

## **Reply to referee #2**

### **\*\*Summary:**

The paper looks at the ability to forecast land-based meteorological drought (defining drought as SPI3, SPI6, SPI12 < -0.8) using the ECMWF System 4 (S4) forecast model and climatological forecasts coupled with 2 reference datasets: GPCC and ERAInterim reanalyses. Their methods follow to a large degree the approaches taken in Dutra et al. 2013 and Yuan and Wood 2013, but using global (often regionalized) analyses and using the S4/GPCC-Climatology models rather than Yuan and Wood's multimodel ensemble. The authors show that the initial conditions (ERAInterim versus GPCC) have significant effects on the forecast statistics, and they characterize the memory time-scale of those initial conditions upon their forecast. Additionally, the authors determine that the S4-GPCC pairing has consistently higher skill than a GPCC climatological forecast alone (although negligibly so when skills are low).

### **\*\*General comments:**

The findings are useful as a means of investigating our ability to forecast meteorological drought (particularly in suggesting that we have some ability to do so at present) and are somewhat provocative in their response to Yuan and Wood's question of whether local drought forecasting is an issue of stochastic forecasting (further discussion of this last point to follow). The methodology is well-suited to the questions being asked, and the data is presented in a clear fashion (the supplementary material is quite important, I would argue, for readers, as many of the interesting results are depicted only as regional plots). I found the paper to be, for the most part, quite straightforward and without overstated claims. The approach is sound for the investigation they are performing, and I would recommend that the paper be published following some minor revisions.

The Authors thank the Reviewer for the very helpful comments and recommendations. In the following we address point by point the Reviewer comments.

With the fairly thorough analysis performed, one thing I found to be somewhat lacking were global evaluations of the forecasts and forecast skills. In particular (unless I missed it), I think it would be useful to quickly plot or correlate the ability to forecast drought (perhaps as the Brier Skill Score) against the frequency (within a particular grid cell) or intensity (averaged within a particular grid cell, presumably as SPI) of droughts. I would assume that locations with very few droughts that have a few severe dry periods would be easiest to forecast with skill, but I do not have much intuition as to whether locations with frequent droughts are harder or easier to forecast. Similarly, I would be interested to know if wetter or drier locations show more drought-forecasting skill. This could be presented as maps, scatter plots, or simply stated in text, but I would recommend including at least some discussion of a) drought frequency, b) drought intensity, and c) climatological precipitation values on forecast skill.

Following the reviewer suggestion, several comparisons were made. For example, plotting the Brier skill score of drought onset forecasts (the map in Figure 4) against the mean frequency of drought onset events (see map in end of the reply, Figure R1) does not display any relation between the skill score and the frequency of drought onset events. Other attempts to compare forecast skill with drought frequency/intensity did not provide any clear information. We should expect that some relations arise when looking at a particular seasons and regions, and these might be also dependent on the seasonal forecasts of precipitation - for example the precipitation forecasts might have a higher skill for below normal conditions than for above normal conditions. Further investigations could be performed by following a similar approach to Yuan and Wood (2013) by conditioning the analysis on El-Nino/La-Nina years but restricting the comparison to particular regions. However, we felt that the additional analysis suggested by the reviewer of linking forecasts skill with regional/season/drought characteristics are out of the scope of this study, and would be the material for a new study.

**\*\*Necessary Edits:**

**Pg 920, line 6-8: "The forecast skill is concentrated on verification months where precipitation deficits are likely to have higher drought impacts..." If I read the paper correctly, I believe you only analyzed droughts for wet periods, so it is not that "skill is concentrated" on those months, but that the "analysis" is focused on those months, correct?**

Yes. The sentence was changed to:

"The forecast evaluation is focused in the periods where precipitation deficits are likely to have higher drought impacts and resumed over different regions in the world."

**Pg 920, line 8-11: "Verification of the forecasts as a function of lead time revealed a reduced impact on skill for: (i) long lead times using different initial conditions, and (ii) short lead times using different precipitation forecasts." The "using different initial conditions" phrasing is a little confusing since we don't know what they are different from. Perhaps, "using a different dataset for initial conditions than for validation," or "using ERAI instead of GPCC." Similar for "using different precipitation forecasts."**

Sentence changed to:

The verification of the forecasts with lead time showed that there is, in general for all regions, the least reduction on skill for: (i) long lead times using ERAI or GPCC for monitoring, and (ii) short lead times using ECMWF dynamical or climatological seasonal forecasts.

