Comments Reviewer 1

MAJOR

L21: Given the lower skill of the analogs (e.g. Fig.1 and 2) but that they are potentially very useful as a tool for making seamless predictions, I think the abstract should make it clear that the skill is lower rather than 'competititve'.

This is a fair point, the text of the abstract (ln. 20-22) was changed to "The analog-based seamless prediction system ~~is~~ shows very similar patterns of skill~~competitive~~ compared to state-of-the art initialised climate prediction ~~systems~~systems and has competitive skill- with initialized numerical prediction systems on annual and biennial forecast ranges. ~~that currently provide forecasts for specific time scales, such as seasonal and multi-annual~~. "

L120: This presumably results in all members having the same trend? If so, this needs a little discussion in the text with pros and cons as you are losing the individual model response to forcing and replacing it with the multimodel mean. Does this also reduce the variance in the ensemble?

The trend adjustment is one of the necessary steps of the analog-based method due to the fact that analogs can be selected from any year in the period 1960-2030 and do not necessarily have the right forcing state. The trend adjustment ensures that if selecting analogs from a different year/forcing state this offset is corrected to actually represent the correct forcing of the year(s) of the predictions. To make this point clear we have modified sentences 132-136: "This is necessary because ~~the observed trend is better represented in the CMIP6 ensemble than in the analog-based predictions without post-processing.~~ the analogs can be selected from any year in the period 1960-2030 and do not necessarily have the right forcing state. The trend adjustment ensures that potential offsets related to selecting analogs from other forcing states are corrected to represent the forcing of the year(s) of the predictions"

L146: Also on trends. The reference forecast R is stated to be a trivial climatological forecast but what does this mean? Is it a constant climatological value for each variable? Why not use a linear trend for Ts? This would seem like a fairer test.

By 'climatological forecast,' we mean the climatological value for each variable in the specified period: December-February (Fig. 1-3), June-August (Fig. 2-3), and annual, biennial, and quadrennial climatology (Fig. 6, 7, and 8, respectively). We believe that besides the skill evaluation against a climatological forecast, which is a common practice in any skill evaluation of climate predictions, we have also presented the skill with respect to a more strict reference than a linear trend: the forced signal (i.e. Figs 1,2,3,6,7,8, panels b and e). The forced signal not only represents a trend that is non-linear, but also the signal stemming from external forcing of unpredictable events such as volcanic eruptions, hence making this a higher benchmark to compare to than a linear trend.

L160: Is it fair to compare ensembles of different sizes? There is plenty of literature on this point and all scores should either be calculated for the same ensemble size or corrected for ensemble size to make them equivalent. Even if large ensembles of analog forecasts are easy to generate, this is important for the comparison and understanding the relative merits of the methods.

We thank the reviewer for pointing this out. We agree that skill is sensitive to ensemble size (See Fig. S1 or Figures below). However, while the initialized seasonal and decadal prediction systems used as benchmarks require substantial computing cost for larger ensembles, the analog predictions can provide large ensemble sizes at no additional cost. Some skill advantage indeed results from the use of larger ensemble size. Just as the examples found in literature where a fair comparison is made between ensembles of different sizes, many others exist in which a larger multi-model ensemble is compared to the single system components (e.g. Hagedorn et al., 2005). Many times the advantage of multi-model ensembles is indeed related to their larger size. We have expanded a bit on this and clarified this in the manuscript in the lines 168-173: "Note that both dynamical prediction systems are limited to 25 members, whereas the analog-based predictions are based on the 149 members from the non-initialized CMIP6 ensemble. A key strength of the analog-based method is its ability to leverage a large-sized ensemble at minimal computational cost as opposed to the significant cost it requires to generate such large ensembles with initialized prediction systems. However, we acknowledge that a fraction of the skill of the analog-based predictions stems from exploiting large ensembles and reducing the ensemble size to match the size of the dynamical prediction systems reduces the skill. This is demonstrated in Fig. S1 which shows that the skill of the analog-based predictions clearly increases with ensemble size, regardless of variable or forecast range. ~~(Fig. S1)~~. ".

To further illustrate here the advantage of a large ensemble, the figures below show that indeed a 25-member analog-based prediction is generally less skillful than the 149 ensemble using the analog method, confirming the what Fig. S1 shows.

