We thank Anne van Loon for the insightful and valuable comments, which we try to address to our best.

1 General comments

It is important to be very clear and consistent in terminology. You use the term “low- flow events” for periods in which streamflow is below a seasonally-varying threshold. Deficiencies in the high-flow season (in the studied catchment the snow melt period) should, however, not be called low-flows as streamflow can be quite high in absolute terms even though it is below the threshold. Therefore, I advise you to use a more appropriate term, such as for example “drought”, “anomaly” or “deficiency” instead of “low-flow”. See Tallaksen and Van Lanen (2004). Please change it throughout the entire manuscript.

A: We appreciate this suggestion but we think that all of the suggested alternatives have some weak points, just as “low-flow”. “Drought” is generally used to describe a phenomenon ranging over many parts of the ecosystem, in our study we focus explicitly on streamflow. “Anomaly” and “deficiency” are, like “low-flow”, only meaningful w.r.t. some reference or threshold. We are aware that the season varying threshold reaches high values in the melt season that would not be critical during the summer season. We interpret low-flow as low w.r.t. what people have adopted to in a certain time of the year. This point and our interpretation of low-flow is addressed in section 2.4..

In this study, the longest drought in the forecast period is taken as the event for further analysis. Why did you use the longest event and not the most severe or most intense or . . .? Choosing the longest event gives problems with truncation of the event by the end of the forecast lead time. This is clearly shown in Fig.2, where you illustrate the threshold level method. The duration of the drought event that is chosen for further analysis (in this case the event on the right-hand side of the figure) is highly influenced by the maximum lead-time of 32 days. The severity is less influenced and the intensity (maximum deviation from the threshold) even less. So taking the most severe or most intense drought, instead of the longest, would give less impact of the used methodology on the results of this study. Can you please indicate what the effect of a different selection criterion for drought would have on the results of your study?

A: About 16% of all low-flow events in the study period are truncated by the limited forecast range and hence not fully captured by the forecast. We are aware of this but, dealing with forecast of up to 32 days, this is something we cannot avoid. However, this percentage might be of interest for the readers and we implemented it in the manuscript (we refrain from using the term “highly influenced”).

To further address this point we performed our analysis based on choosing the most severe event or the event with the highest magnitude in the forecast period, in case there are more events detected during the 32 days forecast period. We show in Response Figure 1 that the results do not change strongly. This can be explained by the high correlation of the length of events with the severity or magnitude. We find a correlation of r^2=0.85 for mean forecast duration with mean forecast severity and r^2=0.79 for mean forecast duration with mean forecast magnitude.

Additionally, the duration of a drought event is, in your study, also highly dependent on whether or not the streamflow signal reaches just above the threshold. The occurrence of small peaks that divide a long drought into two separate droughts strongly determines which drought is chosen and what the characteristics of that event are. You already mention that effect as a problem when you argue the use of a larger catchment (p.6861). A normal procedure in drought research to avoid this strong impact of small peaks is pooling. Various methods are possible, e.g. moving average, inter-event time/volume criterion (Fleig et al. 2006). Please consider applying a pooling method prior to selecting the drought event for further analysis.

In the manuscript, I miss some detailed examples of the performance of drought fore- casting.

A: We appreciate this comment as it helps to clarify the focus of out manuscript. The adjustment of low-flow index time-series for small interruptions is certainly important whenever runoff is considered as representative for large-scale drought events, which also impact other parts of the ecosystem, as runoff itself is more responsive than e.g. ground water. Hydrological droughts, i.e. streamflow drought, as treated in our manuscript, cannot claim this representativeness.

Concerning the suggested smoothing of data in order to prevent short-term interruptions of low-flow events, our forecasts and observations are already daily averages. A further smoothing would in our mind result in a too strong loss of variability, compared to the 32 day forecast range. An argument for not applying a stronger smoothing might be that a low-flow interruption by one day may for some forecast users already be sufficient long to take preventive action. Anyhow, as we do not address a certain type of drought-affected sector or user, we do not think a further smoothing, that would introduce some further subjectivity, would increase the value of this study.

Fig. 7 only gives a very general overview and, according to my view, only limited skill. I would like to see the forecasting skill for two or three cases, for example the 2003-drought (mentioned in the manuscript, but not illustrated) and a drought event/period in the 1990's. For these detailed examples then also the forecasting of timing (i.e. onset and termination) of the drought can be evaluated, because these are quite important features of drought and especially termination is very hard to predict due to the high persistence of drought.
A: Figure 7 actually shows all the cases incorporated in the following verification. Figures 8-10 gives the objective verification results of the forecasts shown in Figure 7. Showing some selected cases is not representative for the overall performance of the forecast system.

