Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2022-60-RC1, 2022

Comment on gmd-2022-60

Anonymous Referee #1

Referee comment on "The Seasonal-to-Multiyear Large Ensemble (SMYLE) prediction system using the Community Earth System Model version 2" by Stephen G. Yeager et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-60-RC1, 2022

General comments

This manuscript describes and evaluates a new ensemble prediction system for lead times up to 2 years. The system is based on the previously documented CESM2 (Danabasoglu et al. 2020) Earth system model, initialized from JRA-55 in the atmosphere, a forced integration following the OMIP2 protocol for ocean and sea ice (FOCI), and a forced simulation of the CLM land surface component. 20-member ensembles are initialized for four initial months per year over the 1970-2019 period, making this dataset a substantial contribution to the climate prediction community.

The paper is well organized, pleasant to read and instructive. I acknowledge the thorough effort led by the authors to evaluate a more diverse range of variables and indices beyond sea surface temperature, circulation indices and precipitation, thereby illustrating the interest of such a system based on an Earth system model and promoting further research with this database. The goals of the paper are clearly stated at the end of the introduction, and in my view, rather adequately fulfilled in the following sections. The number of figures remains quite reasonable with respect to the completeness of the analysis.

Given the quality of this submission, I recommend to accept it for publication in GMD, subject to **minor revisions**. I have two main points I wish to raise, and more specific comments and minor suggestions follow.

1) Although the initialization strategy is described in detail, the authors focus the evaluation on the seasonal-to-multiyear forecast skill. For some variables for which observational data is scarce, or does not cover the entire hindcast period, reconstructions used for initialization are also used as a reference for skill assessments. I understand the reason for this choice but would then have expected more details on the estimated quality of these reconstructions when these haven't been documented elsewhere (which is the

case at least for FOSI). For instance, the authors mention some shortcomings of the CESM2 contribution to OMIP2 that were corrected by tuning parameters and restoring strength for FOSI, but no further details or evaluation of the improvement with respect to independent estimates are provided (it could be included as supplementary information).

With respect to these reconstructions, were any long-term drifts found? I acknowledge several cycles of the forced model have been run but is it enough to avoid spurious effects in the hindcasts?

2) Furthermore, although some figures provide an indication of the ensemble spread, skill is evaluated solely using deterministic scores (anomaly correlation coefficients of the ensemble mean, root mean square error). Having a 20-member ensemble allows for the assessment of other aspects of forecast quality, including reliability and resolution, or other probabilistic metrics of hindcast skill.

Specific comments

1) The authors compare some skill assessments with other reference systems such as the NMME multimodel (for seasonal time scales) and CESM1 DPLE (for November initializations). However these comparisons are only shown in a selection of figures. I'm not necessarily asking for comparisons to be included in each figure, but more discussion on similarities / discrepancies in skill with these two benchmarks could be of interest to the reader.

2) I was confused by differences in lead time values in Figure 5 and in the text (lines 298-312). The shorter lead months don't seem to appear on the plots although they are mentioned in the text (ie: line 300 refers to an ACC of 0.65 I cannot find on the plot). Furthermore using the color and symbol code, lead month 2 for SMYLE-FEB reads as JJA, which isn't consistent at all with definitions provided in the paragraph starting at line 164 – and doesn't make sense. Could you please revise the figure?

3) Correlation / ACC values are often referred to as significant / non-significant, but I found no mention of the significance test and underlying hypotheses (sorry if I missed it!).

Minor suggestions

l. 46: "seasonal protocols call for ensemble simulations lasting 12 months" ☐ not all operational systems go up to 12 months; by WMO standards, seasonal prediction information is provided up to ~ 6 months. I would recommend saying "lasting up to 12 months".

l. 65-69: Some of the potential sources of predictability are associated to a reference, whereas others are not; I would recommend harmonizing this. For snow cover: consider Orsolini et al. (2016) or more recently Ruggieri et al. (2022). For QBO: consider Butler et al. (2016) (QJRMS). For greenhouse gas forcing: Doblas-Reyes et al. (2006)

l. 181: JRA55 (reanalysis) data for precipitation is not an obvious choice; is this due to the hindcast period? Couldn't you use merged precipitation datasets such as GPCP which probably have a higher fidelity to actual observations?

l. 310: Not unrelated to my earlier comment on assessing probabilistic skill and using the 20-member ensemble, did you evaluate the ensemble spread of SMYLE according to target season and forecast time for these ocean indices?

l. 348-364: Was this low (no) NAO skill already found with DPLE-NOV? Another aspect, beyond horizontal and vertical resolution of the atmosphere, is the sensitivity of correlation of NAO to the ensemble size, the length of the re-forecast period (see e.g. Shi et al., 2015) and low-frequency variability of NAO skill during the last century (Weisheimer et al., 2019). They suggest that RMS-based scores are less sensitive estimates of NAO skill.

l. 389: I'm not at all a biogeochemistry expert; there appears to be some variability in skill according to the target season, with summer and fall Zoo C, NPP and carbon export more predictable than winter or spring. Why is this the case? What are the drivers behind what appears to be a return of potential predictability in SMYLE? Some discussion (or references) on this would be helpful!

Figure 9 is a bit blurry: could you increase its resolution?

In figures 10 and 11, correlation and RMSE for CESM2-LE are plotted at lead month 19. I find this choice confusing since CESM2-LE is not initialized; maybe you could use a dotted or dashed line as done for the persistence forecasts?

l. 465-480: Summer (JAS) SIE trends in SMYLE seem different from FOSI, with the ensemble mean generally below FOSI values in the 1970s-1980s, and above after the mid-2000s. Do you have an explanation for what appears to be conditional drift? Did you compare the sea ice thickness fields in SMYLE with FOSI?

Section 3.7: Results are interesting, however some comparison with other recent

evaluations would be nice. Although not focusing on the same period, and using IBTrACS as a reference, Befort et al. (2022) (their Fig. 5) would be a nice comparison for your lead time 1 month results in table 1.

Fig. 15: This figure is quite difficult to read and interpret as it superimposes many time series. I would suggest either presenting a subset of information, or including it in the supplement to the article.

l. 560: missing word ("as")? "as well an experimental system"

l. 574: Out of curiosity: are there any plans to update the system in near real-time? How frequently is JRA55-do updated?

References mentioned:

Befort et al. (2022) doi: 10.1175/JCLI-D-21-0041.1

Butler et al. (2016) QJRMS 142(696):1413–1427

Doblas-Reyes et al. (2016) doi: 10.1029/2005GL025061

Orsolini et al. (2016) Climate Dynamics 47(3–4):1325–1334

Ruggieri et al. (2022) doi: 10.1007/s00382-021-06016-z

Shi et al. (2015) doi: 10.1002/2014GL062829

Weisheimer et al. (2019) doi: 10.1002/qj.3446