Reply on RC2
Stephen Gerald Yeager et al.

Referee comments are repeated below in italics followed by our point-by-point responses. General modifications to note include: 1) elimination of supplemental material (extra figures are now included in appendices) and 2) all main figures are now embedded in the text near where they are first discussed. All line number references below refer to the marked-up version of the revised manuscript.

This submission describes an extensive set of hindcasts from the CESM2 model that enable the performance of initialized predictions in the relatively unexamined multi-year range (out to 24 months in this case) to be extensively explored. Notably, performance over a broad range of Earth system components (atmosphere, sea ice, ocean and land including biogeochemistry) is addressed. The paper is very well organized and written, and criticisms are limited mainly to relatively minor details of description and presentation. Exceptions are items 7 and 17 below, which will require modest additional computation if the authors concur that acting on these recommendations will improve the paper. Overall however the authors are to be congratulated for this interesting and compelling documentation of SMYLE.

Thank you for the overall positive assessment and many helpful suggestions.

Main comments

1) At line 47, suggest replacing “at least 10-years duration” with “up to 10-years duration” because some operational “decadal” systems have a 5-year range, but none that I’m aware of run for >10 years.

Done.

2) Suggest additionally referencing Boer et al. https://doi.org/10.1007/s00382-013-1705-0 and Chikamoto et al. https://doi.org/10.1038/s41598-017-06869-7 in the sentence starting on line 66, possibly as follows: “...volcanic activity (Hermanson et al. 2020), greenhouse gas forcing (Boer et al. 2013), or some combination thereof (Chikamoto et al. 2017).”
Done. Thanks for these suggestions.

3) At line 72 should also reference Ilyina et al. https://doi.org/10.1029/2020GL090695

Done.

4) Line 128 states that forcing is applied cyclically form 1901-1920 to equilibrate the land state. Please go into a bit more detail about the total length of time this cyclic forcing was applied, in relation to expected equilibration times of land variables such as vegetation and soil carbon.

In response to this comment as well as a similar comment from RC1, we have added Figure A4 to the new Appendix A. This figure demonstrates the near equilibration of land carbon stocks over ~4,000 years of spin-up simulation.

Despite some slight imbalance in total ecosystem carbon (primarily associated with soil organic matter carbon), the spin-up was deemed close enough to equilibrium to proceed with the land-only historical simulation. The associated discussion in the text has been expanded (Section 2, starting at line 134).

5) It’s stated that the hindcasts cover 1970-2019. Presumably this is the period covered by the initialization times, and not the simulations themselves which would extend into 2021? If so please be explicit that 1970-2019 spans the initialization times.

We’ve modified the text throughout to clarify that mention of specific historical windows (e.g., 1970-2019) refers to forecast initialization years included in the analysis. We also added a discussion in section 2 clarifying that the verification window is a function of: 1) hindcast initialization time, and 2) observational temporal coverage. Our general approach is to maximize temporal sampling to the extent possible, which means that the actual verification window can vary with lead time. We think this is clear while also avoiding burdening the reader with excessive detail.

6) Regarding “Potentially useful prediction skill (ACC>0.5) is seen for land precipitation over the southwestern US in DJF and MAM (lead month 1)” at lines 208-209, this really should say “southwestern North America” considering that the only DJF grid boxes >0.5 are in Mexico.

In response to a comment from RC1, we have changed the precipitation verification dataset to GPCPy2.3 (spanning 1979-2021). This has resulted in slight changes in skill scores and regions/leads where ACC>0.5. The discussion of Figure 2 has been revised accordingly, and in particular, we now highlight “southwestern North America in DJF (lead month 1)”.

7) Regarding “A more rigorous analysis is needed to definitively demonstrate that SMYLE skill differences from DPLE are statistically significant and not likely explainable by chance” (lines 226-227), this could be done relatively easily by applying the random walk methodology of DelSole and Tippett, https://doi.org/10.1175/MWR-D-15-0218.1, where differences either in the anomaly pattern correlation or the RMSE between the 50 pairs of November-initialized hindcasts could be used as the basis for comparison.

Thanks for this suggestion. In response to RC1 and this comment, we chose to include additional figures in Appendix B (see Figs. B3, B4, B6) that test the significance of local ACC differences between SMYLE-NOV and DPLE-NOV by accounting for uncertainty due to finite ensemble size. The 20-member SMYLE-NOV skill is compared to a distribution of 20-member DPLE-NOV skill scores at each grid point, and overall performance is assessed by comparing the
percentage of global surface area (within 80°S-80°N) associated with skill improvement/degradation. Other methods (such as that proposed by DelSole and Tippett) could be explored in future work, but we think simple skill difference plots will be of particular interest to readers and will suffice for this manuscript.

8) At line 178, please provide a rationale for regridding to a 5x5 or 3x3 degree grid. (Also a small point, but I’m not sure that regridding to a coarser grid qualifies as “interpolation”.)

We have replaced “interpolated” with “mapped” and have added the following rationale (line 198): “This remapping is done to highlight aggregate regional skill, increase the efficiency of skill score computation, and improve the quality of global map plots that include pointwise significance markers (Appendix B).”

9) Below line 360 it would be appropriate to reference Butler et al. https://doi.org/10.1002/qj.2743 in relation to the influence of lid height on skill in forecasting the NAO. (For example could append as “...relative to these baseline SMYLE results, although a robust connection between atmospheric vertical resolution and NAO skill has not been demonstrated (Butler et al. 2015).”)

