We thank the referee for his/her in-depth review and constructive comments. The referee raises some important points (in bold), and we address each of these in a point-by-point fashion below. See the second Supplement for a revised version of the manuscript.

Main issues:

The authors chose for their analyses eight accumulation periods that represent different timescales (1, 2, 3, 6, 9, 12, 24 and 36 months for sub-seasonal, seasonal and annual timescales). These accumulation periods are similar to those that are often used, but still arbitrary. For example, they could have chosen only 1, 3, 6, 12 and 24 months for the same reasons. For the determination of propagation timescales this choice is probably of minor relevance, however, for the applied statistical tests I think it can have quite some impact. The tests used in this study are designed for variables on the interval scale (apart from spearman's rho) but the variables are on ordinal scale. The authors are aware of that problem and state they "assume that the difference between accumulation periods of 12 and 24 months (. . .) to be equivalent to the difference between 1 and 2 months" (p.5, l.5ff). Nevertheless, it is still very relevant to check whether there is an influence of the arbitrary choice of accumulation periods and the related assumption on the results of the statistical tests. Additionally, it needs a strong rationale for using tests designed for interval scaled variables instead of tests appropriate for ordinal scaled variables (e.g. the chi-squared test of independence).

First, we will discuss the choice for certain statistical tests, then we will investigate the sensitivity of the results and conclusions to the (number of) accumulation periods.

As the referee indicated, there are statistical tests that have been designed for ordinal variables, such as Chi-squared and Cramer's V (effect size metric). In the preparation of this study we considered using these metrics and calculated their results. The outcome of Chi-squared, for example, was that difference in SPI-n by climate type is highly significant ($p < 0.001$) for all drought types and seasons. However, an important disadvantage of using Chi-squared and other metrics for ordinal data is that these metrics treat ordinal variables as categorical variables. This means that the relationships between SPI-n are ignored. In the end we chose ANOVA tests because we believe it is important to take the relationship between SPI-n into account and because the SPI accumulation periods are nearly equidistant in log space. In the revised version of the manuscript, we include outcomes of the Chi-squared metric (P12 L1+6) and elaborate on the motivation for using ANOVA tests (P5 L17-21).

The sensitivity of the conclusions to the SPI accumulation periods is a good point. To address this, we recalculated the (significance of) the results using fewer SPI accumulation periods (1, 3, 6, 12 and 24 months, as suggested). As expected, changes to global patterns of SPI-n are minimal when fewer accumulation periods are used. This is shown for summer SPI-n in Figure R1, but is also true for winter SPI-n. In addition, outcomes of Chi-squared and ANOVA tests are still highly significant ($p < 0.001$). The pairwise comparisons using Tukey's honestly significant difference test show minor differences for runoff droughts. To be more specific, the difference in mean rank SPI-n between tropical savanna and dry climates is no longer statistically significant in summer. In winter, the

difference between tropical wet and dry climates is no longer statistically significant. Pairwise t-tests were used to test between summer and winter droughts, and the results for fewer accumulation periods were the same as with more accumulation periods. Therefore, the conclusions of this study are not affected by using fewer SPI accumulation periods. We summarize the results of this sensitivity test in the revised version of the manuscript (P14 L2-5).