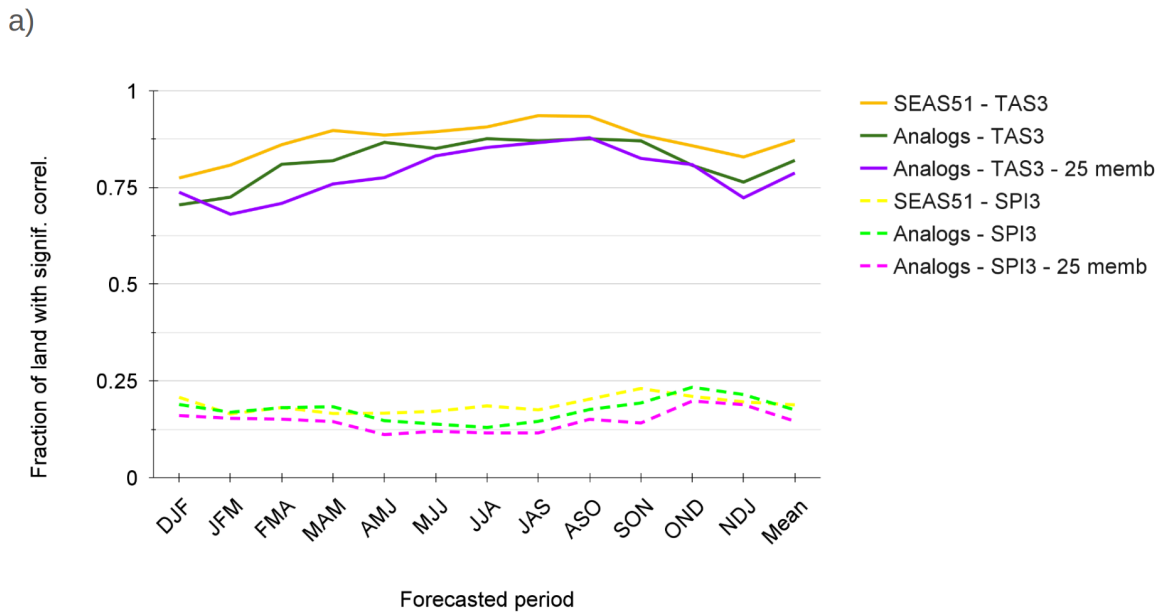

Figure R1 (with 25-member analog ensemble): Fraction of land area with statistically significant positive correlation ($p < 0.1$) between the 3-month TAS (solid lines) and SPI3 (dashed lines) from the analog, a 25-member analog and SEAS51 predictions, and the respective observations. Statistical significance is assessed using a two-tailed t-test. The evaluation period is 1982-2018.