Finally, I would like you to explore any seasonal differences in forecast skill. As winter droughts are caused by very different processes than summer droughts (Van Loon and Van Lanen, 2012), it would be very interesting to see whether one of them is easier to predict. That gives an indication which processes should be improved in the modelling framework to improve forecasting skill of droughts.
A: We explored seasonal differences in predictability but decided not to include the results in the publication du to the large uncertainty in the verification scores associated with the limited number of observation when further stratifying the data.
The verification shown in Response Figure 2 suggests some seasonal dependency of forecast skill with higher scores of forecasts starting in the autumn and winter months as well as June/July. This double peak makes the attribution to any process happening in a specific season difficult. We would rather leave the investigation of seasonal differences in predictability subject of further research.

2 Specific comments

Abstract
Please provide more information on results (including numbers) in the abstract.
A: The main verification results are now mentioned in the abstract.

Introduction
You mention the work of Wood et al. (2002), Luo and Wood (2007) and Li et al. (2008) as studies that use a coupled atmosphere-hydrological model for the long-range prediction of drought. They managed to give reasonable predictions up to several months in advance. You do not clarify what your study adds to these results. What is new in your research? Even more because you evaluate forecasts only up to one month, which is less than in the studies mentioned above.
A: Pointing out the new contribution of a study compared to the already existing research in the field is crucial and, apparently, we needed to improve on that part. Compared to the above-mentioned studies on seasonal drought prediction our study is restricted to a comparatively short forecast range of one month. However, we evaluate runoff forecasts on a daily basis, which has not been done in that forecast range, and the derived forecasts of low-flow scenarios duration, severity and magnitude, which is entirely new. The mentioned studies are restricted to monthly values streamflow, some with a focus on droughts or low-flow. Therefore our study address a partly different user sector and we think that those forecasts, even if they reach "only up to one month", are still of interest. Besides that, our results are based on a comparatively long dataset of 18 years of ensemble reforecasts allowing for a more robust estimation of forecast skill. We now mention these points in the last paragraph of the introduction.

Data and methodology
Section 2.1: Please give more quantitative information about the catchment. What are "relatively cool conditions"? Give yearly average and minimum and maximum monthly temperature.
A: A specification of average temperatures and precipitation in a catchment with large vertical extension is only of limited value. However, we implemented climatological information (Response Figure 3) for both parameters as catchment average and refer to Gurtz et al. (1999) for more information about meteorological variables.

Section 2.3: Please give more information about the model. Provide a short summary of the papers you mention on PREVAH physics, parameterization and downscaling, and the papers on the calibration and verification against observations. Also mention the dates for the co-called “extended reference period” and the details of the “meteorological surface observations” (What? Where? When?). Also describe somewhere in this Section, or in Section 2.1, the runoff gauge in Andelfingen (should not be introduced at p.6867). Is data of this gauge used for calibration? Are observations of state variables used in calibration? And please provide the Nash-Sutcliffe value of logQ besides the mean error, as this metric is much used for evaluating model performance on low-flows.
A: We included the additional suggested information in section 2.1.

Section 2.4: Here, you mention that a seasonally varying threshold was used, but Fig.1 and 2 seem to show a daily varying threshold and in the caption of Fig.7 you mention a monthly threshold. How was the threshold calculated? Did you define seasons by date? In this section you should describe that the quantile used for calculation of the threshold is selected later, based on your results, and what the criteria for selection were.
Furthermore, you state that the “lead-time is no longer a possible source of forecast error”. This is not correct, because the limited lead-time causes a truncation of drought events, and therefore influences prediction of the longest drought (Fig.2).
A: The introduction of the threshold was reworked. We do not use the term seasonal w.r.t. the threshold anymore as this has caused confusion. Instead we only refer to a varying threshold.
A valid point is the possible influence on the forecast error by the truncation of events after 32 days. In our forecast data 28% of the events are still prevalent at a lead time of 32 days (which does not necessarily mean they would also extend further). Of the observed events 19% are prevalent at forecast day 32. We would argue that the skill is less affected by timing errors.

