REVIEWER #1

In the paper “The CMCC Decadal Prediction System”, Nicolì et al. document the CMCC DPS and evaluate its predictive skill with initialized hindcast simulations that have been contributed to CMIP6. The system distinguishes itself by its use of the CMIP6-generation Earth system model CMCC-CM2-SR5 as well as its full-field initialization strategy that covers all components and generates realistic spread in the subsurface ocean. Comparing the CMCC DPS results to uninitialized historical CMCC-CM2-SR5 simulations, the authors demonstrate global skill benefits from initialization that are particularly pronounced in the North Atlantic despite a substantial forecast drift in that region. The authors highlight skillful multiyear prediction of North Atlantic atmospheric circulation variability and attribute modest skill over land to the limited ensemble size. Given the uniqueness and performance of the system, the authors make a strong case that the CMCC DPS is a valuable addition to the suite of existing decadal prediction systems. The data analysis and figures of the paper are clear and innovative, including probabilistic skill measures and skill dependence on lead time and average period. The authors succeeded in keeping the paper short by selecting a few key metrics. Overall, the paper is well and concisely written, the argumentation is sound, and the paper fits well the scope of the journal. I therefore recommend its publication in GMD after having addressed a few minor comments listed below.

A: Thanks for your helpful comments and suggestions. We have revised the paper to implement your suggestions. Revisions are highlighted in the revised manuscript. Our point-by-point replies to the comments follow.

1. L89 “Finally,”

Repeated in the next sentence.

A: We corrected the text.

2. L94 “15-member ensembles”

A bit unclear how the spread for all 15 initial conditions is obtained from two land surface conditions and five ocean conditions. Are the 10 states of Init utilized in addition? Or, did you apply any small initial perturbations? Please elaborate.

A: A summarizing table has been added to better explain the generations of the initial conditions. In this version of the manuscript, we consider 20 members generated by the combinations of: 2 land states, 2 atmosphere states and 5 oceans states.

3. L97 “ERA-Interim (1979–2020)”

To my knowledge, ERA-Interim is only available until August 2019. Considering that you describe CMCC DPS as an operational system, could you please detail which product you use for the initialization in 2020 and later?

A: Thanks to the reviewer for highlighting this caveat. We have used the ERA-interim dataset until November 2018 (since ERA-Interim is available until August 2019). From 2019 onwards, the DPS is operatively initialized using ERA5 dataset.

4. L101 “atmospheric forcing datasets: CRUNCEP version 7 (Viovy, 2016) and GSWP3 (Kim, 2017)”

According to a quick search, CRUNCEP version 7 is available until 2016 and GSWP3 until 2010. Do these products indeed cover the entire 1960¬–2020 period and are they updated in the future? If not the case, how does CMCC DPS initialize the land component beyond their coverage?

A: The reviewer is right. These datasets do not cover the entire period. The CRUNCEPv7 is available until 2016 while GSWP3 covers until December 2014 (here the link to the input forcings repository: https://svn-ccsm-inputdata.cgd.ucar.edu/trunk/inputdata/atm/datm7/atm_forcing.datm7.GSWP3.0.5d.v1.c170516/). Starting from 2015, we used a land-only run forced by ECMWF forcing instead of GSWP3. The CRUNCEPv7 dataset has also been substituted by NCEP forcing. We have now reported this in the new table of the revised manuscript and in the text.

5. L101–102 "provide four instantaneous 2-meter air temperature and humidity, 10-meter winds and surface pressure every six hours"

Do you mean "four" realizations of instantaneous fields (i.e., a mini-ensemble)? Or, do you mean four times per day (i.e., every six hours)? Please clarify.

A: We rewrote the sentence as follows to improve the readability "*These datasets provide to the land model instantaneous 2-meter air temperature and humidity, 10-meter winds and surface pressure every six hours, and 3-hourly-accumulated radiation and precipitation*."

6. L104–105 "An ensemble of 5 ocean initial states is used to initialize the ocean and sea ice components"

Could you add more detail on how the ocean initial state is used to initialize the sea ice component? Does CMCC DPS initialize sea ice thickness in addition to sea ice concentration?

A: The global ocean reanalyses also include estimates of the sea-ice state. To avoid strong inconsistencies we use the same realizations of sea-ice and ocean components. The sea-ice is initialized using direct interpolation on the target grid of sea-ice temperature, sea-ice volume, sea-ice area and snow volume. We state this in the new table (T.1) and in the text of the revised manuscript as follows: "*An ensemble of 5 ocean initial states is used to initialize the ocean and sea-ice components: 3 initial estimates originate from global ocean reanalysis characterized by different assimilation strategies of SST and in-situ profiles of temperature and salinity in a 0.5° configuration of the NEMO ocean model, while the remaining 2 initial states are derived through linear combinations of the former 3 initial states. The ocean initial states provide three-dimensional fields of temperature, salinity and horizontal currents. The sea-ice model has been initialized starting from sea-ice temperature, sea-ice volume, sea-ice area and snow volume*".

7. L116 "The predictive skill for both initialized reforecasts and uninitialized projections is assessed against observational products."

Please specify the temporal coverage of the assessment. Is it always 1961–2021 (for LY1 and LY1–5) and 1966–2021 (for LY6-10)? If the coverage varies for the different validation products, then it should be stated for each product separately (either in the main text or in the figure captions).

A: The temporal coverage of lead year 1 is 1961–2020, since not every observational product used in this study covers from 2021 onwards. Lead year 1–5 and 6–10 considers respectively the periods 1961–2015 and 1966–2020. We are aware that this choice excludes some final start dates. We estimate the ACC considering the same start dates (and consistently the same number of years) independently from the lead year ranges. Second, according to the CMIP6 DCPP-A protocol, starting from 2015 onwards, the DPS uses the same external forcings of the ssp2-rcp4.5 scenario which does not reflect the current climate warming trend. This is clearer in Fig. 1.a of the manuscript: the blue line, which represents the global mean surface temperature according to the historical+scenario, starting from 2015 has a negative trend in opposition to the observed one. We also add that, in terms of ACC/MSSS, no significant changes occur when we consider those years (not shown).

Following your comment we stated this in the method section as follows: "*The temporal coverage of lead year 1 is 1961–2020, since not every observational product used in this study covers from 2021 onwards. Lead year 1–5 and 6–10 considers respectively the periods 1961–2015 and 1966–2020.*"

8. L130 "observed variability over the historical 1960–2014 period targeted by the decadal reforecasts".

Correct? Or should it be something like 1961–2021?

If 2014 is indeed the last verification year, then all hindcasts with start dates in 2014 and later are not included in the analysis.

A: This is right. The period is 1960–2020. We corrected the manuscript.

9. L137 "MSE_HO (MSE_PO) is the mean square error evaluated for the initialized (uninitialized) ensemble mean against observations"

Are the respective climatologies (i.e., the bias) subtracted prior to the MSE computation as in Goddard et al. 2013 who defined anomalies "relative to their respective climatologies (which is equivalent to the removal of mean bias)"?

To my understanding, the MSSS used by Goddard et al. is designed to reflect phase and amplitude errors but not climatological bias.

A: This is correct. The MSSS quantifies the magnitudes between the predicted and observed anomalies. We state this when we introduce this metrics (section 2.2).

10. L159–160 "The statistical significance is assessed with a one-tailed Student's T test (Wilks, 2011), accounting for auto-correlation in the time series (Bretherton et al., 1999)."

Please specify explicitly if local or field significance is tested (I assume "local" but got a bit confused by the citation to Bretherton et al. on the effective number of "spatial" degrees of freedom).

Also, could you please re-state the formula used to account for auto-correlation or cite the equation in Bretherton et al. 1999? Is it the same as used in Vecchi et al. 2017 (https://doi.org/10.1175/JCLI-D-17-0335.1)?

A: We refer to Eq. 30 of Bretherton et al. 1999, which accounts for local significance. We modified the text in section 2.2 as follows: "*The statistical significance is assessed with a*

one-tailed Student's T test (Wilks, 2011), accounting for auto-correlation in the time series (Eq. 30 from Bretherton et al., 1999)."

11. L161 "reference period 1981–2010, in agreement with WMO recommendation"

According to latest WMO recommendation the period for the computation of climate normals is "the most-recent 30-year period finishing in a year ending with 0" i.e. currently 1991–2020 (https://library.wmo.int/doc_num.php?explnum_id=4166). I'd therefore suggest removing "in agreement with WMO recommendation".

A: We completely agree with the reviewer. We have removed that statement from the revised manuscript.

12. L178 "The ensemble spread envelope"

Please state how you defined this envelope (either in the main text or the figure captions). For example, is it defined as min/max or a certain percentile range of the data?

A: We modified the text to define the envelope as follows: "*The ensemble spread envelope of predicted GMST (denoting maximum and minimum range of the members variability, shown in orange) encompasses the observations (...)*"

13. L180–182 "The initialization contributes to the reduction of the Init ensemble spread, which is about half the envelope of the NoInit for lead-year 1. This is not so unexpected since, when a simulation is initialized the observed internal variability is imposed thus reducing the uncertainty related to systematic errors (Doblas-Reyes et al., 2013)"

What do you mean by "systematic errors" here? To me, the sentence seems a bit unclear/confusing and not quite in line with the reasoning of Doblas-Reyes at al. who wrote "The initialization can, in addition to providing information about the phase of internal variability, correct systematic errors in the model response to external forcings".

First, a correction of the model response to external forcings (i.e., the forced climate trend) does not necessarily lead to a reduction in ensemble spread. The forced trend should be the same or very similar for all members given the model parameters are not perturbed.

Second, imposing the observed "internal" variability should not significantly change the forced climate trend. To my understanding, Doblas-Reyes et al. argue that initialization benefit is partly due to imposing the observed forced trend in addition to synchronizing the internal variability.

I would have expected that the reduction in ensemble spread is simply a consequence of the synchronization of internal climate variability. But maybe I have misunderstood, and this is what the authors intended to say. Please clarify.