**Pg 921, line 16: "what is the importance of the monitoring in the forecast skill?" I like your three questions, but this first one could use some rephrasing. Instead of "the importance of the monitoring" you are more specifically comparing two sets of validation data, so the question might be something more like "how sensitive is drought forecasting to the validation data set?" or anything else that you feel is the appropriate question, but make it a bit more specific.**

Point i) was changed to:

i) how sensitive are the drought forecasts to the monitoring dataset used;

**Pg 922, line 12: "the observational dataset" I am assuming that this means GPCC/ERAI, but it would help if you picked one term for the "reference" or "observational" or "monitoring" or "validation and initial conditions" dataset and used that term exclusively throughout the paper. Because of the terminology it is confusing as to whether all validation is performed against GPCC but both GPCC and ERAI are used for initial conditions or whether each are used as both initial conditions and validation.**

Changes were made to clarify this point in several parts of the manuscript. Reference dataset refers to GPCC and is the datasets used for the verification of the forecasts. Monitoring/initial conditions refer to the dataset used as drought monitoring that provides initial conditions to the SPI forecasts, and both GPCC and ERAI are used.

**Pg 922, line 16-18: "Also, the test for drought-like conditions is made by merging and blending the GPCC precipitation observations with forecast precipitation, so that GPCC also serves as an initial condition." A quick (one sentence) explanation of this merging and blending would be useful to readers.**

This was removed, and an explanation of the "merging" of the monitoring with the forecasts is given the following section (2.1.2 - drought indicator), adding: "The merging of the two products is basically a concatenation of the monitoring with the seasonal forecast of precipitation."

**Pg 923, line 5-6: "by merging the seasonal forecasts of precipitation with the monitoring product." Same comment as above. Even though you may describe this in detail in Part 1, a quick hint as to the meaning "merging" would be helpful.**

See the previous reply.

**Pg 923, line 23-25: "In these configurations, all the forecast skill comes from the monitoring period (or initial conditions) and they are used as reference forecasts." This could use some rephrasing to clarify what is meant by the skill coming from the monitoring period or the initial conditions, as well as what "they" refers to.**

The sentence was removed as it was not important in the description of the drought forecasts configurations.

**Pg 924 line 19-21: "The calendar month with 3 months maximum accumulated precipitation is used to verify the SPI-3, while the calendar month with 6 months maximum accumulated precipitation is used to verify the SPI-6 and SPI-12." Using the wet season is, I think, a good move, not just from a water resources perspective, but also to aid in reducing biases that may come from small deviations in precipitation having unduly large impacts in dry locations relative to wet locations (one fewer rainy day in the Sahara will very dramatically affect the SPI and therefore the onset of drought, but that is not the case in India's monsoon region). On the other hand, detection of drought is very difficult if no rain in a 3/6/12-month period is typical. The SPI is, of course, designed to help minimize that bias/sensitivity to some degree, but if you have any thoughts on this comparison across regions with very different absolute values of precipitation, I encourage you to discuss them here.**

The reviewer raises an interesting and important question on the use/application of the SPI, but we feel this is out of the scope of the current manuscript.

We must keep in mind the limitations of a single index, although multi-scalar, to characterize such a complex phenomena as drought. A detailed study of drought characteristics, monitoring and forecasting cannot be accomplished using only the SPI, and such analysis will vary from region to region. This was one of the reasons that lead us to evaluate the forecasts on different regions, and in different periods, focusing on the wet season in each region. However, this approach is limited by the applicability and use of the SPI, which only considers deficits in incoming precipitation, relative normal conditions in each region.