Section 2.5: Please define above which score you regard a forecast to be skillful/beneficial. From Section 3.1 I understand that you denote a forecast as skillful above 0.55 or 0.6.
A: We have reworked the verification scores section, pointing out more clearly the characteristics of the 2AFc score and its interpretation. Values above 0.5 are considered skilful. Depending on the underlying data the reached scores are not significantly above 0.5. In the revised version we changed the colour of those scores to white, showing when the forecast system start to lose its predictive skill.

Results
Section 3.2: Why is the threshold quantile chosen based on forecasting results? For a real forecast this cannot be done, because no observations are available to test which quantile gives best results. Furthermore, in this way it is not related to any user requirements. Please discuss this issue in your manuscript. In the final choice of the 15th quantile, you mention that it is a compromise between the number of drought events and the drought forecasting skills. Why is the number of events such an important issue? If it is so important, mention it already in Section 2.4 and include the number of events as an extra column in Fig.5.
A: Choosing a quantile based low-flow threshold for forecasts based on forecasts is, as has been written, a simple way to correct for potential systematic forecast biases. Basically this is nothing but an additive bias correction depending on the lead-time of the forecast. It could very well be done for real forecast, all that's needed is a training sample of past forecasts. I.e. whenever the requirements for a statistical post-processing are fulfilled, this kind of threshold can be applied as well. In our case this is possible because we study an 18 years long reforecast dataset, giving a sound sample to choose an appropriate threshold.
The number of events is crucial for the robustness of the verification results. The more events can be verified, the more robust the scores will be. More events could easily be obtained by increasing the threshold, this however would make the study less relevant w.r.t. drought related low-flow. We now give the number of observed events associated with each tested low-flow threshold in Figure 5.

Furthermore, you say that the threshold shows a minimum in October/November, when snow accumulation starts. So, in this catchment winter temperatures are just below or around zero in winter, so that occasional melt takes place? Or is this minimum related to seasonality in precipitation and/or evaporation? If temperatures would be far below zero in winter (like in the Scandinavian countries), then the threshold would decrease in winter and show a minimum just before the snow melt peak. In Fig. 1 there is a second minimum in February. Please clarify what the causes are for seasonal changes in the threshold.

A: As the Thur catchment covers a wide range of altitudes, hence climatic zones, it is not straight forward to attribute all maxima/minima in the gliding runoff quantiles to driving processes. Also note that the shown distribution of runoff is based on data from the forecast period 1991-2008 and therefore shows more variability than a long-term climatology using e.g. 30 years or more. We can attribute the maximum in April to snow melt and the minimum from October to February to snow accumulation and less precipitation (see also Response Figure 1). This is now addressed in more detail in the domain section.

Section 3.3: "It is striking how well the duration and severity of the observed low- flow is contained within the range of the ensemble". Be careful! I would not consider the resemblance strikingly well, especially not during the periods without events in observations.
A: We have corrected this and the forecast property of over-predicting is mentioned here, as well as in the following paragraph discussing the rank histograms.

Section 3.4: "For all users, value scores >50
A: We are not sure in what way the manuscript should be changed here.

3 Technical corrections

p.6858, line 17: Do not use the term "indicators" for drought types. Rather use "types" or "processes".
A: In the new version we use "processes".

p.6859, line 28: "although" should read "however"
A: Replaced

p.6860, line 1-4: Sentence not clear, consider revising.
A: The sentence was shortened and should be clearer now.

p.6861, line 18 20: Please explain the difference between the 954 ensemble forecasts and 5 members in the reforecast. I, as a layman in forecasting, cannot understand this.
A: This section was worked over. It should be clear now that the 945 forecasts are from the weekly initialized 18 years reforecast and that each of the 954 forecasts is a 5 member ensemble forecast.

p.6863, line 1-3: Provide references in chronological order (also throughout the rest of the manuscript).
A: Done

p.6863, line 20: "A lead-time dependency of the low-flow threshold was implemented for the forecasts." What do you mean? Please rephrase.
A: This paragraph was edited, also considering the comments of referee #1. We addressed the threshold selection for the forecast already in in the general comments section.

p.6865, line 11-12: Please explain what you mean with "continuous observations" and "dichotomous observed outcomes".
A: We replaced "continuous observations" with "observed runoff" and explained dichotomous ($\in \{0,1\}$)

p.6865, line 18-19: Move these sentences to Introduction.
A: This is already part of the introduction. We like to remind the reader of the motivation and prefer to keep the remarks at this part of the study.

