Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-363-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "Evaluation of global ocean–sea-ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2)" *by* Hiroyuki Tsujino et al.

F. O. Bryan (Referee)

bryan@ucar.edu

Received and published: 28 February 2020

This manuscript is the logical successor to Griffies et al (2016) defining the OMIP protocol and Tsujino et al (2018) describing the construction of the JRA55-do forcing dataset. The collective efforts of the world's leading ocean modeling groups both in preparing for this study (preparing the forcing data, developing and agreeing on the protocol, running the experiments) and in collecting and collating the results is monumental.

The manuscript documents that the massive effort took place, but to be perfectly honest, it is as dull as dirt. Paragraph after paragraph begins "Figure N presents . . .", "Figure N+1 presents . . .", going through the catalog of standard metrics. The authors set a low bar by declaring they they will offer only a "glimpse rather than an in depth view of the many elements of ocean model performance" (line 93). With 150 pages of material and several hundred figures, I believe there is more than can be described as a "glimpse", but agree that an in depth view is not offered. I guess this is the consequence of CMIP-ification of climate science. I am sure many groups will use the figures at some point to calibrate their efforts going forward, and more in depth studies will follow, but I would have hoped that we might have found a little more introspection on the successes and shortcomings of the protocol as well as more on the impacts of the structural changes in the forcing data. For example:

- Does each metric considered add value to the assessment, e.g., Do we need 0-700m heat content and SSH metrics or would one or the other be sufficient to discriminate among the included models?

- Will these metrics be relevant as resolution (and resolved variability) increase? There is already some indication that certain of these metrics become misleading.

- Does a change in ordering among models in various metrics in OMIP-1 vs OMIP-2 suggest the importance or not of different aspects of the forcing? What does the change in spread across the ensemble imply about the forcing?

- Are there any obvious groupings of models (e.g. the NEMO models or the hybrid coordinate models) in model skill metrics or not?

- Did variance in the solutions during in the pre-satellite era change more or less as compared to the later years between OMIP-1 and OMIP-2?

- The Tsujino et al (2018) manuscript calls out several "notable differences between CORE and JRA55-do" (pg 106, first pp) . Are these apparent in the solutions?

- How did the additional variability in runoff included in JRA55-do forcing impact the solutions?

In short, what are the high-level conclusions that we can draw about the value of this exercise? I doubt that this question will be addressed in subsequent studies, so this is the obvious place to address it.

SPECIFIC COMMENTS:

Figures: I find the color bar used for positive definite quantities (e.g. 2b,d,f,h) very difficult to interpret. More contrast would be helpful.

Figure 1 and similar: Some explanation of what accounts for the nearly instantaneous development of the ensemble spread in upper ocean heat content, SST etc would be helpful. Perhaps maps of the year 1 bias in each model and how it compares to the longer term mean bias. What structures are responding this rapidly? What can we learn form experiments integrated for a few years vs 360?

Line 303 and following: A comparison at a subsurface (maybe 50m) depth would be more enlightening to factor out the influence of salinity restoring.

Line 330: Would not a simple broadening of the front (irrespective of the occurrence of recirculation gyres) result in such a dipolar structure?

Figure 9a,b: A nonlinear color scale would be helpful to bring out more than the deep water formation sites.

Line 448 and following: It is notable that the SH mean bias improves more because the worst models get better.

Line 455 and following, Figure 22: I found this to be perhaps the most important figure when considering the limitations of the wash-rinse-repeat OMIP cycling. We really do not capture 60 years of variability with a 60 year cycle. Worth emphasizing more strongly.

Appendix B2: Figure B4 a bit of over kill to make the point (did we really think Drake passage transport might depend on small differences in the properties of moist air?), but oh well, only four more panels among 400!

MINOR TYPOS etc:

Line 100: The four . . . or All four . . .

Line 224: smaller drift

Line 227: “subsurface” not clear what depth range is being described

Line 609: piston velocity

Line 610 : 6 cycles (to be constant with rest of text)

Line 662 (CESM)

Line 730: with OM4 configured

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-363, 2020.