**Section 2.2.2 is a bit of a mess. The definitions and usage of "hit rate" and "false alarm rate" vacillate between being used as rates and as absolute numbers of events. The different cases (false positives, true positives, false negatives, true negatives), while simple and very common notions require frequent re-reading and referencing to be able to finish the section. The "relative operating characteristics" (ROC) acronym is defined twice, but the first definition lacks any explanation and should be moved to the portion of 2.2.2 where ROC is actually defined and discussed. The definition of drought does not become obvious until after much discussion of drought occurrence. I recommend: 1) Using standard terminology for false positives and negatives – these notions are extremely prevalent and standardized (a quick search for either Type I and Type II errors or sensitivity and specificity will get you to much discussion of the topics). There is common notation (alpha and beta for the different error rates...), or you can use your own, but defining some variables, either in-text or using a simple schematic (see the matrix at <http://www.cs.rpi.edu/~leen/misc-publications/SomeStatDefs.html> or many intro statistics texts) will go a long way towards making this section clearer. 2) Not calling anything that is a count of occurrences a "rate." 3) Including some variables and equations. There are only four possible outcomes, so defining the `_number_` of outcomes (a = number of false positives, b = number of true positives, etc.) rather than talking about cases will be simple for the reader to comprehend.**

The definition and use of hit rate and false alarm rate in section 2.2.2 was clarified following the suggestion of the reviewer of defining the a,b,c,d

**Pg 925, line 4-5: "...anomaly correlation of the ensemble mean and the relative operating characteristics (ROC) of the SPI below  $-0.8$ ." You do not at this point define the relative operating characteristics, you redefine the acronym two pages later, and you do not point out that this SPI below  $-0.8$  is your definition**

**of drought. For ROC, either add "(defined below)" or get rid of it here. Move section 2.2.3 to be before 2.2.2 so that we know what you are using as your definition of a drought. I could not understand anything about this sentence until multiple pages later.**

The sentences were changed to:

"...anomaly correlation of the ensemble mean and the relative operating characteristics (ROC, defined below) of the SPI below a drought threshold. The drought threshold is defined as the SPI below -0.8 as suggested..."

**Pg 926, line 15-16: "In one case, called the "hit rate" a forecast of drought is made and drought is, indeed observed (the number of cases for which this holds true: case a)." If it is just a count rather than a ratio, it is not a rate. Use better and more consistent terminology as mentioned previously. This whole paragraph needs to be radically rewritten, preferably with some defined variables and possibly a figure.**

**Pg 927, line 15-17: "The ROC diagram displays the false alarm rate (F ) as a function of hit rate (HR) for different thresholds..." Possibly refer to figure 1. F and HR should have been defined (preferably using standard notation and terminology in the previous paragraph. What are the "thresholds" being discussed? You want to say something about short-term initial condition uncertainty dominating for small lead times and longterm model uncertainties dominating for long lead times, presumably.**

**Pg 927, line 18-19: "The area under the ROC curve..." Do you ever make use of this area in your analysis? It is perhaps interesting, but if interesting enough to mention, perhaps you should talk about that area in your discussion of Figure 1.**

**Pg 927, line 24-25: "The forecasts and verification were transformed into an event (or no event) based on the underlying grid-point distributions." And more importantly, they were transformed into an event by determining if  $SPI < -0.8$ ! Swapping sections 2.2.2 and 2.2.3 will make this clearer, but you might want to state it here again anyway since your definition of a drought is of fundamental importance to your methodology.**

**Pg 927, line 26-27: "...to build the contingency table..." Presumably the contingency table is the accumulated counts of false positives, true positives, etc. If so, define it at some point or use different phrasing. If not, you need to explain what a contingency table is.**

All the above comments were addressed by re-writing the part of section 2.2.2 referring to the contingency tables/ROC/hit,false alarm rates. We thank the reviewer for stressing this point. This section is not clear in the submitted version.

**Pg 928, Equation 5 and line 11: "where  $s$  is the actual score..." What "score" are you referring to? What values do you put into Equation 5?**

Equation 5 was presented in general for any skill score, but now it is presented for the ROC skill score.

"where  $s$  is the ROC score of the forecast,  $s_0$  is the ROC score of a benchmark forecasts and  $s_1$  is the ROC score of a perfect forecast."

**Pg 929, line 10: "lead times 0 and 1.." Months?**

Corrected.

**Pg 929, line 11: "ERA-Interim has higher RMS errors." Why? Does ERA-Interim have higher internal variability? It sounds like GPCP is better, is this true? Are you using each of GPCP and ERA-Interim as validation data, or just GPCP? Have the forecast models just been better calibrated to GPCP (doesn't seem right since you are using the ECMWF model for forecasting and it or its relative did the ERA-Interim reanalysis, right?)? In any case, explain**

**this important point. Pg 929, line 11-14: This sentence will become clearer with clarification of the previous sentence, I believe.**

This clarification was added:

“The forecasts using ERAI for monitoring are penalized when compared with the forecasts using GPCC for monitoring, since GPCC is used as reference dataset (for the forecasts verification). These results do not consider the uncertainties in GPCC that are discussed in more detail in the companion part 1 paper.”

