We would like to thank referee C. Prudhomme for the time and effort spent to review our manuscript and for her thorough and valuable comments. The feedback has helped us improve our manuscript. Below, we respond to each of the comments and indicate what changes have been made to the manuscript. The revised version of the manuscript has been included as a supplement.

The subject is very topical and relevant for publication in HESS. However, I regret that the analysis is done :

1) following climatic lenses (precipitation vs land surface; no analysis of propagation between the different land surface responses; summary/ discussion based on climatic regions without attempt to relate to soil/land surface/ bedrock/ catchment size etc. . . components). This is a shame and a more comprehensive analysis would be more valuable. Note that the title suggest 'effect of climate' but only precipitation (and not temperature/ evaporative losses) are considered, so it is not a full climate analysis that is undertaken ; 2) primarily on a multimodel mean (smoothing out extremely different behaviour; making extremely difficult a physical interpretation of results); 3) without justification of the choice of accumulations periods, which are arbitrary.

1. Undertake a full propagation analysis, by adding correlation between land surface components (soil moisture and runoff; soil moisture and discharge; runoff and discharge), and provide physically-based/ model structure/ parameterisation interpretation of the results. The analysis should also include at minimum catchment size, and if possible information on the land surface fields that should be available for all models.

> We have performed the suggested full propagation analysis by investigating drought propagation from soil moisture to both types of hydrological drought, and from runoff to streamflow drought. Similar to the analysis of propagation from meteorological to streamflow drought, the soil moisture and runoff to streamflow drought analyses used catchment-aggregated values of soil moisture and runoff. In summer, SSMI-1 had the best agreement with SRI and SSFI for most of world (Figure R1). Slightly higher SSMI-n are found in tropical regions, northern Europe, and parts of North America. In winter, longer SSMI-n up to (multi-)annual scales are more common, especially in continental climates. Drought propagation from runoff to streamflow droughts is generally quick in both seasons, with over 90 % of the pixels having a SRI-n less than or equal to 2 months. The similarity between runoff and streamflow drought propagation time scales is consistent with the meteorological drought propagation results (Figure 3 of the manuscript) and the difference in SPI-n (Figure 4 of the manuscript) for these drought types. Furthermore, the regions with positive differences between SPI-n in Figure 4 of the manuscript correspond more or less to the regions where SI-n are longer than 1 month in Figure R1 below, though the magnitudes are not necessarily equal. On the other hand, negative values in Figure 4 of the manuscript usually correspond to SI-n of one month. In these regions, drought propagation to runoff or streamflow is quicker than to soil moisture, which may be linked to a larger surface flow component in total runoff. This results in one-month SPI-n, though in fact this type of drought mechanism largely bypassing soil moisture cannot be captured when analyzing propagation of soil moisture drought to hydrological drought.

The full drought propagation analysis is presented in Figure S3 of the revised manuscript and described in an additional paragraph in Section 4.1 (P13 L14-24).

