Earth Syst. Dynam., 10, 91–105, 2019
https://doi.org/10.5194/esd-10-91-2019

Earth System Dynamics
Open Access
EGU

ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing

Gab Abramowitz[1,2], Nadja Herger[1,3], Ethan Gutmann[4], Dorit Hammerling[4], Reto Knutti[5], Martin Leduc[6,7], Ruth Lorenz[5], Robert Pincus[8,9], and Gavin A. Schmidt[10]

[1]Climate Change Research Centre, UNSW Sydney, Australia
[2]ARC Centre of Excellence for Climate Extremes, Australia
[3]ARC Centre of Excellence for Climate System Science, Australia
[4]National Center for Atmospheric Research, Boulder, Colorado, USA
[5]Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland
[6]Ouranos, Montréal, Québec, Canada
[7]Université du Québec à Montréal, Montréal, Québec, Canada
[8]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA
[9]NOAA Earth System Research Lab, Physical Sciences Division, Boulder, Colorado, USA
[10]NASA Goddard Institute for Space Studies, New York, NY, USA

Correspondence: Gab Abramowitz (gabriel@unsw.edu.au)

Received: 2 July 2018 – Discussion started: 4 July 2018
Revised: 20 December 2018 – Accepted: 7 January 2019 – Published: 13 February 2019

Abstract. The rationale for using multi-model ensembles in climate change projections and impacts research is often based on the expectation that different models constitute independent estimates; therefore, a range of models allows a better characterisation of the uncertainties in the representation of the climate system than a single model. However, it is known that research groups share literature, ideas for representations of processes, parameterisations, evaluation data sets and even sections of model code. Thus, nominally different models might have similar biases because of similarities in the way they represent a subset of processes, or even be near-duplicates of others, weakening the assumption that they constitute independent estimates. If there are near-replicates of some models, then treating all models equally is likely to bias the inferences made using these ensembles. The challenge is to establish the degree to which this might be true for any given application. While this issue is recognised by many in the community, quantifying and accounting for model dependence in anything other than an ad-hoc way is challenging. Here we present a synthesis of the range of disparate attempts to define, quantify and address model dependence in multi-model climate ensembles in a common conceptual framework, and provide guidance on how users can test the efficacy of approaches that move beyond the equally weighted ensemble. In the upcoming Coupled Model Intercomparison Project phase 6 (CMIP6), several new models that are closely related to existing models are anticipated, as well as large ensembles from some models. We argue that quantitatively accounting for dependence in addition to model performance, and thoroughly testing the effectiveness of the approach used will be key to a sound interpretation of the CMIP ensembles in future scientific studies.

Published by Copernicus Publications on behalf of the European Geosciences Union.

1 Characterising uncertainty in ensemble projections

Future climate projections are uncertain for a wide range of reasons, including the following: there is limited knowledge of the future of human behaviour (including greenhouse gases and other emissions associated with them); we have an incomplete understanding of how the climate system works; we have a limited ability to codify what is understood into models; there are constraints on our ability to resolve known processes in models due to computational limitations; there are limitations to the measurements of the state of the climate in the past required for accurate model initialisation; and there are inherent limits of predictability associated with the climate system itself given its chaotic nature. While the first of these issues is addressed by considering projections conditional on specific emission scenarios, the other areas of uncertainty are addressed by the use of multiple working hypotheses (Chamberlin, 1890), in the form of multi-model ensembles of climate projections.

By using an ensemble of climate projections from a range of climate models, as opposed to a single model or single simulation, researchers hope to achieve two related goals. First, they want simulation agreement across models in the ensemble to imply robustness, for any particular phenomenon of interest or, more generally, to gain an understanding of the structural uncertainty in any prediction. Second, and more ambitiously, they would like the distribution of the behaviour of any particular phenomenon across the ensemble to be a reasonable approximation of the probability of its occurrence, conditional upon the assumptions described above.

While we described six sources of uncertainty affecting this likelihood in the first paragraph, understanding the relationship between models and the real world is aided by sorting them into three categories. The first is the unpredictable nature of human behaviour throughout the 21st century, and is typically dealt with using discrete social, political and economic scenarios. The other two categories of uncertainty are conditional on the assumption of a particular emissions scenario. The first of these is "epistemic uncertainty", which relates to our knowledge and understanding of the climate system, and so encompasses uncertainties that are thought to be reducible with more information or knowledge. If for a moment we assume that the climate system is fundamentally deterministically predictable, and that observational records are spatially complete and long enough to characterise any internal variability, an ideal model ensemble distribution would then accurately represent our uncertainty in creating and using climate models. That is, it would represent uncertainty in our understanding of how the climate system works, our ability to codify what is understood in models, and our ability to resolve known processes in models due to computational limitations – as noted above. The existence of epistemic uncertainty in climate modelling is inevitable, as they are low dimensional modelling systems that can never be a perfect representation of the climate system (Box, 1979; Oreskes et al., 1994).

Alternatively, if we assume that we did have a perfect understanding of how the climate system worked and could codify this effectively in models, climate system stochasticity and chaotic behaviour would also mean that for any given (incomplete) observational record and model resolution, an ensemble of solutions would exist, representing the inherent limit of predictability in the climate system – "aleatory uncertainty". The distinction between epistemic and aleatory uncertainty is relevant because the nature of model dependence, and hence how we might attempt to address it, is different in each case.

The use of multi-model ensembles, including those from the widely used Coupled Model Intercomparison Project (CMIP), is common in climate science despite the fact that such ensembles are not explicitly constructed to represent an independent set of estimates of either epistemic or aleatory uncertainty. In fact the ensembles are not systematically designed in any way, but instead represent all contributions from institutions with the resources and interest to participate; therefore, they are optimistically called "ensembles of opportunity" (Tebaldi and Knutti, 2007). The purpose of this paper is to give an overview of approaches that have been proposed to untangle this ad-hoc ensemble sampling, and to discuss assumptions behind the different methods as well as the advantages and disadvantages of each approach. In the end, the goal is to efficiently extract the information relevant to a given projection or impacts question, beyond naive use of CMIP ensembles in their entirety. While we discuss dependence in the context of global climate model (GCM) sampling here, there are clearly many more links in the chain to impacts prediction, such as regional climate model downscaling, and these issues apply equally to other steps in the chain (see Clark et al., 2016 for more on this). Nevertheless, identifying the appropriate approaches for some applications might not only help increase understanding and certainty for projections and impacts research, but it might also help direct limited resources to model development that does not essentially duplicate the information that other models provide.

In the next section we discuss the nature of sampling in multi-model climate ensembles. In Sect. 3, we examine which aspects of models we might want to be independent, as opposed to agreeing with observational data sets, and the relevance of canonical statistical definitions of independence. Section 4 details attempts to define model independence in terms of model genealogy, whereas Sect. 5 discusses definitions based on inter-model distances inferred from model outputs. We discuss the relationship between model independence and performance in more detail in Sect. 6, before considering the role of model independence in estimating aleatory uncertainty in Sect. 7, and how this helps distinguish and contextualise different ensemble interpretation paradigms that are evident in the literature. In Sects. 8 and

9, we examine the critical question of how best to test the efficacy of any post-processing approach to address model dependence or performance differences, including weighting or model sub-selection. We then make recommendations and conclusions in Sects. 10 and 11.

2 Ensemble sampling to address uncertainties

Despite the fact that state-of-the-art ensembles such as CMIP are not constructed to provide independent estimates, they do sample both epistemic and aleatory uncertainty by integrating an arbitrary number of emissions scenarios, global climate models, physics perturbations and initial-condition realisations. This leaves practitioners to interpret the ensemble distribution very much on a case-by-case basis, and decomposing these uncertainties is not necessarily straightforward (see Hawkins and Sutton, 2009, 2011; Leduc et al., 2016a). For instance, if only a subset of the CMIP5 ensemble is considered, the way simulations are selected along the three axes (scenario, model and realisation) will modulate the relative role of each source of uncertainty. This effect is implicit even when considering all available simulations because the broader ensembles such as CMIP are generally constructed based on an the ability of the international modelling centres to contribute, rather than a systematic sampling strategy (Tebaldi and Knutti, 2007).