**Pg 929, line 11: "lead time 2.." Months?**

Corrected

**Pg 929, line 14-15: "In East (Fig. S9, Supplement) and West East Africa (Fig. S9, 15 Supplement) and West Africa (Fig. S20, Supplement)..." Both the figure numbers and the names ("West East Africa"?) seem to be wrong in this sentence.**

Changed to:

“In East Africa (Figure S7) and West Africa (Figure S18) the RMS error for ERAI”

**Pg 929, line 15-16: "...RMS error for ERAI merged with S4 decreases with forecast lead time, which might be contra intuitive..." True. You should probably explain why this happens.**

This is associated with the problems of ERAI in those regions that are discussed in more detail in the companion part 1 paper.

**Pg 929, line 23-Pg 24, line 3: "The comparison of the RMS error of the ensemble mean with the ensemble spread (dashed lines in Fig. 1) suggests that in general the forecasts are slightly under-dispersive. However, we do not consider the observations uncertainty (in this case the GPCC precipitation) that should be added to the ensemble spread when comparing with the RMS error of the ensemble mean. This might be also associated with the deterministic nature of the initial conditions, and the extension of the probabilistic monitoring presented in the companion Part 1 paper could be of potential benefit to increase the spread of the forecasts." Is there a reason not to have done this, since you already completed Part 1? Maybe give some reasons for why the models are under-dispersive, and some quick quantification of the observational error to tell the reader whether observation error is truly enough to correct for this discrepancy?**

We did not included the extension of the probabilistic monitoring, presented in part 1, in this study because the method can only be applied since 2008 (availability of the short-range ensemble re-forecasts), and such a short period does not allow for a robust statistical evaluation of the forecast skill.

Concerning the quantification of the observational error, such development would be very valuable for many applications, not only the present drought work. However, such estimates are not easy to achieve, requiring the extensive processing of many raw stations/satellite data. Furthermore, the observation error also includes the spatial representativeness of the stations when generating the gridded dataset. Therefore, we cannot provide further details on this topic in the current manuscript, but we felt it was important to highlight it within the manuscript.

**Pg 930, line 15-17: "...the ROC scores of GPCC using the S4 forecasts (GPCC S4) are higher than the same S4 forecasts used with ERAI (ERAI S4) during the first few months of lead times..." Again, I think we the readers would like to know why this is! Any explanation?**

This is associated with the use of GPCC as a reference dataset. A clarification was added.

**Pg 933, line 18: "...the brier skill score is used..." Should be capitalized, explained briefly (maybe), and referenced (certainly).**

A reference was added (Wilks 2006).

Pg 936, line 3-4: "Yuan and Wood (2013) questioned whether seasonal forecasting of global drought onset was largely or solely a stochastic forecasting problem only." I think you're stretching that quite a bit. They stated: "This raises the question of whether seasonal forecasting of global drought onset at local scale (e.g., 1\_ in this study) is essentially a stochastic forecasting problem." Point taken that there is skill in your model, but I would argue that to properly refute their claim (which is about LOCAL forecasting), you would want to show an unsmoothed version of Fig 4a minus Fig 4b that demonstrates that skills are strictly positive at nearly all grid points, probably with an accompanying figure showing that drought frequency is not negatively correlated with BSS across all grid points. You should at the very least remove the words "global" in line 3 and "solely" in line 4. I like the direct discussion of this issue, and while I believe that there is more nuance to the discussion of whether this is a problem of non-deterministic stochastics versus deterministic forecasting than you are giving credit, that is probably the topic of a more theory focused paper. I'll let you claim model skill as a partial refutation of their claim for the sake of provoking further discussion on an interesting issue, but I can imagine some raised eyebrows among some readers of this paper.

The sentence was modified to:

"Yuan and Wood (2013) raised the question as to whether seasonal forecasting of global drought onset at local scale (for 1 degree grid cells) was largely or solely a stochastic forecasting problem. Our results are coherent with the findings of Yuan and Wood (2013), but when the grid cells are pooled into a region, our regional analysis highlight that, within several regions in the world, drought onset forecasting is feasible and skilful."