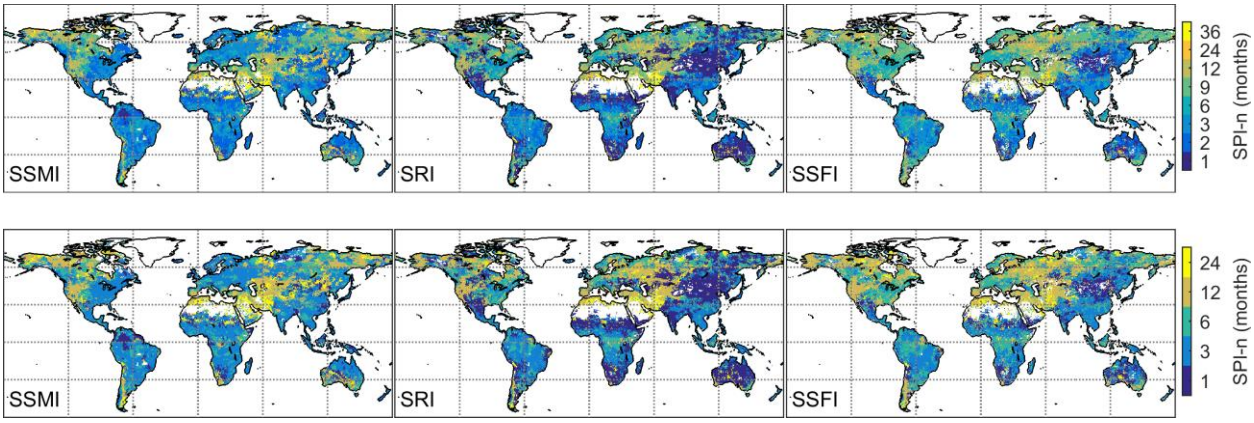

Figure R1. The SPI accumulation period (SPI-n) resulting in the highest correlations with model ensemble mean SSMI, SRI, and SSFI, for summer droughts using the original larger selection of accumulation periods (top) and a smaller selection of accumulation periods (bottom). Pixels where those correlations are not statistically significant ($p < 0.05$) are masked.

The second main issue is about the way the model ensemble mean is calculated: "The model ensemble mean was calculated as the average of the SIs" (p.6, l.29). A very important reason for using standardised indices is to ensure that all time series have the same distribution and are directly comparable (see e.g. Bloomfield and Marchant, 2013; Kumar et al., 2016). Averaging two or more timeseries, that have a standard normal distribution, will lead to a timeseries which distribution has a smaller standard deviation that might favour certain (higher) SPI-n. Moreover, the comparison with the results from the original model time series as it is carried out in chapter 4.3 is not really "fair" anymore, since time series are not directly comparable. The correct way is to average the raw model outputs first and standardize afterwards all seven time series plus the model ensemble mean.

We agree that we have deviated from the usual way of calculating the ensemble mean by averaging SI time series rather than the original model time series. Our motivation for averaging SI time series was that we did not want one or two models with high soil moisture/discharge (variability) to dominate the overall signal. For discharge we expected this to be less of an issue than soil moisture, where total storage (and variability) varies considerably between models (see for example Figure 6 of the manuscript). In addition, though standardization is an important reason for using SIs, we do not use the time series directly in the analyses. Even in the evaluation section, we compare SPI-n and not the time series themselves.

Nevertheless, we recalculated ensemble mean SPI-n using the approach of averaging of the original model time series. Overall, results are similar to when the ensemble mean was based on averaging

SI time series: SPI-n does not change in 60-67% of all pixels, and changes by a maximum of 1 rank SPI-n in 80-85% of all pixels (Figure R2). For all climates and seasons, the mean rank SPI-n changes by less than 1. A closer examination of the pixels showing a change in SPI-n shows that soil moisture droughts are most affected by changing the ensemble mean calculation method. For these droughts, SPI-n tends to be higher when original model time series rather than SI time series are averaged. This is especially the case for summer droughts in tropical wet, dry and polar climates.

