Response to reviewers

We are delighted to read that the reviewers appreciate the content and presentation of our study. We are thankful for their comments and suggestions. We will address them and include the corresponding modifications in an updated version of the manuscript.

First review

General assessment :

This manuscript investigates the main sources of temperature subseasonal forecast skill at the global scale. It evaluates the relationships between potential drivers of three kinds (climate, circulation and land surface) derived from multiple observational/reanalysis datasets, and subsequent temperature forecast errors of the ECMWF extended range reforecasts, at different seasons. Overall, climate drivers tend to prevail, but land surface drivers also greatly contribute to forecast errors. Circulation drivers seem less relevant although not everywhere. Finally, based on correlation strength and forecast error amplitude, the authors highlight regions where subseasonal forecast skill could be potentially improved across seasons.

The scope of this study is both original and of great interest for the S2S community. The underlying rationale is relatively simple, but relevant too, so that I consider it an asset here. I would also like to stress that the paper is well written, well articulated, clear and enjoyable to read. My one main concern is about the evaluation of forecast errors. I feel like the metric used is not well suited for such a study where a distinction is made between seasons (see details below). It may have a limited impact in terms of the Spearman correlations found, but probably not on those of section 3.4. I think the authors should consider either correcting this metric or justify its relevance in the light of my comment below.

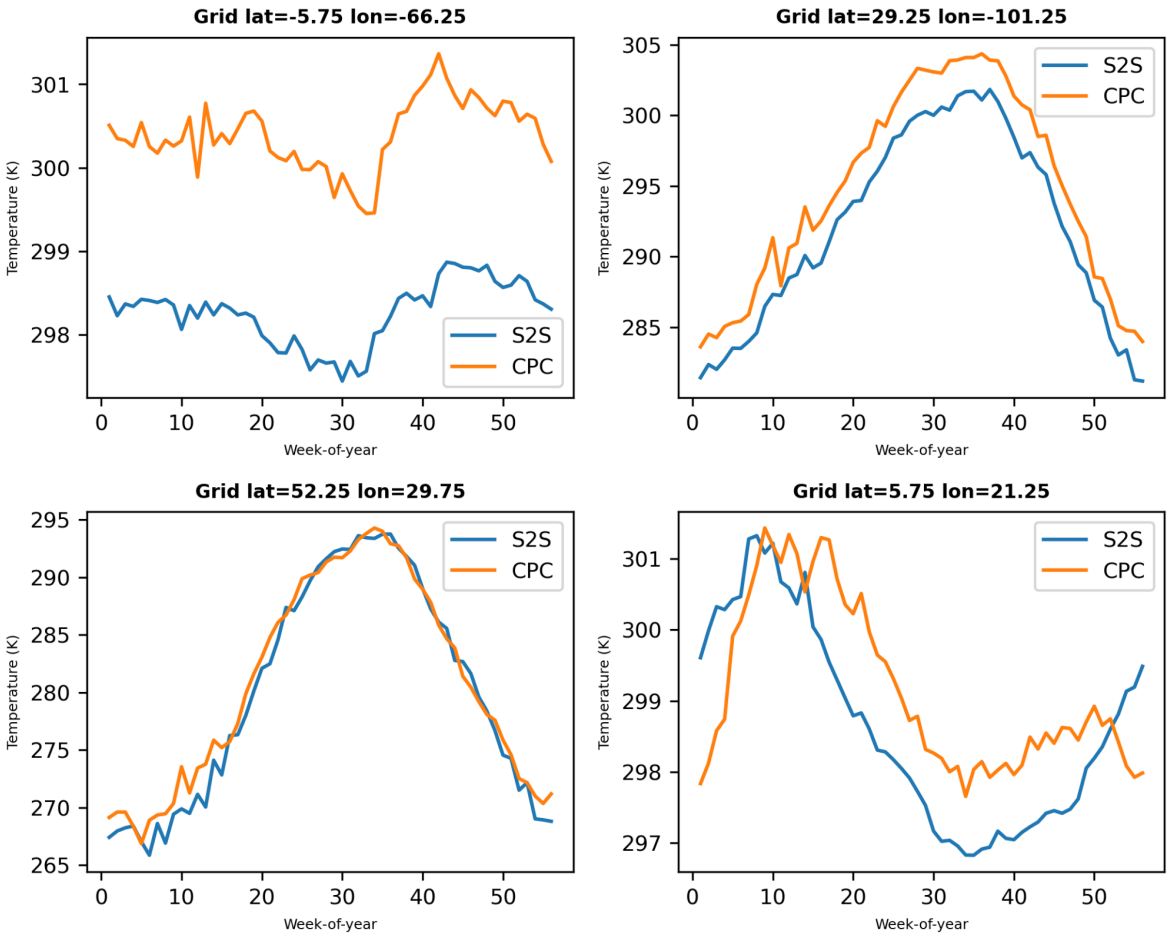
We appreciate the effort and time devoted by the reviewer in writing such constructive remarks. We think that the quality of our manuscript has significantly improved after addressing the reviewer's comments.