In all of these cases, the motivation for using an ensemble rather than a single simulation is to obtain multiple independent estimates of the quantity under consideration. The weather forecasting community has developed metrics that give useful information about ensemble spread interpreted as aleatory uncertainty (e.g. rank frequency histograms or rank probability skill scores); these remain necessary but not sufficient conditions for assessing the independence of aleatory uncertainty estimates (see Hamill, 2001; Weigel et al., 2007). In particular, the effect of epistemic uncertainty on these metrics is difficult to interpret. Generally speaking, there is no established and universally accepted methodology to define whether or not ensemble members are independent. The implications of a lack of independence are clear: agreement within an ensemble might not imply robustness, and similarly, the distribution of ensemble members is unlikely to provide useful information about the probability of outcomes. Indeed, recent work by Herger et al. (2018a) showed that sub-selection of ensemble members by performance criteria alone, without consideration of dependence, can result in poorer ensemble mean performance than random ensemble member selection. Despite this heuristic understanding of the perils of model dependence, a practical definition of model independence is not straightforward, and is problem dependent. In reviewing the literature here, we do not aim to provide a single canonical definition of model dependence, but instead contextualise the range of definitions and applications that have been used to date in a single conceptual framework, and reinforce the need for thorough out-of-sample testing to establish the efficacy of any approach for a given application.

3 What is meant by model independence?

Although the statistical definition of independence of two events A and B is strictly defined as $P(A|B) = P(A)$ (as dealt with in some depth by Annan and Hargreaves, 2017), it is not immediately clear that there is an objective or unique approach to applying this definition to climate projection ensemble members. Indeed there is an obvious way in which models should be dependent – they should all provide a good approximation to the real climate system. Noting that each model is a myriad of discrete process representations makes it clear how complicated any categorical statement about model independence within an ensemble needs to be. For those process representations where models exhibit high fidelity (i.e. where there is sufficient observational constraint to ascertain this), models should be expected to agree in their representation. That is, where we have clear observational evidence, we do not expect a model to exhibit epistemic departures from the true physical system. It is only in the cases where there is insufficient observational constraint to diagnose such an epistemic departure, or those where no model can avoid one, that models should provide independent process representations (e.g. parameterisation of known processes because of scale considerations or empirical approximations for complex or incomplete process representations).

Therefore, it might seem that the limited case of ascertaining whether two models are independent with respect to a specific process representation that is weakly constrained by observational data would be relatively easy to verify. If the two models take different approaches to the underconstrained process, we might argue that they are independent with respect to it. However, this is clearly unsatisfactory. First, we have no context for how different treatments of this process might legitimately be, or the ability to quantify this difference. Next, looking for evidence of the independence of these process representations via the impact they have on simulated climate is also fraught, as we are reliant on the effect they have within a particular modelling system. Two radically different representations of a process might not elicit different responses in a modelling system if the modelling system is insensitive to the process in question. Alternatively, one representation may result in artificially strong model performance (and so misleadingly imply fidelity) if it effectively compensates for other biases within the modelling system. This is an example of epistemological holism well documented by Lenhard and Winsberg (2010). This problem is further compounded by the reality of models being very large collections of process representations, where only a subset of these might be independent. In this context, a cat-

egorical statement about overall model independence seems far too simplistic. It also means that an assessment of dependence will have different outcomes depending on the phenomena and question being investigated, as different parts of a model may affect the problem under consideration.

4 Independence as distinct development paths

Understanding the evolutionary history of models might also seem like a way to characterise model independence (e.g. Fig. 5 in Edwards, 2011). This a priori view of independence is that models that share common ancestry might be deemed partially dependent – similar to the idea of evolutionary cladistics. The analogy is not perfect, as modelling groups have borrowed discrete components from each other over time, but nevertheless the lineage of any particular model snapshot at a point in time might theoretically be traced. To date, with the exception of Boé (2018) who utilises version number as a proxy for differences between model components, we are not aware of any studies that have comprehensively tried to infer independence from the history of model development, most likely because of the paucity of information on each model's history, including the lack of freely available source code. Boé (2018) accounted for the number of shared components by GCMs and quantified how the replication of whole model components, either atmosphere, ocean, land or sea ice, influenced the closeness of the results of different GCMs and could show a clear relationship, where even a single identical component had a measurable effect. This was shown to be true for global as well as regional results, although no component appeared to be more important than the others.

While defining model independence a priori is desirable, it quickly becomes difficult and time consuming for large ensembles such as CMIP, particularly given the lack of transparency regarding precisely what constitutes a given model, the difficulty deciding whether or not components are identical and the role of tuning (see Schmidt et al., 2017). Boé (2018) used version numbers, considering components to be different when the major revision number was different, but not if the minor revision number was different (for example CLM4 and CLM4.5 would be deemed dependent). However, it is unlikely that the approach to version numbering is consistent across modelling centres, meaning that two components might be very different even if they share a major version number, or vice versa. Furthermore, how we might account for the effect of shared model histories within an ensemble if we had all this information available does not seem obvious, beyond the categorical inclusion or exclusion of simulations. As discussed above, an ideal definition of model dependence would only include variability in process representations that are not tightly observationally constrained, so that several models using the Navier–Stokes equations might not represent dependent treatment of process, for example.

5 Independence as inter-model distance

Alternatively, dependence could be defined a posteriori in terms of the statistical properties of model output (perhaps more analogous with Linnaean taxonomy). This is the approach taken by Masson and Knutti (2011) and Knutti et al. (2013), who used hierarchical clustering of the spatio-temporal variability of surface temperature and precipitation in climate model control simulations to develop a "climate model genealogy". Perhaps unsurprisingly, they found a strong correlation between the nature of model output and shared model components. While the family tree of models in the above-mentioned work shows that there is dependence between models, it does not suggest how to account for its effect.

Several studies have defined model independence using a metric that defines scalar distances between different model simulations. Abramowitz and Gupta (2008) proposed constructing a projected model distance space by defining pairwise model distances as the overlap of probability density functions (PDFs) of modelled variables between model pairs. Model behaviour was clustered using self-organising maps (Kohonen, 1989), and the overlap of the model output PDF pairs for each cluster was determined. PDF overlap at each cluster was then weighted by the occurrence of cluster conditions to determine model–model distances. Sanderson et al. (2015a) proposed constructing a projected model distance space by defining pairwise model distances using the rows of an orthogonal matrix of model loadings in the singular value decomposition of seasonal climatological anomaly values of a range of climate variables. Knutti et al. (2017) and Lorenz et al. (2018) used pairwise root mean square distances between model simulations in one or more variables to assess dependence.

All of these approaches allow the definition of distance between different model simulations and observational data sets of commensurate variables to be defined. None verified that the space created met formal metric space criteria in the mathematical sense: each might violate the triangle inequality ($d(a, c) \leq d(a, b) + d(b, c)$), where $d(a, b)$ defines the distance between models a and b, or the identity of indiscernibles ($d(a, b) = 0$ if and only if $a = b$), for example, and describing these measures as "distances" could then potentially be misleading. It is unclear whether these potential issues arise or are relevant in practice.

Inter-model distances may also be problematic as measures of independence because they are holistic. That is, inter-model distances reflect the combined effect of all process representations that affect the chosen metric – including both those processes strongly supported by observational data and those where a lack of observational data might al-

low a departure from the true system behaviour (where we might only want to define independence in terms of the latter, as noted above). Further, by examining model output that is the result of the interaction of all of these process representations (that is, just the impact variable values in model output), they ignore the possibility that their combined effect might lead to equifinality, i.e. different models may arrive at very similar impact variable values through different mechanisms and feedbacks. This situation is an example of the identity of indiscernibles being violated, and so might lead to models being inappropriately deemed dependent, although this is less likely as the dimensionality of a metric increases. This can to some extent also be tested in cases where models are known to share important components.

Conversely, a strength of the inter-model distance techniques is that an observational estimate can effectively be considered as just another model. Specifically, they allow us to measure distances between different observational estimates and compare them to inter-model distances, and perhaps visualise this in a low dimensional projected space (as shown in Fig. 1).