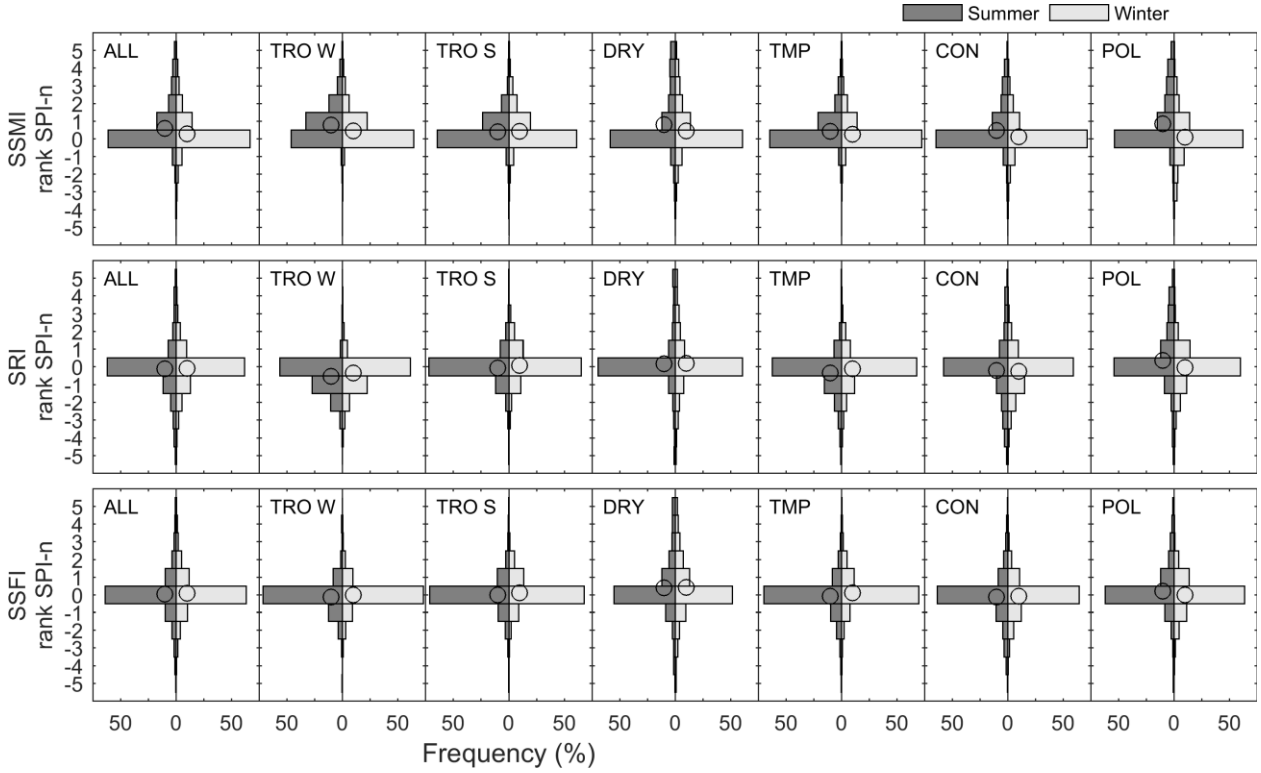

Figure R2. Histograms of the change in rank SPI-n when the ensemble mean is calculated as the average of the original model time series compared to when the SI time series are averaged. Changes in rank SPI-n are shown by climate type, and for summer and winter droughts in SSMI, SRI, and SSFI. Circles represent the mean change in rank SPI-n per climate type and season.

To further investigate the somewhat higher SPI-n for soil moisture droughts, we studied a pixel with a tropical wet climate located in central Africa (Figure R3). ORCHIDEE and SURFEX-TRIP, the models with the highest average soil moisture conditions (largely due to a deeper definition of the root zone), have a much higher soil moisture variability than the other models. The SI time series of these models are also very different from those of the other models. For example, the drought between 2004 and 2007 is much more pronounced in ORCHIDEE and SURFEX-TRIP, and the time series are smoother. As a result of the higher soil moisture variability, SURFEX-TRIP and ORCHIDEE have a larger impact on the average of the original model time series (EnsMean 2), and thus also in the resulting SI time series. Averaging SI time series (EnsMean 1) is a better representation of the 'average' behavior within the model ensemble.

The results shown in Figure R3 are representative of other model pixels where changing the ensemble mean calculation method results in changes of more than 3 rank SPI-n. The underlying cause of these large differences is that the models use different definitions for root-zone soil moisture. In some models this is a fixed depth, in others this varies with vegetation type. Ideally, the root-zone soil moisture time series could be normalized between 0 and the maximum soil moisture content before further analyses. However, the maximum soil moisture content is not always easy to define because vegetation types and rooting depths can vary within pixels.