Done.

10) Please replot Figs. 8d-i using the tick mark values in Fig. 9 which are better aligned with the experiment.

Done.

11) Are there any evident explanations or hypotheses for the strong seasonal dependence of skill in Figs. 8d-i, e.g. high SE US shelf NPP ACC in JJA and SON, and low CA current NPP ACC in DJF?

We think that the seasonal dependence of potential predictability skill in these regions concerns the persistence of subsurface nutrient anomalies (first described in Park et al., 2019). During wintertime mixing, these nutrient anomalies reemerge in the upper ocean and contribute to skill in predicting total summer/fall productivity in these LMEs. We agree that this is interesting and therefore, in response to this comment, we have added a few additional sentences at the end of this paragraph to elaborate on the seasonal dependence of marine ecosystem predictability (starting at line 588).

12) Should mention in the captions for Figs. 8-9 that shading and filled symbols indicate statistical significance.

Done.

13) The OceanSODA-ETHZ aragonite saturation dataset covers 1985 to 2018 according to Gregor and Gruber (2021), so presumably the skill results in Fig. 9 are specific to this period? Or does the verification period leave out the years before 1990 which are much more uncertain according to those authors? Please be explicit about this and any other deviations of verification periods from the 1970-2019 period covered by SMYLE.

The verification against OceanSODA-ETHZ is performed over the window 1985-2018. We have clarified this in the text (line 608). We’ve also made efforts to be clear throughout about verification windows, without getting into excessive detail (see response to comment above).
14) In the captions to Figs. 10 and 11 suggest removing “(see text for details)” since the text doesn’t provide any significant additional detail.

Done.

15) In Fig. 13 the lead times in the legends of the plots disagree with the lead times indicated in the caption.

16) Relating to Figs. 12 and 13, it would be interesting to have a sense of how the correlation and nRMSE values shown compare to values based on comparing OBS to FOSI.

We have corrected Figure 13 to accurately reflect the lead times mentioned in the caption (2, 8, 14, 20), and we’ve removed the sea ice skill scores from plots (Figs. 12, 13) and put them in new tables (Tables 1,2). We think the new tables help improve readability, and they facilitate inclusion of scores between OBS and FOSI.

17) Figure 14 shows JJASON (NH) and DJFMMA (SH) cyclone track densities regressed against annual mean Nino3.4 index. However, because ENSO typically peaks around December and frequently changes phase between about April and August, annual mean Nino3.4 is not a very good indicator of ENSO activity. In addition, this procedure introduces a seasonal disconnect in that January Nino34 is presumed to influence TC activity in the following November (for example). Suggest instead regressing JJASON track densities against JJASON Nino3.4, and DJFMAM track densities against DJFMAM Nino3.4, or else better justifying the original choice made. (Also please be explicit in the caption to Fig. 14 what is the timing of the Nino3.4 index.)

We have updated Figure 14 to show TC track density regressed against the respective seasonal mean Nino 3.4 index (i.e, JJASON track density regressed against JJASON Nino 3.4 index for the Northern Hemisphere, DJFMAM for the Southern Hemisphere). Compared to the annual mean Nino 3.4 index, the seasonal Nino 3.4 index yields a slight decrease of the positive regression coefficients at 10-month lead and longer. The main conclusions remain unchanged.

18) Tables 1 and S1 along with Fig. 15 imply that JJASON and DJFMAM TC predictions are made at 19-month lead time. However, for JJASON this implies initialization on 1 Nov, which in turn implies prediction of JJASO (not JJASON) at 19-month lead by the 24-month hindcast, and similarly for DJFMAM. Although this is a small point it should briefly be acknowledged somewhere (similarly to the 22-mon lead Nino3.4 forecasts in the caption to Fig. 4).

We have added this clarification in the captions of Figs. 14, 15 as well as the TC ACE correlation table (now Table 3).

19) Regarding the RMSE scores, “RMSE” isn’t defined anywhere, and when introducing RMSE in the text should briefly comment on the use of normalized RMSE and introduce the nRMSE notation. Also, is nRMSE defined such that predictions of climatology (zero anomaly) will yield values of 1? If so then briefly mentioning this will help the reader appreciate that nRMSE values <1 indicate that the prediction is more skillful than a climatological prediction.

A new paragraph in section 2 (starting line 202) introduces the skill metrics examined in the paper. We note that nRMSE is defined such that a climatology forecast yields a value of 1.
Technical corrections:

line 203: central America -> Central America

Fixed.

line 207: SAT hasn’t been defined

Replaced with “surface temperature”.

lines 270 and 275: “1998” -> “1997” (as year of a strong El Nino)

Fixed. Note that we’ve modified Fig. 4 so that the x-axis denotes the year of January in the DJF average. The text/caption are modified accordingly.

line 361: Quasi-biennial -> Quasi-Biennial

Fixed.

line 510: Figs. S1, S2 -> Figs. S4, S5

Thanks. These figures are now included in Appendix B (Figs. B8, B9), with text modified accordingly.

line 520 vs 185 vs 591: is it “best track”, “Best Track” or “BestTrack”?

We now consistently used “Best Track”.

line 528: should the 2 in kt2 be a superscript?

Yes, fixed.

line 556: suggest “multi-year skill” -> “multi-year skill or potential skill”

We have modified the sentence accordingly.

line 561: suggest to remove “obviously”

Done.

line 923: ACC map gross -> ACC map for gross

Fixed.