Final Authors Comments to Reviewer 2

Comment on egusphere-2022-449
Anonymous Referee #2

Referee comment on "Seasonal forecasting skill for the High Mountain Asia region in the Goddard Earth Observing System" by Elias Charbel Massoud et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-449-RC2, 2022

Summary

This study evaluates the subseasonal predictions from the NASA GEOS5-S2S hindcasts for 1981-2016 over the High Mountain Asia (HMA) domain with a focus on a set of hydrometeorological variables including 2-m air temperature, precipitation, snow cover fraction, snow water equivalent, soil moisture, and total water storage. The evaluation was done against two reanalyses and other independent datasets, and the evaluation focuses on monthly time scale with lead times up to 3 months. Unbiased root mean square error and anomaly correlation are the major metrics used in this evaluation. Overall, the study provides useful information about the predictive skill of the NASA GEOS5-S2S hindcast over the HMA region. The manuscript is well written and easy to understand, and the quality of the visualizations is generally good. However, the study falls short in several important aspect regarding forecast verification at subseasonal to seasonal time scales. The value and the contribution of this study to our understanding about predictability of the climate system at S2S time scale is very limited. I believe at the minimum a major revision is needed. I list my major concerns and some specific comments below.

Author Comments: We very much thank the reviewer for the time spent on our manuscript. We believe that the revisions detailed below, including the addition of evaluation for specific sub-regions as well as the addition of new text in various parts of the manuscript, should address the reviewer's concerns.

Major issues

- For prediction beyond the typical weather scale (i.e, 1-2 weeks), probabilistic forecast is more appropriate and useful than deterministic forecast given the chaotic nature of the climate system, which is why S2S forecast with numerical models needs to produce ensemble predictions. In this study, only unbiased root mean square error (ubRMSE) and the anomaly correlation (ACC) of the ensemble mean were used, which is useful but only shows very limited aspects of the forecast quality. There are many verification metrics that can be used for ensemble predictions such as those listed at https://www.cawcr.gov.au/projects/verification/#Methods_for_probabilistic_forecasts. I'd highly recommend that a few more meaningful metrics are included in this study.

Author Comments: We thank the reviewer for this comment. We agree that the reliability or the uncertainty in the reported skill of the forecasts is important. For this statement, we point to the ensemble spreads shown in Figures 5-12, which show the spread and therefore the 'reliability' of the forecasts from the model. The higher the ensemble spread in these plots, the less certain the various ensemble members are for each climate variable and lead time. Furthermore, the use of multiple data sources in the evaluation also allows us a look at the uncertainty in our results. We will include a statement in the discussion to elaborate on the verification of our results and the interpretation of their uncertainty. Additionally, we will provide new figures showing the comparison between the ensemble spread and the error. Generally, one can compute the spread/error ratio with the goal of that being close to 1; if it is larger than 1 (more spread than error) this is considered "underconfident", and if it is less than 1 this is considered "overconfident" (Fortin et al., 2014). We will provide new plots and relevant references to provide additional and meaningful verification metrics.

- My biggest concern regarding the analysis is its over-simplified approach to deal with the spatial heterogeneity within the study domain. The study domain is quite large; more importantly, it is very heterogeneous with distinct climates and land surface characteristics including elevation, land cover type, etc. As shown in Figures 7-12, temperature, precipitation and other hydrometeorological variables and model's skill in predicting these quantities can vary drastically across the domain. Spatially averaging them across high mountain ranges, the Tibet Plateau, Taklamakan desert, and the Indian subcontinent does not make much sense, and evaluating the spatially averaged quantities is not very meaningful and insightful. It is not clear what these spatial averages physically mean and how verification at such a level can help us to understand the model deficiency in a meaningful way. Although Figure 7-12 highlight the spatial heteorogeneity, the evaluation is only limited to the ensemble mean, spread, and ubRMSE. I'd suggest that the authors divide the domain into multiple smaller regions that are more homogeneous or multiple watersheds where the spatial averages are more meaningful, and conduct the forecast verification of these regional quantities using multiple metrics (probabilistic and deterministic).

Author Comments: The reviewer makes a good point; different regions may have different forecast skill. Our results in Figures 5-12 show maps of the skill and the ensemble spreads for the whole domain, which can be used to analyze spatial differences in the results and to pinpoint specific regions with higher or lower skills in general. Figures such as these can be used to interpret the skill level for different regions within the domain, such as low- or high-elevation regions. For example, Figure 7 shows that there is a higher ensemble spread as well as a higher error for temperature in the India subregion compared to other regions in the domain. By including the evaluation information spatially, i.e., maps that show ensemble spread and ubRMSE of each grid cell within the whole domain, one can estimate how skillful and reliable the forecasts are for broader regions (such as skill and reliability patterns at the watershed scale) as well as regions that are more local (skill and reliability at individual grid cells). To help further satisfy this concern, we will provide new plots showing the evaluation metrics for smaller sub-regions within the domain and will provide more discussion about these results.

Minor issues

- line 13: "where water resource needs change depending on ..." although this sentence is correct, it could be a little confusing as either "needs" or "change" can be interpreted as the verb, thus resulting in different meanings.

Author Comments: We thank the reviewer for this comment. We will fix this in the paper.

- line 13: how is intensity of the hydrological cycle defined? It was not mentioned in the study.

Author Comments: We will make this more clear in the paper.

- line 30: "a range of factors", the predictability itself is also an important factor.

Author Comments: We thank the reviewer for this comment. We will add this in the paper.

- line 34: remove the comma before "where"

Author Comments: We thank the reviewer for this comment. We will fix this in the paper.

- line 35-36: This sentence reads a little awkward, please consider rephrase.

Author Comments: We thank the reviewer for this comment. We will change this in the paper.