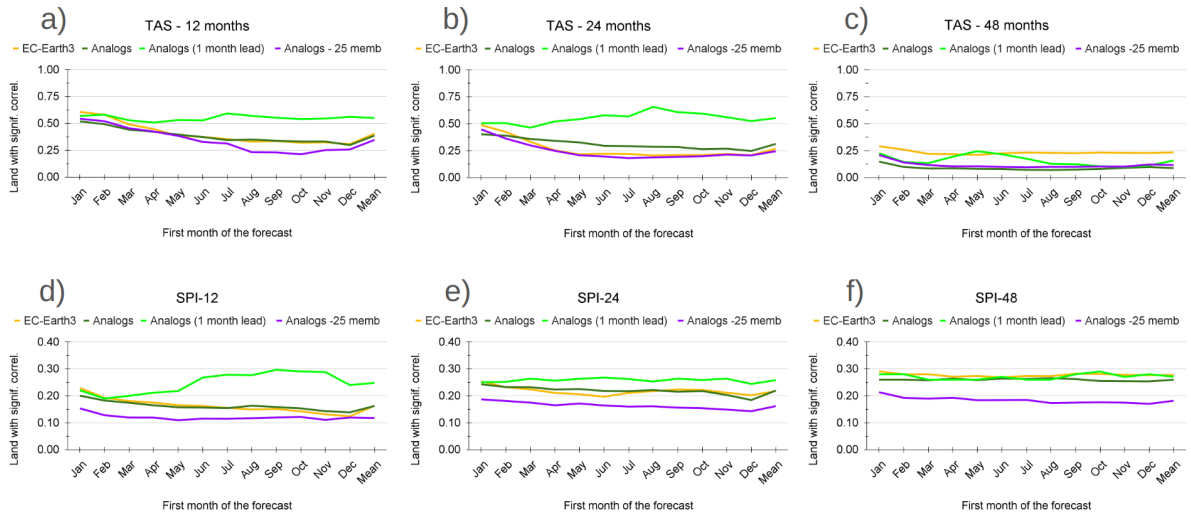

Figure R2 (with 25-member analog ensemble): Fraction of global land area with statistically significant ($p<0.1$) positive residual correlation between TAS predictions and observations for a) 12-month, b) 24-month and c) 48-month forecasts. Panels d), e) and f) are the same as a), b) and c), respectively, but for the statistically significant ($p<0.1$) positive correlation between SPI predictions and observations using a two-sided t-test. The dark green, purple and yellow lines in all panels show the skill of the analog-based, 25-member analog-based

and EC-Earth3 predictions, respectively, initialized every November as the lead-time increases from 2 to 13 months. The light green line shows the skill of the analog-based predictions, initialized always with a one-month lead-time.

Fig. 4 has been modified from its original version to show the area with significant and positive skill for both TAS and SPI3, instead of the area with residual skill. We think that this comparison is more fair and since the skill does not saturate as it happens for longer forecast ranges, makes it a meaningful comparison. The conclusions remain qualitatively the same.

Note also that in response to the comments related to the trend removal, Fig. 11 has been slightly modified to account for the removal of trends using their own model trend. For example, EC-Earth3 residual skill is estimated by removing only the forced signal from EC-Earth3 uninitialized simulations. The conclusions also remain qualitatively the same.

Ref: Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus A: Dynamic Meteorology and Oceanography, 57(3), 219-233.

L170-175, Fig2 and 3, L375: While I am sure readers will be open-minded to this method of forecasting this passage feels somewhat biased in favour of the analog method. The dynamical seasonal forecasts have a better correlation. This discussion needs to be rephrased and a panel of the difference in correlation scores is also needed, perhaps in place of the current panel 1b and panel 3b.

The figure below (Fig. R3) displays the difference in ACC between the analog-based predictions and SEAS51 for DJF and JJA forecasts of TAS and SPI3. Note that panel f in Figs. 1-2 already displays the direct comparison between the analog-based and the SEAS51 predictions with a dedicated discussion and it is comparable with the differences in ACC shown below. We therefore feel that showing the differences in correlation is not necessary. We also think that panels b in Figs. 1-3 are necessary as they show the skill of the forecasts after removing the forced signal. To better address the point of Reviewer 1 and avoid any bias in favor of the analog method we have removed the following sentence in L174-175 "*It is worth noting that skill over land of the analog-based predictions is in general statistically not different ($p<0.1$) between the analog-based and SEAS51 predictions as quantified by MAESS.*", and additionally changed sentence L375 as: "~~The analog-based predictions demonstrate skill on seasonal to multi-annual time scales, in many cases comparable to state-of-the-art numerical prediction systems developed for either seasonal or decadal climate predictions.~~ The analog-based predictions provide skilful forecasts on the seasonal to multi-annual time scales and show in general similar spatial patterns of skill to initialized numerical predictions. Furthermore, the analog-based predictions are competitive with existing annual and multi-annual predictions from initialized numerical predictions."