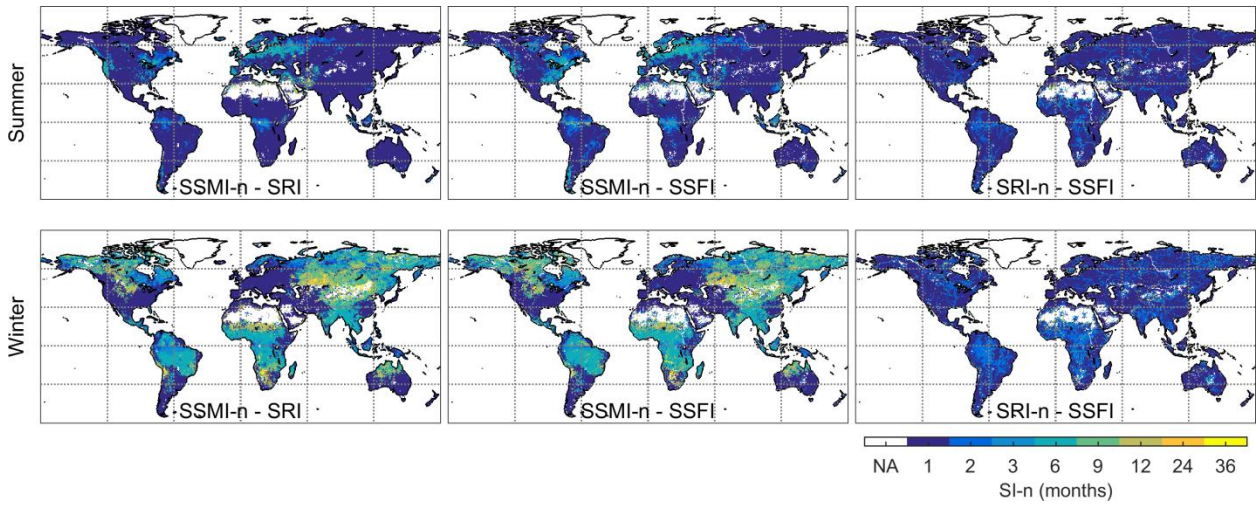

Figure R1. The SSMI and SRI accumulation period (SSMI-n or SRI-n) resulting in the highest correlations with model ensemble mean SRI and SSFI, for summer and winter droughts. Pixels where those correlations are not statistically significant ($p < 0.05$) are masked.

We fully agree that a physically based interpretation of differences between models would be very insightful, but is unfortunately not possible within eartH2Observe. Such an interpretation would require extensive experiments changing a large number of model structures and parameterizations, for example using Monte Carlo analyses. Even then, prescribing different sets of parameterizations is further complicated because choice of a certain parameterization is closely linked to the modeling system. This is also the reason the project did not prescribe a fixed set of static fields. We emphasize the need for comprehensive model structure and parameterization experiments in the Results (P16 L1-3) and Conclusion (P21 L28-30) sections of the revised manuscript.

The "effect of climate" in the title was meant to reflect how many of the analyses in our study focus on differences in drought propagation between Köppen-Geiger climate types. To make this clearer, we have changed the title to "The effect of climate **type** on timescales of drought propagation in an ensemble of global hydrological models".

2. Change the emphasis of the paper to individual models results, with the multi-model mean analysis presented last (if at all) with a justification of what it tells us. I am curious to know how different are the average SIs compared with individual models, and what mean SI represents physically. Understanding how the structure of the models influence drought propagation would be extremely valuable for future analysis. I fully agree with the point made by Referee #1 that there are strong collinearity between the different categories used to divide the models, and this should be considered in the interpretation of the results.

The ensemble mean result gives us an idea of the model consensus on drought propagation time scales globally and how these differ per climate type. We expect that the individual model results

are indeed of interest to the respective modeling groups because they can compare their results to other models and the model ensemble mean, or model consensus. Therefore, we have added the model-specific results to the Supplementary Material (Figures S6 and S7). However, to make individual model results insightful for the larger community we would need to be able to attribute the observed differences to model structures and/or parameterizations. As explained previously in response to the first comment, this is not possible within the current experimental setup.

We agree that we cannot use the categories we used to divide the models to definitively identify the mechanisms underlying differences in SPI-n due to the large number of potential factors and limited number of models (or the collinearity between groups). Instead, we attempted an initial exploration of potential explanations for the differences between models based on our observations and previous work. We have rephrased the introduction to this analysis (P16 L19-21) to better reflect this, and also modified the way we refer to this analysis in the conclusion section (P21 L25-28). In addition, we have removed the (no) reservoir groups from Figure 7 in the revised manuscript (Cohen's d effect size) due to the similarity with the LSM/GHM groups. We use the (no) reservoir group as another example of a factor that was found to be important in previous studies, but which we cannot isolate in our study (P18 L3-7).