p.6866, line 13: "relative absolute error"?
A: Replaced by: predicted runoff at initialization deviates less than 25% from the observed runoff

p.6866, line 15: What do you mean with "beneficial"?
A: Replaced with "contribute to forecast skill"

p. 6867, line 22, "varying quantiles": Please rephrase. Quantile stays the same over the year, but threshold itself varies.
A: (Non-exceedance) probabilities stay the same, quantiles vary. When talking of Q15 or the 15th quantile the varying quantile associated with the 15% non-exceedance probability is meant.

p.6870, line 24: Sect. 2.3 > Sect. 2.4
A: Sorry, no changes as we are not certain what is meant.

p.6870, line 25: by > be
A: Sorry, no changes as we are not certain what is meant.

figures
Fig.1: Emphasize line of Q15 (chosen quantile for threshold)
A: Done

Fig.2: Move information in caption on calculation of severity, magnitude and timing to main text.
A: The text has been reworked, also w.r.t. the comments of referee #1.

Fig.3: Give more information in caption. What do we see? Is it the forecasting score using different quantiles as threshold? And what happened to the ensemble? Is this the ensemble mean? Or the best prediction? Or . . .?
A: By using the term "probabilistic" we thought it is clear that probabilities to exceed threshold are verified. We extended the caption to be more explicative.

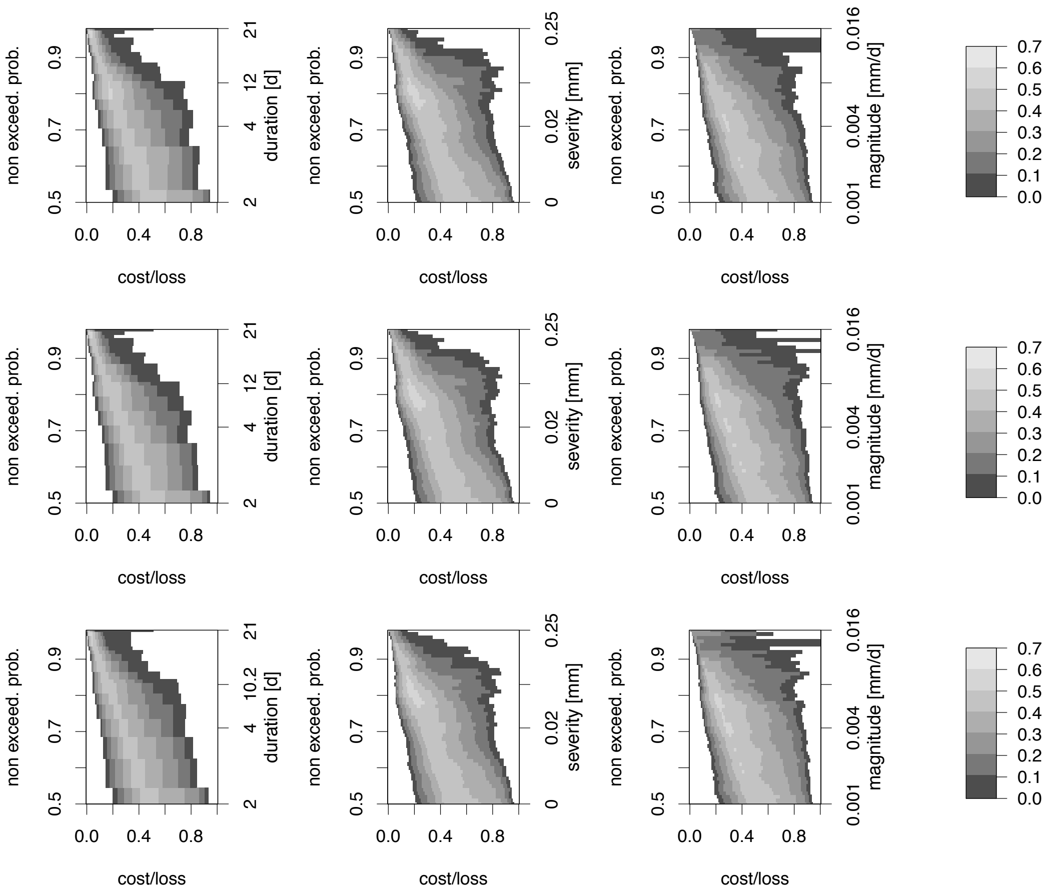
Fig.4: Please show also the lines that are now hidden behind a polygon. Resampling of 1000 times was not mentioned in main text (Methods).
A: Corrected

Fig.8: What is on the x-axis?
A: The x-axes are labelled and are additionally mentioned in the caption. We changed the caption to clarify the rank histogram, for further explanations we would refer to the cited literature.