Main comment:

L. 150-151: I have the feeling that smaller errors found in transition seasons could also be due to the metric used here. This metric is based on departures from annual average temperature (i.e. no annual cycle). Therefore, for a given year, over mid and high latitudes at least, the annual average must be relatively close to fall and spring average temperatures, but substantially higher than winter and lower than summer average temperatures. Consequently, summer and winter forecast departures from annual average temperature must be generally higher (in absolute value) than fall and spring counterparts, both for forecasts and observations. Finally the amplitude range of the resulting forecast errors is probably larger as well. I am puzzled by this choice of annual average temperature to compute departures here. Why not use weekly (or at least monthly) climatologies based on

the 2001-2017 period instead? This would have avoided the issue of seasonality and that of changing the reference every year.

A1. We thank the reviewer for this relevant comment. We now introduce the modification of the forecast metric proposed by the reviewer: we use weekly climatologies (week-of-year averages) based on the 2001-2017 period instead of the annual averages. See in the image below the weekly climatologies for 4 example grid cells, for both reference and model datasets. Our main findings are not substantially affected by the adapted calculation of the forecast errors, maybe because the weekly climatologies between both datasets are similar.



Two more minor comments on this metric or its interpretation are reported below.

Minor points:

L.66-67: To what extent could the use of 2 model versions affect your results? I understand that initialization is unchanged, but readers not aware of the changes between these 2 model versions might wonder if they concern, say, a land/vegetation scheme or/and a key atmospheric parameterization for example. Such changes are prone to modify the relationships studied in this manuscript. If possible, I would suggest defending this particular point.

A2. According to the ECMWF model description (available here: https://confluence.ecmwf.int/display/S2S/ECMWF+model+description), there are no differences in the 2 model versions used in this study that can affect our results from the S2S dataset. The version that introduces differences with respect to the CY46R1 in the S2S dataset is the Cy47R2 which includes more model levels, but our computations are done before this version is implemented. We will add this explanation to the revised manuscript.

L.72: across S2S literature, the definition of leadtime weeks vary: some studies exclude days 1 to 4 after initialization, so that week 1 is defined as the day-5 to day-11 window (e.g. Vitart 2004, de Andrade et al. 2021). Could you specify - and discuss if need be - your method in this respect ?

A3. We use the first 42 days since forecast initialization in our analysis. As we compute weekly averages we use seven days for each week as follows:

- Week 1: day 1 to 7
- Week 2: day 8 to 14
- Week 3: day 15 to 21
- Week 4: day 22 to 28
- Week 5: day 29 to 35
- Week 6: day 36 to 42

We will add this explanation to the revised manuscript. Due to the variations in the literature about the definition of lead time weeks and that some studies exclude the first days after initialization, we focus on the week 3. Even though we do not describe in detail our results for every week after forecast initialization we show in Figure A5 in the manuscript how our results differ between each analysed week. As it can be seen from that figure, the results do not vary strongly with the lead time that we choose.

L.100: I am not sure how T_{for} (annual average) is computed. If I understand well, you have two 6-week forecasts per week, i.e. 104 forecasts x 6 weeks per year. Not to mention the ensemble members. How do you proceed to compute the forecast annual average temperature and ensure it is comparable with observational average (one realization, by essence)? I would recommend to be more specific, and also to state somewhere in the manuscript that forecasts are actually ensembles, and how these ensembles are handled here. I guess you have been dealing with ensemble means, but this needs to be specified somewhere.