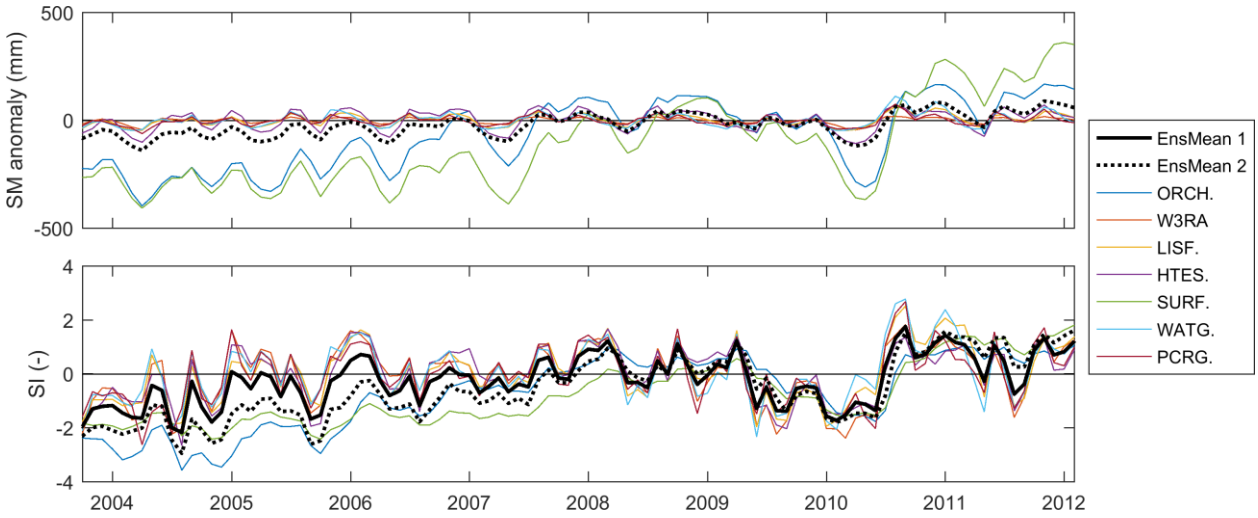

Figure R3. Time series of soil moisture content relative to the multi-year mean (top) and SSMI (bottom) for each of the individual models and two methods of calculating the ensemble mean. EnsMean 1 is based on averaging model SI time series, EnsMean 2 is based on averaging the original model time series.

In summary, though changing the calculation of the ensemble mean can have a large impact on SPI-n for individual pixels, the main conclusions of this study are not sensitive to the ensemble mean calculation method. This is probably because even though averaging does result in fewer extreme values in the ensemble mean, SPI-n are based on correlations between SI time series, which are not as sensitive as other metrics to a narrower range in values. Since the results are similar overall and due to the results of the soil moisture averaging analysis as shown in Figure R3, we have decided to still calculate the model ensemble based on SI time series. However, we have added a statement clarifying why we chose a non-standard method to calculate the ensemble mean (P7 L22-29 and Fig. S1 in the revised manuscript), and that the main conclusions of this study are not impacted by this choice (P14 L5-9).

Finally, the authors use an explanatory analysis to identify relevant model characteristics causing differences in drought propagation timescale (p.15f). They are aware of the difficulties using only seven models for that and the problem of collinearity between the groups. In fact, these limitations inhibit any useful result. For example, the factors GHM/LSM and (no)reservoir are highly correlated. Based on Table 1, it is only the model W3RA that is classified into another group. That means, in a study without this model, the groups would have been identical, similar to what is reported about the

snow scheme. Accordingly, the graphs in Figure 7 of the two groups have a very similar shape. The authors wonder about the reason for the high influence of reservoirs on soil moisture (p.16, l.12), but the real problem is, that both factors (GHM/LSM and (no)reservoirs) represent the combined effect of (no) reservoirs, GHM/LSM, snow scheme and probably several other relevant model structures. As it is not possible to relate the differences of the groups to one model structure we cannot learn much from this analysis.

We completely agree that we cannot attribute the observed differences to investigated model structures and parameterizations. We attempted to make it clear that while we cannot do so, we present an initial exploration only. For example, we think it is important to note the large differences between LSMs and GHMs, even though we cannot pinpoint the exact mechanism(s) responsible for that difference. This exploratory nature of this analysis has been made clearer in the results section (P16 L19-21) and in the conclusion (P21 L25-30).

Our paragraph concerning the simulation of reservoirs was poorly phrased. We included this factor because previous studies have shown that it plays a role in hydrological drought propagation. We agree that reservoirs are not likely to play a role in soil moisture droughts, and therefore by including it we intended to warn that apparent differences can be misleading. In the revised manuscript, we have removed this factor from the effect size figure (Figure 7 in the manuscript). In the text, we instead refer to previous studies that investigated reservoirs and hydrological drought and explain that this distinction is not useful in our case due to the high similarity with grouping by model type (P18 L3-7).

Other points the authors might want to look at: In the introduction the authors acknowledge that an important component of drought propagation is the time lag (p.2, l.13ff). However, time lags are not considered in the analysis but listed to be important for future research (p.19, l.16). Including an analysis of time lags which also might differ for the models would increase the relevance of this study. If lags are not included, there should be at least a rationale for excluding them despite their relevance.