A: We agree with the reviewer that the reduced ensemble spread depends on internal variability and the citation of Doblas-Reyes et al., 2013 is not fully pertinent. Following your suggestion we modified the text as follows: *"The initialization contributes to the reduction of the Init ensemble spread, which is about half the envelope of the NoInit for lead-year 1, due to the beneficial impact of synchronizing observed and model internal climate variability.*"

14. L192 "The added value of initialization"

Regarding the attribution of ACC differences and MSSS to added value of initialization: Did you verify that these metrics are not overly sensitive to the use of different ensemble sizes of NoInit (10) and Init (15)?

A: Since only 10 historical members are available, the 20-member hindcast ensemble has been scaled down using a random sampling with 100 iterations in skill assessment plots where we directly compare the hindcasts with the historical ensemble. It is worth noting that no strong differences occur with the previous assessment as well as with a varying number of iterations (not shown). We specified this in the text as follows:

"*Since only 10 historical members are available, the 20-member hindcast ensemble has been scaled down using a random sub-sampling with 100 combinations in order to allow a fair comparison to the historical ensemble in the skill assessments.*"

15. L199–200 "the skill undergoes a clear deterioration over the tropical and northern part of the Pacific Ocean (Fig. 2c)"

A clear deterioration relative to what? Relative to the first-year skill or relative to the skill of NoInit? Please clarify.

I would say that Fig. 2d does not show a clear deterioration relative to the NoInit skill. That the multiyear skill is deteriorated relative to the first-year skill in the tropical and northern part of the Pacific Ocean is maybe not too surprising given the limited predictability of ENSO and its strong influence on the PNA variability. Or would you expect more multiyear skill in that region?

A: We agree with the reviewer. The deterioration of the multiyear skill over the Pacific Ocean is in comparison to the results of the lead year 1. We also agree that this is so not unexpected since ENSO has limited decadal predictability (as also evident in Fig. 9c). We modified the text as follows: "*In contrast, the skill undergoes a clear deterioration over the tropical and northern part of the Pacific Ocean when multi-year range of prediction skill is considered (Fig. 2c)*."

16. L268 "large ACC values are also found over regions linked to the AMV through remote teleconnections"

For which variables?

A: We refer to the surface temperature. We specified the variable in the revised manuscript as follows: "*It is worth noting that large ACC values of surface temperature are also found over regions linked to the AMV through remote teleconnections*"

17. L278 "Init well captures the NAO variability despite the limited ensemble size (ACC=0.80 applying an 8-year running mean to the model index)"

This is interesting and could indicate that the ratio of predictable components (RPC) is lower than found in other prediction systems (e.g., Scaife and Smith 2018). Could you say anything about how well the amplitude is predicted? A small amplitude (i.e., RPC >> 1) could also be a likely explanation for modest skill over land despite obtaining high ACCs for NAO.

Scaife, A.A., Smith, D. A signal-to-noise paradox in climate science. npj Clim Atmos Sci 1, 28 (2018). https://doi.org/10.1038/s41612-018-0038-4

A: Following your comment, we estimate the RPC (eq. 5 from Scaife and Smith, 2018 hereafter SS18) for the not-smoothed NAO index (ACCmo=0.58, ACCmm=0.27 and RPC=2.15). The RPC is greater than 1 as it is typically the case for decadal predictions of atmospheric circulation anomalies over the North Atlantic (e.g. Athanasiadis et al., 2020). However, it is worth noting that the RPC is lower than that found in other systems, such as in the CESM Decadal Prediction Large Ensemble (CESM-DPLE). In particular, considering an ensemble size of N=20 in Fig. 2d of Athanasiadis et al. (2020) (https://www.nature.com/articles/s41612-020-0120-6/figures/2), we see that for the CESM-DPLE, ACCmm approximately equals 0.15, while the respective ACCmo approximately equals 0.54 (RPC=3.6). Noting that the RPC does not change significantly with the ensemble size (e.g., for CESM-DPLE, RPC=3.5 for N=40), it is reasonable to compare also with the results of Smith et al., 2020, who analyzed a large multi-system ensemble and found an RPC=4.2 for the NAO. Therefore, there is, indeed, a clear indication that despite certain obstinate systematic biases, the signal-to-noise ratio problem (SS18) is less severe in the CMCC DPS. Considering this fact, we have added the following sentence in the manuscript (section 4 and section 5):

"Considering its relatively small ensemble size, the CMCC DPS exhibits an exceptionally high skill for the NAO (ACC=0.58 with 20 members). This is accompanied by a rather low Ratio of Predictable Component (RPC=2.15, estimated following Scaife and Smith, 2018), comparing to NAO predictions by other state-of-the-art decadal prediction systems [Smith et al., 2020; Athanasiadis et al., 2020]. This result suggests that in the CMCC DPS, despite certain obstinate systematic biases, the signal-to-noise ratio problem (Scaife and Smith, 2018) is somehow less severe."

18. L282 "TPI with significant skill peaking at lead-years 4–10"

How would explain the skill emergence for higher lead times? Would you, for example, attribute it to initialization shock, sampling uncertainty (i.e., different end points of evaluation period for different lead times) or a combination of both?

A: This is a very interesting comment. As the reviewer rightly mentioned, initialization shock and sampling uncertainty may affect the predictability of this index. Moreover, we should also take into account that the TPI is computed considering the SST anomalies in three different areas: one located along the equatorial Pacific (exactly 10ºS–10ºN, 170ºE–270ºE) and the other two regions covering the mid-latitudes (25ºN–45ºN, 140ºE–215ºE and 50ºS–15ºS, 150ºE–200ºE). Looking at the spatial ACC maps (Fig. 2 in the manuscript), the Equatorial Pacific is well represented at lead year 1 but there are some significant values also at lead year 6–10. This is clearer focussing on the ENSO 3.4 index (Fig. 10c), in which the predictability is limited to the first lead year ranges. This may likely justify the TPI predictability at lead year 1–1 (ACC>0.30), although it is not significant. The emerging skill at higher lead years is probably due to the inclusion of off-equatorial SST, linked to the predictability of the Pacific Decadal Variability [Henley et al., 2015].

We add this to the text: "*The emerging skill at higher lead years is probably due to the inclusion of off-equatorial SST, linked to the predictability of the Pacific Decadal Variability [Henley et al., 2015].*"

19. L291 "a set of 15-member hindcasts, covering the period 1960–2020"

Maybe "initialized every year during 1960–2020" would be more precise?

If I have understood correctly, the entire forecast period covers 1961–2030 (plus December 1960). Is that right?

A: Yes, it is. The DPS is initialized every year, from November 1960 to November 2020, so that it covers from November 1960 to December 2030. We rewrote the text as follows: "*(...) a set of 20-member hindcasts, initialized every year from 1960 to 2020, (...)*"

20. L302–303 "The poor response in MSSS is also confirmed by a cold bias"

"confirmed" somewhat implies a causal relation between the poor MSSS response and cold bias. Could you elaborate on that? If the relationship is not known, then writing "coincides with" could be more appropriate.

Following Goddard et al. 2013, the MSSS is computed from anomalies "relative to their respective climatologies (which is equivalent to the removal of mean bias)". Hence, the MSSS panelizes amplitude and phase errors but not directly biases. Still, it is plausible that biases in high latitude ocean temperatures and sea ice cover also affect the representation of climate variability there.

A: We deleted this sentence from the manuscript since there was no causal relation between MSSS and bias.

21. L348–349 "inability of the DPS to reproduce the observed variability"

Do you mean inability to reproduce the historical temporal evolution of the observed variability or the inability to reproduce the dynamics and other characteristics of the observed variability? "inability" suggests you anticipate a higher real-world climate predictability over the North Pacific than the hindcasts of CMCC DPS would suggest. If so, why would you think so?

A: The reviewer is right, the text is not clear here. As correctly raised in a previous comment, the poor predictability of the North Pacific ocean likely depends on the limited predictability of ENSO and its strong influence on the PNA variability. We rephrased as follows: "*The predictive skill of the Pacific Ocean rapidly decays with forecast time, especially over the North Pacific, arguably due to the limited predictability of ENSO beyond the first forecast year and the consequent strong influence over the North Pacific (unpredictable ENSO-driven variability).*"

22. Sections 3.4 and 4

Can you provide a rational for why you show and discuss the NoInit results in sections 3.1–3.3 but not in 3.4 and 4. Was it mainly to keep the paper short and/or the benefit of initialization turned out to be less clear for the ROC and climate indices?

A: We have now assessed the climate indices for the historical+scenario runs since they further corroborate the benefit of initialization in synchronizing the internal variability. We modified Fig. 9 and Fig. 10 and section 4 in the new version of the manuscript.

We also plot the ROC score for the historical runs (Fig. R6). Compared to the ROC score of the DPS (Fig. 8 of the revised manuscript, also attached below), the historical assessment shows overall less skill especially at lead years 1 and 1–5, likely due to the lack of initialization. Most of the difference occurs over the oceans (e.g. over the North Atlantic ocean), the realm which is

most sensible to the initialization at decadal time scale. We have decided to put Fig. R6 in the supplementary materials (as Fig. S7) to keep the paper short.

