, and to what might be expected by chance with independent samples from a Gaussian distribution that summarises the full distribution, they introduce a threshold at which model differences may be considered appropriately large. However, the epistemic nature of their resulting ensemble is unclear and the resulting reduced ensemble is still only described in terms of *reducing* rather than eliminating dependency.

To summarise, the literature presents a strong consensus that the models are not independent, but does not appear to present such a clear viewpoint concerning what to do about this, or even the meaning of this term. Given this lack of clarity, it is perhaps unsurprising that the IPCC does not address this topic in detail, while nevertheless acknowledging its importance (Cubasch et al., 2013, 1.4.2). Thus, we see not only the opportunity, but also the necessity, of making further progress.

2.2 The literature concerning the independence of observational constraints on climate system behaviour

The value of the equilibrium climate sensitivity S is one of the fundamental questions of climate change research. A wide range of approaches have been presented which attempt to estimate this number. Typically, a Bayesian approach (which will be discussed in more detail in following sections) is used in which some prior assumptions are updated by means of an observationally-based likelihood function to form a posterior estimate. The observations commonly relate to the warming observed during the instrumental period (Tol and De Vos, 1998; Forest et al., 2006; Skeie et al., 2014), but analyses have also been presented which use longer-term climate changes seen during the paleoclimate record (Annan et al., 2005; Köhler et al., 2010), or short-term variations seen at seasonal to interannual time scales (Wigley et al., 2005; Knutti et al., 2006). In each case however, the observations are not a direct measure of the sensitivity S per se but must be related to it through the use of a climate model or models, which may be simple or complex. Collins et al. (2013, Box 12.2) and Annan (2015) survey and discuss some recent analyses which use a variety of observational data sets and modelling approaches, and Rohling et al. (2012) covers the paleoclimate field in some detail.

The question naturally arises as to whether these different constraints could, and should, be synthesised. In most of the Bayesian analyses, the prior is typically chosen to be vague, though there is some debate concerning this choice (Annan and Hargreaves, 2011a; Lewis, 2014). Irrespective of this, the prior then typically becomes substantially narrower when updated with observations to form a posterior. One might reasonably wonder what the results would look like if this resulting posterior

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

was then used as the prior in a new analysis in which it was updated by a *different* data set. This question was first explicitly raised by Annan and Hargreaves (2006), who argued that an assumption of independence between the constraints allowed for sequential updating which would result in a substantially tighter constraint than had previously been obtained. Hegerl et al. (2006) also updated a posterior arising from the 20th century, with a separate data set relating to climate changes over earlier centuries under an assumption of independence. However, the validity of these analyses is not immediately clear, as the independence of different constraints has not been clearly explained or demonstrated. Nevertheless, we always expect to learn from new observations (Lindley, 1956), so it is reasonable to expect that an analysis which accounts for multiple lines of evidence will generate a more precise and reliable result than analyses that do not. It is therefore surprising that there has been so little subsequent discussion of this topic in the scientific literature, and very few recent attempts to combine diverse data sources.

3 The statistical context for independence

In probability theory, independence has a straightforward definition. Two events A and B are defined to be independent if the probability of A, $P(A)$, is not affected by the occurrence of B, so that $P(A|B) = P(A)$ (e.g. Wilks, 1995, Section 2.4.3). Since the joint probability of both events, $P(A, B)$, is given by $P(A|B)P(B)$ we see that two events are independent if their joint probability is equal to the product of their individual probabilities, i.e. if $P(A, B) = P(A)P(B)$. Independence is therefore a symmetric property: A is independent of B if and only if B is independent of A. The concept of independence can also be generalised to the case of conditional independence: two events A and B are conditionally independent given an event S, if their joint probability conditional on S, $P(A, B|S)$, is equal to the product of their individual probabilities conditional on S, $P(A|S)P(B|S)$. Independence and conditional independence generalise naturally both to continuous distributions $p()$, which is more appropriate for the scenarios considered in this paper, and also to more than two events.

As we have seen in Section 2.1, much research on model independence either ignores or explicitly disavows any direct link to this mathematical/statistical definition. Conversely, the primary goal of this manuscript is to argue that this definition must be central to any usable, quantitative theory.

Bayes' Theorem tells us that we can update a probabilistic estimate of an unknown quantity $p(S)$ in light of some observations or evidence A via the equation

$$p(S|A) = p(A|S)p(S)/p(A).$$

$p(A|S)$ is known as the likelihood function (particularly when A is considered fixed, and S allowed to vary) and may be written as $\mathcal{L}(S|A)$.

If we have two events A and B then the corresponding equation is

$$p(S|A, B) = p(A, B|S)p(S)/p(A, B).$$

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics Discussions Open Access EGU

The first term on the right hand side of this equation can be expanded by the laws of probability, resulting in the equivalent equation

$$p(S|A,B) = p(B|S,A)p(A|S)p(S)/p(A,B).$$

Either of these two equations can in principle be used to calculate the posterior probability with two different observations or lines of evidence. While some might have found it strange that the Intergovernmental Panel on Climate Change (IPCC) stated that the literature contained no consensus on such a method (Collins et al., 2013, Box 12.2), it is possible that they merely meant that it is not clear how to calculate these terms, most specifically $p(B|S,A)$.