3. Better justification of the choice of accumulation periods, which are very arbitrary: how different would be the results if different / additional accumulation periods were used? Ideally, a sensitivity analysis should be conducted. Are the statistical metrics used appropriate? (Point also raised by Referee #1) Whilst I understand the rationale, I struggle very much with the analysis of the 'difference in ranks' as they are really arbitrary. For example I very much like fig 3 but find fig 4 might be greatly dependent on the arbitrary accumulation periods.

We apologize that we did not make our choice of SPI accumulation periods clearer. The accumulation periods were based on those commonly used, where possible adding intermediate values to allow more subtle differences to be observed. These accumulation periods are furthermore nearly equidistant in log space.

Additional analyses were performed to determine whether using fewer SPI accumulation periods (1, 3, 6, 12 and 24 months) would impact the main conclusions of this study. As shown in Figure R2, the global patterns of SPI-n are not greatly impacted by this choice. In addition, the global patterns of the difference in rank SPI-n (Figure 4 of the manuscript) are very similar when fewer accumulation periods are used (Figure R3 below). The values and range of the difference in rank SPI-n change due to the smaller number of accumulation periods, but overall the direction and relative magnitude are similar. In this way, reducing the number of accumulation periods does not have a large effect on the conclusions of this study. We have added a summary of the results of the sensitivity analysis to the revised manuscript (P14 L2-5).

Finally, we address the point related to the statistical tests. As the referee indicated, there are statistical tests that have been designed for ordinal variables, such as Chi-squared and Cramer's V (an effect size metric). However, these metrics treat ordinal variables as categorical variables, which

means that the relationships between SPI-n are ignored. In the preparation of this study, we did calculate Chi-squared and found highly significant results ($p < 0.001$) for all tests analyzing SPI-n for different drought types by climate type or season. In the end we chose ANOVA tests because ignoring the relationship between SPI-n seemed unrealistic. This explanation has been added to the methods section (P5 L17-21). In addition, we report outcomes of Chi-squared tests (P12 L1 + 7). See also our response to referee 1's first comment.

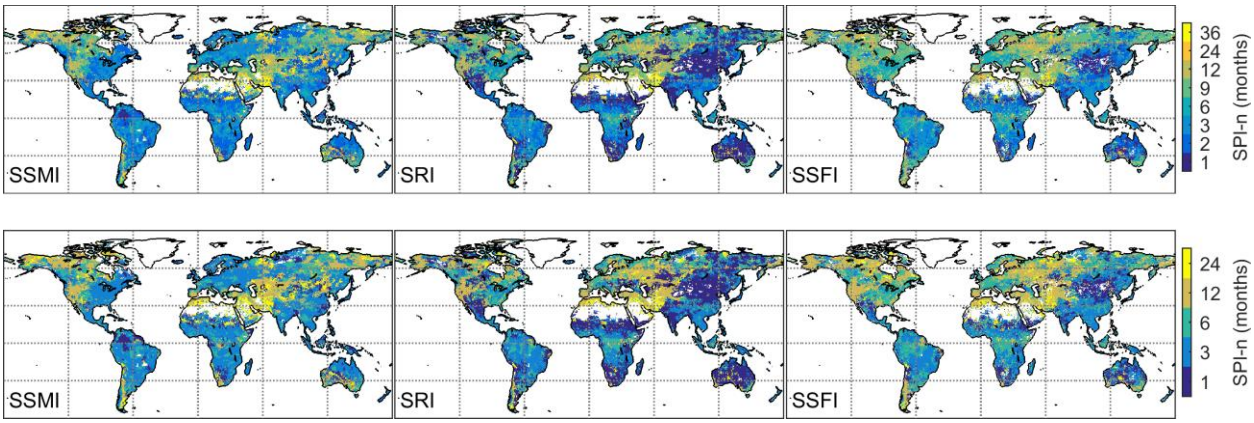

Figure R2. The SPI accumulation period (SPI-n) resulting in the highest correlations with model ensemble mean SSMI, SRI, and SSFI, for summer droughts using the original larger selection of accumulation periods (top) and a smaller selection of accumulation periods (bottom). Pixels where those correlations are not statistically significant ($p < 0.05$) are masked.