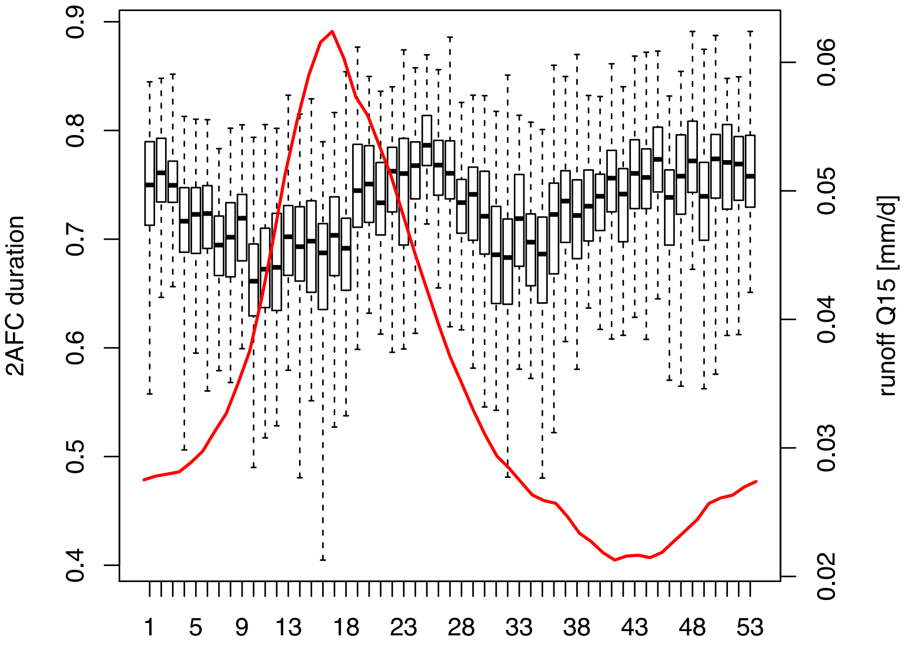
Fig.10: What are the white and grey dots?
A: There are no white dots. Can this be an optical illusion due the changing background colour?

4 References

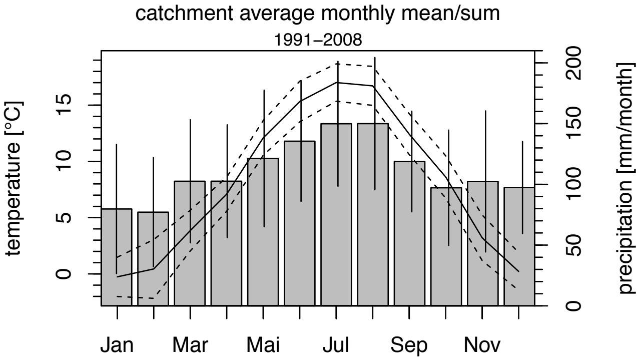
Fleig, A. K., Tallaksen, L. M., Hisdal, H., and Demuth, S.: A global evaluation of stream- flow drought characteristics, Hydrol. Earth Syst. Sci., 10, 535–552, doi:10.5194/hess- 10-535-2006, 2006.
Tallaksen, L. M. and Van Lanen, H. A. J.: Hydrological drought: processes and esti- mation methods for streamflow and groundwater, Developments in Water Science 48, Elsevier Science B.V., The Netherlands, 2004.
Van Loon, A. F. and Van Lanen, H. A. J.: A process-based typology of hydrological drought, Hydrol. Earth Syst. Sci., 16, 1915–1946, doi:10.5194/hess-16-1915-2012, 2012.



Response Figure 1: Comparison of the results given in Figure 9 of the original manuscript using different approaches to detect low-flow events in the forecast time-series. Top row: the original approach using the longest event within the forecast period in case that more than one event occurs. Middle row: using the most severe event. Lowest row: using the event with the highest magnitude.



Response Figure 2: Variation in the forecast quality of low-flow duration using the 2AFC score. Each box comprises data from a week±1 of the years 1991-2008. The box ranges are found by 100 times resampling with replacement of the underlying data. The red line shows the applied low-flow detection threshold.



Response Figure 3: Thur catchment average mean monthly temperature (solid line) and precipitation (bars) sum in the study period 1991-2008. The dashed lines and vertical bars show one standard deviation.