If A and B are conditionally independent given S, then $p(A,B|S)$ can also be decomposed as $p(A|S)p(B|S)$. In practice, the term 'independent' is frequently used to refer to conditional independence, especially when A and B are being discussed primarily as observations of, or evidence concerning, some unknown S. The practical value of this conditional independence is that if we have likelihoods $p(A|S)$ and $p(B|S)$, then conditional independence allows us to directly create the joint likelihood by multiplication, rather than requiring the construction of $p(B|S,A)$ as an intermediate step.

Conditional independence of A and B given S is therefore equivalent to the condition that $p(B|S,A) = p(B|S)$. This equation states that the predictive probability of B, given both S and A, is equal to the predictive probability of B given S. In other words, if we know S, then additionally learning A does not change our prediction of B. This interpretation can be a useful aid to understanding when independence does and does not occur.

3.1 The Bayesian perspective

The above elementary probability theory applies to both the frequentist and Bayesian paradigms. Within the Bayesian paradigm, the probability calculus may be used to describe the subjective beliefs of the researcher. In the remainder of this manuscript, we exclusively adopt this paradigm, since all the relevant uncertainties discussed here are epistemic in nature (relating to imperfect knowledge) and not aleatory (arising from some intrinsic source of 'randomness'). Thus, rather than considering 'the pdf of S' it is more correct to refer to 'my pdf of S' or perhaps 'our pdf for S' in the case that many researchers share a consensus view. While of course we anticipate using S here primarily to refer to the equilibrium climate sensitivity, the discussion applies equally to any unknown parameter (or vector of parameters) of the climate system.

While the subjective nature of the prior $p(S)$ has been extensively discussed in the literature, it is less widely appreciated that the likelihood $p(A|S)$ is also a fundamentally subjective matter. Even assuming that the parameter S is well-defined in the real world (which, given a sufficiently precise definition, may be possible) there is no world in which S takes a different value, with which we could check to see how the observations change. Therefore, while the likelihood should give a reasonable prediction of the observations A when the correct value of S is used, there is no objective constraint or check on what it might predict for some alternative incorrect S. The only practical way in which a likelihood can be constructed is via some model in which S varies, either as an explicit parameter in a simple model, or an emergent property of a more complex GCM. There can be no 'correct' way to vary S, again because there is no world in which S takes a different value against which to validate

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

our choices. Furthermore, it is trivial to create entirely artificial models with arbitrary behaviour ('if $3.21 < S < 3.22$ then do X, else do Y'). Within the Bayesian paradigm therefore, the likelihood can only reflect the researcher's subjective beliefs and modelling choices rather than any physical truth. Different models will in principle lead to different likelihoods, though in practice there may be a reasonable level of agreement between researchers.

As we have seen, the question of (conditional) independence boils down to the question of whether $p(B|S,A)$ is equal to $p(B|S)$. Our discussion of the subjective nature of likelihood within the Bayesian probabilities should make it clear that there is not an objectively correct answer to this question, but rather it depends on the subjective view of the researcher in question. Posing the question presupposes that the researcher already has likelihoods $p(A|S)$ and $p(B|S)$ in mind, or else the observations A and B would not be considered useful evidence on S. Would knowing A change their predictive distribution (likelihood) $p(B|S)$? If it would not, then A and B *are* conditionally independent given S, for this researcher. That is, if the researcher does not know how to use the additional information A in order to better predict B, then A and B are conditionally independent to that researcher. Thus, ignorance implies independence. If, conversely, A does provide helpful information in addition to S, then their improved prediction is the new likelihood function $p(B|S,A)$, which directly enables the joint likelihood $p(A,B|S)$ to also be created.

As an alternative to this uncompromisingly subjectivist viewpoint it is common to use modelling experiments (such as perturbed parameter ensembles) to generate likelihoods. In the following Subsections 3.2 and 3.3, we will demonstrate how the question of independence as it relates first to climate models, and then to constraints on the climate sensitivity, can be addressed in a more concrete manner using model ensembles.

3.2 Model independence in the Bayesian framework

We now explore how this Bayesian framework can be applied to the question of model independence. We first consider the 'truth-centred' hypothesis which is perhaps most clearly presented in Tebaldi et al. (2005). In that work, the outputs of the models, M_i (where $1 < i < n$ indexes the different models) are assumed to be samples from multivariate Gaussian distributions centred on the truth T. The likelihood for each model $p(M_i|T)$ is therefore a Gaussian of the same width centred on the model outputs. Furthermore, in this scenario the joint likelihood for multiple models is the product of the individual likelihoods, which as we have seen above is equivalent to considering that the models are independent conditional on the truth (i.e., that they are independent estimates of the truth). The joint likelihood is centred on the ensemble mean and will narrow in proportion to the square root of the number of models, which is the mathematical justification for the expectation that the ensemble mean will converge to the truth. As we have already mentioned, this behaviour is contradicted by analysis of the model outputs (Knutti et al., 2010). Thus, although such an interpretation of model independence can be presented in terms of the statistical definition of independence, it does not describe the behaviour of the models adequately because the models do in fact commonly share biases.