Lag is indeed an important characteristic of drought propagation. We have recalculated SPI-n using SPI with different accumulation periods as well as lags up to 12 months for the model ensemble mean. Results show that, as expected, taking lag into account has a larger impact on winter drought propagation than on summer drought propagation. In summer, the best drought propagation result has a lag of 0 months in around 70 % of the pixels (Figure R4). In winter, 0-month lags are still the most common, but account for only about 40 % of the pixels. Shorter lags (1–3 months) are more prevalent than longer lags (8–12 months).

The frequent occurrence of 0-month lags for summer drought propagation means that SPI-n also remains unchanged in a majority of pixels. However, even for winter droughts SPI-n are not affected by taking lags into account in about 60 % of the pixels. This means that for about 20 % of pixels, taking lag into account does not change SPI-n, but does improve the correlation between SPI-n and the winter drought SI of interest. Changes in SPI-n in both seasons tend to be small, with changes in rank SPI-n larger than 2 occurring in less than 10 % of pixels. Positive changes in SPI-n are more

frequent than negative changes, meaning that overall taking lag into account leads to longer SPI-n. This is the opposite of what we hypothesized in the previous version of the manuscript. We had expected that lag would be especially important in areas with significant snow cover in winter, and that including lags would lead to lower SPI-n.

These results show that lag does play a role in drought propagation, though this is more so in winter than in summer. Even so, SPI-n are not very sensitive to whether lags are included in the analysis or not. We could add this figure to the supplementary material, but the Supplement is already large, containing seven figures. Since this was not one of the major comments, we suggest not adding this figure to the supplement. However, we leave the final decision to the referee and editor.

