Review of “Streamflow drought: implication of drought definitions and its application for drought forecasting” by Sutanto and Van Lanen.

The Study of Sutanto and Van Lanen compares different drought identification approaches: 1) the fixed threshold level method, 2) the variable threshold level method and 3) the threshold level method applied on SSI time series, for simulated river flow at the pan-European scale. They show that (average) drought event characteristics differ based on the used drought identification method. Consequently, they show that drought event forecasts differ, depending again on the used drought identification method. Overall, the main recommendation of the paper is strong and relevant, i.e., droughts differ depending on the used method and streamflow drought forecasters and stakeholders should agree which type of drought should be forecasted. In addition, I believe that Figure 6 provides an informative message for the users and developers of hydrological drought forecasting systems.

However, given that this paper focusses on the definitions of drought and methodology of drought identification, it sets an example which types of drought identification approaches can be used for drought forecasting applications (and how). Therefore, it should be extra “sharp” in its drought definition and identification approaches as well. At this stage, this is not the case and there are several methodological concerns that should be addresses carefully. In addition, the comparison of the results is far from straight forward. The used drought identification approaches do not only vary in overall method, but also in: 1) threshold (<10 percentile for the fixed and variable threshold approaches and around <50th percentile threshold for the SSI), 2) data accumulation period (1 month for the fixed and variable threshold based approaches vs. 6 months for the SSI), and 3) temporal resolution (daily vs. monthly). Finaly, the most novel part of this paper, which deals with the implications for drought forecasting, is rather limited and deserves more attention in my opinion.

Major comments:

Methodology

SSI computation:

Why SSI-6? For me, it makes sense to aggregate meteorological drought indices (SPI, SPEI) to differentiate between slow and fast responding (hydrological systems), e.g., catchment with small and large storage components. However, riverflow already encompasses the accumulation and delay of the meteorological signal caused by e.g. delayed groundwater flow. From a riverflow drought perspective, it is often important to know what is currently happening in the river (SSI-1) and not what happened in the past 6 months (SSI-6). Also, the SSI-6 is not at all comparable to the 30-Day moving window used for the FT and VT approaches. This makes the interpretation of the comparison between both approaches less straight forward. Finaly, the reasoning to choose the SSI-6 over the SSI-1 because the SSI-1 results in many minor drought events does not compensate for the advantages of the SSI-1.

Why an SSI threshold of zero to identify drought? I would not term something that happens 50% of time drought. Please note that the original SPI paper of Mckee (1993) uses a similar threshold, but has the additional requirement that the SPI should at least reach a value of -1 over the course of the drought event. In addition, an SSI threshold of zero is far from comparable to an FT or VT of Q90 used for the threshold level approaches.

Why the gamma distribution to derive the SSI? I agree that is hard to find a suitable distribution to fit to riverflow time series (line 150-151). However, that is not a good argument to simply use the Gamma distribution. There are likely to be better alternatives for your pan-European dataset (See e.g. Svensson et al., (2016), Tijdeman et al. 2020).

Why no goodness of fit testing? The studies above concludes on different suitable candidate distributions for the SSI (other than the gamma distribution) that might be applicable for the current study. However, that does not mean that they can be applied on your dataset of simulated streamflow series by default, as your dataset might exhibit different properties as compared to the observed riverflow timeseries. Careful evaluation which distribution is most suitable for your set of rivers is required.

Which distribution fitting method was use?

For the forecasted SSI: Did you use the parameters of the population distribution derived from historical monthly flow values to derive the SSI for forecasted values? Or did you replace the historical values with forecasted values and than recalculated the population distribution to derive the SSI? And why, e.g., what should a forecaster do?

Threshold approach:

Line 123-143: Many different smoothing procedures have been applied in combination with the threshold level method. This has been done for good reason, however, sometimes resulting in an (unwanted) increase/decrease in drought occurrence, especially for the VT method. For me, a 10th percentile implies that 10% of the time series is in drought and that drought occurrence is equally distributed over the year in case of the VT method. However, by first deriving the the threshold from daily streamflow data, and then smoothing both the threshold and riverflow timeseries seperately, this is not necessarily the case anymore. This might be solved relatively easily, i.e., first apply the moving average and then derive the threshold. Or you could use monthly data.