We modified the section 3.4. as follows: "*Comparing the ROC score for NoInit, most of the difference occurs over the oceans (e.g. over the North Atlantic), the realm which is most sensible to the initialization at decadal time scale.*"

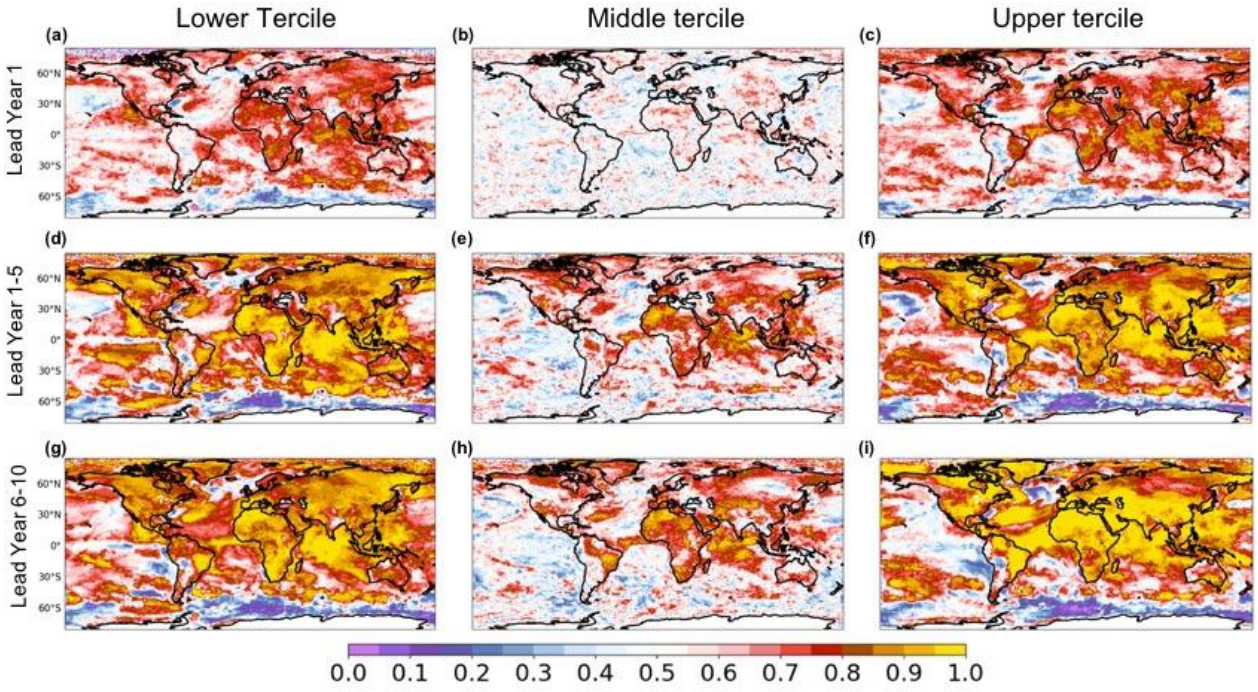


Fig. R6: Relative Operating Characteristic (ROC) from historical runs for near-surface temperatures (SST/TAS) for lead years 1, 1–5 and 6–10, considering three tercile categories: lower tercile (left column), middle tercile (central column) and upper tercile (right column).

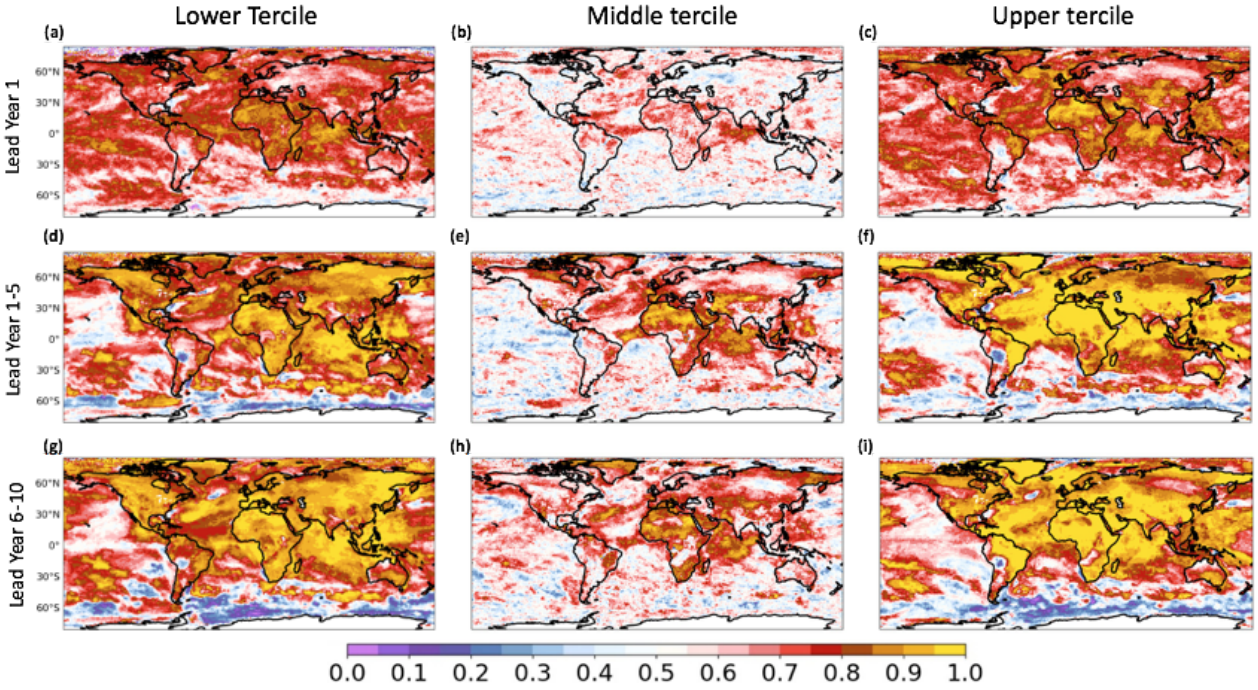


Fig. 8 of the revised paper: Relative Operating Characteristic (ROC) from DPS for near-surface temperatures (SST/TAS) for lead years 1, 1–5 and 6–10, considering three tercile categories: lower tercile (left column), middle tercile (central column) and upper tercile (right column).

REVIEWER #2

General comments

In this manuscript the authors present an overview of the CMCC decadal prediction system (CMCC DPS), a contribution to CMIP6 and DCPP, with a focus on the evaluation of CMCC DPS's decadal retrospective forecasts (hindcasts) of surface air temperature. After an introduction to the topic, the authors present a basic description of the DPS setup, go into the details of their verification metrics, and then present their results mainly in terms of surface parameters and the evaluation of some climate indices. The manuscript closes with a summarizing discussion.

I think that CMCC DPS and its skill assessment should definitely being published in GMD. However, with the current manuscript the authors fall a bit short of getting my recommendation. (1) In my opinion the skill assessment is just too much focussed on surface parameters, a short detour into the ocean would make the picture perfectly round, especially when considering that the system is working on decadal time scales. (2) Also, the authors' approach toward their full-field initialization scheme could be better explained and motivated. (3) And last not least I am missing some major conclusions in the abstract or at the end of the manuscript. Is the DPS' skill good enough to warrant the DPS use as an operational system? What is it that CMCC DPS can do very well?

A: We would like to sincerely thank reviewer #2 for her/his constructive comments and insightful suggestions. All comments and suggestions have been taken into account in our revision. Below are our point-to-point responses.

Major comments

(1) On the focus on surface parameters: I would like to ask the authors to include an assessment of upper ocean heat content prediction skill and an assessment of the state of the AMOC. I am aware that reliable AMOC observations only exist since 2004 (at 26°N), so a figure of the mean AMOC cell and a timeseries at 26°N may suffice.

A: We thank the reviewer for providing these thoughtful suggestions. In response, we have assessed the DPS AMOC mean cell at different lead years (Fig. R1), monthly time series for maximum AMOC at 26.5°N from RAPID array and DPS (Fig. R2) and annual time series for maximum AMOC at 26.5°N from RAPID array and DPS at different lead years (Fig. R3). We also plot the ACC (Fig. R4) and MSSS (Fig. R5) for the ocean heat content integrated over the first 300-meter depth (OHC300).

The mean AMOC cell in the DPS is quite well reproduced in terms of structure although the maximum at lead year 1 (Fig. R1) is located more south (below 20°N) compared to other AMOC reconstructions based on different oceanic reanalysis [Karspeck et al., 2017]. At lead year 1– 5 and 6–10 the maximum moves northwards, due to the model adjustment towards its own climatology, resembling the structure reported in many studies [e.g. Tsujino et al., 2020]. The slightly negative trend of the observed AMOC occurring during the last decades is reproduced just at lead year 1 in the hindcasts (Fig. R2 and Fig. R3), while the simulated low-frequency variability is consistent with the observed one also at longer lead years.

Our analysis on DPS has highlighted that the initialization shock may lead to the AMOC slowdown up to lead year 2 (Fig. R2 and Fig. R3), underestimating the maximum by about 2 Sv at 26.5°N in the period covered by Rapid array. We are aware that the DPS uses a horizontal grid with 1-degree resolution, which limits the ocean dynamics. A coarse grid may be the cause of a

weaker AMOC compared to the observed one in the RAPID array, independently from the adopted model [Tsujino et al., 2020].

Other similar studies [e.g. Bilbao et al., 2021] show results consistent with our findings, with a sharp weakening of the simulated AMOC during lead year one. The Norwegian DPS [Bethke et al., 2021] shows a high sensitivity of the AMOC to the details of the initialization approach with considerable impact on local surface temperatures. Moreover, the initial condition may affect AMOC predictability and, in general, the North Atlantic Ocean, due to a mis-representation of the impact of the atmosphere–ocean feedbacks [Brune and Baher, 2020].

The ACC and MSSS patterns computed for the OHC300 anomalies (Fig. R4 and R5, respectively) are consistent with the results obtained for the SST (Fig.2 and Fig.4 in the manuscript). At lead year 1, significant ACC covers most part of the oceans, except for the Eastern Atlantic and Southern ocean. The anomalies' values are also well captured north of 30°N. The OHC300 prediction skill is spatially reduced when higher lead-time ranges are considered. At lead year 1–5 and 6–10, the ACC is significant over the tropical Pacific, excluding the equatorial band due to the poor long-term predictability of ENSO, as also found in other DPSs (e.g. Bilbao et al., 2021). Positive ACC values also cover part of the Indian Ocean, South and North Atlantic regions. The lack of skill over the subpolar gyre may be partly due to the erroneous AMOC representation in the DPS, altering the local ocean circulation and heat content. The MSSS shows positive values mainly localized over the midlatitudes in the North Atlantic (Fig R5), quite consistent with SST MSSS (Fig. 4).

Following the reviewer's suggestion, we added the OHC300 ACC figure (Fig. R4) and relative discussion to the manuscript. Furthermore, the OHC300 MSSS and AMOC patterns and timeseries have been included in the supplementary material. In this way, we think we have strengthened the assessment of the DPS performance by also analyzing subsurface ocean parameters.

