Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2020-66-AC3, 2020

Earth System Dynamics
Discussions

Interactive comment on "Assessment of a full-field initialised decadal climate prediction system with the CMIP6 version of EC-Earth" *by* Roberto Bilbao et al.

Roberto Bilbao et al.

roberto.bilbao@bsc.es

Received and published: 3 December 2020

Review of "Assessment of a full-field initialised decadal climate prediction system with the CMIP6 version of EC-Earth", by Bilbao et al.

Reply to Dr. Steve Yeager:

Specific Comments:

My specific recommendations for improvement:

1) Much of the paper elaborates on the negative effects of an "initialisation shock" in the subpolar Atlantic, and this term is even included in the abstract. While the authors

offer a definition of what this phrase means ("abrupt changes that occur soon after initialisation as a result of the adjustment of the climate model to the initial state"), I felt that the precise meaning of this term (and its usefulness for understanding system behavior) faded as I read. Certainly, there is a pathological adjustment to initialization going on in this system, but the distinction between shock and drift is not clear, nor is it clear that the initial shock (enhanced Labrador Sea convection) causes the longer term drift (towards reduced convection and sea ice expansion, AMOC decline, etc.). Is the shock really the essential problem in EC-Earth, or is it the drift towards ice-covered Labrador Sea? I suspect the latter is the more fundamental problem. I recommend a reconsideration of the phraseology used throughout.

Reply: This is a really good point. We agree that both the initialization shock and the mean drift are closely related in our predictions, and that is not possible to disentangle from our analysis if the problem is caused by the processes behind the initial adjustment or by those related to the long-term drift, which might be related. We have tried to improve the clarity of their definitions in the introduction and explain how they might relate with one another. We also specify now that initial adjustments or shocks, when they occur systematically across start dates, can be regarded as the initial stage of the model drift, and even condition its later evolution. We have also carefully revised the rest of the paper to mention both the shock and the drift as the ultimate causes of the lack of skill in the central SPNA.

2) Related to above, the skill improvement with lead time for NASPG-OHC300 (Figs. 5k and S1k) is interpreted as reflecting initialization shock behavior. However, the later figures (in particular, Fig. 7) make me question whether the relatively high skill for later lead times (e.g., LY7-10) is real skill. I note that HIST_NoConv exhibits a reasonably high correlation with RECON for Western SPNA-OHC300 (Fig. 7c) which is almost certainly spuriousâĂŤit appears to relate to a post-1990s spinup of the NASPG in those members (Fig. 6b) which in turn appears related to a transition from fully ice-covered Lab Sea to only partially-covered Lab Sea, with associated increase in

convection (Fig.7). This mechanism for reproducing the late 20th century warming of the SPG is unequivocally unrealistic, even though it might yield higher correlation scores for NASPG-OHC300 than HIST itself (could you check this?). At long lead times, PRED seems to show similar behavior as HIST_NoConv (as noted in the text, but also in terms of Lab Sea transition from ice-covered_no-convection to partially-ice-covered_some convection), suggesting that the better NASPG-OHC300 “skill” at long lead times is a spurious artifact of an unrealistic warming mechanism. If true, this changes the interpretation of what is happening in the prediction system (i.e., it is not “initialization shock” followed by skill recovery via better representation of real mechanisms). If not true, how do the authors explain the increase in NASPG-OHC300 skill with lead time (Fig. 5k)?

Reply: This is a really interesting hypothesis, which we had not thought of. To check if the hypothesis is true we have looked in more detail at the East and West SPNA OHC in the upper 300m (see Supporting Figure 1). As suggested, we have divided the historical ensemble into those members which exhibit convection (Hist_Conv, 7 out of 10 simulations) and those with suppressed convection (Hist_NoConv, 3 out of 10). The reviewer correctly pointed out that the historical members with no convection have higher ACC values than those with convection (Figure 1c), at least in the West-SPNA (note that to avoid differences in skill due to differences in the verification period we have set a common verification period: 1971-2008). This is because the members with no convection show a long-term warming trend which happens to coincide, in large part, with the observed one, even if it happens for the wrong reasons (that is the melting of Labrador Sea ice allowing for open ocean convection).

Even though PRED at the later forecast times reaches comparable ACC values to Hist_NoConv, these do not seem to be explained by the same mechanism explaining the spurious OHC300 warming in Hist_NoConv (Figure 1c). Indeed, the timeseries in Supporting Figure 1b shows that the large improvement in skill at the longest forecast range with respect to the first forecast years comes from a good representation of the

decadal variability, including the quick transition from cold to warm OHC anomalies that occurred during the mid-90s, with a radically different long-term evolution than in Hist_NoConv. The lack of skill in PRED at the initial forecast times comes from a poor representation of the observed inter-annual variability, which in the forecast shows some ‘spikes’ (or abrupt transitions) that might well result from the initialization shocks.