Another way of understanding how this lack of independence relates to shared model biases is to consider the subjective conditional likelihoods that a typical researcher might adopt. If we believe that the equilibrium climate of a typical model usually has RMS temperature errors of around σ°C at the local scale, and are furthermore told the true annual mean temperature

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics Discussions
Open Access
EGU

in London is 10.4°C, and, then an obvious prior prediction $p(M_n|T)$ for this output of an arbitrarily selected model M_n would be $N(10.4, \sigma)$. If we then learn that model M_1 actually simulates a temperature of 11.6°C at this point, then it would be reasonable to update the prediction for M_n to $N(10.4 + \delta, \sigma)$ for some $\delta > 0$, also expecting it to over predict to some extent (since we know that models tend to have biases of the same sign) and possibly also adjusting the uncertainty. Thus, $p(M_n|T, M_1) \neq p(M_n|T)$ and the models are not conditionally independent given the truth, so the truth-centred interpretation fails.

In light of the failure of the truth-centred approach, we now present an alternative application of statistical independence as follows. Consider that the outputs of models $M_1, \ldots, M_{n-1}$ (that is, the ensemble with one member omitted) can be summarised by the multivariate Gaussian distribution $N(M, \epsilon)$ where $M = \sum_{i=1,\ldots,n-1} M_i/(n-1)$ is now the sample mean of the reduced set and ϵ also measures the spread of this reduced ensemble. While the subjective nature of Bayesian probability does not allow for a definitive assertion, it seems reasonable that many researchers, if provided with such an M and ϵ and asked for a probabilistic prediction of the outputs of an additional model M_n, would answer with the distribution $N(M, \epsilon)$. Would additionally knowing the outputs from one named member of the ensemble, M_1 say, change this prediction? In most cases, it seems likely that the prediction would in fact be unchanged. For example, knowing the output of the HadGEM model in addition to the ensemble statistics is unlikely to provide much guidance to help in predicting the outputs of a previously unseen model from GISS. In this case, the two models could be considered independent, *conditional on the ensemble distribution* $N(M, \epsilon)$. However, if the two models were known to have some particular relationship, such as being two model versions from the same research centre, or even two instances of the same model structure with merely different parameters, then knowledge of M_1 would probably provide information concerning the target M_n and allow for an improved prediction of its outputs. This, we believe, encapsulates many of the same ideas as the model similarity analyses of Abramowitz and Gupta (2008); Knutti et al. (2013); Sanderson et al. (2015) and others. However, it has the advantage that the independence here can be defined in absolute terms (albeit conditional on $N(M, \epsilon)$) and is not merely a measure of relative difference. If a researcher does not know how to improve their prediction of M_n, in light of the outputs of M_1, then the models are in fact independent to them.

To provide a concrete demonstration of the previous ideas, we analyse the models which contributed to the CMIP3 database. Several modelling centres contributed more than one model version and we expect, based on the existing literature such as Knutti et al. (2013), that these may be noticeably more similar to each other, than models from different randomly-selected centres. In total, we use the outputs of 25 climate model simulations, and analyse two-dimensional climatological fields of surface air temperature (TAS), precipitation (PREC) and sea level pressure (PSL) for their pre-industrial control simulations.

We use as a simple distance metric the area-weighted root mean square difference between the climatological data fields (of commensurate variables) after regridding to a common 5 degree Cartesian grid. Considering firstly the temperature fields, the distance between a single target model and the mean of the remaining ensemble is just under 2°C. A randomly-selected model from the ensemble is of course usually (though not always) rather further from the target, with these distances averaging around 3°C. However, 6 climate centres provided two models and one (GISS) contributed three to the database, giving us 9 pairs of models that we might expect to be unusually close. Indeed, the pairwise distances within this set are clearly lower than they were for random model pairs, averaging around 2.1°C. Note that these distances are still greater than for the ensemble mean,

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics
Discussions
Open Access
EGU