We add this paragraph to the text:

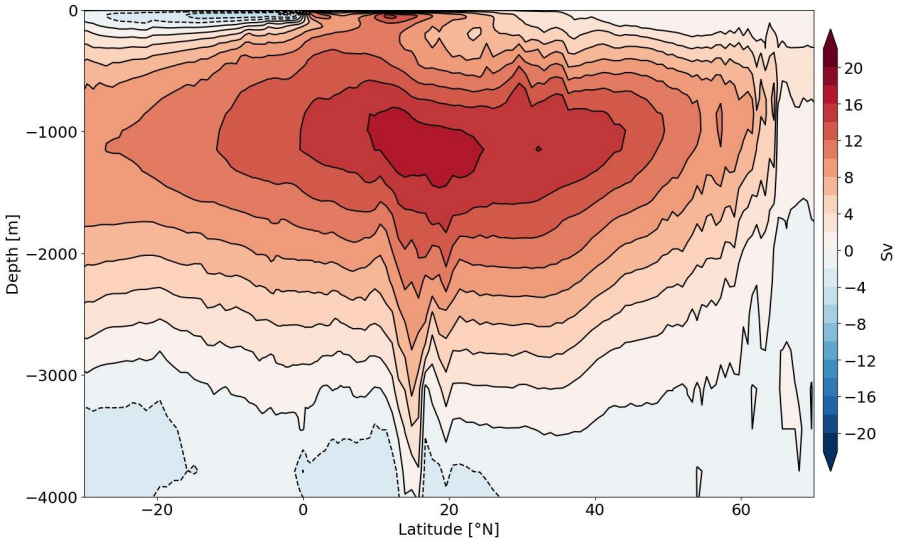
"*To corroborate the skill analysis of surface temperature at decadal time scales, we assess the skill for the ocean heat content integrated over the top 300 meters of the water column (hereafter OHC300). The ACC pattern computed for the OHC300 anomalies (Fig. 5) is similar, and thus consistent, with the results obtained for the SST (Fig. 2). At lead year 1, significant ACC covers most part of the oceans, except for the Eastern Atlantic and Southern ocean. The anomalies' values are also well captured north of 30°N. The OHC300 area exhibiting significant skill is reduced when higher lead-time ranges are considered. At lead year 1–5 and 6–10, the ACC is significant over the tropical Pacific, excluding the equatorial band due to the poor long-term predictability of ENSO, as also found in other DPSs (e.g. Bilbao et al. 2021). Positive ACC values also cover part of the Indian Ocean, South and North Atlantic regions. The MSSS shows positive values mainly localized over the midlatitudes in the North Atlantic (Fig. S6) and is found to be quite consistent with SST MSSS (Fig. 4). The lack of skill over the subpolar gyre may be partly due to the erroneous representation of the Atlantic Meridional Overturning Circulation (AMOC) in the DPS, altering the local ocean circulation and heat content. A complementary analysis reveals that the mean AMOC cell in the DPS is quite well reproduced in terms of structure although its maximum is located too far south (below 20°N) at lead year 1 (Fig. S3) compared to other AMOC reconstructions based on different oceanic reanalyses [e.g. Karspeck et al., 2017]. At lead year 1–5 and 6–10 the maximum moves northwards, due to the model adjustment towards its own climatology, resembling the structure reported in other studies [e.g. Tsujino et al., 2020]. The initialization shock may lead to the AMOC slowdown up to lead year 2 (Fig. S4 and Fig. S5),*

underestimating the maximum by about 2 Sv at 26.5°N in the period covered by Rapid array. The slightly negative trend of the observed AMOC occurring during the last decades is reproduced just at lead year 1 in the hindcasts (Fig. S4 and Fig. S5) while the simulated low-frequency variability is consistent with the observed one also at longer lead years. ”

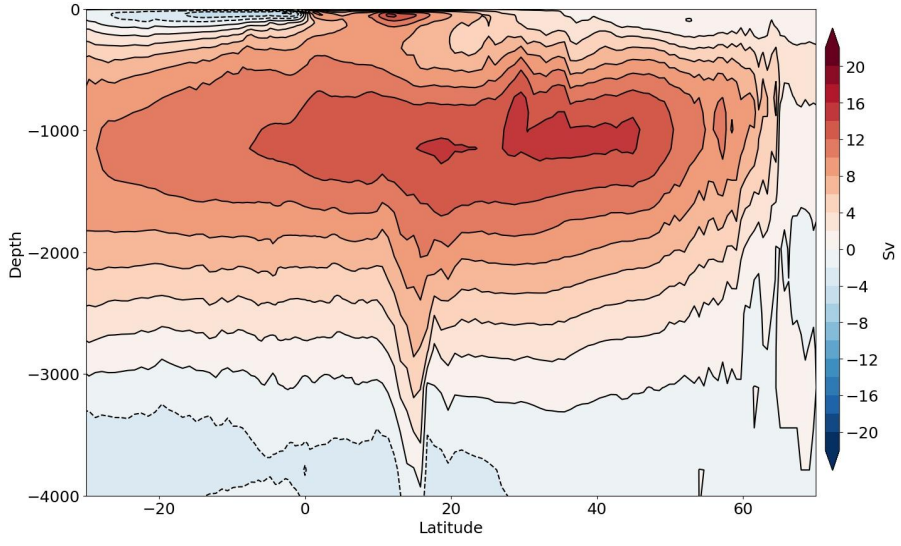
REF: Tsujino, H., Urakawa, L. S., Griffies, S. M., Danabasoglu, G., Adcroft, A. J., Amaral, A. E., ... & Yu, Z. (2020). Evaluation of global ocean–sea-ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2). Geoscientific Model Development, 13(8), 3643–3708.

REF: Karspeck, A. R., Stammer, D., Köhl, A., Danabasoglu, G., Balmaseda, M., Smith, D. M., ... & Rosati, A. (2017). Comparison of the Atlantic meridional overturning circulation between 1960 and 2007 in six ocean reanalysis products. Climate Dynamics, 49(3), 957–982.

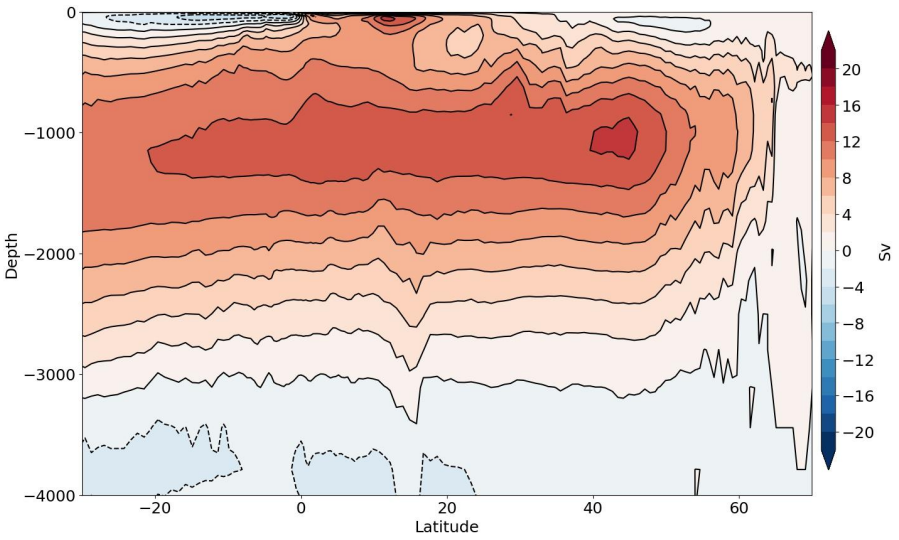
REF: Brune, S., & Baehr, J. (2020). Preserving the coupled atmosphere–ocean feedback in initializations of decadal climate predictions. Wiley Interdisciplinary Reviews: Climate Change, 11(3), e637.



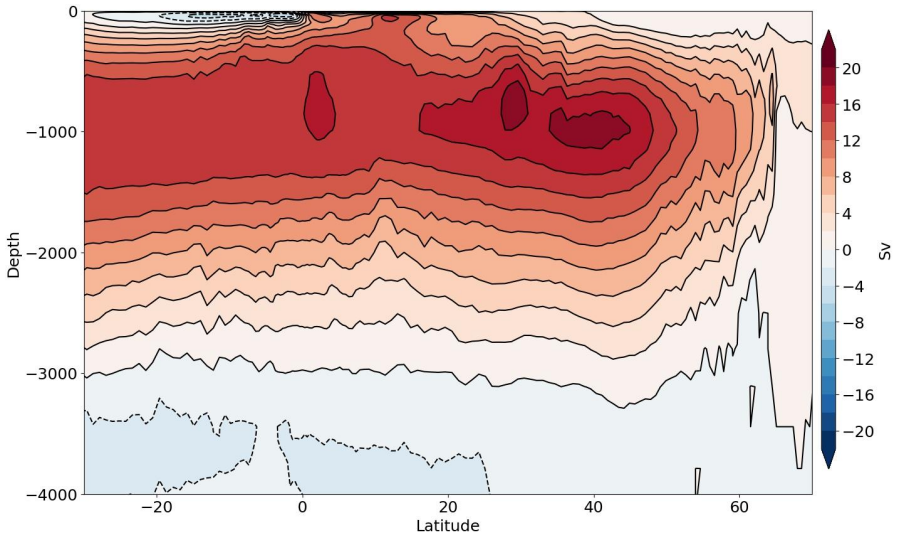
a)



b)



c)



d)

Fig R1: Mean annual AMOC cell in the DPS for lead year 1 (a), lead year 1–5 (b), lead year 6–10 (c) and for the historical run (d). We consider all the start dates (1960–2020).