- line 40: Part of the study domain is heavily populated, but the majority of HMA do not have much population, such as Tibet Plateau and dessert.

Author Comments: We thank the reviewer for this comment. We will edit this in the paper.

- line 43: The term "water tower" of the Earth have been used for many years among researchers in Asia, so some earlier literature needs to be cited here to be more appropriate.

Author Comments: We thank the reviewer for this comment. We will add to this in the paper.

- line 144: This is only over the real-time forecast period, isn't it? Please clarify that these 6 additional members are not available in the hindcast period and thus not used in the evaluation.

Author Comments: We thank the reviewer for this comment. We will clear this up in the paper.

- line 147: "a long period for forecast validation" "validation" and "verification" are

different terms although they are related. One can verify if a forecast is correct or wrong, but you cannot validate a forecast when the forecast is wrong. So it would be more appropriate to say "forecast verification" or "forecast evaluation" here.

Author Comments: We thank the reviewer for this comment. We will fix this in the paper.

- line 233: remove "in our evaluation" as it is redundant with "in this study" at the beginning of the sentence.

Author Comments: We thank the reviewer for this comment. We will remove this in the paper.

- line 234-235: Does this mean the dataset is heterogeneous in space and time? If that is the case, how does this affect the evaluation? Please explain.

Author Comments: The product itself is a blended version of multiple different products. So, although the different measurements that are synthesized have varying depths, the output product has a consistent 5cm depth. We will make this more clear in the paper.

- line 282: It would be useful to give the equation for R_anom too. Does this includes both space and time dimensions?

Author Comments: Yes, the equation should include both space and time dimensions. We will add this to the paper.

- line 291-292: Is the ensemble spread also lead-time dependent?

Author Comments: Yes, we show in Figures 5-12 that each lead time has a different ensemble spread map. Generally, the ensemble spread increased with lead time for most variables and in most regions, except for precipitation in the Indian Subcontinent (Figure 8F and 8J). We will make a note of this in the paper.

- Section 3.2: Since the evaluation metrics are based on anomalies, what purpose does this section serve in the paper?

Author Comments: The anomalies are created by subtracting the mean value for each variable in each respective month (e.g., mean January T2M subtracted from all the January T2M's, mean February T2M subtracted from all the February T2M's, etc.). Section 3.2 shows the annual cycle, which portrays how each variable changes through the course of the year.

- Figure 4 and others: Since the gridded model forecast is spatially averaged over the large domain with different masks for different variables, it would be useful to show the masks for these variables in Figure 1 so that readers know how the spatial average is calculated.

Author Comments: We thank the reviewer for this comment. These masks are shown in Figures 9 for fSCA, Figure 10 for SWE, and Figure 11 for SM. Grid cells that are white are the ones that are masked out. This is also mentioned in the captions of Figure 9-11. We will make this clearer in the paper.

- Section 3.3: This section is about the absolute error. Because of the seasonality discussed in the previous section, it is not surprising that errors are generally larger during the season when the absolute value of variable is also large. So it will be necessary and more informative to discuss the relative errors beyond the absolute error.

Author Comments: We thank the reviewer for this comment. Note, however, that the seasonality of the errors is not totally dependent on the absolute value of the variable. For example, Figures 5A and 6A show that error for T2M is higher in the winter months, when T2M is in fact lower. These types of figures are useful to show this seasonality in the forecast error for each variable and how they differ between the different evaluation products. Furthermore, the ubRMSE calculates the error based on the anomaly forecasts, and so in that sense the measure of error is a relative error since it is based on anomalies and not absolute values.

- Figure 6: For each panel,the y-axis should be set to the same range as that in the corresponding panel in Figure 5.

Author Comments: We thank the reviewer for this comment. This might be helpful if the figures were being used solely to compare the errors based on the evaluation products, but the main point here is to show the seasonality of the errors for each variable. Since there is a large range for some variables between Figures 5 and 6 (e.g., for TWS), we do not believe it would be useful to make the y-axis the same in these figures, because this would make it difficult to visually notice the seasonality in some plots. We will write a statement in the paper to explain the difference in the y-axis for Figures 5 and 6.

- Line 489-490: The results in this study do not seem to back up this statement.

Author Comments: We thank the reviewer for this comment. The results that support this statement are shown in Figures 2A and 2B and explained on Lines 483-492. However, to make the explanation not sound too speculative, we will change the wording of this sentence.

- Line 518-520: This statement is speculative. It would be more appropriate to provide justifications.

Author Comments: We thank the reviewer for this comment. The results that support this statement are shown in Figures 4E and 4F and explained in Section 3.2. As for the role of the monsoon, the statement might sound speculative, and we will change the wording in the next paper.

- Line 529: How is 4% cold bias calculated? Using different units such as Kelvin, Celsius

will certainly result in different percentage change? So a statement like this does not make much sense.

Author Comments: We thank the reviewer for this comment. These results are reported from a different paper (Hsu et al., 2021). We will edit these comments to remove any text that can be confusing.

- Line 603-604: This statement assumes that the ensemble spread of the forecast is informative. The assumption may or may not be true. Linking a smaller forecast spread with higher skill is unjustified and questionable.

Author Comments: We thank the reviewer for this comment. We mention that this 'might' be an indication of more reliability, which is true in many cases. However, to avoid any erroneous claims, we will edit this sentence to include the comments that the reviewer has brought up.

- Line 633-634: It is not clear how this study achieve this as it does not provide much insights that can guide model improvements.

Author Comments: We thank the reviewer for this comment. Although we believe that the results and corresponding conclusions are explained in the text (e.g., TWS having high errors compared to GRACE since the model does not have groundwater pumping whereas the GRACE signature includes this process), we will add these specific details in this section of the paper to avoid any confusion.