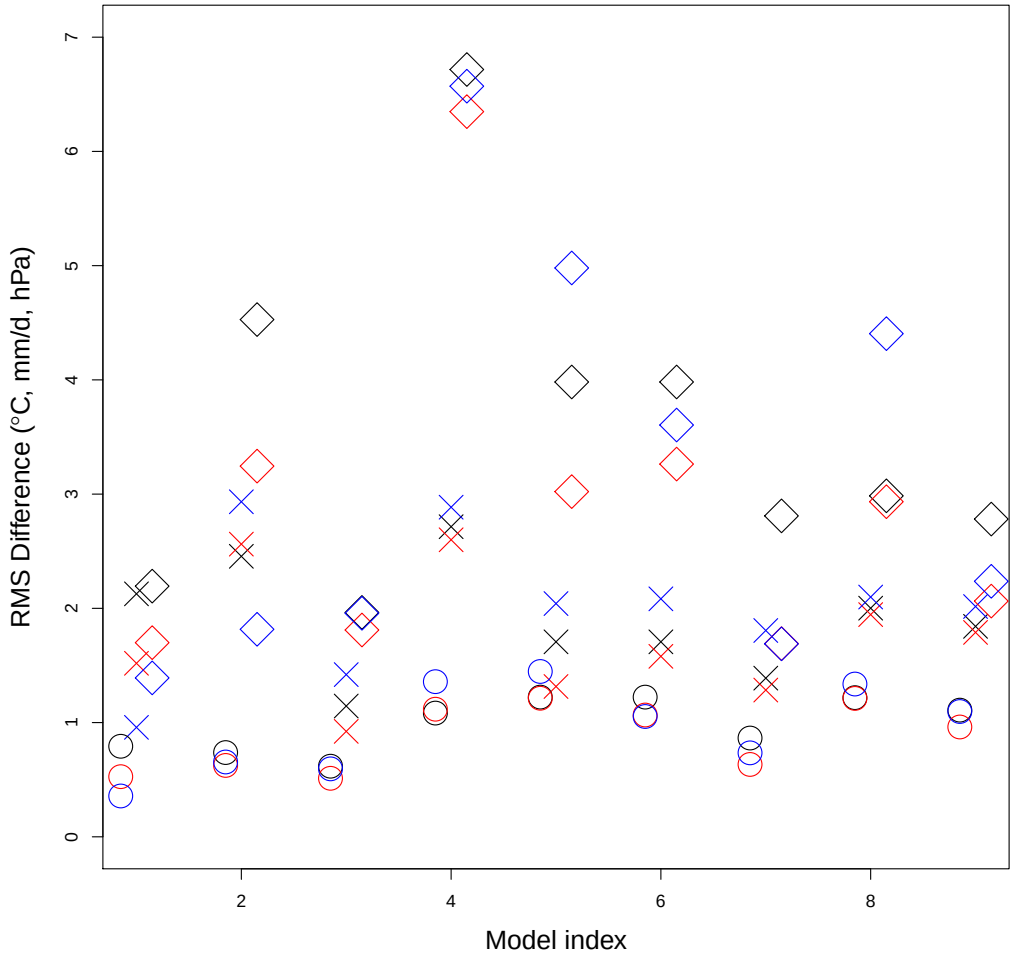

Figure 1. Analysis of CMIP models. Crosses represent TAS, circles PREC and diamonds PSL. Black symbols indicate distance of target model from mean of residual ensemble, blue symbols indicate distance of target model from plausibly related model, and red indicates distance of target model from interpolated prediction. Interpolation (red) generates smaller distances than black for almost all cases, and overall.

but only marginally. Therefore, using each model as a predictor for its colleague would actually lead to a worse prediction, compared to just using the ensemble mean. Qualitatively similar results are obtained for the PREC and PSL fields, with the related models being unusually close to each other compared to random model pairs, but generally not closer than they are to the ensemble mean of the other 24 models.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics Discussions
Open Access
EGU

However, even though the 'related' climate model is usually a little further from the target than the ensemble mean is, trial and error finds that a simple interpolation between these climate fields to $M' = 0.6M + 0.4M_1$ generates a noticeably lower prediction error, by roughly 10 – 15% across the three data fields used, than the original ensemble mean M. That is, we can replace our original prediction $N(M, \epsilon)$ which was based on the ensemble summary, with $N(M', 0.9 \times \epsilon)$, to give a better prediction of the unseen M_n. Thus we have demonstrated empirically and numerically that two models contributed by a single research centre are not conditionally independent given M and ϵ, according to the standard statistical definition. The results for each pair of related models and for each of the three climate fields considered here are presented graphically in Figure 1.

We must bear in mind that more sophisticated analyses might also be possible, which could give further improvements. For instance, experts with knowledge of the model structures might be able to predict more detailed similarities between the outputs of model pairs. However, our result here is sufficient to illustrate how the concept of statistical independence can be directly applied to the question of model independence, while encapsulating much of what is discussed in the literature.

An important point to note is that this interpretation of independence is entirely unrelated to model performance. Reality (e.g. observations of the real climate system) does not enter into any of the calculations or definitions above. Thus, the two concepts of performance and independence as used here are entirely unrelated. It remains a challenge to develop some useful interpretation of (conditional) independence which *does* use real data and which is informative regarding both model performance and pairwise similarity. However, the definition as presented here does have obvious applications in terms of interpreting and using the model ensemble. It suggests that we may be able to usefully reduce the full CMIP ensemble to a set which are independent (conditional on the ensemble statistics, as above). This will provide a smaller set of models for analysis and use in downstream applications including downscaling to higher resolution regional simulations of climate change. This is likely to be increasingly important and necessary given the heterogenous nature of simulations which will likely be submitted to future CMIP databases.