Following the reviewer suggestion, we plotted an unsmoothed version of Fig4a minus Fig4b (see below Figure R2). Figure R2 was not smoothed, and displays the original 1x1 data. We can see that this difference is positive in most of the globe; further validating our claim that drought onset forecast is skillful regionally.

**\*\*Minor edits:**

Pg 921, line 9: "combined to" -> "combined with"

Pg 924, line 9: forecasts -> forecasts'

Pg 925, line 1: "is" -> "are"

Pg 926, line 20: "...for which is drought..." -> "...for which a drought..."

Pg 927, line 18: "...has the attribute to discriminate between..." -> "...has the ability to discriminate between..."

Pg 927, line 19: "statistics" -> "statistic" Pg 928, line 4: "vales" -> "values"

Pg 929, line 16: "contra intuitive" -> typically "counter-intuitive"?

Pg 929, line 22: "...over South Africa when compared with East of West Africa" Should this be "East or West"?

Pg 931, line 7: "(FS)" This acronym was already defined once.

Pg 931, line 8: "(FS)" This acronym was already defined TWICE now.

Pg 932, line 20: "in the order" -> typically "on the order"

Pg 933, line 8: "(POD)" already previously defined

Pg 933, line 8: "exchanging" -> "exchange"

Pg 934, line 17: "overplayed" -> "overlaid"? Or I don't understand the terminology.

Pg 935, line 29: "evaluate forecasts" -> "evaluate a forecast's"

Corrected

**References**

Wilks, D., S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 648 pp.

Yuan, X., and Wood, E. F.: Multimodel seasonal forecasting of global drought onset, *Geophys. Res. Lett.*, 40, 4900-4905, doi: 10.1002/grl.50949, 2013

**Figures**

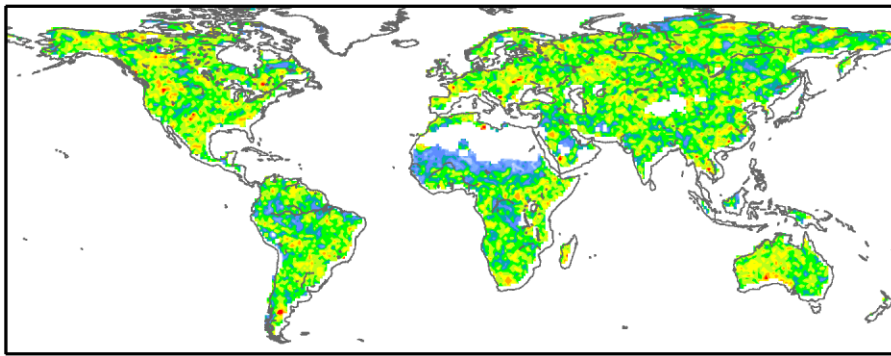


Figure R1. Mean annual frequency of drought onset events using the SPI-6 of GPCP. For example a 0.2 value stands for 1 event every 5 years on average for the period 1979-2010.

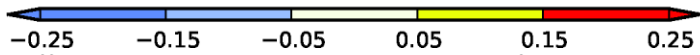
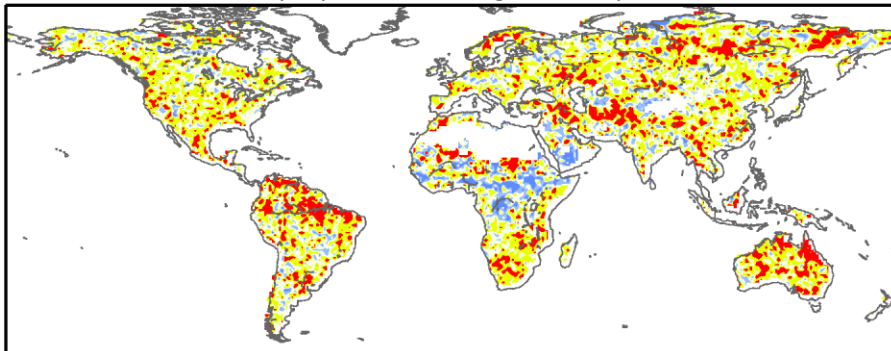


Figure R2. Differences between the Brier skill score of drought onset forecasts between GPCP S4 - GPCP CLM.