Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-3-AC1, 2016

Geoscientific Model Development Discussions

Interactive comment on "Evaluating Statistical Consistency in the Ocean Model Component of the Community Earth System Model (pyCECT v2.0)" *by* A. H. Baker et al.

A. H. Baker et al.

abaker@ucar.edu

Received and published: 17 April 2016

We very much appreciate the thorough review and suggestions for improvement. We address the four main issues below.

(1) Model configuration concerns

We agree that POP-ECT cannot test for "climate consistency" as specified by the reviewer for low-frequency mode ocean events. Indeed, many works emphasize that the ENSO and low-frequency modes do not meet the climate consistency as defined here. A well-known example for our CESM framework is the CESM Large Ensemble Project (LENS) which perturbs SST at O(-12) round-off level (Kay et al., 2015 - see

below) and takes advantage of this inconsistency driven by the small initial differences to establish the large ensemble bases. However, the purpose of POP-ECT is to test for consistency with an established ensemble, and its main application is to identify potential problems associated with software or hardware configurations, not primarily scientific exploration.

We agree that our design minimizes the natural variability introduced by the surface boundary conditions and other potential forcing by using the climatological data-driven forcing. ENSO or low-frequency variability simulations will fail the POP-ECT if coupled simulations are conducted because of the chaotic behaviors in the atmosphere model. We also agree that the current setup is the "minimal" requirement to detect if the new run is consistent with the ensemble or not because the oceanic model is much more dissipative than the atmospheric model, which is why the approach of Baker et al. (2015) cannot be directly applied to the ocean component (see our discussion in section 1). In our current setup, a failing result from the POP-ECT can inform the user that a code or environment (hardware/software) change made by the user may be problematic. We note that the purpose of both Baker et al. (2015) and this manuscript is not to evaluate the climate consistency in the coupled climate production simulation. We are instead trying to identify potential errors induced during the software development lifecycle, such as porting to a new machine architecture, optimizing the code, changing compilers, or modifications to the machine hardware or software stack, etc. We have clarified our intent in the revision by adding text to the abstract, introduction, and conclusion.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S., Danabasoglu, G., Edwards, J., Holland, M. Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and M. Vertenstein (2015), The Community Earth System Model (CESM) Larg Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability, Bulletin of the Amer-

ican Meteorological Society, doi: 10.1175/BAMS-D-13-00255.1, 96, 1333-1349. http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-13-00255.1

(2) Test criteria

We completely agree that the test criteria is a function of the setup, and it is not obvious that the same criteria would be appropriate for tenth degree resolution, for example. This same comment applies to our previous CAM-ECT work. That said, because the tool is specifically intended to verify a new CESM hardware setup environment or a minor code modification intended to result in a consistent climate (e.g. reordering operations in a stable and mathematically equivalent way for optimization purposes), in most cases a single test configuration will be sufficient. One cannot hope to test every possible configuration and resolution, but an error in the software or hardware will likely manifest itself regardless of the configuration. If, however, one made a code change that only affected high resolutions, then POP-ECT with a low resolution would not catch such an error. We would argue, though, that such testing should be done in the context of software unit testing, however - not via this ensemble consistency evaluation. Upon re-reading the manuscript, we realized that we did not make the intent of the tool clear enough and have updated the text accordingly.

(3) Test cases

We agree that additional tests that evaluate a change in the hardware or software stack are of interest. In the revision, we have added a new section that includes results from the following modifications:

-changing the compiler on the same machine (e.g., GNU or PGI instead of Intel) -changing the machine (e.g. NERSC's Edison machine) -changing the compiler version or optimization flag

(4) Section 2.2

We agree that there are errors in this section. We have decided to remove most of this

section, as it is tangential to our focus. We left only the text describing work done in Hu et al. 2015 for background.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-3, 2016.