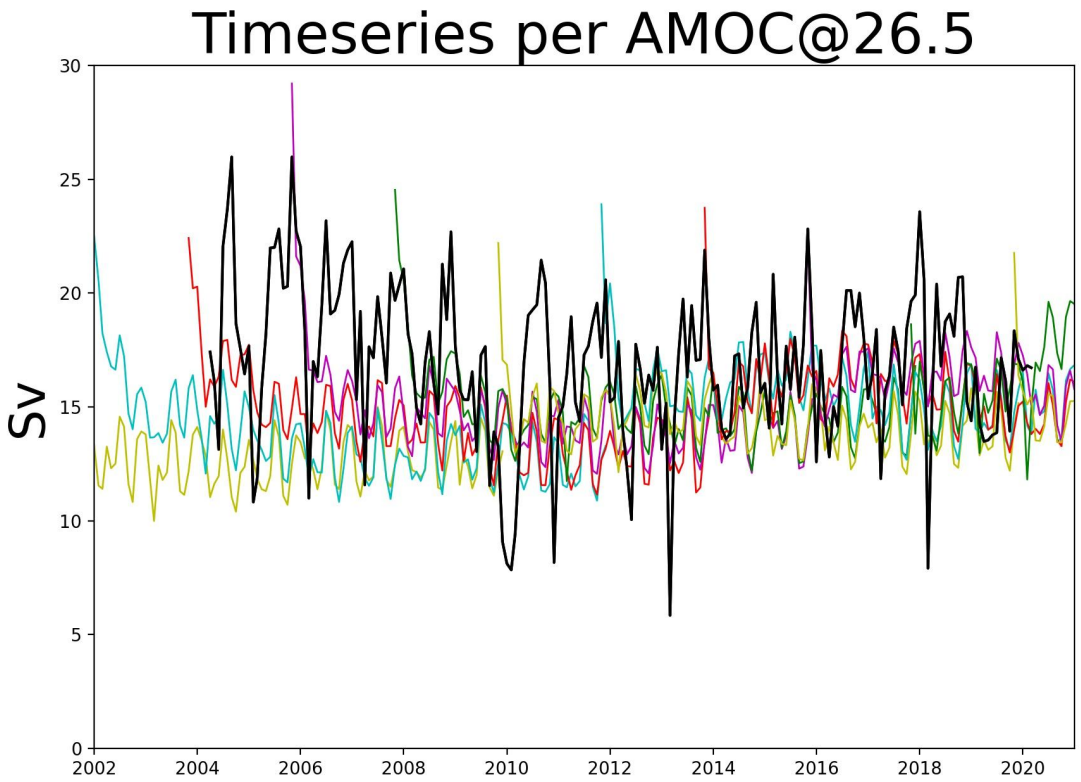
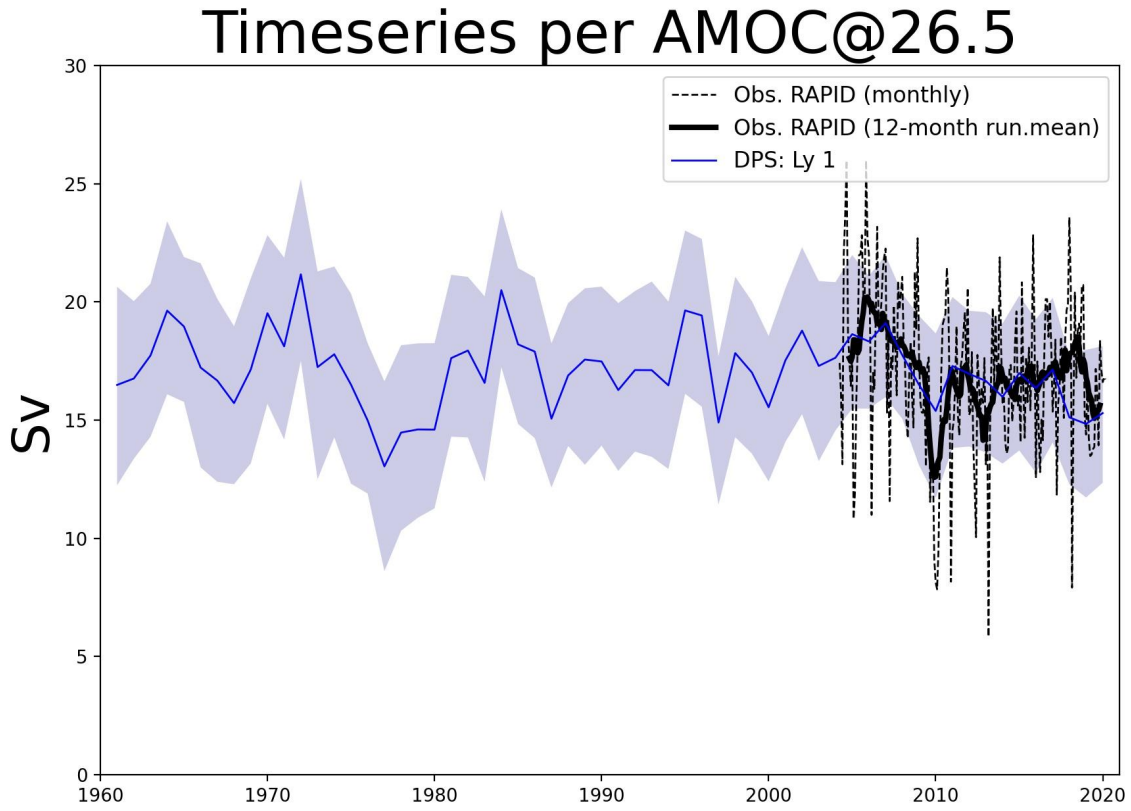
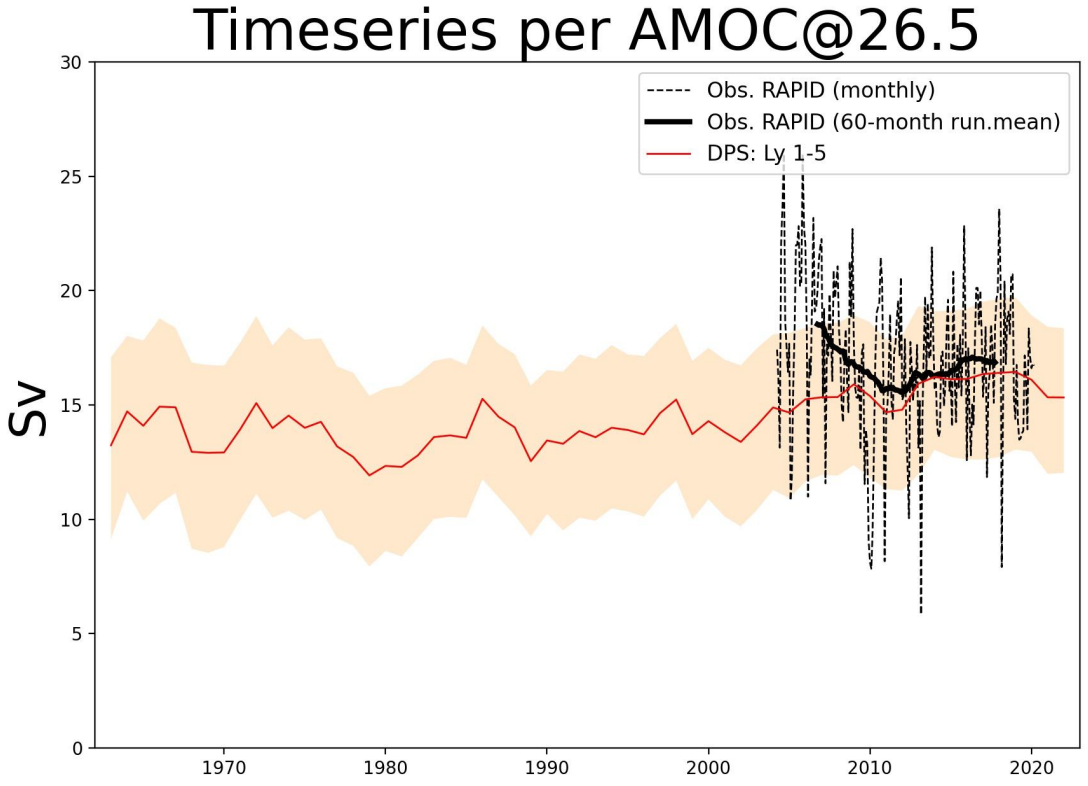


Fig R2: Monthly time series for maximum AMOC at 26.5°N from RAPID array (in black, Moat et al., 2022) and from DPS (each colored line represents a 20-member ensemble mean). For illustrative purposes, we plot ensemble means every two years.

a)


Timeseries per AMOC@26.5
Obs. RAPID (monthly)
Obs. RAPID (12-month run.mean)
DPS: Ly 1
Sv
30
25
20
15
10
5
0
1960
1970
1980
1990
2000
2010
2020

b)


Timeseries per AMOC@26.5
Obs. RAPID (monthly)
Obs. RAPID (60-month run.mean)
DPS: Ly 1-5
Sv
30
25
20
15
10
5
0
1970
1980
1990
2000
2010
2020

c)