Line 366-367: You encourage using monthly streamflow data for drought forecasts but use daily streamflow in your own analyses. I would have find it logical to do this as well in this study, e.g., instead of the FT and VT approaches applied on daily data, it could be applied monthly averaged data. This also increases the comparability with the SSI. Further, is there really merit in forecasting streamflow drought duration and deficit at a daily resolution, especially for the longer lead-times? Is this being done somewhere? Can this be done with any skill? If not, wouldn't it be better to just stick to monthly data for which at least some skill might be achieved?

Results and discussion:

Section 3.2. The forecasting section, which is the most the novel part of this paper, would benefit from some more attention. Figure 6 provides a nice illustration, even though it might be a little obvious at this point in the papers that drought characteristics derived with different methods will vary, given that you apply a different threshold on the same forecast data. However:

- I disagree that the drought of 2003 in the river Rhine started in August 2003. According to the SSI-1, river levels dropped to below normal anomalies much earlier. I suggest to start earlier in the year.
- Why not add the observed hydrograph to the plot?
- Isn't the fact that the VT method does not forecast a drought a good thing? According to this method, there was also no drought in the observed hydrograph (Fig. 4a) – how could this method have "performed better" (line 340).
- Why not show the SSI-1 here?

Given the focus of the paper on river flow forecasts, I would expect more focus on the latter, and not only an examplary timeseries river flow forecasts for one river / event. It would be interesting to include.

- At least, an evaluation and discussion of the spread in streamflow forecast and especially in the spread in streamflow drought forecast, and (i.e., not only the evaluation of the median forecast). What are the ranges in drought characteristics derived from the forecast ensemble?
- Consequently an evaluation or discussion of the streamflow (drought) forecasts skill, i.e., can certain “types of droughts”, e.g., FT vs. VT vs. SSI, be forecasted better?

The above evaluation would benefit the consideration of multiple rivers, drought events, or start months.

Again, I would avoid the SSI-6 here, due to the strong autocorrelation of this index, which makes it relatively easy to forecast on short lead times. For example, for a forecast with a lead-time of 1 month, 5 out of 6 months are already known. Rather, I would look at the SSI 1.

Finaly, some (non-committal) suggestions for Section 3.1 that could further improve the manuscript:

Section3.1.1 Next to showing the amount of streamflow droughts, you could consider showing other characteristics such as the average duration, deficit volume, or the number of minor drought events. This provides valuable insights in differences between methods, and further makes the notions in 3.3.1 about regions with more minor drought quantitative. In addition, you can derive a proxy for deficit volume from standardized time series. The units are meaningless and not comparable with the deficit volumes derived with FT and VT method. However, the relative difference over Europe should pop-up.

Section 3.1.2 In addition to discussing when most drought starts, it might be interesting to see when most drought occur in difference climates. This can be presented as a series of histograms for each climate, with the month on the x-axis and the fraction of drought months that occurred in that month on the y-axis.

Minor comments:

Line 2: “… the term streamflow drought forecasting, rather than streamflow forecasting …” You could briefly explain difference between the two here.

Line 5: “within” Correct?

Line 6: Be careful with terming these extreme events. They are anomalies, but something that happens on average at least once every year, as is the case in your study, is not an extreme event.

Line 7, 8: “observed” might be “observations”

Line 7: “a LISFLOOD model“… are there more?

Line 10: add method to VT and FT, e.g. variable threshold level method.

Line 10: You also apply a threshold based approach on SSI time series. Mention this here.

Line 16: “Eliminate”. Not true. You can still have 1-day droughts with these TL approaches.

Line 24: “IPCC” should be “The IPCC”.

Line 34: This sentence slightly contradicts with Line 1, where you state that drought forecasting is a key element of DEWS. I would expect there to be some examples. Which contemporary “DEWS” include streamflow drought forecasting, using the approaches as described in the paper (FT, VT and SSI), not just streamflow forecasting)?

Line 41: “evaporation” should be potential evapotranspiration

Line 47: “used” should be “be used”

Line 49: Mention that you specifically focus on simulated streamflow drought.

Line 75: “There” should be “There is”

Line 85: “Proxy” should be “Proxies”

Line 89: “proxy observed streamflow“ could just be “simulated streamflow”

Line 112: “reforcasted data 2003” should be ”re-forecasted data of 2003”

Line 119: “in” should be “for”

Line 128: “were moving averaged” rephrase

Line 134: “For the threshold” …this refers to variable threshold approach I guess? In this section, make the clear distinction between FT and VT and seperately explain how both are derived.

Line 138-140: add here that MA introduces a significant amount of auto-correlation, which affects the skill of the river flow forecast for the first 30 days significantly.

