Review of manuscript GMD-2021-131 entitled "Systematic global evaluation of accuracy of seasonal climate forecasts for monthly precipitation of JMA/MRI-CPS2 by comparing with a statistical system using climate indices" by Yuji Masutomi et al.

OVERALL RECOMMENDATION

Reject

SUMMARY

This study proposes to compare the performances in monthly precipitation prediction between a newly-released dynamical seasonal forecasting system and a statistical model based on 18 climate indices. The analysis is carried out at the global scale for 12 initialization dates (at the end of each month) over a 30-year reforecast period (1981-2010). Lead months 1 to 6 are considered. The authors conclude that the dynamical system is more accurate for month-1 lead time but is superseded by the statistical model from month-2 onward (except in the 20°S-20°N equatorial region).

Note: There is trouble with the line numbering of the manuscript. The line counter is reset to 0 at section 2.2. Here, I refer to the line numbers as they appear, although they are erroneous.

MAJOR COMMENTS

The idea of using a statistical model as a benchmark is quite relevant for the evaluation of dynamical seasonal forecasts, and the concept of this study could have led to a valuable contribution to this field. However, the manuscript suffers from a lack of clarity and substantial flaws, so it fails to fulfill the promises it bears in the abstract. I encourage the authors to carry on this study and re-submit a new, enriched version, but I think these modifications are beyond a major revision. This why my recommendation is to reject the current manuscript.

Here are my major concerns:

1) The test that is used to define a significant ACC at the 95% level should be named and described, as it is a key component of the results.

2) The definitions of the "Ratio of sig." and "Ratio of higher ACC" indicators that are represented in Figures 5 and 8 are not clear at all, while these indicators play a central role in the interpretation of the results. Therefore, they should be explained in more details in Section 2.1 or in Section 3.1 (l. 40-47).

3) The construction of the statistical forecasts is very unclear too:

a) I do not understand why there is a separation between Step 2-1 and Step 2-2. From what I understand, leave-one-out cross-validation must be applied in the model fitting from the outset, otherwise it seems the statistical model is fit by including unknown data to be predicted.
b) I do not understand either how a single statistical forecast is obtained from the 18 statistical models (Step 2-3). I know it is the purpose of lines 84-88 (Section 2.1) to explain it, but they are actually very confusing.

4) The abstract claims that the statistical model can be used to diagnose slow dynamics that are not well reproduced by the dynamical model. This would be the most important contribution of this article, but the authors do not address it extensively while they have 18 climate indices available.
I guess Section 3.3 and Figure 9 are meant to illustrate this point, but they fail to convince. Indeed, to my understanding, the figure is purely observational and there is no analysis of the model behavior relative to the relationship between Nino 3.4 and precipitation in Paris. Then, I cannot see how it is possible to conclude "the slow dynamics (...) are not reproduced in JMA/MRI-CPS2" (l. 79-80).

5) Section 4 (Discussion) should be thoroughly re-organized, re-written and possibly merged with Section 5 (Conclusion). In its present form, I feel it only rephrases Section 3 and does not bring any additional insight.

MINOR COMMENTS

l. 19-22: The sentences should be switched: mention the comparison at the global scale first, before going into details about the 10°S-10°N equatorial band.

l. 30-31: "which is implemented in most SCF systems". Unnecessary, I suggest removing.

l. 37-38 vs l. 91-92:
"The mean square skill score is often used to evaluate the forecast accuracy of SCF systems" (l. 37-38)
"The anomaly correlation coefficient (ACC) between the forecast and observed values was used to evaluate the forecast accuracy" (l. 91-92).
It is strange that the MSSS is mentioned in the introduction while the whole assessment of accuracy is based on the ACC. I suggest mentioning the ACC from the outset, while removing the MSSS.

l. 34-37: "For dynamical SCF (…) with a smaller cost." This sentence is quite long and intricate, I suggest splitting and/or rephrasing for the sake of clarity.

l. 48: "**and the** Madden-Julian Oscillation"

l. 50: I would rather say “The predictability **in** St-SCFs” rather than “The predictability of St-SCFs”.

l. 76-79: For the sake of clarity, I suggest trimming and rephrasing the sentences.

l. 89: “the forecast accuracy of JMA/MRI-CPS2 ~~forecasts~~” Avoid repetition

l. 91: The “ACC” term is ambiguous, as it has different meanings across various studies. It might as well designate the correlation of spatial patterns or the temporal correlation between time series. From the results in the manuscript, I assume that it corresponds to a temporal correlation. Then I suggest using another expression.

l.93: “A significance level of 0.05 was used to evaluate statistical significance of ACC” Please mention the significance test here (see Major comment #1).

l. 93-94: “Forecasts with 1 to 6 lead months were evaluated”. Note that there might a conflicting naming convention of lead times with other works on seasonal forecasting. For instance, in the operational Copernicus C3S seasonal forecasts (https://climate.copernicus.eu/seasonal-forecasts), if we consider forecasts initialized on September 1st, the month of September is lead time 0-month, while October is lead time 1-month. Although it is of minor importance, I am unsure what you designate by month-1, month-2, etc.

l. 109-110: “The hindcast data included five ensembles with different initial conditions (...). There were two forecasts starting in the middle and end of each month.”
Something is unclear about the forecasting system setup: do you mean your ensemble forecast is a lagged-ensemble with two forecasts initialized in the middle of the month and two forecasts at the end? If so, where does the fifth member come from? And if not, do you have 5 members launched in burst mode at the end of the month (e.g September 28)?
→ Suggestion: The last two remarks (l. 93-94 and l. 109-110) could be clarified with a simple diagram for a representative start date.

l. 65, 67: “**above** 20°**N** and **below** 20°**S**”

Figure 5, caption:
“ratio of higher ACC with significant between JMA/MRI-CPS2 and St-SCFs”
I do not understand the sentence, some words must be missing or jumbled up.