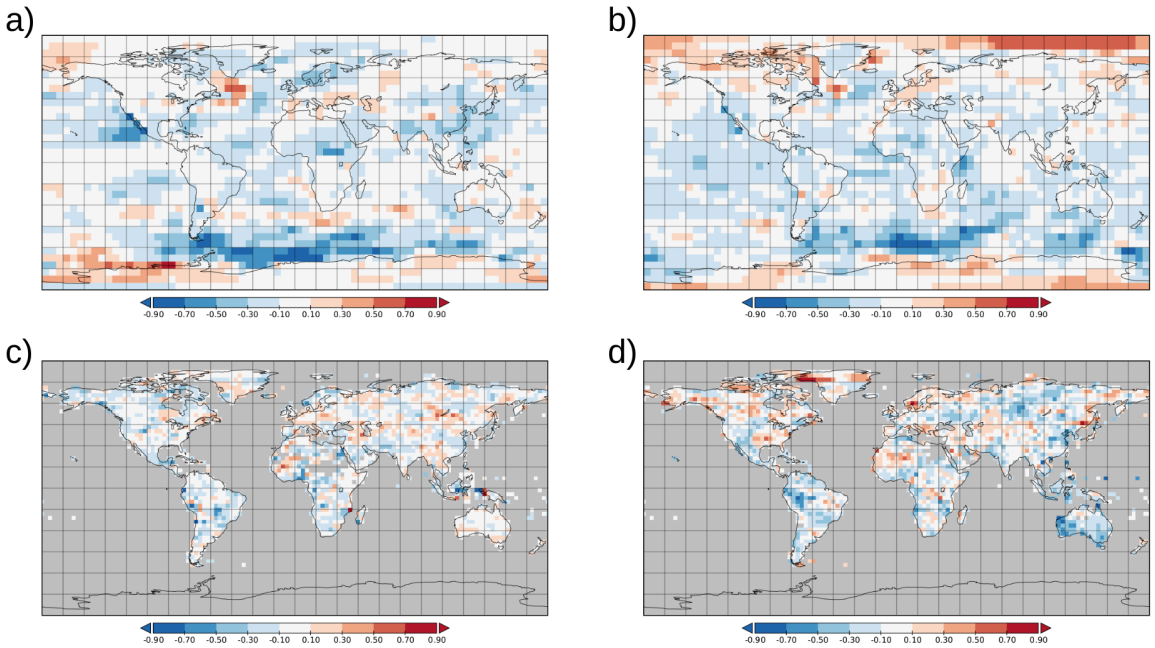

Figure R3: Difference in TAS skill (anomaly correlation coefficients) between analog-based ensemble mean predictions and SEAS51 in a) December-February and b) June-August. Panels c) and d), show the same as a) and b), respectively, but for SPI3.

Fig.4 and Fig.11: I think it is important that these metrics are changed to the *average correlation skill over land where it is significant*, rather than just the *area that is significant* because the current metric does not reflect the higher skill of SEAS5 in many regions and this is important for the value of the forecasts.

We disagree with the suggested method for displaying a summary of skill. Showing the *average of correlation coefficients only where it is significant* could yield biased results, but more importantly averaging correlation coefficients is not mathematically valid. For example, having a prediction with very little skill and only a few locations with high, but statistically significant correlation would give a high skill metric, whereas a prediction with widespread statistically significant correlation, but with low values would give a much lower metric, suggesting that is worse than the former prediction. In any case, we have recomputed skill in Figs. 4 and 11 (Figs. R4-R5 below) with the method suggested by the reviewer and see that both methods for computing the average skill yield qualitatively similar results. For these reasons we have decided to keep the original Figs. 4 and 11.