Line 147: “median” should be “expected median”.

Line 155-160: Add here that it is quite easy to forecast the SSI-6 for short lead times, given the strong autocorrelation of the timeseries. E.g., for 1-month lead-times, you already know five months and only have to forecast one.

Line 162-164: Please explain how you classify an event with varying SSI values into one category.

Line 162-177: Did you derive the climate classification yourself using the approach described in Peel et al (2007)? Or did you use their dataset?

Line 179: “definitions“ … “drought identification approaches” might be better.

Line 188: “Lower than median streamflow” … Not nescecairely true. Technically, above median streamflow can still be a negative SSI and vice versa. Depends on the sample and (goodness of fit) population distribution to derive the SSI.

Line 189: Figure 3 does not show that streamflow droughts occur every year.

Line 200: This is comparing apples and pears, as the thresholds are completely different.

Line 203-206: Could this not be compensated by a higher number of drought in winter for the VT?

Line 221: “drought that has” should be “droughts that have”

Line 228. “(Coincides with hydrologic years in most of Europe)” remove: unneeded repetition

Line 264-266. Is the last part, i.e., about the lowest and n-day minimum flow, needed? Interrupts flow.

Line 266-267. Looking at Fig. 5a, I find the SSI-1 timeseries much more informative about drought in the river Rhine. Rhine drought reaches is maximum in summer 2003, and recovers in winter 2004. For me, this make much more sense than the SSI-6 timeseries. Was the drought in the river Rhine really a multiyear event? Were there impact directly related to Rhine river flows over the course of 2004?

Line 270. For me, this description of drought in the river Rhine makes much more sense. It would make even more sense if you would use a more appropriate drought threshold (maybe SSI-1 < -0.84, corresponding to the 20th percentile). I don’t see the problem of having 2003 split up in different events and question why it is better to use an SSI-6 and thereby inflate the event to a multiyear drought.

Line 285: “C” should be century.

Line 295-302. Why limit yourself here to the four case study Rivers and the limited time window? You could directly compare the number of drought events & their deficit volumes over a longer time period and for all the catchments (starting by deriving the difference between Fig. 2a and b).

Line 312-329. According the definition of drought according to VT and the SSI, droughts are expected to occur for an equal amount of time over the year. Please provide an explanation for the distinct temporal differences in drought occurrences. Or is this still referring to the start month of the drought?

Line 309: “(except for the Rhine River)” this contradicts with the discussion in the paragraph above.

Line 337: Not only meteorological drought, also streamflow drought according to the SSI-1 (Fig. 5a).

Line 354. Which is good, because there was no drought according to the VT, or?

Line 382: “eliminate” … not correct as minor droughts can still occur.

Line 361: “rare extreme drought events” … extreme events are by definition rare. Rephrase.

Line 368: “the FT method produces higher drought deficit volumes and duration than VT” not shown for the pan-European dataset.

Line 375: “occurred” should be “started”.

Line 377: “what being identified by” rephrase

Figure 1: Nice. What is the difference between light and dark grey in e.g., the Alps?

Figure 2: You could add the upper boundary, e.g. 30-xx instead of >30.

Figure 3: “The timing for drought was determined based on the first month of each drought event.” This is the same as what is said in the beginning of the caption.

Figure 4: Some droughts are hardly visible (e.g. in Figure 4a). It might work to use a log-scale

Figure 4: Axis lables: m^3 sec^{-1} or m^3 / sec instead of m3/sec

Figure 4: Are the grey vertical lines the hydrological years?

Figure 4. You might consider using a different color when VT and FT overlap.

Figure 5. Add grey vertical lines here as well.

Figure 6. Same comments as for Figure 4 and 5.

Table 1. Would be interesting to also compare average deficit volume and timing.

References:

Mckee, T. B., Doesken, N. J., and Kleist, J. (1993): “The Relationship of Drought Frequency and Duration to Time Scales.” *AMS 8th Conference on Applied Climatology*

Svensson, C., Hannaford, J., and Prosdocimi, I. (2016): “Statistical Distributions for Monthly Aggregations of Precipitation and Streamflow in Drought Indicator Applications.” *Water Resources Research*, https://doi.org/10.1002/2016WR019276

Tijdeman, E., Stahl, K., and Tallaksen, L. M. (2020) “Drought Characteristics Derived Based on the Standardized Streamflow Index – a Large Sample Comparison for Parametric and Nonparametric Methods.” *Water Resources Research*, https://doi.org/10.1029/2019WR026315