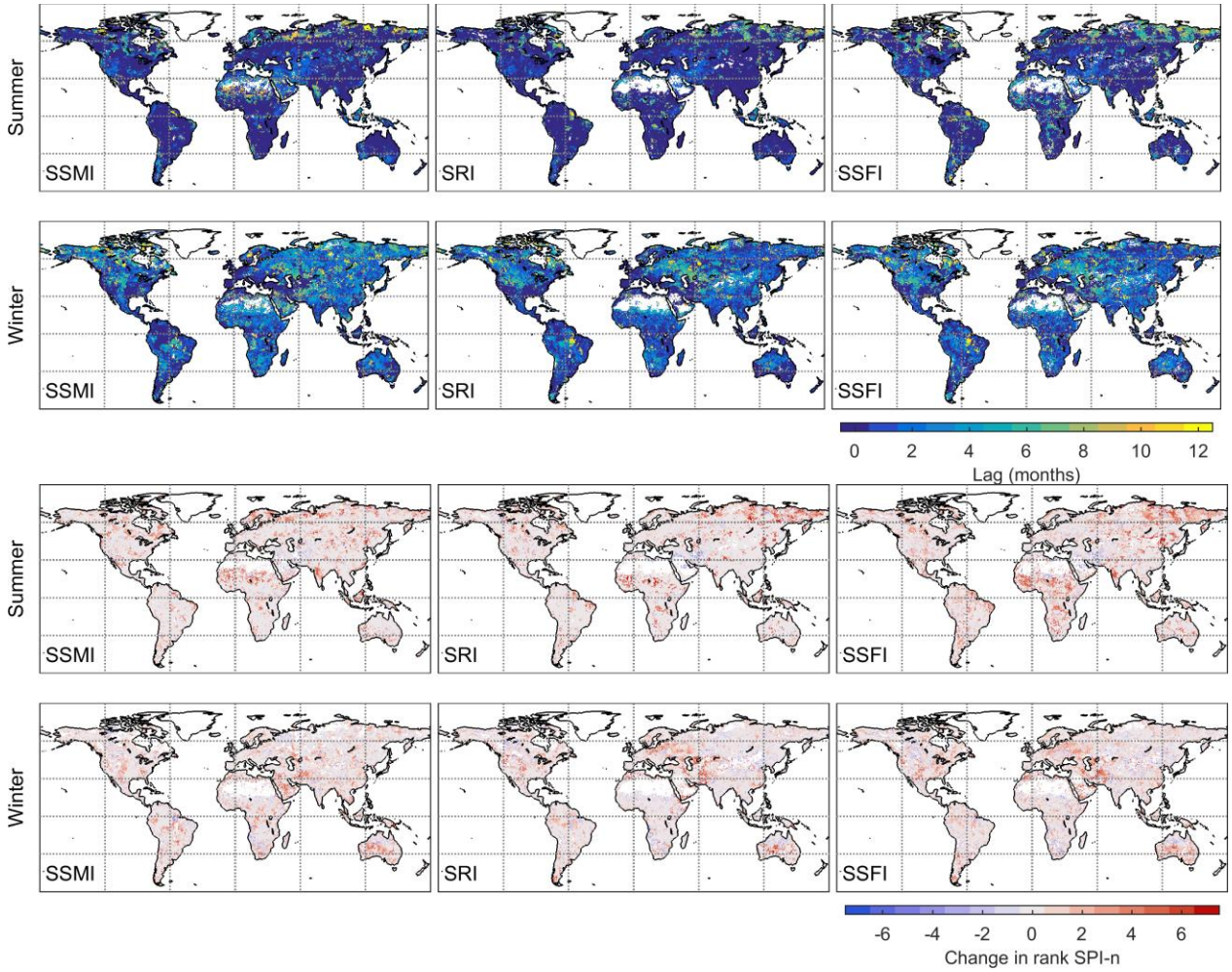

Figure R4. The SPI-n lag in months leading to the best correspondence with SSMI, SRI and SSFI for summer and winter droughts (top) and the change in rank SPI-n compared to when lags are not taken into account (bottom). Pixels where the correlations between lagged SPI-n and SI time series are not statistically significant ($p < 0.05$) have been masked.

In chapter 4.1 analyses of the "mean SPI-n" are presented (e.g. p.10, l.17; caption of Figure 3). For me it does not become clear, whether this is really the arithmetic mean of the SPI-n or rather the mean of

the ranks. For example in Figure 3: If there were the two accumulation periods of 1 and 36 months, is "mean SPI-n" (1+36)/2=18.5 or rather (1+6)/2=3.5? This is quite relevant for the plotted circles. If they are calculated as an arithmetic mean, it might be very hard to read the values from the plot due to the very non-linear y-axis.

This should indeed all be mean rank SPI-n. We have changed "mean SPI-n" to "mean rank SPI-n".

Moreover, it is important to report somewhere the 'sample size', i.e. the absolute number of cells which are not masked for the different climates and drought types. Otherwise it is for example hard to understand, that the t-test leads to significant different SPI-n means for winter and summer in runoff of TMP (Figure 3).

Agreed, we have added the number of pixels for each climate and drought type to the panels in Figure 3 of the revised manuscript.

On page 10, l.16 the authors describe the results of the ANOVA: "The means of SPI-n for winter hydrological droughts in continental and polar climates are not significantly different". Again, for me it is not clear whether the rank mean or the arithmetic mean is meant here. However, more important is the fact that it sounds like two categories were directly compared to each other. In this case, it would have been a t-test rather than an ANOVA what was used. Please clarify, which variables were used for the ANOVA and in which cases a t-test was used.

This should indeed be mean rank SPI-n, and the text has been revised to reflect this. An ANOVA test was used for the comparison over multiple groups. A statistically significant ANOVA test result was followed by Tukey's honestly significant difference tests to compare each pair of group means. This test is very similar to a t-test, but corrects for family-wise error rates. This correction is needed because the chance of making a type 1 error (false positive) increases when comparing multiple groups. The use of Tukey's tests has been added to the Methods section (P5 L21-23) and the Results section (P12 L2-4).

The stations used for the evaluation against observations are distributed very uneven (as the authors write on p.18, l.7). In Figure 8 it looks like there were very few to no stations in the climates polar and tropical wet. However, the authors state that "errors between models and observations are not related to climate". To enable the reader to comprehend this important finding, I think it is necessary to give more information on the number of stations per climate zone, the test used to reach this conclusion as well as the results of the test.

We agree that the link between GRDC stations and climate type is not clear. Therefore, we have included the stations in Figure 1 of the manuscript (the global map of Köppen Geiger classification used). In addition, we have included the number of stations falling within each class in the legend of the same figure.

The relationship between error in SPI-n and climate zone was based on ANOVA tests ($p < 0.05$). In the previous version of the manuscript, ANOVA results were not significant for any of the models. In

the revised version of the manuscript, however, some of the results have changed because we included an additional criteria for GRDC stations based on the agreement in upstream catchment area (see referee 2, comment 4). ANOVA results are now statistically significant for two out of seven models in summer and four models in winter. For the model ensemble mean, a statistically significant result is only found for summer droughts. The explanation of the test used and its results has been added to this section (P19 L7 – P20 L4, Fig. 9).