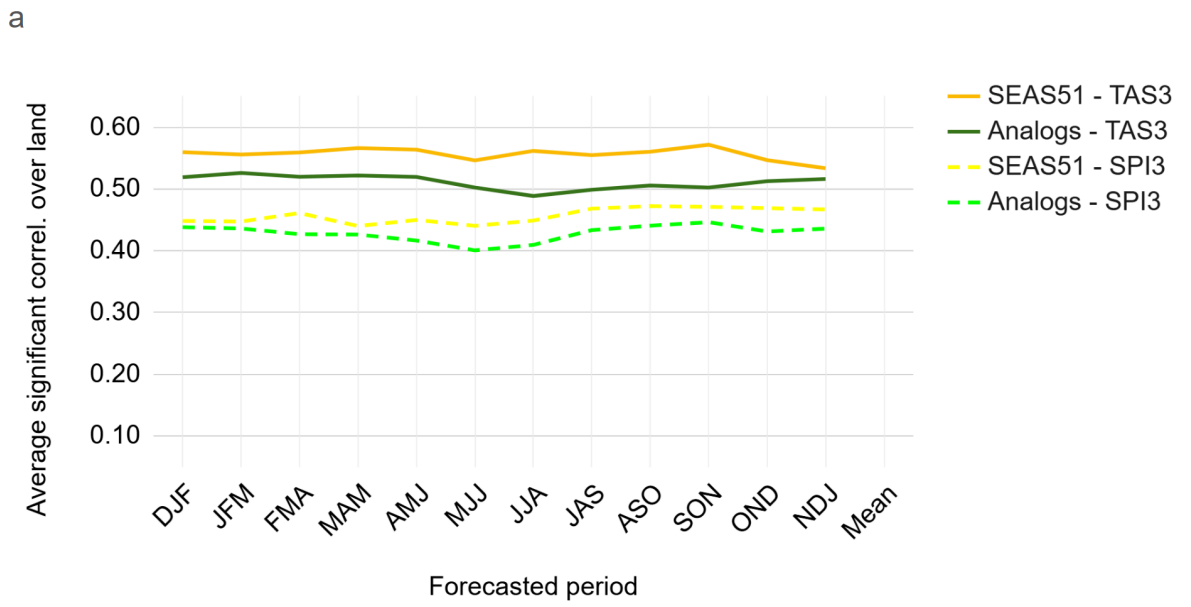

Figure R4: Spatial average of skill over land that is positive and statistically significant (p<0.1), measured with the correlation of the 3-month TAS (solid lines), the SPI3 (dashed lines) from the analogs, and SEAS51 predictions, and the respective observations using a two-tailed t-test. The evaluation period is 1982-2018. The displayed values are based on residual correlations for TAS to remove the impact from the forced signal.

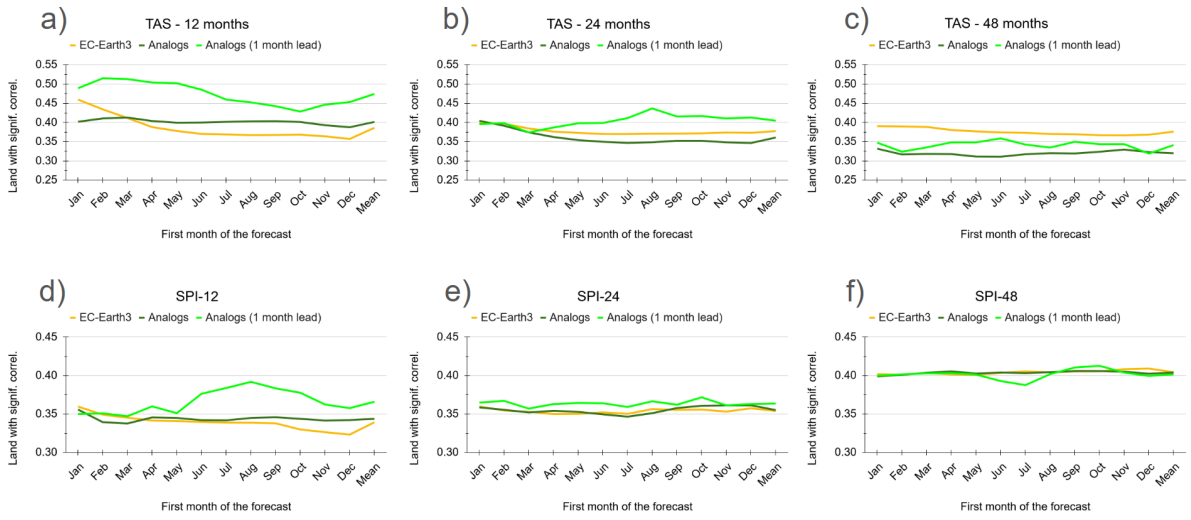

Figure R5: Spatial average of skill over land that is positive and statistically significant (p<0.1), measured with the residual correlation between TAS predictions and observations for a) 12-month, b) 24-month and c) 48-month forecasts. Panels d), e) and f) are the same as a), b) and c), respectively, but for the statistically significant (p<0.1) positive correlation between SPI predictions and observations using a two-sided t-test. The dark green and yellow lines in all panels show the skill of the analog-based and EC-Earth3 predictions,

respectively, initialized every November as the lead-time increases from 2 to 13 months. The light green line shows the skill of the analog-based predictions, initialized always with a one month lead-time.

