Summary

In this study, the authors evaluate the accuracy and hydrologic utility of subseasonal to seasonal (S2S) forecasts generated from the S2S project. The paper is well written and provides a good overview of the performance of the current state-of-the-art S2S forecasts.

However, the paper, as it stands, is lacking in detail, and would especially benefit from better discussion of results.

Major Comments

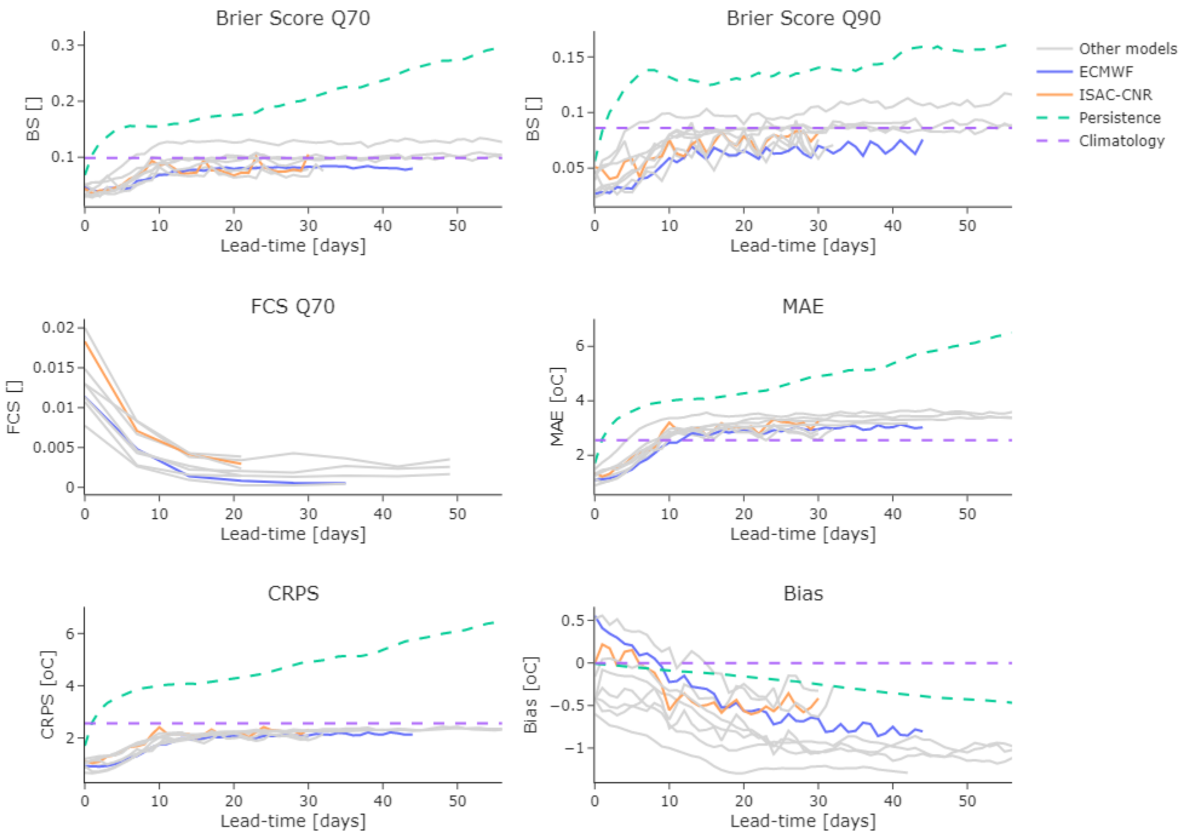
1. **R1: Why were these two study regions selected? The total area is 4100 km2. However, the model grid size is 1.5 x 1.5 deg (~20,000km2) but the authors mention that the study regions encompass 4 grid points. I am not sure how this number was arrived at.**

Authors: The two catchments were selected according to existing models and data, and previous experience of the authors in the areas. Regarding the model grid, the Upper Main catchment spans four grid points, but only partially. The averaging for each parameter takes this into consideration. We have altered the text to clarify this.

Line 72: "… the whole catchment spans four grid points of the S2S raster grid (see the grid delineation in Figure 1)."

2. **R1: I do not fully understand the reason behind not showing the performance of temperature. As NWP modules use different physics for simulating precipitation and temperature, similarity in performance metrics should not preclude the inclusion and discussion of the temperature results as they pertain to different aspects of the NWP model.**

Authors: We have added the temperature results (Figure 3) and discussed them in the text.



Line 195: "Figure 3 presents the same result metrics for temperature. The results for both variables are generally alike, albeit S2S models had constantly better absolute performance for temperature than precipitation, as temperature is generally easier to forecast. Notably, the baseline models are also relatively better at forecasting temperature events, therefore the S2S does not have significantly higher skill for temperature than precipitation. Indeed, skill of temperature predictions in relation to precipitation is highly dependent on the model and inconsistent over the dataset; for each model, this is expected to vary depending on the parameter. Nonetheless, the general yield time of 10 days also holds for temperature."