The respective plots for the eastern SPNA OHC300 can help explain the origin of the skill recovery at forecast years 7-10 over the whole SPNA. Indeed, they show that the eastern side of the region has rather constant predictive skill, comparable in PRED and both Historical ensembles, which implies that most of the skill might be forced. In terms of the mean gyre circulation, the eastern SPNA is upstream of the western side, and therefore the mean flow might be advecting the (skilfully) forecasted anomalies from the eastern into the western region, eventually substituting the unrealistic OHC anomalies generated in the first forecast years by the labrador convection collapse. If we take into account that the eastern SPNA maintains a similar level of skill all along the forecast, and that the western SPNA recovers it at the very end, we could then explain the increase in skill with lead time for the whole SPNA in Figure 5k. This is, of course, just one plausible hypothesis, but exploring it further would require extending the paper in a new direction, something that we would prefer not to do given that it is already lengthy and dense.

A reduced version of Supporting Figure 1 has been included in the Supplement, and is now discussed in the text to explain our hypothesis behind the whole SPNA skill recovery.

Supporting Figure 1: Timeseries and ACC skill of the West and East SPNA-OHC300 anomalies (with respect to 1971-2018). The timeseries and ACC have been computed for the common period for all forecast ranges (i.e. 1970-2018). The first two columns show the observed (grey bars) and predicted (PRED in red, HIST in blue, HIST_Conv in green and HIST_NoConv in purple) timeseries for the 1-4 and 7-10 forecast years respectively. The third column shows the ACC for PRED (red), HIST (blue), HIST_Conv (green) and HIST_NoConv (purple). Statistically significant ACC values (at the 95% confidence level) are shown as empty circles.

3) Figures 8 and 10 have many small thin lines of various colors and hues that are very hard to distinguish (this reviewer is slightly color blind). Can a revised version be developed that is easier to see, particularly Fig. 8? I recognize that "easy to see" is quite subjective, and that these figures contain lots of information that is hard to display any other way. Perhaps the answer is "the figures are as clear as they can reasonably be" and I am in a small minority that has trouble viewing them, but if others (reviewers, coauthors, colleagues) also have difficulty with these figures then please make an effort to improve them.

Reply: The resolution and quality of the figures have been improved. In figures 8 and 10 the profiles now go down to 500m rather than 800m to allow for a better visualization of the near-surface differences. And in Figure 8 we now include two rows, one comparing PRED with RECON, and another comparing PRED with HIST, which reduces the amount of lines, and allows to see better the differences between RECON and PRED.

Additional Comments (by line number)

63: This is not a complete sentence.

Reply: corrected.

80: The meaning of "biases in the predictions" is not clear. Model mean bias is to be expected when using anomaly initialization. Do you mean "time-dependent biases in the predictions" (i.e. drift)?

Reply: We have rephrased it now as "skill degradation in the predictions".

111: ORCA has not been defined

Reply: Added information.

124: There is no mention of how the land model component is initializedâĂŤcan you

please clarify?

Reply: Added information of the land model component (HTESSEL) and the initialisation.

205-207: Since sentence paragraphs are not advisable.

Reply: the sentence has been deleted as it was not necessary.

243: "signal" instead of "trend" to avoid awkward phrasing?

Reply: suggestion accepted.

264: "associated with" instead of "to"

Reply: corrected.

271: It would help to interpret Fig. 3 if the breakdown of MSSS into correlation and conditional bias terms were given explicitly (perhaps in section 2.3), and the corresponding relationships between Fig. 3 panels clarified (e.g., is panel a = panel d + panel g?).

Reply: The description of the MSSS in section 2.3 has been extended. We now include two equations, the one we used for the computation, that it's taken from Goddard et al. (2013; Eq. 5), and a more compact version that represents the numerator as the difference between two terms, one based on ACCs and another on conditional biases. Both equations are reproduced below. The re-arrangement in Equation 2 has allowed us to see that the middle and bottom plots in the former version of Figure 3 did not represent direct contributions to the MSSS. It is actually the differences in the squared ACCs/conditional biases that determine the final MSSS, and therefore those are the quantities that we now represent in the middle and bottom rows of Figure 3 to guide the interpretation of the MSSS. If we disregard the denominator (which is just a scaling factor with no impact on the sign to produce a skill metric that goes from -1 to 1), we can interpret the values in the upper row as the difference between the values on the second and the third row. The discussion on Figure 3 has been modified according to the changes.

Also, because positive values in the difference between the squared ACC in PRED and the squared ACC in HIST do not necessarily correspond to a beneficial effect of initialization on skill (e.g. if the ACC in PRED is negative, and positive in HIST) we have decided to keep the plots on the differences in ACC from the former figure, which have been placed as a third row in Figure 2.

286: Missing "(Figure"

Reply: corrected.

302: There also seems to be noteworthy skill in the western tropical Pacific which should not be ignored.

Reply: added to the text.

315: I'm confused by this statement. Since both PRED and HIST show SER<1 in the first few months (Fig. 5c), aren't they both overconfident (under-dispersed)?

Reply: This has been corrected.

