

## Review – Annan and Hargreaves (model independence)

The issues (plural) that are discussed in this paper are important, but the discussion is muddled by the consideration of independence in two (admittedly related) problems – model independence and the independence of constraints on climate system properties. I agree with the first reviewer that it might not be helpful to attempt to discuss both in the same paper.

### Model independence:

From a probabilistic perspective, independence is discussed in terms of “events”, which are collections of elements (subsets of a sample space) that describe the multiple ways in which an “event” might occur. From a statistical perspective, this translates into a discussion in terms of random variables (functions defined on the sample space) where equations such as  $X = x$  or  $X \in (x-\epsilon, x+\epsilon)$  for some small  $\epsilon > 0$  describe events. Here  $X$  indicates a random variable (a function that maps the sample space onto the space is observed), and  $x$  indicates a realization of that random variable (the particular value that is observed).

The only way I can conceive of the question of model independence is to start with the notion that we have available an ensemble of realizations  $\{m_1, \dots, m_n\}$  of random variables  $\{M_1, \dots, M_n\}$  where

- $m_i$  is the model (the entire model, not just a simulated temperature or whatever) that is the end-point of a model development process or the end point of an effort to set model parameters, and
- $M_i$  represents all of the different outcomes that would have been possible as a result of the  $i$ 'th model development / parameter selection process.

There are constraints on  $M_i$  that originate from the laws of physics (thus realizations of  $M_i$  cannot simply be random collections of code), but within those constraints, one could in principle consider whether random variables  $M_i$  are independent or perhaps even identically distributed.

The general question is not tractable in my view (whether you are a frequentist or a Bayesian) because while we may have a general notion of how to construct the sample space, we are unable to describe how that space is sampled by the model development / parameter selection process, and thus we are unable to describe the distribution of  $M_i$ . We might only suspect that these random variables are not all independent, or at least, that they are not identically distributed, since priorities, resources and stopping rules for the model development process differ between modelling centres. These efforts also sometimes lead to multiple versions by “branching” the model development process close to the time when the development process ends, presumably leading to dependence and common biases. Evidence that we see in the CMIP experiments that is suspected to be due to lack of model “independence” because it is associated with structural commonalities

between models, might actually be evidence that the  $M_i$  are not identically distributed. That is, the two  $i$ 's in "iid" might be confounded.

More specific questions, where something is known about the sampling process, such as in perturbed parameter experiments using latin hypercube sampling, are more tractable. In this case lack of independence presumably arises due to changes in nonlinear interactions between parameterized processes when parameter values change.

Independence of constraints:

While the Bayesian formalism used to explore this question is the same as that used by the authors to explore the question of model independence, the question is rather different in that it concerns the independence of observables. Thus conceptually, the source of randomness that imparts distributions on the observables is rather different. I agree with reviewer 1 that this would more appropriately be discussed in another paper.

Some specific comments (page number, line number):

1,16: This statement supposes that identical initial conditions are used each time – which does not reflect the way in which multi-simulation ensembles with a given model are constructed. Typically, initial conditions are varied between runs so that different realizations of internal variability can be sampled across the ensemble.

1, 19: Why would the use of known, accepted physics be a problem?

2, 11: "magnitude" → "potential magnitude".

3, 14-15: This statement implies that we know the truth – but reality as we know it is subject to a substantial amount of observational uncertainty and the effects of unforced internal variability. Whether truly independent or not, I think it is reasonable to think that internal variability is effectively beaten down to very small levels in the multi-model ensemble mean. In contrast, this is a source of uncertainty in observations that can only be reduced by increasing the length of the observational record.

4, 20: There seems to be a small unintentional double entendre here. "Value" can be interpreted in two different ways – you could debate whether users are overly fixated on the equilibrium climate sensitivity (a property of the climate system that is not directly observable), or you could be concerned about actually estimating this number.

- 5, 1-5: It might be worth pointing out that sequential updating does not necessarily require independence, but does require less than full dependence to be useful, and an understanding of the dependence structure.
- 5, 19-20: The description on lines 13-19 is not specific to discrete or continuous distributions. Rather, the way in which this translates into statements concerning discrete or continuous distributions depends upon the nature of the random variable (i.e., the function that maps points in a sample space onto the values that are observed).
- 6, 6-9: It seems to me that Box 12.2 was very clear – they described the state of the literature by writing “The peer-reviewed literature provides no consensus on a formal statistical method to combine different lines of evidence”, and by providing reasons for that assessment in the preceding discussion. That doesn’t strike me as being “strange” at all. Consensus would imply an approach that is used broadly across the community – and evidently that did not yet exist at the time when the AR5 report was written.
- 10, 2: This looks like a classic bias/variance trade-off.
- 10, 32-33: This begs the question of why this hasn’t been done in this paper for CMIP5.
- 12, 8-9: Some justification for these priors would be appropriate.