There is also a growing amount of work in the statistical literature that has not been applied to large model ensembles such as CMIP, but to conceptually related problems, which could more comprehensively define model–model and model–observation similarities (e.g. Chandler, 2013; Smith et al., 2009). These approaches view model outputs and observations as realisations of spatial, temporal or spatio-temporal random processes; hence, they can make use of the powerful framework of stochastic process theory. Combined with a Bayesian implementation, these approaches could in principle address many of the shortcomings of relatively simplistic techniques, such as those in the paragraphs above, like the need to use a single metric or to apply a single weight for each model or simulation. They could also allow for more comprehensive and transparent uncertainty quantification as part of the formulation of dependence. However, there are practical questions related to computation and details of the application, which are yet to be addressed.

6 Independence and performance

Figure 1 also highlights why an assessment of model independence can be at least partly conditional on model performance information. Suppose, for example, that a "radius of similarity" was used to identify dependent models in one of the inter-model distance spaces defined in the section above, as illustrated in Fig. 1 by the red shaded regions around models (this idea is raised in both Abramowitz, 2010 and Sanderson et al., 2015a, b). In Fig. 1a, models 1 and 4, by virtue of being relatively close together might be deemed dependent, and so somehow down-weighted relative to models 2 and 3.

In Fig. 1b the model positions are identical but observational data sets now lie between models 1 and 4, making the picture less clear. Models 1 and 4 both appear to perform very well (as model–observation distances are relatively small), and as they are spread around observational estimates, might be considered to be independent. In this sense, inter-model distances alone in the absence of observational data are an incomplete proxy for model independence. Both Abramowitz and Gupta (2008) and Sanderson et al. (2015a, b) address this issue by proposing model independence weights that scale cumulative model–model distances for an individual simulation by its proximity to observations. However, neither study explicitly addressed how multiple observational estimates might be incorporated, although there is also no theoretical barrier to this. Note that there is no implicit assumption here that observational estimates are close together, just that categorical statements about model dependence are less clear if they are not.

An alternative approach that combines model distance and performance information is to define model dependence in terms of model error covariance or error correlation (e.g. Jun et al., 2008a, b; Collins et al., 2010; Bishop and Abramowitz, 2013). This has the advantage that "error" only reflects deviations from an observational product (rather than similarity in model outputs per se), and while it still suffers from the integrative holism noted in the section above (that is, that error covariances are sensitive only to the integrated effect of all process representations), differences in the structure of error between models are likely to reflect differences in the sections of model representation that are not tightly constrained by observations. Incorporating different observational estimates in this case, unfortunately, is more complicated.

A little thought about the values of error correlation that we might expect between independent models reveals how problem-dependent accounting for model dependence can be. If, for example, we examine gridded climatological (time average) values of a variable of interest, then under the (flawed) assumption that an observational estimate is perfect, and the period in question is stable and long enough to define a climatology, departures from observed climatology might reflect a model's inability to appropriately simulate the system and so represent epistemic uncertainty. In this case, we might suggest that independent simulations should have pairwise zero error correlation, as is the case for independent random variables, since we might a priori believe climatology to be deterministically predictable (that is, that a perfect model should be able to match observations). Just as the mean of n uncorrelated random variables with variance 1 has variance $1/n$, we should expect that the ensemble mean of independent models defined in this way would (a) perform better than any individual simulation, and (b) asymptotically converge to zero error as the size of the ensemble of independent models (with zero error correlation) increases. This is illustrated in Fig. 2a, which shows 30 yellow lines, each of which is comprised of 50 draws from the normal distribution

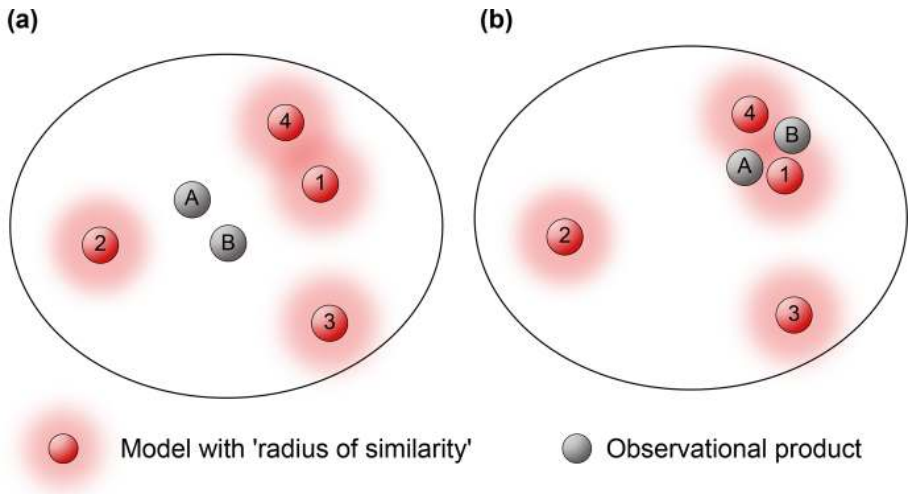

Figure 1. A two-dimensional projection of an inter-model distance space, showing different models and observational estimates, with a radius around models that could be used to determine model dependence. The radius around observations might be related to the uncertainty associated with a given observational estimate. Panels **(a)** and **(b)** illustrate how the relative position of observational data sets in this space could complicate this definition of model dependence.

$N(0, 0.125)$. Each of these lines is a conceptual representation of error from a different model, where zero error would be shown as the horizontal black line (imagine, for example, that the horizontal axis represented different points in space). The red line shows the mean of these 30 model error representations, and clearly has significantly reduced error variance. This understanding of independence within an ensemble has been dubbed the "truth-plus-error" paradigm (see Annan and Hargreaves, 2010, 2011; Knutti et al., 2010b; Bishop and Abramowitz, 2013; Haughton et al., 2014, 2015), and has often been assumed rather than explicitly stated (e.g. Jun et al., 2008a, b).

7 Independence and aleatory uncertainty

But the truth-plus-error framework is not always appropriate. Knutti et al. (2010b) noted that the ensemble mean of the CMIP3 ensemble did not appear to asymptotically converge to observations as ensemble size increased. Gleckler et al. (2008) also show that the variability of the ensemble mean is much less than individual models or observations – and so does not represent a potentially real climate. If we wish to consider ensemble simulations where unpredictability or aleatory uncertainty is an inherent part of the prediction, we no can longer expect that the system might be entirely deterministically predictable. This includes, for example, any time series prediction where internal climate variability between models and observations is out of phase (e.g. CMIP global temperature historical simulations or projections from 1850 initialisation), or climatology (mean state) prediction where the time period is too short to be invariant to initial state uncertainty. In these cases we accept that some component of the observational data is inherently unpredictable, even for a perfect model without any epistemic uncertainty. Ensemble spread in this case might ideally give an indication of the amount of variability we might expect from the chaotic nature of the climate system given uncertain initial conditions, and could be investigated using initial-condition ensembles of climate change projections (Kay et al., 2015; Deser et al., 2016) as well as in the context of numerical weather prediction ensembles (e.g. Hamill et al., 2000; Gneiting and Raftery, 2005).

A simple illustration of the role of aleatory uncertainty is shown in Fig. 3, taken from the Technical Summary of WG1 in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report. Internal variability within each climate model simulation (yellow lines) and observations (the black line) is out of phase, so that the variance of the multi-model mean (the red line) is significantly less than individual models or the observations. While ensemble spread here represents a combination of both epistemic and aleatory uncertainty, it should be clear that the lack of predictability caused by internal variability removes the expectation that the model ensemble should be centred on the observations.

