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

General Comments

This manuscript presents an overview assessment of the performance of the EC-Earth prediction system, contributed by BSC to CMIP6 DCPP. The skill assessment is focused on surface temperature (a focus that the authors might consider highlighting in the title), and includes assessment of modes of surface temperature variability and related ocean fields (heat content, AMOC, etc.). The second half of the study investigates the reasons for curiously low skill in the central subpolar North Atlantic and presents some interesting analyses that shed light on initialization shock and drift in this key region. The quality of the writing is very good and clear, with ample references to recent literature, and the quality of the figures is high (with some exception—see comment below). This study will be of interest to many in the decadal prediction community as it nicely documents the overall behavior of a single high-profile system (one of the WMO's Global Producing Centres for annual-to-decadal climate predictions). I therefore recommend publication after (mostly) minor revision.

Specific Comments

My specific recommendations for improvement:

1) Much of the paper elaborates on of 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.

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 spurious—it 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)?

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.

Additional Comments (by line number)

63: This is not a complete sentence.
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)?
111: ORCA has not been defined
124: There is no mention of how the land model component is initialized—can you please clarify?
205-207: Since sentence paragraphs are not advisable.
243: "signal" instead of "trend" to avoid awkward phrasing?
264: "associated with" instead of "to"
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?).
286: Missing "(Figure"
302: There also seems to be noteworthy skill in the western tropical Pacific which should not be ignored.
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)?
Fig. 5: It's unclear from the caption whether purple line (persistence forecast) is an ACC or MSSS score.
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.
360: "black" should be "green"?
396: Fig. 7f is mislabelled as "e)"
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.
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.
451: Incorrect reference to figure 10 within this sentence.
Fig. 11: I think the last sentence of caption should be "dark green cross"?
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?)