L250: in fact all the indices are of weak amplitude (even Nino3.4) so this needs to be stated with some comments about the ability to recalibrate the amplitude.

This is a good point, we have included the following sentence in L250 to acknowledge this point: "This implies that a recalibration of the ensemble could render the analog-based forecasts more valuable."

Fig.6: The analogs are clearly more competitive on this longer timescale and the striking similarity with the dynamical model is impressive, at least with EC-EARTH. However, I am not convinced EC EARTH is the best decadal prediction system. Does this result hold for other models? Either way, I think the abstract should reflect the benefit of analogs may be greater for the longer timescales.

The scope of the study was to show that the method has a comparable skill to an operational forecasting system. Whether or not EC-Earth is the best decadal forecast system is arguable, but from the skill analysis presented by the WMO lead center on decadal predictions (https://hadleyserver.metoffice.gov.uk/wmolc/), EC-Earth (BSC and SMHI/DMI) is ranked in the top half of models for both temperature and precipitation and forecasts of 1 and 5 years (Figures R6-R7 below). Additionally, the analog method is better than any single system for TAS forecasts of year 1 and years 1-5, and almost as high as the multi-system ensemble from all the producing centers. For precipitation (SPI12 and SPI60) the analog method is better than most single systems, but not as high as the multi-system ensemble (See figures below). We added the following sentence in lines in the summary and conclusions section "We have chosen EC-Earth3 as representative model of the typical decadal prediction system. It is possible that other decadal prediction systems perform better in particular regions and timescales, but EC-Earth3 forecasts quality metrics reveal it to be a good representative of these systems. For a thorough evaluation of several decadal prediction systems including EC-Earth, the reader is referred to Figures S7-S11 in Delgado-Torres et al., (2022). "

New reference:

Delgado-Torres, C., Donat, M. G., Gonzalez-Reviriego, N., Caron, L., Athanasiadis, P. J., Bretonnière, P., Dunstone, N. J., Ho, A., Nicoli, D., Pankatz, K., Paxian, A., Pérez-Zanón, N., Cabré, M. S., Solaraju-Murali, B., Soret, A., & Doblas-Reyes, F. J.: Multi-Model Forecast Quality Assessment of CMIP6 Decadal Predictions. *Journal of Climate*, *35*(13), 4363-4382, https://doi.org/10.1175/JCLI-D-21-0811.1, 2022.