While the question of model similarity and ensemble member selection has already been considered by others (eg Sanderson et al., 2015), the work here provides a more clear-cut definition of what it means to be independent, which is directly testable. If researchers can demonstrate dependence (in terms of an improved prediction of model outputs as illustrated here) then independence is violated, and if not, it may be reasonably assumed. Another important difference between the approach presented here, and that of many other authors, is that independence is determined *a priori* in terms of the anticipated outputs of a model, rather than *a posteriori* in light of the model outputs. Pairwise similarity between model outputs may arise through convergence of different approaches to simulation, and not merely through copying of ideas. For example, one pair of models which exhibit unusually similar temperature fields in our analysis consists of the model from CNRM and one of the GFDL models, which do not share any particularly obvious genealogy. We do not believe that a coincidentally similar behaviour should be penalised by downweighting of these models, as it may represent a true 'emergent constraint' on system behaviour. An obvious test of our ideas would be to apply this analysis to the CMIP5/6 climate model ensembles, to check whether the interpolation and dependence ideas presented here apply generally to ensembles of climate models rather than being an example of over-enthusiastic data mining.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics Discussions
Open Access
EGU

3.3 Independence of constraints in the Bayesian context

To paraphrase a statement made in Section 3.1, ignorance is independence. Given a likelihood $p(A|S)$ we ask ourselves, how can we change this by additionally including B to form $p(A|B,S)$? If the answer is that B provides no additional information on A, then A and B are conditionally independent given S. This answer may seem a little unsatisfactory, as it relies on a dogmatically subjectivist and personal interpretation of probability. While we emphasise that Bayesian probability is fundamentally subjective, it is quite usual to use models as a tool to represent and understand our uncertainties. We must however remember that as discussed in Section 3.1, even the likelihood is itself a subjective function: there is no such thing as the "correct" likelihood $p(A|S)$.

In this section, we explore these ideas in a little more detail, to sketch out how we may be able to provide a credible basis for our judgements. Typically, a likelihood $p(A|S)$ is generated not as a purely subjective matter of belief, but instead justified via a model or ensemble of models. For example, if the equilibrium sensitivity is varied across an ensemble of energy balance models (along with other input parameters: S here may be used as a shorthand for a vector of relevant uncertainties) then we will find that in simulations of the 20th century, the warming observed will vary across the ensemble. This can then be used as the basis for the likelihood function (eg Tol and De Vos, 1998; Forest et al., 2006; Skeie et al., 2014). Similarly, for another observable B, which may require another set of simulations using the same ensemble. We now outline how it is possible to test whether B is conditionally independent of A given S, in the context of this model.

We firstly form the likelihood $p(A|S)$, and use this to generate the appropriate predictions of A for each ensemble member, given its known value of S. There will of course be residuals, the magnitude of which indicate the limited information which S provides concerning A. We can then explore whether an additional observable B is informative regarding these residuals. If it is not, then we may reasonably conclude that B provides no additional information on, and is conditionally independent of, A given S. Conversely, if B is informative regarding the residuals, then this is proof that it is not independent of A.

A simple example is used to illustrate the point graphically. We use a zero dimensional energy balance model to simulate the changes of both the 20th century and the Last Glacial Maximum. For simplicity, we only consider a subset of relevant uncertainties: the equilibrium sensitivity S, the planetary effective heat capacity C, the uncertainties in radiative forcing due to aerosol forcing over the 20th century F, and atmospheric dust and the large ice sheets which existed during the Last Glacial Maximum, D and I respectively.

For the warming of the 20th century, we assume a linear forcing ramp from 0 in 1900 to $2-F$ in 2000 (using a value of $2Wm^{-2}$ to approximately represent the sum of all other forcings other than aerosols, which are dominated by greenhouse gases). The zero dimensional energy balance model satisfies the equation

$$C \cdot dT/dt = (G - T)/S, \quad (1)$$

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

where $T = T(t)$ is the temperature anomaly (relative to 1900) at time t, and $G = G(t)$ is the total forcing. Our first observable A, the change in global mean surface air temperature over the 20th century is taken to be the linear trend over this interval. For the LGM, the temperature change B is calculated as

$$B = (3 + D + I)/S \tag{2}$$

where the total forcing $3+D+I$ is the sum of greenhouse gases ($3Wm^{-2}$), the uncertain dust forcing (D) and the uncertain effective forcing of the ice sheet (I) respectively. The ice sheet forcing uncertainty term here implicitly accounts for the nonlinearity of how this combines with the other forcings. For simplicity, we do not consider observational uncertainties for either the LGM or 20th century temperature changes, though accounting for these would be straightforward. We use the following priors:

$S \sim U[0.5, 6]$

$C \sim U[10, 30]$

$F \sim N[1, 0.5]$

$D \sim N[1, 0.5]$

$I \sim N[3, 1]$

A plot of the simulated 20th century warming A versus sensitivity S is shown in Figure 2a. The relationship shown demonstrates the basis for a likelihood function $p(A|S)$: for any specified sensitivity we can predict the resulting temperature (with uncertainty), and therefore for any specific warming A we can calculate how the probability of this being observed varies with S. In this example, the linear regression provides a good fit to the data, though the errors clearly grow towards larger sensitivity values. Similarly, the LGM cooling B is also linked to S (Figure 2b) and this relationship can be used as the basis for a likelihood function $p(B|S)$.

