Interactive comment on "Seasonal predictions of agro-meteorological drought indicators for the Limpopo basin" by F. Wetterhall et al.

Anonymous Referee #2
Received and published: 18 September 2014

Review of: Seasonal predictions of agro-meteorological drought indicators for the Limpopo basin Authors: Wetterhall et al. Recommendation: Accept subject to major revisions

General Comments: The paper presents a relatively simple post-processing approach for ensemble seasonal forecasts that have agricultural significance, and I do not doubt the basic conclusions that the post-processing improves the forecasts. The paper as a whole is relatively brief compared to most HESS papers, and lacks the typical depth of analysis and detail, but the results are practical and likely to be of interest to the forecast applications community. To reach a standard that is acceptable for publication, however, the authors must provide additional detail for various aspects of their approach that are unclear. In particular, the QM must be explained in more detail, and I strongly suggest that the authors either bolster their rational for applying the thresholds as they do, or evaluate a suggested variation (see comments below). In addition, further diagnosis of the results is warranted, given that some aspects (such as the disparity in performance in Fig. 5) are difficult to understand.

I would strongly urge the authors to bring the paper up to the density of a typical HESS article through the addition of further visualizations of the results (including, perhaps, timeseries comparing each years predictions and observations for the main predictands, among other analyses) – such displays offer different insights than those captured in the summary metrics alone. The more general discussion of the value of forecasts is useful and appropriate, and probably could be left as is.

We thank the reviewer for the comments provided. Regarding the length of the paper, it is relatively short because if the special issue on droughts that it belongs to. This paper is a case study on the benefits of using bias correction on the forecast performance to show the usefulness of applying seasonal forecasts as opposed to using the climate as proxy for forecasts. We agree that the method section needs revising, also the application of the thresholds. We will include a figure showing the annual time series of the dry spells, as that will give a good overview of the general performance of the system. The specific comments are answered below.

Specific Comments to the Authors:

865-23: 'for shorter lead times, such as . . .' (clarify)

The statement relates to the fact that the predictability is higher with shorter lead times, but it does not add anything to the hypothesis, since all possible lead times are investigated. Therefore, "and for shorter lead times." was deleted from the sentence

Table 1: Comparisons might be better presented as %bias, so that the reader doesn't have do the arithmetic in his/her head.

We will change table 1 considerably (see also comment to reviewer 1 and include biases.

Figure 3: This figure might be improved by an X-axis that shows increments of month?

Thanks for the suggestion, we will add that to the figure.

867-10: The detail on the modeling system is perhaps more than needed, but as long as detail is being given, please include the strategy to create ensemble members. Are their atmos components initialized at various lags from time zero, eg?

The description of the model system was shortened since it is already published elsewhere. A sentence on the creation of the ensembles were added: "The 51 ensemble members are made up with one control member initialised by ERA-Interim and 50 ensembles in which the initial conditions (ocean and atmosphere) combined with stochastic schemes in the model physics of the atmospheric model."

868-28: The refs given for the QM bias-correction approach all relate to climate change applications. Perhaps better refs would be Voisin et al (2010), where it was applied to medium range forecasting, or an earlier seasonal forecasting paper, Wood et al JGR (2002).

** Voisin, N., Schaake, J. C., and Lettenmaier, D. P.: Calibration and Downscaling Methods for Quantitative Ensemble Precipitation Forecasts, Weather Forecast., 25, 1603–1627, 2010.
** Wood, A. W., E. P. Maurer, A. Kumar, and D. Lettenmaier, Longrange experimental hydrologic forecasting for the eastern United States, J. Geophys. Res., 107(D20), 4429, doi:10.1029/2001JD000659, 2002.

Thanks for the suggestions, the references will be added to the paper

868-24: 'a simple mean monthly bias correction would not correct biases' – I disagree – if all daily rainfall values in a month were scaled in the process of correcting monthly means, the step would alter daily rainfall amt distributions and could indirectly improve such biases.

The statement is a bit unclear. A simple mean bias correction would not alter the number or length of dry spells. The sentence was changed to: "In this study we evaluate dry spells based on daily precipitation, and a simple mean monthly bias correction would not correct biases in the distribution of dry events"

Also, please clarify: Are quantile mappings applied to each grid cell separately or to are the CDFs created for all cells in the domain? Explain whether the forecast and obs rainfall distributions are uniform enough across the domain to justify this choice if so. Could there be spatial variations in bias that would cause one correction applied across the whole domain to have suboptimal performance?