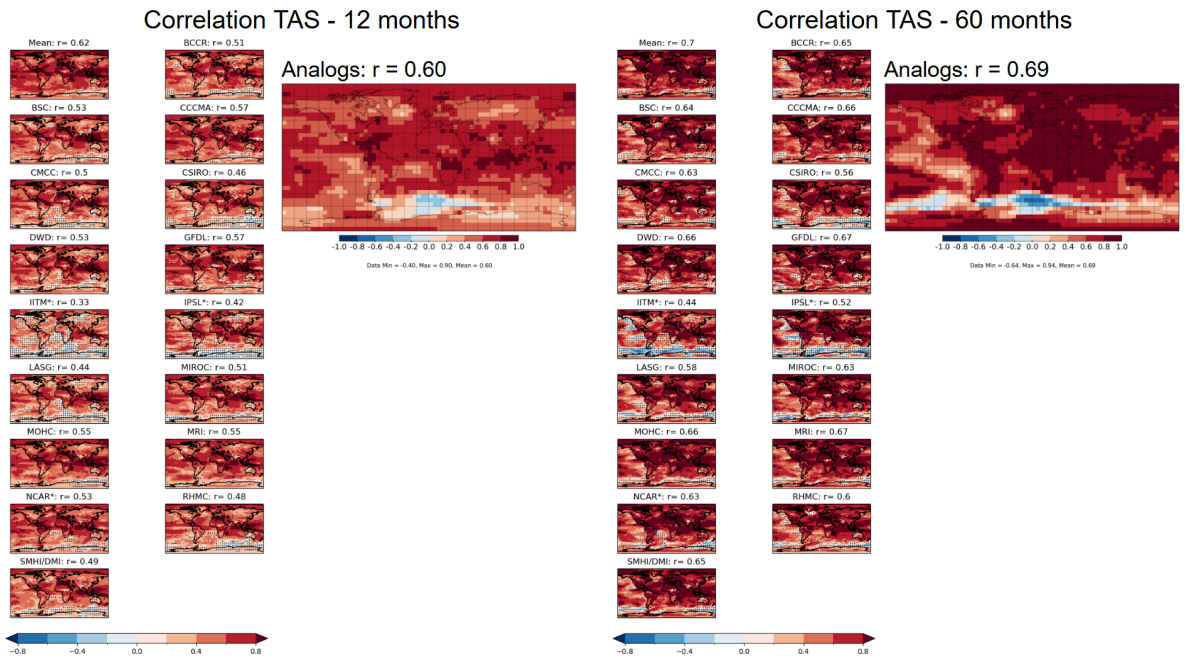

Figure R6: Anomaly correlation between different decadal prediction systems and observations contributing to the WMO decadal predictions for 12-month (left) and 60-month TAS predictions (right). The decadal multi-model and the 149-member analog-based skill maps are shown on the top left and right, respectively.

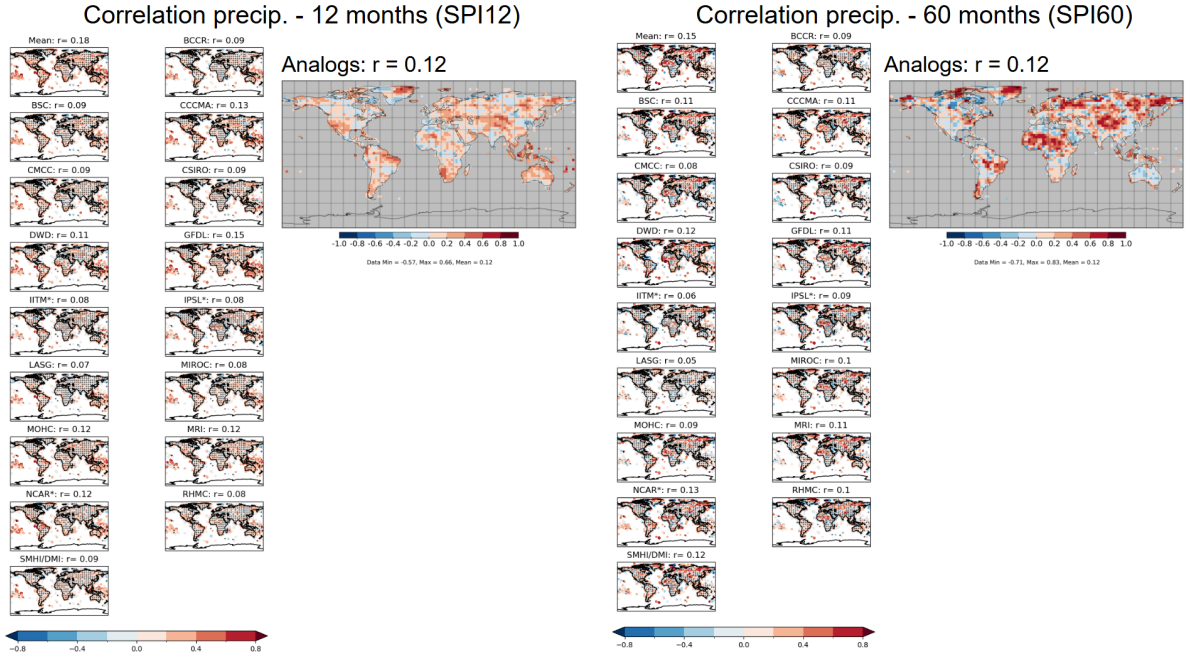

Figure R7: The same as Fig. R5, but for precipitation (decadal prediction systems) and SPI12 and 60, for the analog-based predictions.

Fig.8e: Presumably this result comes from the fact that the analogs can be selected from any year? Does it improve if the analogues have to be selected e.g. from the same decade as the target? Or is this already accounted for by the removal and replacement of the forced trend?