By construction, we already know that the two constraints are independent given S, since the other uncertain parameters that relate to each observations are disjoint. However, if we did not know this analytically *a priori* but were instead merely able to use the model as a black box, we could check in the following manner. In order to check for the independence of these constraints, we need to determine whether $p(A|S, B)$ differs from $p(A|S)$. In order to do this, we create an ensemble with an arbitrary but fixed value of $S = 3.5°$C, say, and simulate both the 20th century warming and the LGM state for each member of this ensemble. The likelihood function arising from Figure 2a gives us a predicted warming of $A = 0.85°$C (with uncertainty of 0.4°C) for these ensemble simulations. We now check the prediction errors to see if they exhibit any relationship with B. Figure 2c indicates that they do not, with the regression coefficients being insignificantly different from zero. The conclusion is that the additional knowledge of B, once the sensitivity S is known to be 3.5°C, does not provide any additional help in predicting A. A and B are therefore independent, conditional on $S = 3.5$C. This experiment can be repeated for as many

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

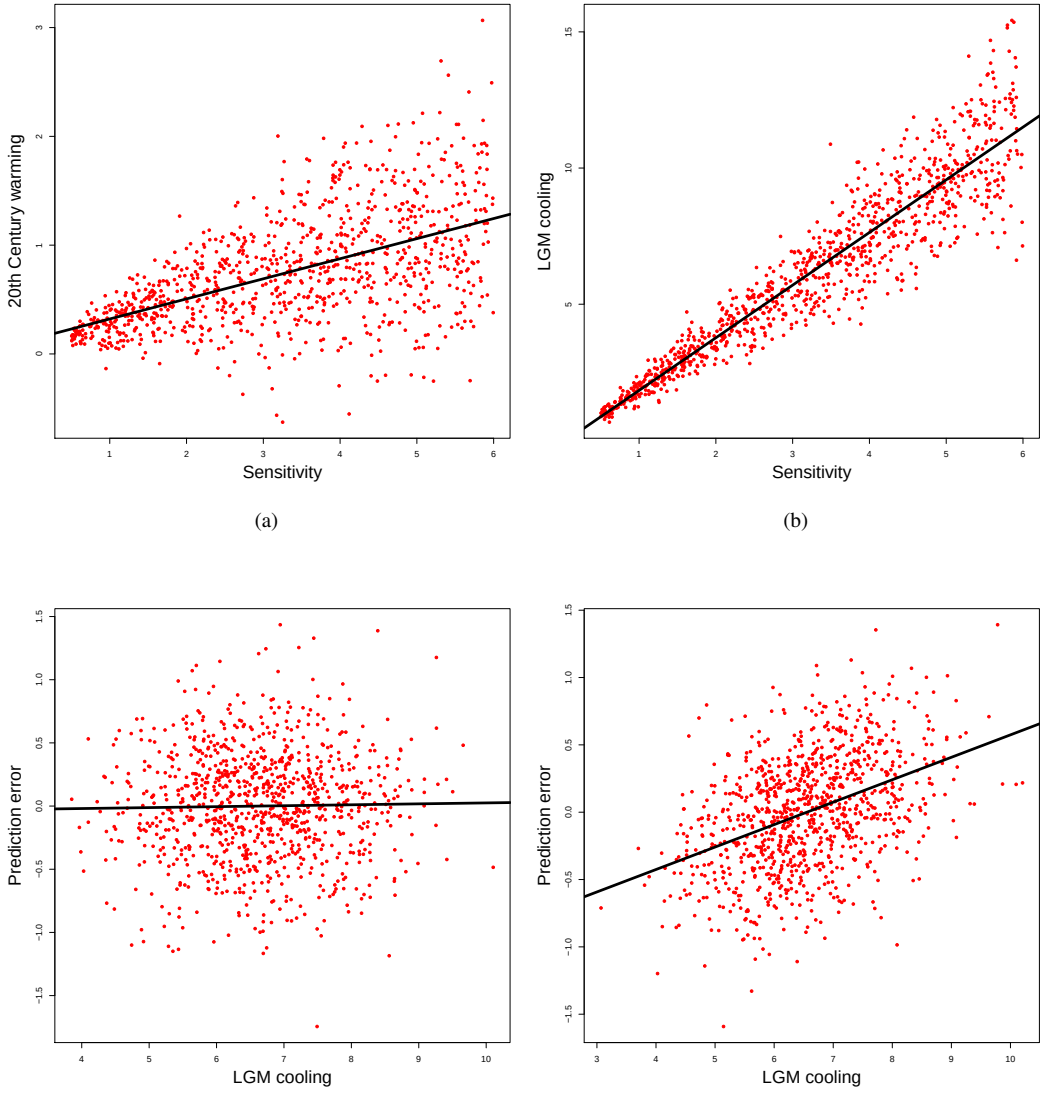

Figure 2. Outputs of ensemble simulations: (a) 20th Century warming versus equilibrium sensitivity (b) LGM cooling versus equilibrium sensitivity (c) 20th century prediction residuals versus LGM cooling, dependent case (d) 20th century prediction residuals versus LGM cooling, independent case.

different values of S as is desired, and the same negative result will be found. This is of course not surprising, as the model has been constructed in this way.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

Earth System Dynamics Discussions Open Access EGU

We now make a small change to the model, and substitute D with F in Equation 2. This modified model now makes the assumption that the effective dust forcing at the LGM is the same as the aerosol forcing during the 20th century. This is of course again a very simplistic approach but it is not completely unreasonable to assume a link, as both forcings relate to the effects of condensation nuclei on clouds. Importantly, the univariate likelihood functions $p(A|S)$ and $p(B|S)$ are unchanged by this substitution, as D and F are identically distributed. Therefore, we can generate the same prediction for A, conditional on a known $S = 3.5°C$. However, in this case the prediction errors are strongly correlated with B, as is shown in Figure 2d. Therefore, a new distribution function $p(A|S, B)$ can be created which makes a more precise prediction of A given knowledge of both S and B. Thus it can be diagnosed from the model outputs alone, without direct knowledge of the model's internal structure, that A and B are not independent conditional on S. This result is of course easily interpreted in terms of the known model structure: for a given sensitivity, a smaller than expected cooling at the LGM suggests a low dust/aerosol forcing, which then implies that the 20th century warming will be greater than originally expected.

The linear regressions are not necessarily the best way to represent a relationship that may in practice be more complex. The point of these numerical experiments is to demonstrate that this dependence can be diagnosed from model outputs directly, in the case that the model structure is imperfectly understood or known. Furthermore, the conditional likelihood $p(A|S, B)$ can be generated from the ensemble outputs. This then enables us to generate the joint likelihood $p(A, B|S) = p(A|B, S)p(B|S)$ as required for a Bayesian inversion.