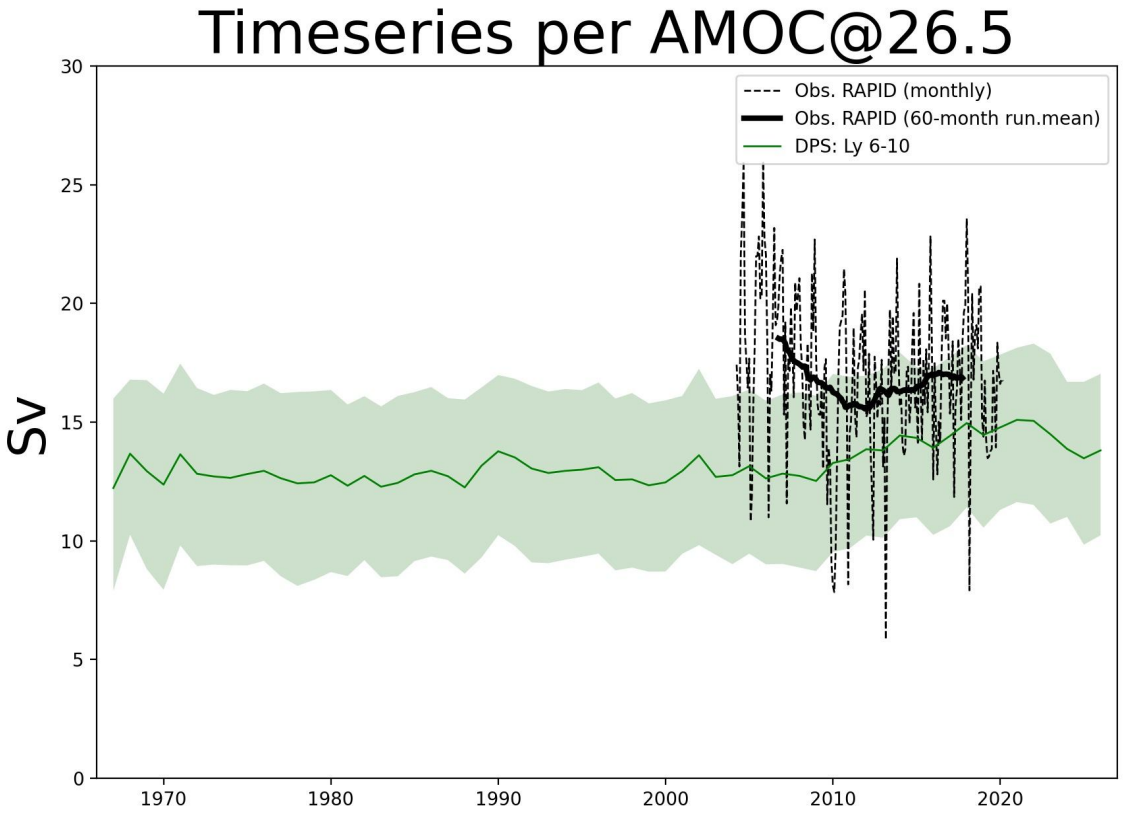


Fig R3: Annual time series for maximum AMOC at 26.5°N from RAPID array (in black) and DPS (colored lines) and its intra-ensemble spread (colored envelopes) for for lead year 1 (a), lead year 1–5 (b) and lead year 6–10 (c). The black dashed line represents monthly observed RAPID values while the black solid line is the 12-month (60-month) running mean for lead year 1 (lead year 1–5 and lead year 6–10) for a consistent comparison between the timeseries.

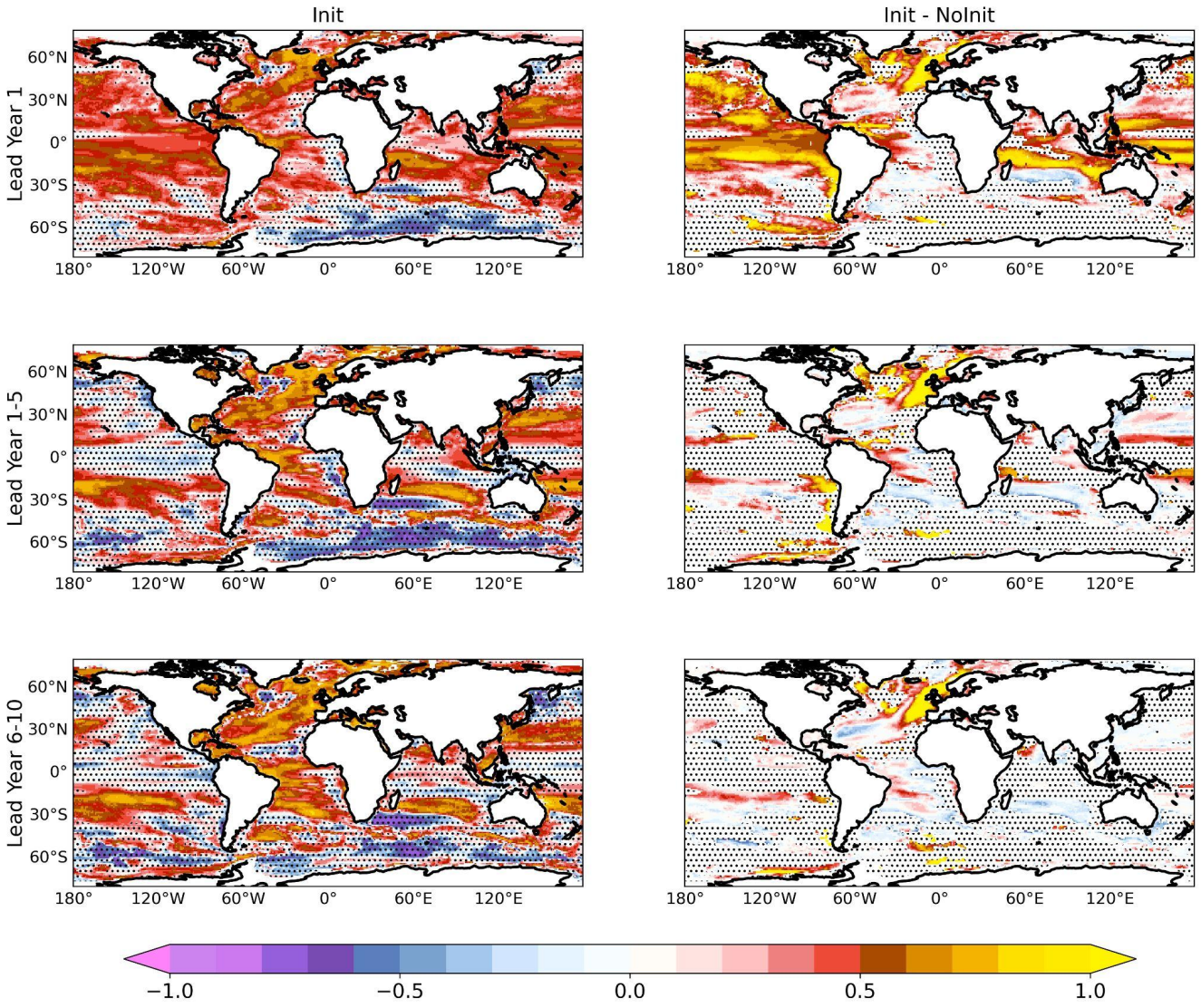


Fig. R4: Anomaly Correlation Coefficient (ACC) of the hindcast ensemble (“Init”) and its difference with the historical ensemble (“Init - NoInit”) for Ocean Heat Content integrated over the top 300 meters for lead years 1 (top panels), 1–5 (middle panels) and 6–10 (bottom panels). Only “Init” significant values are plotted in “Init - NoInit” plots. Stippling denotes points where 95% statistical significance is not reached, according to a one-tailed t test. Effective degrees of freedom have been considered [Bretherton et al., 1999]. Reference dataset is ORAS4 [Balmaseda et al. 2013].

Balmaseda, M.A., Mogensen, K. and Weaver, A.T. (2013), Evaluation of the ECMWF ocean reanalysis system ORAS4. Q.J.R. Meteorol. Soc., 139: 1132–1161. https://doi.org/10.1002/qj.2063

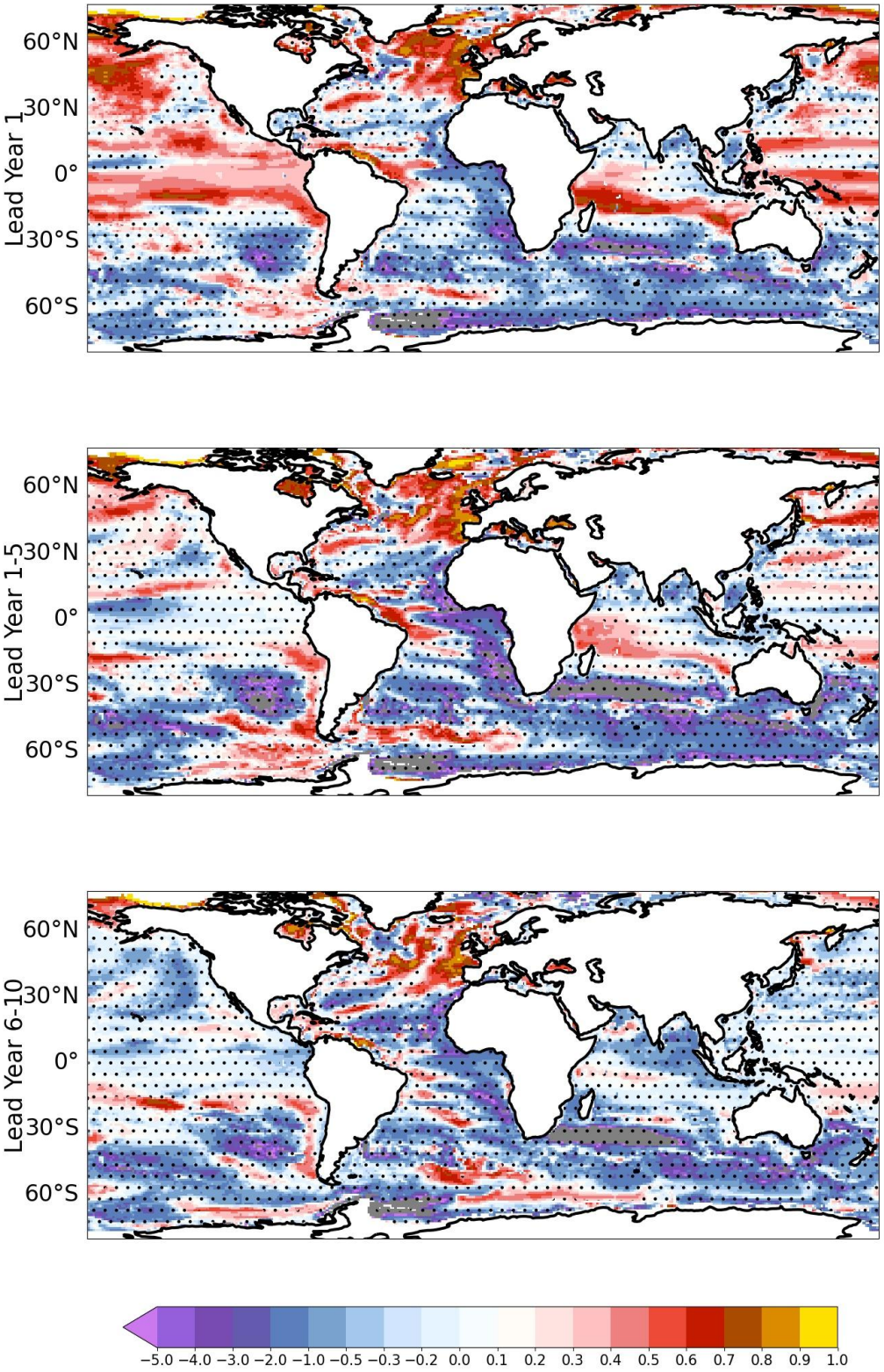


Fig R5: Mean Squared Skill Score (MSSS) for the Ocean Heat Content integrated over the top 300 meters computed for the hindcasts (Init) using NoInit runs as the reference forecast. Note that the colorbar is not symmetric around zero. Stippling is used to indicate points where 95% statistical significance is not reached, according to a one-tailed *t*-test. Accounting for auto-correlation, the effective number of degrees of freedom has been estimated following Bretherton et al., 1999.

