Review of: „Spatiotemporal development of the 2018-2019 groundwater drought in the Netherlands: a data-based approach“ by Brakkee et al.

The authors use an impulse-response time series modeling (TSM) approach to estimate groundwater heads over a larger region. By doing so, they were able to provide a much denser network of point-based groundwater drought estimates over a larger region, compared to using solely groundwater observation. The paper covers various interesting topics, including groundwater data quality assessments, groundwater time series grouping, groundwater drought analyses, and the „prediction“ of groundwater drought. In addition, it suggests that some of the dominant controls on groundwater drought development are the spatial variability in precipitation and differences in geological-topographic conditions.

The TSM approach applied on the regional scale to increase the spatial density of groundwater drought information is novel (as far as I know) and I am impressed by the large amount of data that is collected and used for the analyses. That being said, there are quite some methodological aspects that should be better, explained, justified, or evaluated (listed below). In addition, all the proposed analyses could make for an interesting paper. However, at this moment, few of these analyses are thorough enough to generate some more complete conclusions. Below, I provide various suggestions how the analyses could be elaborated, but the authors may have some different ideas. If the additional analyses make the paper to lengthy, the authors may choose to drop e.g. the prediction part. Besides the major remarks, I have various minor comments that should be addressed.

Methodological aspects (listed in order of appearance).

- Line 159: What is „short-term“ (days, months, years?). And why are these not relevant. Please explain.
- Line 162: Why are long term disturbances relevant for drought studies? Please explain.
- Line 166: Why are time series with long-term changes less reliable for drought analyses (also given the prev. point). Or does this refer to their simulations?
- Line 170-71: I would prefer a bit more detailed description of the PASTAS approach here (I can read it from the mentioned publication, but a short summary of the approach and the used parameters would be easier for readers that do not want to go over all the details).
- Line 173: describe the used non-linear approach in more detail (or delete / move to discussion that you tested a non-linear approach and how).
- Line 176: why a cutoff at >20 cm. Is the reason that GW head can exceed surface level related to inaccuracies in the used elevation data?
- Line 177-178: the use of thresholds like 20% of data and 50% of the measurement ranges should be better justified or evaluated. Also, by focusing on both the wet and dry outliers at the same time, you might either: disregard some GW wells that are suitable for drought but have a lot of variance in the wetter domain or include some less suitable wells that have a high variance in the dry domain and a low variance in the wet domain. This should be discussed.
- Line 180: what is „basic model settings of PASTAS“
- Line 185: better explain A, f and d, and what these metrics mean.
- Line 202 (and other places): The term „aggregated” alone can mean various things. Be precise.
- Line 205: why randomly select 120 GW timeseries and then select 56 from those 120. Why not just select 56 randomly?
- Line 205-206: it would be helpful to show some exemplary timeseries with and without visual errors, preferably in the results (maybe in the supplement).
- Line 209-210: please provide abbreviations of true positive etc. here. Otherwise, few people will understand table 2.

- Line 218-219: please provide more detail about the PASTAS model as applied in this study here or in the previous section.
- Line 232: what are these „practical consequences“
- Line 238: which distribution is used for the SPEI?
- Line 240: why not simply the mean of all stations in a region?
- Line 250**: „**probably insufficient to do this reliably“ either test this or delete.
- Line 256: the original SGI paper of Bloomfield and Marchant (2013) used 1 / (2n) and not 1 / n. Please explain why you differentiate.
- Table 1: values are not correct. E.g., if the lowest SGI is based on 1 / n (1 / 30), the lowest SGI value would be >-2.
- Line 269-272. When comparing observed vs simulated SGI, did you match their record lengths. For example, if the simulated SGI covers 1990-2019, but the observed SGI 1993-2019, did you recalculate the simulated SGI for 1993-2019?
- Line 294: did you combine simulated and observed data here?
- Line 360-361. This should be presented in the method section and more clearly explained.