The quantile mapping is applied to each grid cell individually to reflect the observed rainfall pattern. The observed pattern EGCP has a larger spatial variability than the forecast, so the bias is not uniform across the domain, although it is spatially correlated. We have not looked into the possibility of applying a uniform bias correction across the whole area, but we will look into the bias of the forecast.

869-4: perhaps you mean quantiles 0.5-99.5? I don't believe there is a unique 100th quantile in theory, at least if it is taken to mean non-exceedence percentile.

The quantile matching was done for the quantiles from 0 to 1 in bins of 0.02. This was corrected in the revised version.

869-6: filter -> 'rainy day threshold' or some other more descriptive term; filter is a little unclear.

good suggestion, filter was changed to "rainy day threshold"

869-7: describe the problem that this step (bootstrapping followed by averaging) solves – ie, a single empirical CDF mapping can be jagged due to inadequate sampling?

Yes, the reason for doing boot-strapping was to minimize the sampling error

869-7: what is done with SYS4 forecast values that lie beyond the extremes of the SYS4 CDF? Or because the mapping is done only for forecasts that contributed to the SYS4 CDF, this does not occur? In real-time application, this might have to be addressed unless the CDFs were recalculated each forecast time to include the latest.

In the current setup they would fall in the highest percentile, which does cap the precipitation amounts. However since we are interested in dry spells, this would not be a problem.

869-29: it would be helpful to add one sentence that describes why biases in features like dry spells and precip intensities are biased in NWP or climate forecast model predictions.

We added a sentence to make this clearer: "Precipitation is a non-linear and intermittent process, and many atmospheric general circulation models (AGCMs) are not able to correctly resolve these processes, for example the number of rainy days and heavy precipitation events, due to constraints in resolution and how the processes are implemented in the model. With increasing resolution and better descriptions of model physics the modelling of precipitation will improve in future model versions (Haiden et al, 2014).

870-5: grammar – phrase starting with 'Firstly' is not a sentence; perhaps remove 'the fact that'

The sentences were changed to: The threshold accounts for a number of factors, for example that part of a day's rainfall is intercepted by canopy, understory, litter and by the very top few centimetres of soil; and therefore evaporates before infiltrating down to the root zone (Savenije, 2004; de Groen and Savenije, 2006; Gerrits et al., 2007).

870-13: evaporative cooling from intercepted canopy moisture and increased humidity do, however, influence the crop – though the roots do not receive water. Minor point.

Thank you for pointing that out

870-20: If a flow chart of the processing can be given, including both the forecasts & obs, and the BC and the thresholding, that would be helpful in illustrating the experiment.

A schematic flowchart of the processes and models were added to the paper.

Also, "This results for each lead time and for each year of available data, in. . ." – perhaps also for each threshold? Are the thresholds applied identically to forecasts and obs pcp? If so, clarify the logic. I would think a more obvious strategy is to apply it only to the forecasts as a 'calibration' step for dry-days/spell lengths, to be verified against the obs (EGPCP). If at the same time the obs are adjusted, the dry-day calibration has a moving target, and the skill of the SYS4 may not be optimized. Clearly the EGPCP at large scales will have biases in dry day and other pcp characteristics relative to local observations and crop impacts, but massaging and interpreting that relationship is a different topic and objective.

Yes, we agree with the points raised. The threshold was only applied to the forecasts, not EGCP. This was clarified in the revised version

871-12: 'poor man's ensemble' is a somewhat jargony way of saying a climatological ensemble, composed of members drawn from the historical observations for the forecast calendar period. I have seen the phrase used in other ways in forecasting (ie, to denote an ad-hoc collection of single-value forecasts such as control runs from different NWP models). Perhaps it's not quite correct here?

We agree, using poor man's ensemble is ambiguous; this was changed to climatological ensemble throughout the paper.

872-1: 'distribution' might be clarified by the qualifier 'magnitude' or 'amount', since in the experiments, temporal distribution is also a feature of interest.

The term "distribution of precipitation amounts" was added to better describe this

3.1: I did not expect the QM alone to affect the dry spells because the text earlier gave the impression that only forecast rainy day amplitudes were corrected (thus zero rainfall forecast days remained zeros, and the precipitation frequency would be unadjusted).

The QM alone does not affect the dry spells. In figure 5, you can see that for no threshold, the skills of QM-matched and raw forecasts are identical. The results where they differ is when a threshold was used. We will make it clearer in the paper when the threshold was used.

The next few comments go back to the earlier section to suggest further corrections, the first of which is organizational.

