Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2019-34-SC1, 2019

Earth System
Dynamics
Discussions

Interactive comment on "Emulating Earth System Model temperatures: from global mean temperature trajectories to grid-point level realizations on land" *by* Lea Beusch et al.

Robert Link

robert.link@pnnl.gov

Received and published: 10 July 2019

This paper does some interesting work toward systematizing the way we construct climate model emulators, which could be very useful for comparing different kinds of emulators and for designing interoperable components for emulating climate models.

I would have liked to see a little more depth in section 6.3, "Quantitative verification". The authors show plots comparing the quantiles of the emulator-generated ensemble to the corresponding quantiles of the CMIP ensemble, for three regions, and they remark that "the median [of the CMIP ensemble] is successfully emulated, but the emulations are a bit underdispersive", but this assessment seems to be based entirely

on visual inspection of Figure 8. This analysis would be a lot more compelling if it included quantitative statistical tests, such as a t-test for equality of the means and the Kolmogorov-Smirnov test for equivalence of the overall distributions. If underdispersion is a particular concern, tests for equality of variances could also be applied. Better still would be to develop measures of differences in key properties of the distribution and to derive confidence intervals for those difference measures. Such measures would give prospective users the tools they need to evaluate whether an emulator is fit for whatever use they intend to put it to.

In addition to concerns about how these marginal distributions are evaluated, the marginal distributions appear to be the only dimension along which the authors evaluate the emulator performance. There is no mention at all of testing the spatial correlation or time correlation properties of the emulator. This is a significant omission because the marginal distributions are surely the easiest part to get right when designing an emulator. Capturing the space and time correlations is the true test of the algorithm. In particular, we know that both ESMs and the real climate system display long-range teleconnections and quasi-periodic oscillatory behavior with periods ranging from years to decades. In order to truly evaluate the emulator algorithm, the authors need to explore its ability to produce these phenomena.

The authors' choice to do out of sample validation was interesting, but I am unsure as to whether I agree that it's a useful step in this sort of work. Out of sample validation is normally done when developing models that provide point estimates of the system they are modeling. The theory is that the fitting data is a combination of features that are a deterministic function of the covariates and random features that are idiosyncratic to the sample data. Out of sample validation provides a way to ensure that the model is capturing the former and ignoring the latter.

The goal of this kind of emulator, however, is something different. Instead of trying to provide a point estimate that reflects the influence of certain covariates, we are trying to simulate random draws from the probability distribution implicitly defined by the ESMs, including all components, both random and deterministic. Therefore, it is not clear what it is that we are trying to exclude by doing out of sample validation. In other words, normally overfitting is caused by the presence of noise (i.e., random response) in the fitting data, but if the noise itself is what we are trying to fit (i.e., we are trying to produce a stochastic variable with similar properties to the noise), then what is it that we are potentially overfitting?

In equation (3) the authors split the global mean temperature time series into a deterministic component and a stochastic variable component. Their purpose in doing this is to allow the local temperature to respond differently to the two components, an innovative approach that makes some sense theoretically. However, they do not take the next step of evaluating the local mean temperature model to see whether the additional coefficient is supported by the data. Either the deviance information criterion (DIC) or Watanabe-Akaike information criterion (WAIC) would be a good choice for such an analysis.

The more I read of the literature in the this area of including variability in climate model emulators, the more I am convinced that designing a plausible emulation algorithm is the easy part of this kind of research. What is hard is proving that the statistical properties of the distribution of the emulator outputs are consistent with those of the emulated system. The big frontier in this research area lies in finding ways to characterize similarities and differences between the joint probability distribution of the variables produced by the emulator and that of the system being emulated. Such methods should be fully quantitatvie (i.e., they should produce a measurement of how much the emulator distribution might deviate from the distribution in the real system). Determining what properties of the joint distribution should be reproduced will be an important step in this sort of evaluation. These properties should include, at a minimum, not only marginal distributions, but also space and time correlation properties.

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2019-34, 2019.