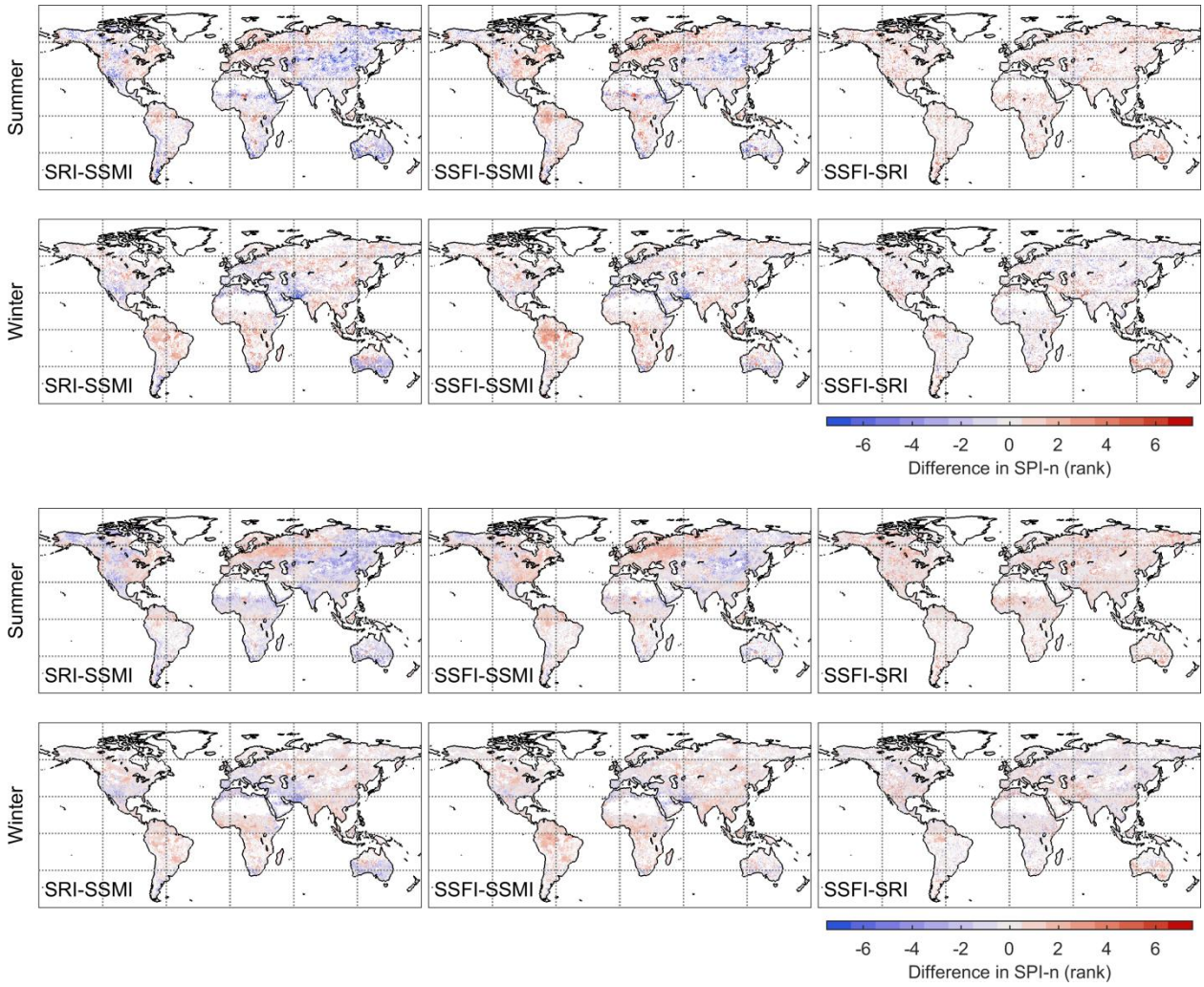

Figure R3. The difference in the rank of SPI-n for SRI and SSMI, SSFI and SSMI, SSFI and SRI for the original SPI accumulation periods studied (top, same as Fig. 4 of the manuscript) and for a smaller selection of accumulation periods (bottom). Pixels where the difference between accumulation periods are not statistically significant ($p < 0.05$) are masked.