(2) On the initialization approach: I did not fully understand how the initialization approach in CMCC DPS works. I am fine with the listing of all the products used in the process of initialization. But still: is there any "classical" data assimilation involved? Are initial fields

directly derived from reanalyses or other reference datasets? How does the initialization compare to other initialization approaches within DCPP?

A: We appreciate this question and take the chance to better explain our initialization strategy. We used 3D states of the Earth system components directly derived from different reanalyses (now summarized in table T.1 in the manuscript), following the same approach used in the operational CMCC seasonal forecasting system for oceanic, sea-ice and land components. In particular, the ocean component is initialized using CHOR/CGLORSv7 reanalysis produced with a three-dimensional variational data assimilation system, surface nudging and employing a bias correction scheme [Storto and Masina, 2016]. it is worth noting that the reanalysis is performed with the same ocean model used in the CMCC DPS (i.e. NEMO v3.6). Similarly, to initialize the land component, we use a land-only model (CLM4.5) run forced with atmospheric fluxes obtained from the ECMWF reanalysis (ERA5). Also in this case, it is important to note that the land-surface model used to produce the land-surface initial conditions is the same as the one used in our DPS as the land-surface component. Then, an interpolation on the respective DPS grid has been performed to the source data for each component.

Our full-field initialization approach is similar to that adopted by other DPSs [e.g. Bilbao et al., 2021; Sospedra-Alfonso et al. 2021], in which an a-posteriori correction (lead-year-dependent mean bias correction) is applied to account for the model drifting towards its own climatology. Other DPSs [e.g. Bethke et al., 2021, Kataoka et al. 2020] have been initialized using the observed anomalies superimposed to the respective model climatology, so as to avoid the initialization shock and the subsequent drift Both of these strategies/approaches are deemed valid in CMIP6 DCPP protocols [Boer et al., 2016], each presenting some drawbacks: in the former, the dynamics associated with the inevitable drift may interact with the dynamics driving the actual variability to be simulated, degrading predictability, while the benefits of the latter depend on the assumption that the model variability is independent from the model mean state and it realistically represents the observed variability. Nevertheless several studies have shown that differences in skill are rather small and localized in space and lead time [Smith et al., 2013; Bellucci et al., 2014; Volpi et al., 2016].

We modified the text at section 2.1

(3) On the conclusions: I appreciate the author's state to "only" report on the prediction system and its skill. However, some sharper conclusions (at the end of the manuscript and in the abstract) could improve the manuscript considerably. They could go along the line of these questions: Is the system good enough to be used for decadal predictions? What is it that the system can do exceptionally well?

A: We modified the manuscript, in “conclusions” and “abstract” sections as follows:

5. Summary and Conclusions

In this paper we analyzed the predictive capabilities of the CMCC DPS, using a set of 20-member hindcasts, initialized every year from 1960 to 2020, performed with the CMCC-CM2-SR5 coupled model and a full-field initialization strategy.

The study has highlighted the following main findings:

- *the DPS skilfully reproduces the observed variability of surface temperature (T2M over land and SST over the oceans) and upper-ocean heat content (Fig. 2 and Fig. 5, respectively), with a large fraction of the total skill stemming from long-term trends*

associated to changes in the external forcings (Van Oldenborg et al., 2012). The North Atlantic Ocean is the region that benefits the most from initialization (ACC difference up to 0.80 in Fig. 2.b,d,f and MSSS close to 0.6 in Fig. 4), with the largest skill enhancement (compared to historical simulations) over the subpolar gyre region. As still typical in decadal predictions, a lack of skill characterizes the whole Pacific Ocean, over which significant ACC values are bound to lead year 1.

- *Some climate variability patterns in the North Atlantic sector feature significant predictability. In particular, the observed AMV signal is skillfully predicted (ACC=0.91, Fig. 8a), and this likely contributes to obtaining significant skill also in remote areas through downstream influence, circulation changes and teleconnections. A remarkable prediction skill is also found for the NAO index, with maximum correlations obtained in the 1–9 lead-year range (ACC=0.58). This promising result has been demonstrated also for other decadal prediction systems (Athanasiadis et al., 2020; Smith et al. 2020), yet always with significantly larger ensembles. The predicted NAO index can be used to improve forecasts of poorly simulated variables influenced by the observed NAO changes (Dunstone et al., 2022). Over the tropical Pacific Ocean, ENSO variability exhibits some predictability up to year 3, with highest values of ACC bound to the first winter after initialization (ACC=0.95). Moreover, the TPI shows higher predictive skill on longer timescales despite the low skill that the DPS exhibits over most of the Pacific Ocean.*

- *Regarding precipitation, the CMCC DPS shows limited skill, with statistically significant correlations only over specific areas, a feature shared with other state-of-the-art decadal prediction systems. Indeed, significant skill is only found over Sahelian Africa, Northern Eurasia and over the Western and Central Europe, with ACC values up to 0.50 (Fig. 3). This spatially confined skill may derive from the AMV, which is known to be a source of predictability for these regions, influencing rainfall variability on annual-to-decadal timescales (Doblas-Reyes et al., 2013; Ehsan et al., 2020; Ruggieri et al., 2020). On the other hand, no significant skill for precipitation is found over the rest of the globe, probably also due to the relatively small size of our ensemble (Yeager et al., 2018) and to the very high spatial variability (Goddard et al., 2013) and absence of a strong trend in precipitation (Gaetani and Mohino 2013; Bellucci et al., 2015). Improvements compared to the historical simulations are bound to some regional features, suggesting no substantial benefits from initialization in terms of precipitation skill.*

Systematic errors also affect the forecast quality. The strong cold bias occurring over the Northern Hemisphere tends to produce an initialization shock and a subsequent drift, typical of the full-field initialization (He et al., 2017) approach used in the CMCC DPS. Admittedly, AMOC is particularly affected by the initialization strategy (Fig. S3), with the full-field approach inducing a long-term adjustment due to the bias in the representation of the large-scale ocean circulation (Polkova et al., 2014). In this context, another sensitive region is the Equatorial Pacific in which full-field initialization seems to give the strongest benefit in skill compared to anomalous initialization (Bellucci et al., 2015). From another perspective, model errors may be mitigated by enhancing spatial resolution (both horizontal and vertical) in the oceanic and atmospheric model components, since coarse resolution limits a realistic representation of key physical processes (e.g. realistic SST front in the Gulf Stream region), impacting the atmospheric circulation downstream (Athanasiadis et al., 2022; Paolini et al., 2022). For instance, an eddy-permitting ocean model (i.e. 0.25° horizontal resolution) in a fully-coupled system led to improved decadal predictions over the whole equatorial zone (Shaffrey et al., 2017). Moreover, increasing the

ensemble size is expected to further increase the skill by allowing the predictable signal to emerge more clearly from the chaotic variability (Athanasiadis et al., 2020).

Interestingly, the results obtained from the CMCC DPS are broadly consistent with similar assessments from other CMIP6 decadal prediction systems (Bethke et al., 2021; Bilbao et al., 2021; Kataoka et al., 2020; Robson et al., 2018; Sospedra-Alfonso et al., 2021; Xin et al., 2019; Yang et al., 2021; Yeager et al., 2018) and multi-model studies (Borchert et al., 2021a, 2021b; Delgado-Torres et al., 2022). In particular, most of the DPSs feature high predictive skill over the Atlantic Ocean, the Indian Ocean and continental areas, where a large fraction of predictability stems from the external forcings. The added value of initialization is most noted over the subpolar gyre and the subtropical Atlantic (in most of the DPSs), confirming these areas as those in which decadal predictions benefit the most from realistic initialization. The non-significant skill found over the Southern Ocean is considered to be, at least in part, due to the lack of oceanic observations in that region that prevents the accurate estimate of the initial state of the ocean. The predictive skill of the Pacific Ocean rapidly decays with forecast time, especially over the North Pacific, arguably due to the limited predictability of ENSO beyond the first forecast year and the consequent strong influence over the North Pacific (unpredictable ENSO-driven variability).

We remind that the CMCC DPS provides decadal forecasts (operationally since 2021) for the annual release of the WMO (World Meteorological Organization) Global Annual-to-Decadal Climate Update, a multi-model assessment of the near-term climate for societal applications (Hermanson et al., 2022). At the same time, the CMCC DPS makes a significant contribution the grand ensemble CMIP6 DCPP-A hindcasts.

The encouraging results obtained in this study indicate that climate variability over land can be predictable over a multi-year range, as well they demonstrate that the CMCC DPS is a valuable addition to the current generation of DPSs. This stresses the need to further explore the potential of the near-term predictions, further improving future decadal systems and initialization methods, in the perspective to provide a reliable tool to inform decision makers on how regional climate will evolve in the next decade.

***Abstract**. Decadal climate predictions, obtained by constraining the initial condition of a dynamical model through an accurate estimate of the observed climate state, provide the best available assessment of climate change in the near-term range and a useful tool to inform decision-makers on future climate-related risks.*

Here we present results from the CMIP6 DCPP-A decadal hindcasts produced with the operational CMCC decadal prediction system (CMCC DPS), based on the fully-coupled CMCC-CM2-SR5 dynamical model. A 20-member suite of 10-year retrospective forecasts, initialized every year from 1960 to 2020, is performed using a full-field initialization strategy.