2.3: this section presents both forecast data description and a method description – it should stick to the former and end on line 14.

Thanks for the suggestion, the section was split in two parts.

868-15: the remainder of section 2.3 should be moved into one devoted to 'forecast improvements' or some appropriate title – which would also include the dry day threshold text. It could be a two part section on 'amplitude correction' and 'calibration of dry day threshold'. A possible third or final subsection could be 'experimental design', basically including material after 870-20, describing the various trials that you assessed, and including

the schematic of the flow of data & adjustments and end-point assessments (dry day, dry length).

Thanks for the suggestion, Chapter 2 was the revised version was divided into three main chapters: study basin, data description and experimental setup. some parts of chapter 2 was also moved to the introduction.

869-5: the statement that only rainy days were corrected appears to be misleading, as it somehow also changed the frequency of rainy days, hence dry spell length. The CDF mapping must have been allowed to translate rainy to dry days or vice versa, to change their intermittency, correct? The typical problem is that the forecast model drizzles, and the frequency of non-zero days is too small. This is simply handled by applying QM to the entire distribution including zero-precip days – in which case some of the model's rainy days map to zeros in the lower quantiles of the obs CDFs. In the alternate case, in which the model is too frequently dry, zeros with a non-unique quantile (eg every quantile below obs (1-prob. of precip)) must be mapped to a quantile range that contains both zero and rainfall amts. This can be done with random estimates of quantiles within that range. In any case, it's not clear from the text how this issue is handled, and the authors must supply more detail. Related to this point, if only rainy days are mapped, why do the CDFS in Figure 3 (top row) not include the full probability range from (0,1)? My guess is that the QM quantiles were established using all days, but only the rainy day portion is shown, which would also explain the altered frequencies.

No, only rainy days were corrected, so the intermittency was untouched by the QM. The threshold was applied in the secondary step to adress the effect on the length and number of dry spells. This was more clearly explained and addressed in the text, as it was pointed out by both reviewers. CDFs in Figure 4 shows only the part of the cdf which contain precipitation values. The results indicate a better forecast if using thresholds, but also the CRPSS values are quite low, so the interpretation of the improvement is not straight-forward. Adding mode skill scores, like a contingency table will bring more clarity to this issue.

873-5: there is something missing in this sentence. "after the performance is comparable to the other areas"??? Can the authors give more diagnosis of the spatial variation in results? What is it about the distributions or forecasts in different areas that gives such a variation?

The text was rephrased to "The forecasts has the lowest skill over area 4, which also is the area least sensitive to droughts, but after the bias correction the skill scores are comparable to the other areas."

Table 1 – can the authors explain why the frequency of dry spells forecasted becomes worse with the correction for the first 3 lead times, despite the fact that their length forecast has improved? Perhaps plotting the mean of the forecast ensemble for each metric versus the obs as a scatter plot would illustrate some basic features of the impact of the correction and the thresholding.

The number of dry spells increase in general with the application of the bias correction, therefore the "worsening" in the mean as shown in table 1. As can be seen in Figure 5, the skill of the ensemble increases, which is the most important thing, especially for longer lead times. More statistics will be added to table 5 to be able to better discuss the effects of bias correction on the mean.

Fig 5 – I don't see intuitively why the skill scores for the raw forecasts with a correction fall so dramatically as the threshold rises, and this should be diagnosed and explained. The CDFs suggest that rainfall amounts are not badly biased (most correction factors not that far from 1). Please diagnose this result more completely. In addition, it would be worth comparing to the case in which the threshold is applied only to the forecast (both with and without QM) for verification against the non-thresholded obs. That, I think, would resemble most other applications of post-processing that I've seen.

The skill scores does increase with threshold, and to fully discuss this we will show more the effects of the thresholds on the mean of the ensemble, and we will also add to Figure 5 the forecast with just the threshold with and without QM to show the effect of bias correction by itself. We will also add more skill scores to the paper based on a contingency table to further aid the discussion. Figure 6 will be dropped since it does not really add anything to the results. The areal filtering will be mentioned but not discussed.

References

Haiden, T., Magnusson, l., Tsonevsky, I., Wetterhall, F., Alfieri, L., Pappenberger, F., de Rosnay, P., Muñoz-Sabater, J., Balsamo, G., Albergel, C., Forbes, R., Hewson, T., Malardel, S., Richardson, D., 2014, ECMWF forecast performance during the June 2013 flood in Central Europe, ECMWF Technical Memoranda, 723, 34 pp, Reading, United Kingdom.