4. I find difficult to understand the rationale and use of the evaluation section, as there are no real links with the rest of the analysis/ discussion/ interpretation. I think it is great to have it, but it should be more prominent. Moreover, as the authors mention, the analysis is extremely skewed with a very unequal distribution of catchments geographically. A filtering, with much fewer catchments in US and western Europe should be done. The drainage area of the model extracted points should also be compared with the catchment one. How do the stations relate to the climate zones?

This section evaluates whether drought propagation from meteorological to streamflow drought in the models is similar to observations. Therefore, this is a first reality check for the results shown in the previous sections. We have added the rationale for the evaluation section (P3 L7-8) and added a link to the results in Sections 4.1 and 4.2 (P18 L16-17). The number of GRDC sites does vary considerably over the study sites. While this is unfortunate, removing stations to ensure there are

an equal number of stations per climate type would result in not using more than half of the observational data available.

We did not compare the model and GRDC upstream catchment areas in the first version of the manuscript. In the current version, we applied an additional criteria specifying that the model upstream catchment area may not deviate more than 25% from the size of the GRDC catchment area. This has resulted in significantly fewer stations (126 instead of 297). Not surprisingly, the mean absolute error and Spearman correlations between modeled and GRDC rank SPI-n tend to improve (see revised version of Figure 8). The additional criteria based on the GRDC catchment area has been added to the methods section (P8 L10-11) and the results in section 4.3 have been updated to reflect the new selection of GRDC stations.

5. The method section needs to be re-written, especially the section on timescale propagation, and the rationale and description of the difference analysis p5 l9 to 20; what does mean 'statistical significance test does not reflect the relevance of differences between groups'? What is the group mean (mean correlation? something else?) in equation 1 and 2?

We try to distinguish between "statistical significance" and "relevance" of the difference between groups. That is to say that with a large number of observations as we have in this study, even very small differences between group means can be statistically significant. This sentence has been rephrased in the revised manuscript (P5 L25-27).

The group mean in equation 1 and 2 is the mean rank SPI-n for a specific climate type. This is now specified in the revised manuscript (P6 L1).

The section on evaluation of drought propagation also needs clarifying. Are the RMSE done on daily or monthly streamflow? How well the drainage area of the pixel matches that of real catchment? What model results have to be recalculated and why?

We agree that this section could use some clarification. The RMSE based on monthly streamflow data is used to assign a GRDC station to a model pixel. The streamflow data have been evaluated in previous work (Beck et al., 2017; Schellekens et al., 2016), therefore our evaluation focuses only on the drought propagation, or SPI-n. As stated in response to the previous comment, we have added an additional criteria for GRDC site selection to ensure that the model catchment area is within 25% of the GRDC catchment area. This information has been added to the Methods and Data sections.

The model SPI-n have to be recalculated in the evaluation section because there can be missing values in the observational time series. To ensure we compare like with like, we recalculate the model SI time series, and resulting SPI-n, using only months in which observational data are available. We have rephrased the sentence to make this clearer (P6 L19-21).

References

Beck, H. E., Van Dijk, A. I. J. M., De Roo, A., Dutra, E., Fink, G., Orth, R. and Schellekens, J.: Global evaluation of runoff from ten state - of - the - art hydrological models, Hydrol. Earth Syst. Sci., 21, 2881–2903, doi:https://doi.org/10.5194/hess-21-2881-2017, 2017.

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W. and Weedon, G. P.: A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset, Earth Syst. Sci. Data Discuss., 1–35, doi:10.5194/essd-2016-55, 2016.