A synthetic example illustrates this point. If we assume that observations of global mean temperature anomalies in Fig. 3 are well approximated by the sum of a linear trend and random samples from $N(0, 0.125)$ – the black line in Fig. 2b – then an ensemble of independent models that adhered to the truth-plus-error paradigm might look like the yellow lines in Fig. 2b. Each of these are the same "models" shown in Fig. 2a, but this time they are presented as a time series and shown as random deviations about the "observations" (rather than the zero line; "models" are shown instead of the model error). It is perhaps no surprise in this situation that the mean of this 30 member ensemble (the red line) very closely ap-

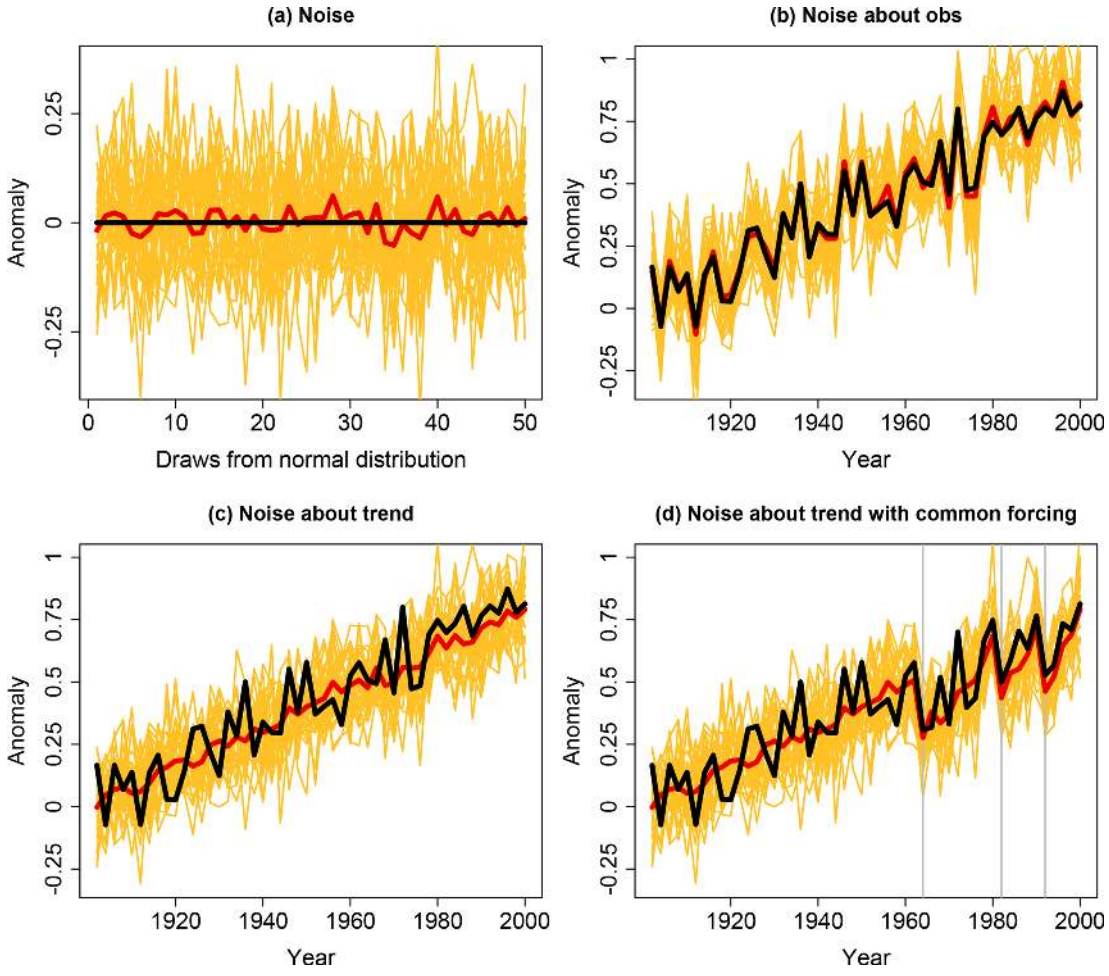

Figure 2. **(a)** The "truth-plus-error" paradigm illustrated using random samples from $N(0, 0.125)$ (yellow lines) as a proxy for error in models in an ensemble, with "observations" in black and the multi-model mean in red. **(b)** The same "models" shown as deviations from "observations" approximated by a noisy linear trend. In contrast, panel **(c)** illustrates model error and observation time series as draws from the same distribution, both shown as noise about a linear trend, and **(d)** displays the effect of applying common forcing perturbations to both models and observations. The same draws from $N(0, 0.125)$ are used in **(a)**–**(d)**.

proximates the "observations". This is clearly very different to the role of the ensemble mean we see in Fig. 3 (red curve).

An alternative to the truth-plus-error paradigm is to consider observations and models as being draws from the same distribution (e.g. Annan and Hargreaves, 2010, 2011; Bishop and Abramowitz, 2013; Abramowitz and Bishop, 2015). Figure 2c shows the same "observations" as Fig. 2b, but this time represents models in the same way as observations – the deviations from the linear trend (instead of deviations from the observations, as in 2b). In this case we can see that the ensemble mean (again in red) has much lower variability than observations, as seems evident for the first half of the 20th century in Fig. 3. By introducing external forcing common to the representations of models and observations in Fig. 2 – three step deviations that gradually return to the linear trend, intended to approximate volcanic forcing at the locations shown by grey lines in Fig. 2d – we can produce an entirely synthetic ensemble that very closely approximates what is shown in Fig. 3.

There are of course many reasons why what is shown in Fig. 2d is not an appropriate representation of models or observations. Collections of simulations such as CMIP are in reality a mix of epistemic and aleatory uncertainty, not just the aleatory uncertainty shown in Fig. 2. The nature of the perturbation that results from external forcing (such as the faux volcanoes in Fig. 2d), as well as the nature of internal variability itself, are also likely functionally dependent upon forcing history, and models exhibit different trends. Nevertheless, this simplistic statistical representation of ensemble spread closely approximates the nature of the CMIP ensemble.

Bishop and Abramowitz (2013) argued that independent climate simulations should have the statistical properties of the "models" in Fig. 2d. Specifically, as error-free observa-

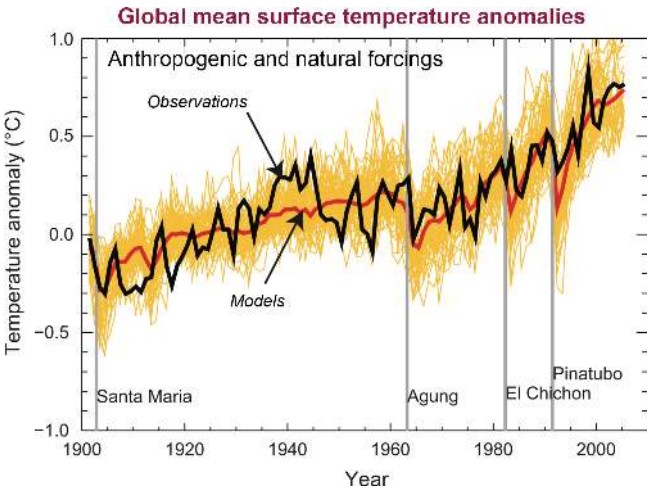

Figure 3. Figure TS.23 taken from the Technical Summary of WG1 in the IPCC Fourth Assessment Report, showing the multi-model mean's decreased variability (red) relative to individual models (yellow) and observations (black), as well as the effect of volcanic forcing on ensemble behaviour.

tions and perfect models (i.e. without epistemic uncertainty) would both be draws from the same distribution (they named samples from this climate PDF "replicate Earths"), they should both approximate the same level of variability about the mean of this distribution (which represents the forced signal), given enough time. They attempted to partially account for epistemic uncertainty in the CMIP ensemble by offering a transformation of models so that the transformed ensemble strictly adhered to two key statistical properties of this distribution defined entirely by aleatory uncertainty. These were, given a long enough time period, that (a) the best estimate to any particular replicate Earth is the equally weighted mean of a collection of other replicate Earths (that is, CPDF mean), and (b) the time average of the instantaneous variance of this distribution (the CPDF) across replicate Earths should approximate the variance of any individual replicate Earth about the CPDF mean over time. By treating observations as the only true replicate Earth, they transformed the CMIP ensemble to be replicate-Earth-like, with respect to these two properties.

Annan and Hargreaves (2010) also proposed that observations and models are best considered as draws from the same distribution. The meaning of ensemble spread in the "statistically indistinguishable" paradigm they propose, however, is not immediately clear, and is not explicitly stated in Annan and Hargreaves (2010). They do not discuss internal variability, but in a later blog post suggested spread represented "collective uncertainties about how best to represent the climate system", which seems to imply epistemic uncertainty. Both Annan and Hargreaves (2010) and Bishop and Abramowitz (2013) suggested that pair-wise error correlation between independent model simulations should be 0.5, as the observations are common to both.