Results:

- You used plenty of outlier removal approaches, but never really evaluated how these different approaches affected (improved) the simulations. For example, did removing the lowest and highest 20% of values for some wells result in better GW / SGI simulations. I strongly recommend repeating the evaluation presented in table 3 based on data that did not receive (various steps of) preprocessing. Here, it would be also interesting to look at model performance for the different subsets of wells (e.g. discarded, deep, atypical, typical). Finally, it might be interesting to investigate model performance as a function of the available record length. All in all, this would provide a more quantitative assessment of your data quality and outlier removal approach.
- You note that groundwater drought (propagation) characteristics appeared to be dominantly governed by the spatial distribution of rainfall and the geological-topographic setting (abstract). This statement would benefit from some more quantitative analyses. Would it somehow be possible to group wells based on common aquifer characteristics (if not, possibly just the wells in the highlighted areas vs. the wells in the other regions, Fig. 1) and investigate whether different subsets of wells respond differently to the drought of 2018-2020? Such analyses could be linked to the maps of figs. 5-7, possibly comparing the density functions of the different subgroups (like in Figure 4). Here, it might also be interesting to compare the different types of wells again.
- For the usefulness analyses, I strongly suggest evaluating the usefulness of your approach for the different groups of wells (typical, atypical, deep, discarded). Further, I would not purely focus on the drought of 2018-2019, but also on the entire period. In addition to deriving the correlation with SGI observations for all stations at once (Fig. 8), it would be interesting to see the (spread in) correlation of each individual well (and again see if these correlations vary among the different categories of wells).
- The prediction analysis is not really a prediction analyses but more a model evaluation (like you split a hydrological model experiment in a calibration and evaluation period). The section is generally ok, besides the last paragraph (406-411), which contains an interesting statement about the limited value of including the initial conditions for prediction. This statement is contrary to what I would expect from often highly autocorrelated groundwater timeseries. Therefore, it deserves more attention! I assume the stated values are averages? I think providing ranges here would be important, as I assume that there are definitely some wells that benefit from the inclusion of the initial condition. Grouping based on well type might work again (deep wells might possibly be predicted for longer lead-times). Possibly a Figure would be nice here

as well, with on the x-axis the lead time and on the y-axis the „improvement“, showing both the average (e.g. bold line) as well as the ranges (background shade).

Minor comments:

- **Line 1:** „development“: it is not just the development.
- **Line 1:** „2019“ you include 2020.
- **Line 2:** „a data-based approach“. This does not add a lot. I think all GW drought assessments are data based. You could be more specific with regard to the used method (or delete this part).
- **Line 5:** „northwestern“. Parts of central Europe were affected as well.
- **Line 55:** „has so far mainly been studied from a meteorological perspective“. I do not think that this is the case. The reference to Bakke includes hydrological observations and previously mentioned refs. study tree response to the drought of 2018. In addition, a reference of Brunner et al. (2019) would fit here nicely, as it was one of the first studies to investigate the 2018 drought (from different perspectives).
- **Line 66-67:** „few measurement wells“ the mentioned study of Van Loon et al. (2017) used quite a lot of wells.
- **Line 66-67**: a reference to Kumar et al. (2016) might fit here.
- **Line 98-99:** please check if all mentioned references focus on GW forecasting. Also, I am a bit surprised that physical based models can only predict two months ahead, given the often strong autocorrelation in GW signals.
- **Line 119:** please briefly note why these regions are highlighted (higher vulnerability).
- **Line 120:** please provide (ranges in) average precipitation (surplus).
- **Figure 1b.** I would separate based on the four considered regions and not the provinces. Then, you can remove the map from Fig. 3.
- **Figure 1b.** you could show weather and precipitation stations on the map.
- **Line 131:** „overlapping timeseries were combined“ and checked for consistency?
- **Line 134:** „precipitation“ I would define precipitation (P) as you also do this for reference evapotranspiration (ET_{ref}). You could even go for E_{ref}, to be consistent with HESS guidelines.
- **Line 197:** I cannot judge what a negative value of f is and $-0.05 < f < -1.95$ is more concise.
- **Line 205:** would split these 44 further in groups with outliers (1) abnormal long-term behavior (2) or both.
- **Line 211:** „some errors“ this is not very precise…
- **Line 224:** I do not like the use of the term mean error (ME) to describe bias. Why not simply refer to it as (mean) bias?
- **Line 225:** „the mean error“ you just introduced an abbreviation for this term. Use it.
- **Line 226:** „mean(simulated - observed)“ do not really like this. Why not use proper symbols (e.g. G_{obs} or GW_{obs}) and put them into an equation.
- **Line 226:** why was rank correlation not considered. For the SGI, especially the non-parametric one, it does not really matter if you get the absolute values right but rather the ranking of values is important.
- **Line 253:** but non-parametric approaches can have greater uncertainties (Tijdeman et al., 2020).
- **Line 254:** „aggregated“ see above.
- **Line 258-260.** I like this note and support the use of the SGI. But just wanted to mention that, if the probability is not important but just the rank, why do the probability transformation? Why not simply use ranks? Or, to take into account differences in record length, percentiles expressing the historical non-exceedance frequency in a given time period (Tijdeman et al., 2020)?
- **Table 1:** „dryest“ driest.
- **Line 273:** „prediction“. I do not think it is really prediction but more evaluation over the uncalibrated period.
- **Line 282: „**mean error ME“ just ME