The predictive skill for key variables is assessed and compared with the skill of an ensemble of non-initialized historical simulations so as to quantify the added value of initialization. In particular, the CMCC DPS is able to skilfully reproduce past-climate surface and subsurface temperature fluctuations over large parts of the globe. The North Atlantic Ocean is the region that benefits the most from initialization, with the largest skill enhancement occurring over the subpolar gyre region, comparing to uninitialized historical simulations. On the other hand, the predictive skill over the Pacific Ocean rapidly decays with forecast time, especially over the North Pacific. In terms of precipitation, the CMCC DPS skill is significantly higher than that of the historical simulations over a few specific regions, including Sahel, Northern Eurasia and over the Western and Central Europe.

The Atlantic Multidecadal Variability is also skilfully predicted, and this likely contributes to the skill found over remote areas through downstream influence, circulation changes and teleconnections. Considering the relatively small ensemble size, a remarkable prediction skill is also found for the North Atlantic Oscillation, with maximum correlations obtained in the 1–9 lead-year range.

Systematic errors also affect the forecast quality of the CMCC DPS, featuring a prominent cold bias over the Northern Hemisphere, which is not found in the historical runs, suggesting that in some areas the adopted full-field initialization strategy introduces significant perturbations to the model in respect to its equilibrium state.

The encouraging results obtained in this study indicate that climate variability over land can be predictable over a multi-year range, as well they demonstrate that the CMCC DPS is a valuable addition to the current generation of DPSs. This stresses the need to further explore the potential of the near-term predictions, further improving future decadal systems and initialization methods, in the perspective to provide a reliable tool to inform decision makers on how regional climate will evolve in the next decade.

(4) On the 10 historical members for verification: I understand that the authors would rather use 15 prediction members than only 10. Nevertheless the more consistent approach would be to use only 10 members in the prediction ensemble when it is going to be verified by a 10 member historical ensemble. Please elaborate more on that you can actually use the 15 members.

Since only 10 historical members are available, and now the 20 members for the hindcasts are considered, the hindcast ensemble has been scaled down using a random sub-sampling with 100 iterations in skill assessment plots where we directly compare the hindcasts with the historical ensemble. It is worth noting that no strong differences occur with the previous assessment as well as with a varying number of iterations (not shown). We specified this in the method section 2.2 as follows:

"*Since only 10 historical members are available, the 20-member hindcast ensemble has been scaled down using a random sub-sampling with 100 combinations in order to allow a fair comparison to the historical ensemble in the skill assessments.*"

Minor comments

l. 10 "Abstract"

The abstract is missing any conclusion.

Please include some of your main conclusions, they are also still missing in "Summary and conclusions"

A: We modified the manuscript, in "conclusions" and "abstract" sections. See major comment #3

l. 15 "full-field initialization"

Please be precise in terminology. In the summary it is named "full-value". From what I understood the term "full-field and full-value" might actually be describing the setup best. But "full-field" would be accurate enough, I guess.

A: We agree. Now we use consistent terminology (i.e. "*full-field*") in the revised manuscript.

l. 27 "Model systematic errors"

Since these errors may not be present in an uninitialized historical or control simulation, I would suggest to call them not "model systematic" but only "systematic" errors.

A: We agree with the reviewer. We have corrected the text following her/his suggestion.

l. 31 "For a long time"

Please be more specific. What do you mean with "long"? 10 years, 50 years?

A: We modified the text as follows: "*Prior to year 2000*"

l. 94 "All members are initialized on November 1st, starting from full-field estimates of the observed state of the ocean, sea-ice, land surface and atmosphere."

I am puzzled how the 15 ensemble members are initialized. Just by simply taking reanalyses fields as the starting point? In the following, the authors name different initial conditions for atmosphere, ocean etc. Could you please summarize this in a table so that the 15 different initial conditions are shown together?

A: Following your comment, we have added a summarizing table on the initial conditions in this version of the manuscript. We have also adopted a suite of 20 members in this version. The 20 members are generated by the combinations of 2 initial conditions from the land component, 2 from the atmosphere and 5 from the ocean.

	Data Source	Number of ICs	Procedures
LAND	Land-only analyses forced by 2 different atmospheric datasets: CRUNCEPv7 [Viovy, 2016] and GSWP3 [Kim, 2017]. Note: from 2015 onwards, the atmospheric fluxes to force the land-only analysis are taken from NCEP reanalysis (instead of CRUNCEPv7) and from ECMWF ERA5 (instead of GSWP3).	2 ICs (2 runs forced by 2 different datasets, providing instantaneous 2-meter air temperature and humidity, 10-meter winds and surface pressure every six hours and 3-hourly-accumulated radiation and precipitation)	Direct interpolation on target grid from land restarts
ATMOSPHERE	ERA40 [Uppala et al., 2005] for 1960–1978 start dates, ERA-Interim [Berrisford et al., 2019] for 1979–2018 start dates and ERA5 [Hersbach et al., 2020] from 2019 onwards.	2 ICs (derived from time-lagging perturbations, using the 1st and 2nd November)	Direct interpolation on target grid from atmospheric 3D state of temperature, specific humidity and horizontal wind components
OCEAN	CHOR [Yang et al., 2016] for 1960–2010 start dates and CGLORSv7 [Storto and Masina, 2016] for 2011–present start dates.	5 ICs (from 3 realizations of the global ocean/sea-ice reanalysis + 2 ICs from linear combinations of the former 3 ICs)	Direct interpolation on target grid from 3D state of temperature, salinity and horizontal components of the ocean currents.

	Data Source	Number of ICs	Procedures
SEA-ICE			Direct interpolation on target grid of sea-ice temperature, sea-ice volume, sea-ice area and snow volume.

It remains unclear, how much of data assimilation has been used for initialization. Please include a statement on that topic.

If you have not been using data assimilation, could you please elaborate more on why not?

A: We did not use any assimilation runs for the generation of initial conditions. We have discussed this in major comment #2.

If you have been data assimilation, how many of the reanalyses used for this DPS have been done with the same or similar model? This question goes along the line of trying to use the very same coupled ESM for data assimilation, ensemble generation, and prediction, see e.g. Brune and Baehr 2020.

A: As mentioned in comment #2 in the “Major” section, no data assimilation has been applied to the initial conditions.

l. 113 "We use a 10-member ensemble of historical simulations initialized"

Unfortunately only 10 historical members are used for verification. I strongly suggest to use 15 historical members as well, or scale down the hindcasts ensemble to 10 members. If that's impossible, please elaborate more on the discrepancy between the two ensembles and the possible downsides in terms of skill assessment.

A: As stated in major comment #4, since only 10 historical members are available, the 20 -member hindcast ensemble has been scaled down using a random sub-sampling with 100 combinations in skill assessment plots (i.e. the Init skill is equal to the averaged kill of 100 different combinations) where we directly compare the hindcasts with the historical ensemble. It is worth noting that no strong differences occur with the previous assessment as well as with a varying number of iterations (not shown). We specified this in the text as follows:

“*Since only 10 historical members are available, the 20-member hindcast ensemble has been scaled down using a random sub-sampling with 100 combinations in order to allow a fair comparison to the historical ensemble in the skill assessments.*”

l. 290 "Summary and conclusions"

This is rather a summary discussion. A proper discussion is missing. For a summary it is quite long.

Please include at least a paragraph at the very end with the conclusions, incorporating your two or three main findings.

A: We modified the manuscript, in “conclusions” and “abstract” sections (see major comment #3).

l. 292 "full-value initialization"

See above, please use consistent terminology.

A: Fixed

l. 295 "different variables"

Please be more specific here: are you referring to "surface parameters", or "surface temperature", actually the results are about "surface parameters" and "surface climate indices".

A: Following your comment, we corrected the text using "*surface temperature*"

l. 297, l. 302, l. 314 "NoInit"

I suggest using "non-initialized historical simulation" or "historical simulation" here to not overfreight the paragraph with abbreviations.

A: We agree with the reviewer. We changed the text in the mentioned lines.

l. 297 "The ROC score index in Init highlights the DPS ability"

Please use "initialized prediction" or "prediction" instead of "Init" here. Also, the sentence sounds a bit awkward. Please rephrase, e.g. "The DPS correctly discriminates (...) intervals, with ROC scores of <value> (Fig.?)."

A: We changed the text as follows: "*The DPS correctly discriminates the occurrences of below-tercile and above-tercile surface temperature anomalies throughout different lead-year intervals, with ROC scores close to one over land (Fig. 7).*"

l. 308f "significant skill is only found over Sahelian Africa ..."

Perhaps it is worthwhile here to include skill values or the range of them.

A: We added the ACC values.

l. 317 "AMV signal is skillfully predicted"

Same as above, a skill value would help.

A: We changed the text adding the ACC value of the AMV.

l. 325 "Model systematic errors"

see above comment for the abstract, I would suggest to name this only "systematic errors".

A: We deleted the "*Model*" word from this sentence.

l. 327 "Admittedly, AMOC is particularly affected by the initialization strategy,..."

I suggest including a figure illustrating the initial AMOC state and the resulting long-term drift (at 26°N). With the direct full-field/full-value initialization, strange things could happen

to AMOC and come back later on the time scale of 10 to 30 years, which is characteristic for AMOC disturbances. This figure could go into the supplement.

A: Following your suggestion, we have added AMOC figures to the supplementary material. We have discussed the figures in the major comment #1 and in the manuscript at section 3.2.

Technical comments

l. 41 "(CMIP5)(Smith et al. ...)

Please use correct citation for CMIP5 AND please avoid brackets)(. Here I would suggest: "(CMIP5, Taylor et al. 2012; Smith et al. ...)"

A: Fixed

l. 44, l.64 Same as above, please avoid brackets)(.

A: Fixed

l. 332 "course" Should read "coarse".

A: Fixed

References

S. Brune and J. Baehr, “Preserving the coupled atmosphere-ocean feedback in initializations of decadal climate predictions,” WIREs Clim. Change, vol. 11, no. 3, Art. no. 3, 2020, doi: 10.1002/wcc.637