Thus, categorical separation of epistemic and aleatory uncertainty is challenging, as it requires an accurate quantification of internal variability. While we have some tools that can help us estimate internal variability, ultimately we have measurements of just one realisation of a chaotic Earth system, and internal variability is affected by the state of the Earth system and forcing conditions (Brown et al., 2017). There is also evidence that the internal variability in some modelling systems (i.e. initial conditions ensembles – see Collins et al., 2001; Deser et al., 2012) may not be a good representation of internal variability in the climate system (e.g. Haughton et al., 2014, 2015). Each of the techniques that give an indication that ensemble spread is similar to internal variability, such as rank histograms (Hamill, 2001), spread-skill scores in forecasts, the Brier skill score (Brier, 1950; Murphy, 1973) and reliability diagrams (Wilks, 1995), also have the potential for misinterpretation (e.g. regional biases in an ensemble appearing as under-dispersiveness). In addition, timescales of internal variability are difficult to ascertain from our sparse and short observational record, but there is some evidence that it may operate on very long timescales (e.g. James and James, 1989; Ault et al., 2013; PAGES 2k Consortium, 2013). Therefore, while we have techniques for assessing and accounting for model dependence of epistemic uncertainty that try to nullify aleatory uncertainty by averaging over time, the potential for unquantified aleatory uncertainty to compromise this strategy remains real.

8 Robust strategies for addressing model dependence

Given that a priori measures of independence have yet to prove robust and that aleatory uncertainty could confound the ability to interpret model–observation distance as purely epistemic uncertainty, how might proposals to account for independence be interpreted? Recent experience suggests caution: accounting for dependence or performance differences within an ensemble can be very sensitive to the choice of variable, constraining observational data set, metric, time period and the region chosen. Herger et al. (2018a), for example, detail an approach that optimally selected subsets of an existing ensemble for properties of interest, such as the root mean square (RMS) distance of the sub-ensemble mean from observations of a variable's climatology. The resulting subsets are sensitive to nearly every aspect of the problem, including the following: which variables are considered; whether the weighting is inferred from climatological fields, time and space variability, or trends; and whether subsets are chosen before or after bias correcting model projections. Furthermore, for a given variable notably different sub-ensembles are obtained when using different constrain-

ing observational estimates; this is even found for relatively well-characterised quantities such as surface air temperature.

These types of results are familiar to researchers who utilise automated calibration techniques, and reinforce that post-processing to account for dependence or performance differences within an ensemble, whether by weighting or sub-selecting ensemble members, is essentially a calibration exercise. It also reinforces that thorough out-of-sample testing is needed before one might be confident that weighting or ensemble sub-selection will improve climate projections or an impacts assessment. It is clear that most post-processing approaches improve ensembles as intended utilising the data used to derive them (that is, they work well in-sample, typically using historical data). But how can we have confidence that this is relevant for the projection period?

In the context of climate projections we have at least two mechanisms to assess whether the observational data and experimental set-up used to derive the weights or ensemble subset provide adequate constraint for the intended application. The first is a traditional calibration–validation framework, where available historical data is partitioned into two (or more) sets, with the first used to calibrate weights or a sub-ensemble, and the second used to test their applicability out-of-sample (e.g. Bishop and Abramowitz, 2013). For most regional to global climate applications, this will often be limited to 60 or fewer years of high quality observational data – depending on the region and variable – and if the intended application period is far enough into the future, the nature of climate forcing in the calibration and validation periods might not be sufficiently representative for the application. For some quantities, palaeoclimate records also offer the potential for calibration–validation testing but their application to weighting and/or sub-selection has been limited (see Schmidt et al., 2014).

The second approach is model-as-truth, or perfect model experiments (e.g. Abramowitz and Bishop, 2015; Boeì and Terray, 2015; Sanderson et al., 2017; Knutti et al., 2017; Herger et al., 2018a, b). This involves removing one of the ensemble members and treating it as if it were observations. The remaining ensemble is then calibrated (that is, weighted or sub-selected) towards this "truth" member, using data from the historical period only. The calibrated ensemble can then be tested out-of-sample in the 21st century as the "truth" member's projections are known. The process is repeated with each ensemble member playing the "truth" role, and in each case, the ability of the sub-selection or weighting to offer improvement over the original default ensemble is assessed. A weighted ensemble can be compared to the equally weighted original ensemble mean; in the case of ensemble sub-selection, comparison can be with the entire original ensemble, or a random ensemble of the same size as the subset. Results are synthesised across all model-as-truth cases to gain an understanding of the efficacy of the particular approach being tested.

Perfect model tests are most informative when the similarity of ensemble members is approximately equal to the similarity between observations and ensemble members, in metrics that are relevant to the calibration process and application. For example, if a model-as-truth experiment were performed using all CMIP ensemble members, including multiple initial conditions members from the same model, the ensemble calibration process could fit the "truth" simulation much more closely than models are likely to be able to fit observational data. That is, weighting or sub-selection would favour any simulations from the same model as the truth ensemble member, so that the experiment's success might be misleading. This suggests eliminating obvious duplicates before the perfect model tests (see e.g. Fig. 5 in Sanderson et al., 2017). It is also worth emphasising that the motivation for this process is not to test the weights or ensemble subset as far out-of-sample as possible, but rather to ensure that the calibration process is appropriate for its intended application. Note that biases shared among models, especially those which affect projections, will increase agreement among models relative to observations, so that model-as-truth experiments should be treated as a necessary but not sufficient condition for out-of-sample skill.

Thorough out-of-sample testing is important for a number of different reasons. The first, and perhaps most obvious, is to ensure against overfitting due to sample size. We need to make certain that the weighting or sub-selection approach we use has been given enough data to appropriately characterise the relationship between the models we are using, especially if there are many of them, and the constraining observational data. A naive rule of thumb for any simple regression problem is roughly 10 times the number of data points as there are predictors (models in this case). While covariance between data points can complicate this rule, it should give an indication of whether any poor performance in out-of-sample testing is simply due to a paucity of observational data.

A second reason is "temporal transitivity", making sure that the time period and timescale used to calibrate the weights or ensemble subset provides adequate constraint on the intended application period and timescale. For example, Herger et al. (2018a) found that selecting an ensemble subset to minimise climatological surface air temperature bias in the historical period (1956–2013) provided good out-of-sample performance in 21st century (2013–2100) model-as-truth experiments. When this was repeated using linear surface air temperature trend instead, good in-sample improvements were not replicated out-of-sample. That is, the biases in climatology had high temporal transitivity, or predictability out-of-sample, while the biases in trend did not. This example illustrates why temporal transitivity is particularly important in the case of future projections. It is possible to have two models that have similar behaviour in current climate, for example because the models have both been developed with the same observational data sets for comparison, yet have very different climate sensitivities. As well as temporal pe-

riod transitivity, one might also consider transitivity between one timescale and another (e.g. the relevance of calibration using monthly data for daily extremes).

The third aspect of out-of-sample testing to consider is "metric transitivity". That is, ensuring that the metric used to weight or sub-select the ensemble constrains the quantity of interest that the ensemble will ultimately be used for. There are many examples of published work where metric transitivity was simply assumed. Abramowitz and Bishop (2015) assumed that historical RMS distance in gridded time and space fields of surface air temperature and precipitation informed global mean temperatures and end-of-21st-century projections. Sanderson et al. (2015) assumed that optimising for dependence in a multivariate seasonal climatology provided a constraint on climate sensitivity. There is of course no a priori reason why these assumptions should be valid, and indeed they could be tested with appropriate model-as-truth analysis.