Fig. 5: It's unclear from the caption whether purple line (persistence forecast) is an ACC or MSSS score.

Reply: It has been indicated in the caption that the purple line refers to the persistence based on the ACC.

326, 340: It's not clear to me that the HIST spread is "excessive" and "too large" (although it is certainly larger than PRED) since I'm unsure how the concept of reliability applies to uninitialized ensembles that aren't expected to be able to predict internal variability.

Reply: The spread-error-ratio is a measure of reliability that has been typically applied both to initialised and non-initialised forecasts (Ho et al. 2013; Robson et al. 2018).

It evaluates if the typical distance between ensemble members is comparable to the typical distance between the individual members and the observations. Both terms are small at the beginning of an initialised forecast (because of initialization itself) and are expected to grow as the forecast progresses, although their ratio could vary, and converge to the one in the uninitialised experiments. That's why we show them both.

References:

Ho CK, Hawkins E, Shaffrey L, Bröcker J, Hermanson L, Murphy JM, Smith DM, Eade R (2013) Examining reliability of seasonal to decadal sea surface temperature forecasts: the role of ensemble dispersion. Geophys Res Lett 40(21):5770–5775

Robson, J., Polo, I., Hodson, D. L. R., Stevens, D. P., and Shaffrey, L. C.: Decadal prediction of the North Atlantic subpolar gyre in the HiGEM high-resolution climate model, Climate Dynamics, 50, 921–937, https://doi.org/10.1007/s00382-017-3649-2, 2018.

360: "black" should be "green"?

Reply: corrected.

396: Fig. 7f is mislabelled as "e)"

Reply: corrected.

Fig. 8: I find it very hard to make out the relevant details in this figure even after magnifying to 400%. Can you devise a better graphic that is more legible for color challenged individuals? Same comment applies to Fig. 10. One simple option might be to just plot upper 400m to magnify the key region of interest. Another might be to plot as differences from HIST.

Reply: To improve the visibility of the relevant details we have increased the size of the figure, zoomed it to the upper 500m, and duplicated the panels to compare separately PRED and RECON (for which the lines are closer to each other), and PRED and HIST.

431: Please double check the sign of the restoring freshwater fluxes. Fig. 8 suggests that RECON is saltier at the surface than HIST (less stratified by salinity) which implies that a positive SALT flux (ie, negative freshwater flux) is used in the restoring.

Reply: This has been corrected. The freshwater fluxes are defined in the model as going out of the ocean, while the heat fluxes into the ocean. The figure has been modified so both fluxes are into the ocean and the text adjusted.

451: Incorrect reference to figure 10 within this sentence.

Reply: corrected.

Fig. 11: I think the last sentence of caption should be “dark green cross”?

Reply: corrected.

501: Here and elsewhere, the distinction between “initialization shock” and model “drift” could be clarified. (also, what is the “expected trajectory”? a skillful one? one towards the model mean climatology?)

Reply: The sentence and all the other mentions to the drift and initialisation shock have been rewritten according to this comment and the first one.

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2020-66, 2020.

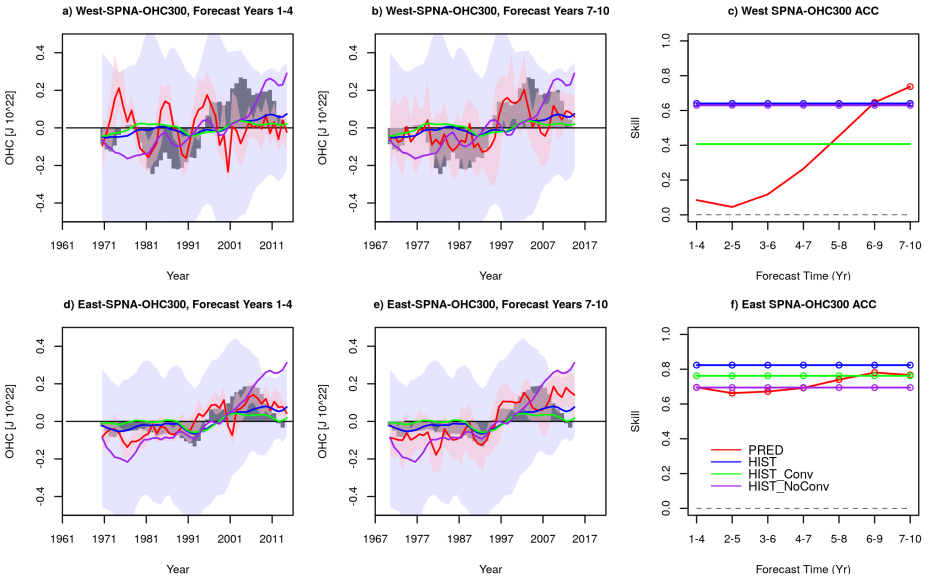

Fig. 1. Timeseries and ACC skill of the West and East SPNA-OHC300 anomalies (with respect to 1971-2018).