Such analyses may be impractical for the outputs of small ensembles such as those arising from the CMIP multi-model experiments which explore structural uncertainties. However, they are plausible for larger ensembles where parameters are varied within a single model structure. The key requirement is that the simulations relating to different observables are performed with the same ensemble members, in order that any dependence between constraints can be explored. The results will of course depend on the model used, but this is as expected: the likelihood does not reflect reality, but rather, the modelling assumptions, as discussed in Section 3.1.

4 Conclusions

We have discussed and presented a coherent statistical framework for understanding independence, and explained how this applies in two distinct applications. Climate models cannot sensibly be considered independent estimates of reality, but fortunately this strong assumption is not required in order to make use of them. A more plausible, though still optimistic, assumption, might be to interpret the ensemble as merely constituting independent samples of a distribution which represents our collective understanding of the climate system. This assumption is challenged by the near-replication of some climate models within the ensemble, and therefore re-weighting or sub-sampling the ensemble could improve its usefulness. We have shown how the statistical definition of (conditional) independence can apply and how it helps in defining independence in a quantifiable manner. The definition we have presented is certainly not the only possible one and we expect that others may be able to suggest improvements within this framework.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

When considering the use of observational evidence in constraining climate system behaviour (including the specific example of the equilibrium climate sensitivity), observational uncertainties themselves can generally be regarded as independent. However, the independence of the resulting likelihood functions is not so immediately clear, as it typically also rests on a number of modelling assumptions and uncertainties. Here we have shown how the question of independence can be readily interpreted and understood in terms of the conditional prediction of observations. These ideas may be useful in the design and analysis of ensemble experiments underpinning the analysis of observational constraints.

While our analyses may not provide complete solutions to the questions raised, we have shown how the statistical framework can be usefully applied. Further, we see little prospect for progress to be made unless it is underpinned by a rigorous mathematical framework. Therefore, we hope that other researchers will be able to make use of these ideas in their future work.

5 Acknowledgements

We acknowledge the modelling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-34, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 19 August 2016

References

Abramowitz, G. and Gupta, H.: Toward a model space and model independence metric, Geophysical Research Letters, 35, L05 705, 2008.

Annan, J. D.: Recent Developments in Bayesian Estimation of Climate Sensitivity, Current Climate Change Reports, pp. 1–5, 2015.

Annan, J. D. and Hargreaves, J. C.: Using multiple observationally-based constraints to estimate climate sensitivity, Geophysical Research Letters, 33, 2006.