Given observational constraints, spatial transitivity may also be relevant to out-of-sample testing, depending on the particular application. This could apply both to calibrating using one region and testing at another, and calibrating at one spatial scale and applying at another. For example, Hobeichi et al. (2018) tested the ability of the weighting approach outlined in Bishop and Abramowitz (2013) to offer performance improvements in locations not used to derive weights, as well as the ability of site-scale measurements to improve global 0.5° gridded evapotranspiration estimates. One might also evaluate CMIP simulations at coarse spatial scales as a way of deciding which subset of simulations to regionally downscale, making the assumption that low resolution performance will translate to downscaled performance.

The need for out-of-sample testing to ensure that overfitting and transitivity are not issues for a given application applies equally to the use of bias correction techniques (e.g. Macadam et al., 2010; Ehret et al., 2012; Maraun, 2016), emergent constraints (Nijsse and Dijkstra, 2018) and indeed to the entire chain of models used in downstream climate applications (e.g. Clark et al., 2016; Gutiérrez et al., 2018), although there are very few examples of this being done.

9 Towards generalised ensemble calibration

The sensitivity of weighting and sub-ensemble selection to metric, variable, observational estimate, location, time and spatial scale, and calibration time period underscores that model dependence is not a general property of an ensemble, but is application-specific. Dependence is also not a property of a model simulation per se, as performance is, but is rather a property of the simulation with respect to the rest of the ensemble. Nevertheless, in instances where a very specific variable and cost function are known to be the only properties of interest, it is quite likely that, with an appropriate out-of-sample testing regime, a solution to improve projection reliability can be found using existing techniques.

However, if we decided that application-specific calibration was not generally satisfactory, and that we wanted to try to calibrate a given CMIP ensemble for model dependence without knowing the intended application, how would we do this in a way that would be defensible? Given the increasing number and range in quality of CMIP contributions, it might be useful to suggest a strategy for general ensemble pre-processing for a range of applications.

We propose that using one model from each modelling institution that submitted to CMIP is the best general-purpose selection strategy. This strategy has proved a reasonable approximation to more detailed quantitative approaches that account for model dependence in the CMIP5 ensemble (Abramowitz and Bishop, 2015; Leduc et al., 2016b). This "institutional democracy" approach requires two important caveats, namely care in excluding models that are near-copies of one another submitted by different institutions and equal care in including models from the same institution with significantly different approaches or assumptions. Given due diligence, institutional democracy is a simple but reasonably effective approach to accounting for model dependence which, we argue, provides a better basis on which to calculate a naive multi-model average for generic purposes such as projection best estimates in IPCC reports.

Institutional democracy as an a priori approach is not bound by any particular statistical metric, variable or observational estimate. However, as institutions increasingly copy or co-develop whole models or components, there is no guarantee that such an approach will remain effective in the future.

The approach is similar in spirit to one proposed by Boé (2018) to account for the number of shared components by GCMs. Boé's approach quickly becomes difficult and time consuming for large ensembles such as CMIP, given the lack of transparency regarding precisely what constitutes different models and the role of tuning. Using this information to account for dependence would also likely be difficult, as categorical inclusion or exclusion of simulations seems the only option. Also we note that shared history as it pertains to dependence should only include process representations that are not tightly observationally constrained (so that Navier–Stokes equations might not represent dependent process treatment, for example), as discussed above – model convergence might well imply accuracy, rather than dependence. We note that these issues apply equally to version democracy (as per Boé, 2018) and institutional democracy.

A more comprehensive a posteriori approach to generalised calibration might be to simultaneously optimise for all of the variables, metrics and observational estimates believed to be informative. The simplest way to do this is to combine all the relevant cost functions into a single cost function for optimisation, resulting in a single optimal ensemble, or set of weights for model runs in the ensemble. This solu-

tion, while tractable, has at least two potential disadvantages. First, it makes assumptions about the relative importance of each term in the final cost function that are hard to justify. Different variables and metrics have different units, so these need to be standardised in some way. Given that the shapes of the distributions of different variables can be very different, the approach to standardisation will impact the nature of the final weights or ensemble sub-selection.

Second, this approach risks underestimating uncertainty. If calibration against one key cost function (for example, surface air temperature climatology) gives a very different ensemble subset or weights to calibration against another (say precipitation extremes), the discrepancy between these optimised outcomes is important information about how certain our optimal estimates are. If, for example, there was a universally independent subset within the larger ensemble, it would be the same subset for both variable optimisations. The discrepancy between them is an indication of the degree to which our model ensemble is not commensurable with the observations of the climate system we are trying to simulate. As noted above, one way to try to minimise overconfidence (underestimated uncertainties) is to use model-as-truth tests to test for variable and metric transitivity, although this cannot avoid the issue of shared model assumptions.

These difficulties may be avoided if we take a more expansive view of what optimisation means. A broader approach could use multiple criteria separately (e.g. Gupta et al., 1999; Langenbrunner and Neelin, 2017) in the context of ensemble sub-selection. One might, for example, examine the spread of results from both the surface air temperature climatology optimal ensemble and the precipitation extremes optimal ensemble when considering use of either variable. More generally, we propose that the optimal ensembles obtained from optimising against each relevant data set–variable–metric combination of interest are all effectively of equal value. A generalised ensemble calibration would utilise all of these ensembles for projection – that is, it would use an ensemble of ensembles – and ideally employ a Pareto set of ensembles. This would give a far better description of the uncertainty involved with projection, as uncertainty due to models' inability to simultaneously simulate a range of aspects of the climate system can now be expressed as uncertainty in a single variable and metric. This remains a proposal for thorough exploration at a later date.

10 Recommendations and next steps

As we have discussed, it is unlikely that model dependence can be defined in a universal and unambiguous way. In the absence of easy and agreed-upon alternatives, many studies still use the traditional "model democracy" approach; indeed, it seems unlikely that a "one size fits all" approach or list of good and independent models would be meaningful. However, we argue that users do have a suite of out-of-sample testing tools available that allow the efficacy of any weighting or sub-sampling approach to be tested for a particular application. These should be applied to understand the validity of both the technique itself and the intended application in terms of metric, temporal and spatial transitivity, as discussed above. If out-of-sample results are robust, they give an indication of the degree to which dependence affects the problem at hand, by way of contrast with the equally weighted status quo. As many of the studies discussed above have shown, accounting for dependence can give markedly different projections. This is particularly important for the upcoming CMIP6, where the model dependence issue is expected to increase. Even the somewhat naive approaches of institutional democracy (Leduc et al., 2016b) or component democracy (Boeì, 2018) are likely to be less biased approaches to ensemble sampling. Nonetheless, we discourage the use of weighting or sub-sampling without out-of-sample testing, as the risks may well outweigh the potential benefits (Weigel et al., 2010; Herger et al., 2018b).

For most applications, questions of how best to select physically relevant variables, domains and appropriate metrics remain open. This should ideally be done by considering the relevant physical processes for the phenomena in question. For the cases where a specific variable and scale is clear, a comparison of existing approaches, a discussion of the circumstances in which they should be used and the construction of an appropriate out-of-sample testing regime would help guide users' choices. If a more holistic ensemble calibration is needed, further exploration of the idea of multi-objective optimisation is required, which results in the novel concept of an ensemble of model subsets or weighted averages.

There has also recently been a push for more transparency regarding models' development history. This is relevant for the a priori approaches, which are based on similarity of model codes. In terms of those approaches, there is also a need to explore the impact of tuning on dependence (Schmidt et al., 2017; Hourdin et al., 2017). It has been shown that parameter perturbations based on otherwise identical code bases (such as in the climateprediction.net exercise; Mauritsen et al., 2012) can lead to notably different projections. Better documentation from the modelling institutions regarding the standard metrics used to judge a model's performance during development and the preferred observational products used for tuning is needed. This information can help determine the effective number of independent models in an ensemble in relation to the actual number of models for a given application.

We stress again that the simulations made for CMIP do not represent a designed ensemble. In particular the simulations do not span the full uncertainty range for GCM projections or systematically sample the set of all possible model configurations. This is something to keep in mind for any subset-selection or weighting approach. Moreover, it is currently unclear how to deal with the situation when models

start to converge on the true climate state, which might occur as models resolve more and more processes. In such a situation, despite considering the models to be interdependent, we do not want to eliminate them. This might get even more complex when dealing with observational uncertainty.