Line 201: "It is important to note that, apart for ECMWF and ISAC-CNR, most models showed significant bias in predicting both precipitation and temperature, though a positive bias for precipitation was found, while temperatures had to a tendency to be under forecasted."

3. **R1: Why does the performance improve after the yield point? This is very surprising and the authors should provide some explanation as to why performance improves with increase in lead times.**

Authors: Performance after the yield time indeed does not increase, it remains approximately stable, as stated in Chapter 4.1 and 4.2. Noise exists partially because of the different issuance frequency of the S2S models, and the yield point is rather difficult to pinpoint for a few scores (e.g. Bias). Regarding performance improvement over time, it also remains approximately constant after the yield point, as shown in Figure 5.

Line 214: “Interestingly, BS of NCEP for Q70 seems to not stabilize, most likely because it’s yield point is above 40 days, which is also seen for two other models with longer maximum lead-times. Persistence and Simple Persistence stabilize at around 50 days and BS scores around 0.2 (not shown in Figure due to scaling).”

4. **R1: The manuscript does not discuss the results comprehensively. What is the impact of ensemble member size on forecast performance? Why do ECMWF models perform better? Have other studies found out the same?**

Authors: We have extended the results chapter thanks to your and Reviewer 2’s suggestions. Besides the temperature paragraph (as answered in your second point), we have added comments on the model factor significance and quality evolution over time, along the minor changes requested. I add here those significant improvements:

Line 253: “Though there are other factors to consider, such as the model physics and the region of interest, we can infer that a model with a large ensemble size and high resolution should perform well. Indeed, ECMWF has the biggest number of ensembles and the finest original resolution. Other studies show similar findings: Phakula et al. (2020) found that ECMWF is better at predicting minimum and maximum temperatures than CNRM and UKMO in South Africa; Guimarães et al. (2021) concluded that ECMWF forecasts precipitation anomalies in Brazil better than other S2S models; Deoras et al. (2021) reported that ECMWF has the best ensemble spread-error relationship among all S2S models when predicting Indian monsoon low pressure systems. On the other hand, ISAC-CNR has a higher original resolution and less ensembles (though still more than half of the other models), but its score might be influenced by the small maximum lead-time. As previously discussed, Li et al. (2019) found that KMA and UKMO fared better than ISAC-CNR in south China, which means the good performance of ISAC-CNR might be strongly dependent on the region investigated.”

Line 281: “By expanding the evaluation range, the current work found that gains in hydrological performance continue from 14 days to 30 days. It is important to note, however, that the 30 days value is somewhat artificial due to the interpolation methodology and the low averaged BSS scores in 2015. The initial and final BSS curves run parallel from lead-time 14 days on, which means the interpolated final BSS is bigger than any initial BSS, and the computed improvement is consequently constant. Note that computed gain points are very noisy, but do go below the initial BSS. One may thus take conservatively the 14 days as total gain for the system over the 5 years.

Another important point is that due to data availability constraints, the 5 years evaluated are not enough to completely capture the region climatology. This means the improvement seen may be artificially increased by an easy to forecast weather in the more recent years, thus more data would be needed to confirm these findings. Still, the increase in performance over time may be explained by improvements in the model resolution, increases in the number of ensemble members, and changes in parametrization schemes. For example, in March 2017, JMA has increased their model ensemble size from 24 to 49 members; in March 2016 ECMWF doubled their grid resolution; and in July 2019 ECCC upgraded their parameter perturbation methodology. For a complete list of model changes, please see ECMWF (2022). Indeed, we expect that investments in models should result in increased model performance for the meteorological variables produced.”

Minor Comments

1. **R1: Apart from being a project, S2S is generally used to refer to a specific forecast horizon in the forecasting community. I request the authors to explicitly mention that they are referring to the project.**

Authors: Thank you pointing this out. We have clarified this in the abstract and in the introduction:

Abstract: "Recently, projects such as the Sub-seasonal to Seasonal Prediction Project (S2S)…"

Line 25: "In that regard, recently projects such as the Sub-seasonal to Seasonal Prediction Project (S2S) …"

Moreover, we added an acknowledgement to the project:

L321: "**Acknowledgements**

This work is based on S2S data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). The original S2S database is hosted at ECMWF as an extension of the TIGGE database."

2. **R1: Abstract: 'Results show that the S2S models have skill at the catchment-scale, particularly for lower threshold levels …'. What does 'lower threshold levels' mean here?**

Authors: We have clarified this point in the abstract:

Line 12: "Results show that the S2S models have skill at the catchment-scale, particularly for less extreme parameter thresholds such as Q70 (30% percentile) …"

3. **R1: Line 30: What is 'hydrological horizon of skillful predictability'?**

Authors: We have changed our wording to improve clarity.

Line 30: "…in the hydrological horizon of skilful predictability (i.e., the maximum lead-time when forecasts are skilful) are…"

4. **R1: ECMWF models are referred to as 'ecmf' and ISAC-CNR as 'isac' in the figures. Please maintain consistency.**

Authors: Those are the model codes, but they are indeed less clear than using the actual model names. We have thus changed the figure legends to use the model names.