A4. As mentioned before in the response to the main comment, now we use a modified forecast error metric based on week-of-year averages instead of the annual averages. Also, we only use one forecast per week, instead of two. We compute these week-of-year averages as follows:

1. We bin all the daily data according to the week lead time they belong to from week 1 to week 6. The next steps are computed independently for every week lead time
2. We compute weekly averages from the daily data
3. We compute the multiannual average of the same week-of-year

About the comment of the ensembles, we use the S2S reforecasts produced only by the ECMWF. These are global ensembles that simulate *i*) initial uncertainties using singular vectors and ensembles of data assimilation and *ii*) model uncertainties due to physical parameterizations using a stochastic scheme. The ensemble is based on 51 members and we use the average of this ensemble that is available in the S2S dataportal (https://apps.ecmwf.int/datasets/data/s2s-reforecasts-daily-averaged-ecmf/levtype=sfc/type=cf/). We will include this information in the manuscript.

L. 156-158: Agreed, but could it also be related to the greater temperature variability over mid-latitudes? I mean that if you would compute the seasonal average of $(T_{i,for} - T_{for})$ absolute values, for each grid cell, I expect these values to be lower at low latitudes, and therefore, the forecast errors end up lower as well (see e.g. Extended Data Fig. 7 and 8 in Tamarin-Brodsky et al. (2020)). I am not 100% affirmative but since you have not normalized temperature anomalies with their standard deviation, this could explain some (most?) of the meridional gradient depicted in Fig. 2.

A5. We agree that forecast errors in the extratropics might be higher than in the tropics due to a higher temperature variability, and will add this argument to the corresponding paragraph in the manuscript.

Table 1 layout: I would suggest to make it clearer that when the column “Source” is empty, it means “similar source as above”. Maybe a double quote could do ? And also, if possible and if allowed by the editor, try to reduce the font to have less line breaks. This would ease the reading.

A6. It will be adapted as suggested.

Figure 6, NA region, DJF season: by eye, significant correlation seems quite unlikely although this may be due to a “Pearson correlation oriented” perception instead of Spearman. Why not apply a lighter color shade, or a dashed style for instance, to the smoothing lines corresponding to pixels without significant correlation? Alternatively, you could indicate in the subplots the percentage of pixels of each region with significant correlation.

A7. It will be adapted as suggested. We will include the percentage of grid cells in this figure that have a significant correlation between forecast error and the Earth system variable depicted.

L.159: typo: parenthesis issue

A8. It will be adapted.

L.292: I am not sure it is correct to describe NAO and MJO as “ocean phenomena”

A9. These two phenomena are part oceanic, part atmospheric phenomena. We will rephrase the sentence as weather phenomena.

Figure 6 : last row (SA region): the x-axis tick labels are arguably wrong (no positive values)

A10. It will be adapted.

Another question that comes to mind when reading your conclusion is the extent to which the same Earth system variables would contribute to explain temperature forecast errors in the same regions for other S2S forecast systems. For example the spatial patterns of subseasonal temperature forecast skill show similarities between models and some predictability drivers are known to impact the same regions for different models (e.g. Ardilouze et al. 2021). I understand this would go way beyond the scope of this study, but I mention it for consideration.

A11. This is a very interesting point. As mentioned by the reviewer, this point is beyond the scope of our study, but still it is a potential follow-up analysis. We can only speculate that other forecasting models would exhibit similar patterns in sources of predictability when using the same drivers that we explore here. In addition, machine learning based studies in temperature forecasts at the subseasonal scale have highlighted some of the same drivers found here as important sources of predictability (Rasp et al., 2020; Herman and Schumacher, 2018). We will include and discuss this point in the conclusion section of the manuscript.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. Journal of Advances in Modeling Earth Systems, 12, e2020MS002203. https://doi.org/10.1029/2020MS002203

Herman, G. R., & Schumacher, R. S. (2018). Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests, *Monthly Weather Review*, *146*(5), 1571-1600. https://doi.org/10.1175/MWR-D-17-0250.1

References:

Tamarin-Brodsky, T., Hodges, K., Hoskins, B.J. *et al.* : Changes in Northern Hemisphere temperature variability shaped by regional warming patterns. Nat. Geosci. 13**,** 414–421, 2020

Vitart, F.: Monthly forecasting at ECMWF, Mon. Weather Rev., 132, 2761–2779, 2004

de Andrade, F. M., Young, M. P., MacLeod, D., Hirons, L. C., Woolnough, S. J., and Black, E.: Subseasonal Precipitation Prediction for Africa: Forecast Evaluation and Sources of Predictability, Weather Forecast., 36, 265–284, 2021

Ardilouze, C., Specq, D., Batté, L., and Cassou, C.: Flow dependence of wintertime subseasonal prediction skill over Europe, Weather Clim. Dynam., 2, 1033–1049, 2021