Model-as-truth analyses are essential to test the skill of any weighting or sub-setting approach out-of-sample. However, they are necessary but not sufficient tests and have the potential for overconfidence given that many climate models are based on similar assumptions and are thus not truly independent. While some steps can be taken to ameliorate this issue, due to the central role of such analyses, the limits of these tests should be explored in future studies and guidelines provided regarding how to best set them up.

11 Conclusions

With model component and process representation replication across nominally different models in CMIP5, and the anticipation of more to come in CMIP6, the need for an effective strategy to account for the dependence of modelled climate projection estimates is clear. Perhaps the biggest obstacle to doing this is that the manifestation of model dependence is problem-specific, meaning that any attempt to address it requires an approach tailored to individual projection impact analyses. We presented a holistic framework for understanding the diverse and apparently disparate collection of existing approaches to addressing model dependence, noting that each addresses slightly different aspects of the problem.

Critically, we reinforce that the efficacy of any attempt to weight or sub-select ensemble members for model dependence or performance differences, or indeed bias correction, must be tested out-of-sample in a way that emulates the intended application. Calibration–validation with different time periods within the observational record, as well as model-as-truth experiments were discussed as two approaches to doing this.

Universal calibration of an ensemble for model dependence that is not specific to a particular application remains elusive. In that context, preselecting simulations based on an a priori knowledge of models, using institutional democracy (one model per institute, with the additional removal of any supplementary simulations that are sourced from known model replicates at different modelling institutes and/or the addition of clearly distinct variants from within a single institution, see Leduc et al., 2016b), or component democracy (as detailed in Boé, 2018) is more defensible than naive use of all available models in many applications.

The final step of relating dependence in model output to similarities in model structure can only be achieved once we have a transparent system for documenting and understanding the differences in the treatment of processes, and tuning, between different climate models. While there are some ad-hoc examples of attempts to do this (e.g. Fig. 5 in Edwards, 2011; Masson and Knutti, 2011), a formal requirement to document the nature of model structure, parameter evolution and freely available source code would be a welcome step that would spawn new areas of enquiry in this field (Ince et al., 2012). This would ultimately result in a more effective investment in model components that provide independent projection information and bring the community a step closer to producing well calibrated ensembles for climate projection.

Data availability. No data sets were used in this article.

Author contributions. GA prepared the paper with contributions from all co-authors. NH and GA developed the conceptual illustration in Fig. 1, and GA developed the conceptual illustration in Fig. 2.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We wish to thank the Australian Research Council Centre of Excellence for Climate System Science for funding the workshop in December 2016 at the National Center for Atmospheric Research (NCAR) that seeded this paper, as well as the NCAR IMAGE group for hosting the workshop and providing administrative support. NCAR is sponsored by the US National Science Foundation. Thanks also to the World Climate Research Program Modelling Advisory Council that endorsed the workshop. Extensive discussions at a second workshop in July 2017 at the Aspen Global Change Institute on Earth System Model Evaluation to Improve Process Understanding also contributed significantly to the content of this work. Gab Abramowitz is supported by the ARC Centre of Excellence for Climate Extremes (grant no. CE170100023), and Nadja Herger is funded by the ARC Centre of Excellence for Climate System Science (grant no. CE110001028). Ruth Lorenz is supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 641816 (CRESCENDO). Tom Hamill, James Annan and Julia Hargreaves provided valuable feedback.

Edited by: Somnath Baidya Roy
Reviewed by: two anonymous referees

References

Abramowitz, G.: Model independence in multi-model ensemble prediction, Aust. Meteorol. Ocean., 59, 3–6, 2010.

Abramowitz, G. and Gupta, H.: Toward a model space and model independence metric, Geophys. Res. Lett., 35, L05705, https://doi.org/10.1029/2007GL032834, 2008.

Abramowitz, G. and Bishop, C. H.: Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections, J. Climate, 28, 2332–2348, 2015.

Annan, J. D. and Hargreaves, J. C.: Reliability of the CMIP3 ensemble, Geophys. Res. Lett., 37, L02703, https://doi.org/10.1029/2009GL041994, 2010.

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

Annan, J. D. and Hargreaves, J. C.: On the meaning of independence in climate science, Earth Syst. Dynam., 8, 211–224, https://doi.org/10.5194/esd-8-211-2017, 2017.

Ault, T. R., Cole, J. E., Overpeck, J. T., Pederson, G. T., St. George, S., Otto-Bliesner, B., Woodhouse, C. A., and Deser, C.: The continuum of hydroclimate variability in western North America during the last millennium, J. Climate, 26, 5863–5878, https://doi.org/10.1175/JCLI-D-11-00732.1, 2013.

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dynam., 41, 885–900, https://doi.org/10.1007/s00382-012-1610-y, 2013.

Boé, J.: Interdependency in multi-model climate projections: component replication and result similarity, Geophys. Res. Lett., 45, 2771–2779, https://doi.org/10.1002/2017GL076829, 2018.

Boé, J. and Terray, L.: Can metric-based approaches really improve multi-model climate projections? The case of summer temperature change in France, Clim. Dynam., 45, 1913–1928, 2015.

BOX, G. E. P.: Robustness in the Strategy of Scientific Model Building, in: Robustness in Statistics, edited by: Launer, R. L. and Wilkinson, G. N., Academic Press, Inc., New York, 1979.

Brier, G. W.: Verification of forecasts expressed in terms of probabilities, Mon. Weather Rev., 78, 1–3, 1950.

Brown, P. T., Ming, Y., Li, W., and Hill, S. A.: Change in the Magnitude and Mechanisms of Global Temperature Variability With Warming, Nat. Clim. Change, 7, 743–748, https://doi.org/10.1038/nclimate3381, 2017.

Chamberlin, T. C.: The method of multiple working hypotheses, Science, 15, 92–96, 1890.

Chandler, R. E.: Exploiting strength, discounting weakness: combining information from multiple climate simulators, Philos. T. R. Soc. A, 371, 20120388, https://doi.org/10.1098/rsta.2012.0388, 2013.

Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., Fowler, H. J., Prudhomme, C., Arnold, J. R., and Brekke, L. D.: Characterizing Uncertainty of the Hydrologic Impacts of Climate Change, Curr. Clim. Change Rep., 2, 1–10, https://doi.org/10.1007/s40641-016-0034-x, 2016.

Collins, M., Tett, S. F. B., and Cooper, C.: The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments, Clim. Dynam., 17, 61–81, https://doi.org/10.1007/s003820000094, 2001.

Collins, M., Booth, B. B., Bhaskaran, B., Harris, G. R., Murphy, J. M., Sexton, D. M. H., and Webb, M. J.: Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles, Clim. Dynam., 36, 1737–1766, https://doi.org/10.1007/s00382- 010-0808-0, 2010.

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: The role of internal variability, Clim. Dynam., 38, 527–546, https://doi.org/10.1007/s00382-010-0977-x, 2012.

Deser, C., Terray, L., and Phillips, A. S.: Forced and internal components of winter air temperature trends over north america during the past 50 years: mechanisms and implications, J. Climate, 29, 2237–2258, 2016.

Edwards, P.: History of climate modeling, WIRES Clim. Change, 2, 128–139, https://doi.org/10.1002/wcc.95, 2011.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: HESS Opinions "Should we apply bias correction to global and regional climate model data?", Hydrol. Earth Syst. Sci., 16, 3391–3404, https://doi.org/10.5194/hess-16-3391-2012, 2012.

Evans, J., Ji, F., Abramowitz, G., and Ekstrom, M.: Optimally choosing small ensemble members to produce robust climate simulations, Environ. Res. Lett., 8, 044050, https://doi.org/10.1088/1748-9326/8/4/044050, 2013.

Gleckler, P., Taylor, K., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, https://doi.org/10.1029/2007JD008972, 2008.

Gneiting, T. and Raftery, A. E.: Weather forecasting with ensemble methods, Science, 310, 248–249, https://doi.org/10.1126/science.1115255, 2005.

Gupta, H. V., Bastidas, L. A., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L.: Parameter estimation of a land surface scheme using multicriteria methods, J. Geophys. Res., 104, 19491–19503, 1999.