Selecting the analogs from the same decade as the target would likely improve the trend but it also would reduce reduce the pool of analogs to select from, for example if the observations show a strong positive ENSO state, having only 10 years in the model would largely limit the available states and very possibly the part of the skill that is not externally forced. The figure below shows the metric for analog selection (left) and the years of selection of the analogs for each member based on that metric for an example prediction around the center of the period (year 2000) (See Fig R8). It is indeed accounted for by the removal and replacement of the forced trend (See answer to 2nd point above).

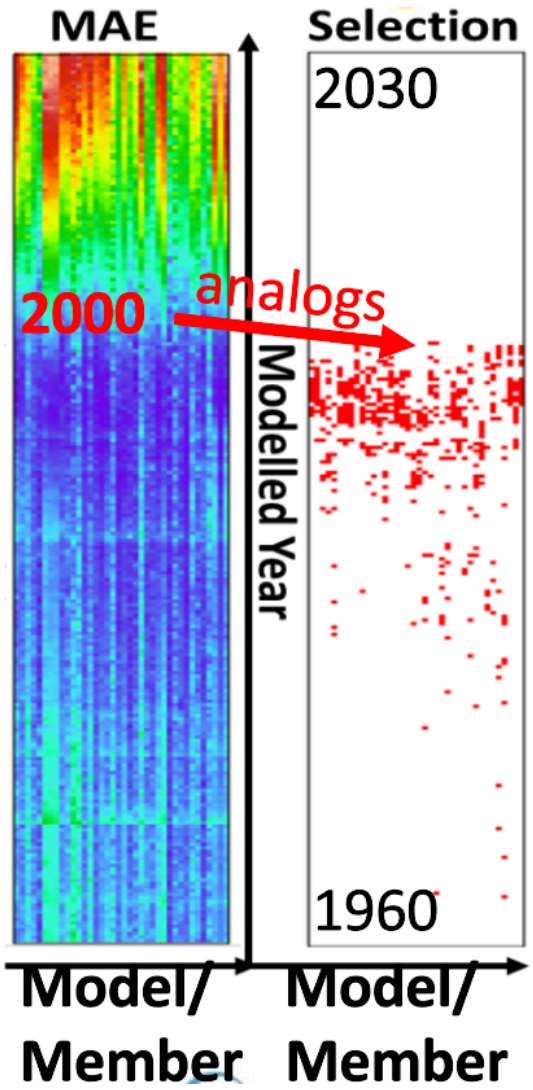

Figure R8: Example of the construction of an analog forecast for year 2000. Left) Globally-averaged mean absolute error of monthly model SSTs (vs. observations) as a function of ensemble member (x-axis) and model year (y-axis). Right) Red dots indicate the

year of the 5 analogs selected for each member used to construct the analogs. Lowest values in purple and highest values in red.

MINOR

L52: 'is meant to constitute a pool…' of course it does not always achieve this

Thank you, suggestion added.

L55: the number is not very small as it is now over 10 on subseasonal, seasonal and decadal scales. See for example Kumar et al, 2024, BAMS. Suggest to say "limited number"

The sentence L55 was changed to: "thus being produced only by a limited number of~~few~~ institutions around the world."

L64: '…of a more sophisticated'

Suggestion added.

L64: it is stated earlier that models drift to their own climatology and that this reduces skill. However L64 states that the analog method is not subject to drift because the model is in its own climate. This seems very one sided in favour of the analog approach and so it needs to be rephrased.

In reality the initialization shock is the largest concern, for that reason the word drift has been deleted from the sentence.

L70 Kushnir et al., 2019, Nat. C.C. is an important missing reference on the operationalisation of decadal predictions.

Included.

L80-85: please state the total sample size (in years), is it really greater than the decadal hindcast size?

Added. 10579 years in total.

L104: constraint

Changed.

L215: there is a long literature on Sahel forecasts so please add some references here.

References added.

Fig.10: please reduce the vertical scale to better show the variability.

Figure 10 has been changed accordingly.

L340: Smith et al 2018 specifically examined the ability of GCMs to predict global temperature: Smith et al, 2018. Predicted chance that global warming will temporarily exceed 1.5C. Geophys. Res. Lett.

We don’t see how this reference fits well in line 340, for this reason it has been omitted.