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.*"

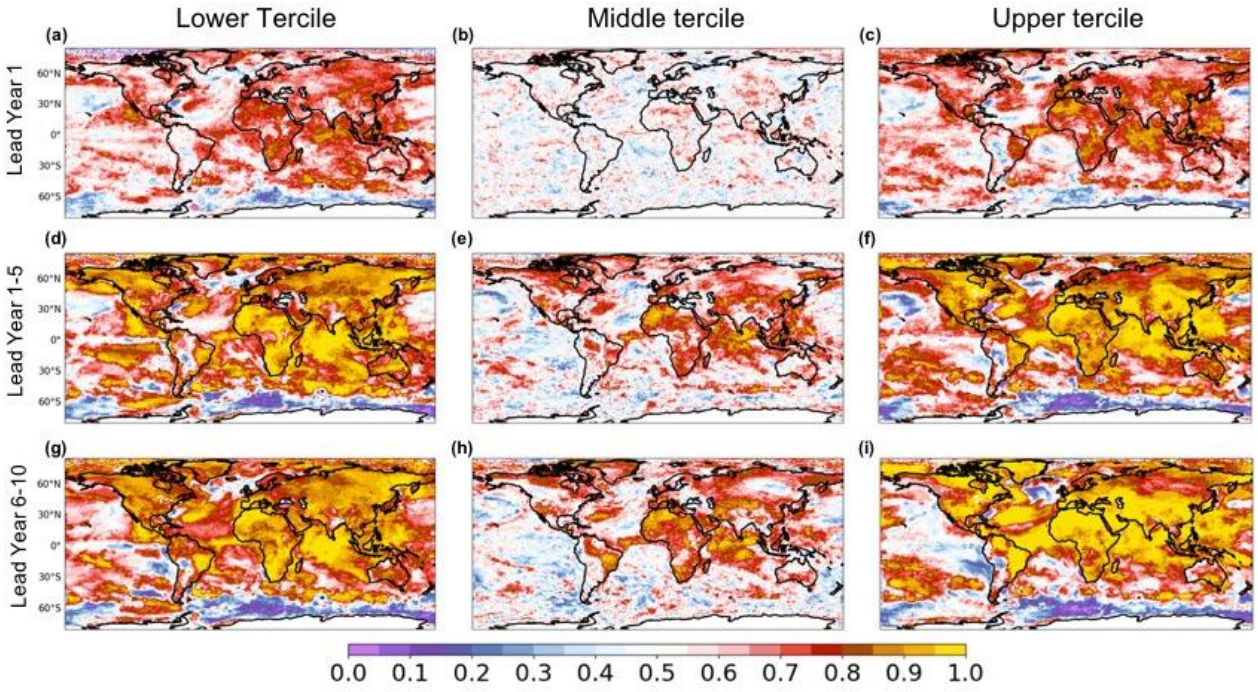

Fig. R6: Relative Operating Characteristic (ROC) from historical runs for near-surface temperatures (SST/TAS) for lead years 1, 1–5 and 6–10, considering three tercile categories: lower tercile (left column), middle tercile (central column) and upper tercile (right column).

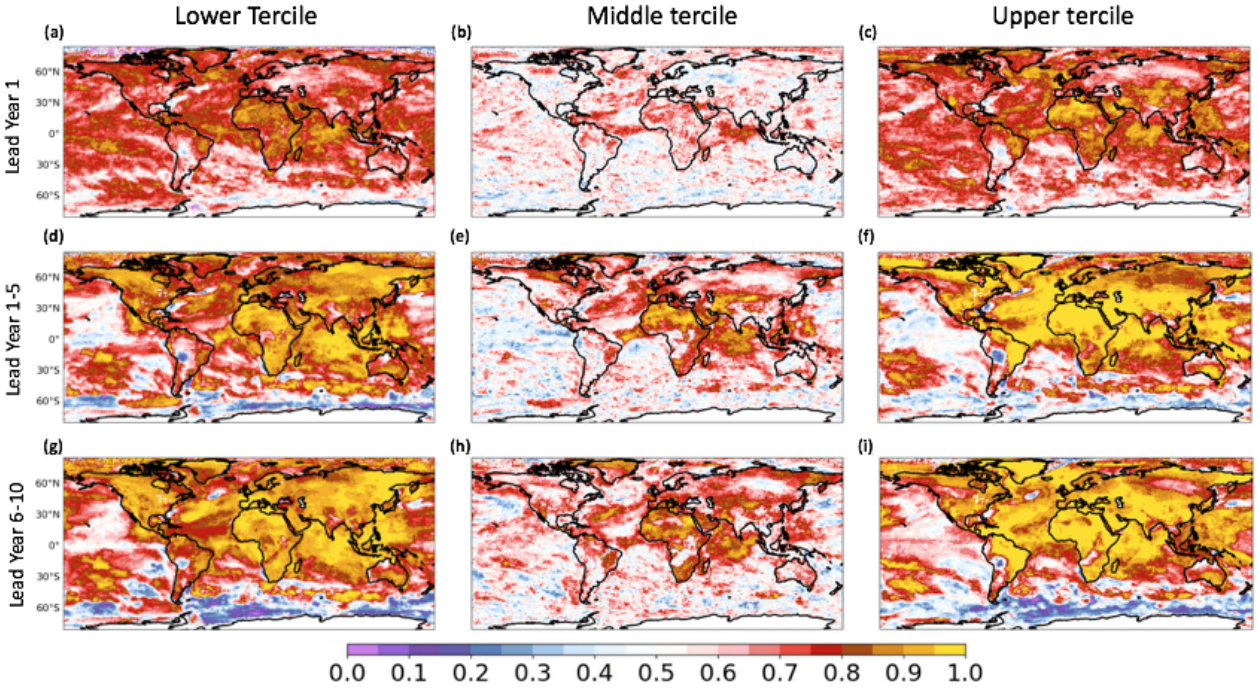

Fig. 8 of the revised paper: Relative Operating Characteristic (ROC) from DPS for near-surface temperatures (SST/TAS) for lead years 1, 1–5 and 6–10, considering three tercile categories: lower tercile (left column), middle tercile (central column) and upper tercile (right column).