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Roessler, O., Wibig, J., Wilcke, R., Kotlarski, S., San Martín, D., Herrera, S., Bedia, J., Casanueva, A., Manzanas, R., Iturbide, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Pórtoles, J., Räty, O., Räisänen, J., Hingray, B., Raynaud, D., Casado, M. J., Ramos, P., Zerenner, T., Turco, M., Bosshard, T., Štěpánek, P., Bartholy, J., Pongracz, R., Keller, D. E., Fischer, A. M., Cardoso, R. M., Soares, P. M. M., Czernecki, B., and Pagé, C.: An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment, Int. J. Climatol., https://doi.org/10.1002/joc.5462, 2018.

Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Mon. Weather Rev., 129, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Hamill, T. M., Mullen, S. L., Snyder, C., Baumhefner, D. P., and Toth, Z.: Ensemble forecasting in the short to medium range: Report from a workshop, B. Am. Meteorol. Soc., 81, 2653–2664, https://doi.org/10.1175/1520-0477(2000)081%3C2653:EFITST%3E2.3.CO;2, 2000.

Haughton, N., Abramowitz, G., Pitman, A., and Phipps, S.: On the generation of climate model ensembles, Clim. Dynam., 43, 2297–2308, https://doi.org/10.1007/s00382-014-2054-3, 2014.

Haughton, N., Abramowitz, G., Pitman, A., and Phipps, S. J.: Weighting climate model ensembles for mean and variance estimates, Clim. Dynam., 45, 3169–3181, https://doi.org/10.1007/s00382-015-2531-3, 2015.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in regional climate predictions, B. Am. Meteorol. Soc., 90, 1095–1107, https://doi.org/10.1175/2009bams2607.1, 2009.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in projections of regional precipitation change, Clim. Dynam., 37, 407–418, https://doi.org/10.1007/s00382-010-0810-6, 2011.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, Earth Syst. Dynam., 9, 135–151, https://doi.org/10.5194/esd-9-135-2018, 2018a.

Herger, N., Angélil, O., Abramowitz, G., Donat, M., Stone, D., and Lehmann, K.: Calibrating climate model ensembles for assessing extremes in a changing climate, J. Geophys. Res.-Atmos., 123, 5988–6004, https://doi.org/10.1029/2018JD028549, 2018b.

Hobeichi, S., Abramowitz, G., Evans, J., and Ukkola, A.: Derived Optimal Linear Combination Evapotranspiration (DOLCE): a global gridded synthesis ET estimate, Hydrol. Earth Syst. Sci., 22, 1317–1336, https://doi.org/10.5194/hess-22-1317-2018, 2018.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Klocke, D. J. D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson, D.: The art and science of climate model tuning, B. Am. Meteorol. Soc., 98, 589–602, https://doi.org/10.1175/BAMS-D-15-00135.1, 2017.

Ince, D. C., Hatton, L., and Graham-Cumming, J.: The case for open computer programs, Nature, 482, 485–488, https://doi.org/10.1038/nature10836, 2012.

James, I. N. and James, P. M.: Ultra low frequency variability in a simple global circulation model, Nature, 342, 53–55, https://doi.org/10.1038/342053a0, 1989.

Jun, M., Knutti, R., and Nychka, D.: Spatial analysis to quantify numerical model bias and dependence: how many climate models are there?, J. Am. Stat. Assoc., 103, 934–947, 2008a.

Jun, M., Knutti, R., and Nychka, D. W.: Local eigenvalue analysis of CMIP3 climate model errors, Tellus, 60A, 992–1000, https://doi.org/10.1111/j.1600-0870.2008.00356.x, 2008b.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., and Bates, S. C.: The community earth system model (CESM) large ensemble project: a community resource for studying climate change in the presence of internal climate variability, B. Am. Meteorol. Soc., 96, 1333–1349, 2015.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple models, J. Climate, 23, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1, 2010b.

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

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophys. Res. Lett., 44, 1909–1918, https://doi.org/10.1002/2016GL072012, 2017.

Kohonen, T.: Self-Organization and Associative Memory, Springer, New York, 1989.

Langenbrunner, B. and Neelin, J. D.: Pareto-optimal estimates of California precipitation change, Geophys. Res. Lett., 44, 12436–12446, https://doi.org/10.1002/2017GL075226, 2017.

Leduc, M., Matthews H. D., and de Elia, R.: Regional estimates of the transient climate response to cumulative CO_2 emissions, Nat. Clim. Change, 6, 474–478, 2016a.

Leduc, M., Laprise, R., De Elía, R., and Separovic, L.: Is Institutional Democracy a Good Proxy for Model Independence?, J. Climate, 29, 8301–8316, https://doi.org/10.1175/JCLI-D-15-0761.1, 2016b.

Lenhard, J. and Winsberg, E.: Holism, entrenchment, and the future of climate model pluralism, Stud. Hist. Philos. M. P., 41, 253–262, 2010.

Lorenz, R., Herger, N., Sedlaìček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America, J. Geophys. Res.-Atmos., 123, 4509–4526, https://doi.org/10.1029/2017JD027992, 2018.

Macadam, I., Pitman, A. J., Whetton, P. H., and Abramowitz, G.: Ranking climate models by performance using actual values and anomalies: Implications for climate change impact assessments, Geophys. Res. Lett., 37, L16704, https://doi.org/10.1029/2010GL043877, 2010.

Maraun, D.: Bias Correcting Climate Change Simulations – a Critical Review, Curr. Clim. Change Rep., 2, 211, https://doi.org/10.1007/s40641-016-0050-x, 2016.

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

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H., and Tomassini, L.: Tuning the climate of a global model, J. Adv. Model. Earth Syst., 4, M00A01, https://doi.org/10.1029/2012MS000154, 2012.

Murphy, A. H.: A new vector partition of the probability score, J. Appl. Meteorol., 12, 595–600, 1973.

Nijsse, F. J. M. M. and Dijkstra, H. A.: A mathematical approach to understanding emergent constraints, Earth Syst. Dynam., 9, 999–1012, https://doi.org/10.5194/esd-9-999-2018, 2018.

Oreskes, N., Shrader-Frechette, K., and Belitz, K.: Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences, Science, 263, 641–646, 1994.

PAGES 2k Consortium: Continental-scale temperature variability during the past two millennia, Nat. Geosci., 6, 339–346, https://doi.org/10.1038/ngeo1797, 2103.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, J. Climate, 28, 5171–5194, 2015a.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties, J. Climate, 28, 5150–5170, 2015b.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geosci. Model Dev., 10, 2379–2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

Schmidt, G. A., Annan, J. D., Bartlein, P. J., Cook, B. I., Guilyardi, E., Hargreaves, J. C., Harrison, S. P., Kageyama, M., LeGrande, A. N., Konecky, B., Lovejoy, S., Mann, M. E., Masson-Delmotte, V., Risi, C., Thompson, D., Timmermann, A., Tremblay, L.-B., and Yiou, P.: Using palaeo-climate comparisons to constrain future projections in CMIP5, Clim. Past, 10, 221–250, https://doi.org/10.5194/cp-10-221-2014, 2014.

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., Molod, A., Neale, R. B., and Saha, S.: Practice and philosophy of climate model tuning across six US modeling centers, Geosci. Model Dev., 10, 3207–3223, https://doi.org/10.5194/gmd-10-3207-2017, 2017.

Smith, R. L., Tebaldi, C., Nychka, D., and Mearns, L. O.: Bayesian Modeling of Uncertainty in Ensembles of Climate Models, J. Am. Stat. Assoc., 104, 97–116, 2009.

Tebaldi, C. and Knutti, R.: The use of the multimodel ensemble in probabilistic climate projections, Philos. T. Roy. Soc. A, 365, 2053–2075, https://doi.org/10.1098/rsta.2007.2076, 2007.

Weigel, A. P., Liniger, M. A., and Appenzeller, C.: The discrete Brier and ranked probability skill scores, Mon. Weather Rev., 135, 118–124, 2007.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of Model Weighting in Multimodel Climate Projections, J. Climate, 23, 4175–4191, https://doi.org/10.1175/2010JCLI3594.1, 2010.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences: An Introduction, Academic Press, San Diego, 467 pp., 1995.