- **Line 291:** „the decay parameter of the noise model“ I would prefer that this is briefly explained here so the reader does not necessarily have to go back to the mentioned study to understand the outcome of this equation.
- **Line 308:** „over-smoothing“ to what does this refer?
- **Table 2: TP, TN** etc. are nowhere defined!
- **Line 320:** „20 % error“ is this the average? Provide ranges?
- **Table 3** “20p“ is not needed and neither is (mod-obs). Precision of units is not consistent, with some having only 1 digits behind the point and others 3.
- **Table 3**. Please add rank correlation.
- **Figure 3.** I don‘t like the map here (better in Fig 1), but if it is kept, I would rather show regions instead of center points.
- **Figure 4.** Nice! Just density instead of “kernel density” on the y-axis.
- **Figure 5 and 6.** Ok, but in my opinion, differences between regions / subsets of wells do not come out strongly from the maps (e.g difficult to recognize the different symbols). What could work is adding density functions split by region or split by well subset (deep, atypical etc.).
- **Line 359.** It should be emphasized that response time is not typical for the drought of 2018 but rather for the entire timeseries.
- **Figure 7.** Adding density plots (like Fig. 4), splitting response time by region or type of well (atypical etc.), might reveal some interesting results.
- **Figure 8.** Some redundant text in caption e.g. what the axis describe.
- **Figure 9.** comparison with simulations is difficult, not only because they are on another page but also because 2018-11 is only shown for the observations. Comparing density functions (possibly grouped by the different type of wells) of simulated and observed SGI might be a suitable alternative.
- **Line 392-394:** it is not clear where these numbers come from.
- **Table 4.** MAE etc. should be defined in the methods.
- **Table 4, line 401-405.** But does the performance maybe relate to well type (typical, etc)?
- **Line 459:** the comparison with Van Loon is not fair as the latter study compares the overall correlation between SGI and SPI / SPEI (not just during few drought months).
- **Line 490-491:** I really do not like this generalization (see above) and more analyses is required to make this claim, even within the discussion.
- **Line 500:** and the probabilistic uncertainty in the SGI estimation itself.
- **Line 509:** 30-year records in certain months. If you want to make a statement of breaking 30-year records overall, you should compare e.g. annual minima or duration.
- **Line 518-526:** the discussion in this paragraph, e.g., the importance of winter recharge, implies the importance of considering initial conditions. This contradicts some that is discussed before.

References:

Bakke, S. J., Ionita, M., and Tallaksen, L. M.: The 2018 northern European hydrological drought and its drivers in a historical perspective, Hydrol. Earth Syst. Sci., 24, 5621–5653, https://doi.org/10.5194/hess-24-5621-2020, 2020.

Bloomfield, J. P. and Marchant, B. P.: Analysis of groundwater drought building on the standardised precipitation index approach, Hydrol. Earth Syst. Sci., 17, 4769–4787, https://doi.org/10.5194/hess-17-4769-2013, 2013.

Brunner, M. I., Liechti, K., and Zappa, M.: Extremeness of recent drought events in Switzerland: dependence on variable and return period choice, Nat. Hazards Earth Syst. Sci., 19, 2311–2323, https://doi.org/10.5194/nhess-19-2311-2019, 2019.

Kumar, R., Musuuza, J. L., Van Loon, A. F., Teuling, A. J., Barthel, R., Ten Broek, J., Mai, J., Samaniego, L., and Attinger, S.: Multiscale evaluation of the Standardized Precipitation Index as a groundwater drought indicator, Hydrol. Earth Syst. Sci., 20, 1117–1131, https://doi.org/10.5194/hess-20-1117-2016, 2016.

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
Van Loon, A. F., Kumar, R., and Mishra, V.: Testing the use of standardised indices and GRACE satellite data to estimate the European 2015 groundwater drought in near-real time, Hydrol. Earth Syst. Sci., 21, 1947–1971, https://doi.org/10.5194/hess-21-1947-2017, 2017.