Annan, J. D. and Hargreaves, J. C.: On the generation and interpretation of probabilistic estimates of climate sensitivity, Climatic Change, 104, 423–436, 2011a.

Annan, J. D. and Hargreaves, J. C.: Understanding the CMIP3 multimodel ensemble, Journal of Climate, 24, 4529–4538, 2011b.

Annan, J. D., Hargreaves, J. C., Ohgaito, R., Abe-Ouchi, A., and Emori, S.: Efficiently constraining climate sensitivity with paleoclimate simulations, Scientific Online Letters on the Atmosphere, 1, 181–184, 2005.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A., and Wehner, M.: Long-term Climate Change: Projections, Commitments and Irreversibility, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., chap. 12, pp. 1029—-1136, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, doi:10.1017/CBO9781107415324.024, www.climatechange2013.org, 2013.

Cubasch, U., Wuebbles, D., Chen, D., Facchini, M., Frame, D., Mahowald, N., and Winther, J.-G.: Introduction, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., book section 1, p. 119–158, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, doi:10.1017/CBO9781107415324.007, www.climatechange2013.org, 2013.

Forest, C. E., Stone, P. H., and Sokolov, A. P.: Estimated PDFs of climate system properties including natural and anthropogenic forcings, Geophys. Res. Lett, 33, 2006.

Hegerl, G. C., Crowley, T. J., Hyde, W. T., and Frame, D. J.: Climate sensitivity constrained by temperature reconstructions over the past seven centuries, Nature, 440, 1029–1032, 2006.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., et al.: The art and science of climate model tuning, Bulletin of the American Meteorological Society, 2016.

Jun, M., Knutti, R., and Nychka, D.: Local eigenvalue analysis of CMIP3 climate model errors, Tellus, 60, 992–1000, 2008a.

Jun, M., Knutti, R., and Nychka, D.: Spatial analysis to quantify numerical model bias and dependence: how many climate models are there?, Journal of the American Statistical Association, 103, 934–947, 2008b.

Knutti, R., Meehl, G., Allen, M., and Stainforth, D.: Constraining climate sensitivity from the seasonal cycle in surface temperature, Journal of Climate, 19, 4224–4233, 2006.

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

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophysical Research Letters, 40, 1194–1199, 2013.

Köhler, P., Bintanja, R., Fischer, H., Joos, F., Knutti, R., Lohmann, G., and Masson-Delmotte, V.: What caused Earth's temperature variations during the last 800,000 years? Data-based evidence on radiative forcing and constraints on climate sensitivity, Quaternary Science Reviews, 29, 129–145, 2010.

Lewis, N.: Objective Inference for Climate Parameters: Bayesian, Transformation-of-Variables, and Profile Likelihood Approaches, Journal of Climate, 27, 7270–7284, 2014.

Lindley, D. V.: On a measure of the information provided by an experiment, The Annals of Mathematical Statistics, 27, 986–1005, 1956.

Lloyd, E. A.: Model robustness as a confirmatory virtue: The case of climate science, Studies in History and Philosophy of Science Part A, 49, 58–68, doi:10.1016/j.shpsa.2014.12.002, http://dx.doi.org/10.1016/j.shpsa.2014.12.002, 2015.

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett, 38, L08 703, doi:10.1029/2011GL046864, 2011.

Parker, W. S.: When Climate Models Agree: The Significance of Robust Model Predictions, Philosophy of Science, 78, 579–600, 2011.

Pennell, C. and Reichler, T.: On the Effective Number of Climate Models, Journal of Climate, 2010.

Randall, D. A., Wood, R., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R., Sumi, A., and Taylor, K.: Climate Models and Their Evaluation. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 8, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 2007.

Rohling, E., Sluijs, A., Dijkstra, H., Köhler, P., van de Wal, R., von der Heydt, A., Beerling, D., Berger, A., Bijl, P., Crucifix, M., et al.: Making sense of palaeoclimate sensitivity, Nature, 491, 683–691, 2012.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, Journal of Climate, 28, 5171–5194, doi:10.1175/jcli-d-14-00362.1, http://dx.doi.org/10.1175/JCLI-D-14-00362.1, 2015.

Skeie, R., Berntsen, T., Aldrin, M., Holden, M., and Myhre, G.: A lower and more constrained estimate of climate sensitivity using updated observations and detailed radiative forcing time series, Earth System Dynamics, 5, 139–175, 2014.

Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and (eds.), P. M.: IPCC 2013: Summary for Policymakers, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2013.

Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 2053, 2007.

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

Tol, R. S. and De Vos, A. F.: A Bayesian statistical analysis of the enhanced greenhouse effect, Climatic Change, 38, 87–112, 1998.

Wigley, T. M. L., Amman, C. M., Santer, B. D., and Raper, S. B.: Effect of climate sensitivity on the Response to Volcanic Forcing, Journal of Geophysical Research, 110, 2005.

Wilks, D. S.: Statistical methods in the Atmospheric sciences, Academic Press, London, 1995.