From the provided text, it is evident that the referee is engaged in a detailed commentary on the manuscript titled "Improved understanding of regional groundwater drought development through time series modelling: the 2018–2019 drought in the Netherlands" by Esther Brakkee et al. The reviewer expresses some reservations about the novelty and potential advantages of the approach used in the manuscript. Specifically, the referee highlights the need for a clearer justification of the impulse-response time series modelling approach, particularly in its application to drought analysis. The reviewer also comments on the substantial amount of data preprocessing involved, which is a critical aspect of the methodology. The manuscript is praised for its comprehensive analysis of groundwater drought, but the reviewer suggests focusing either on the analysis of the drought event or on drought prediction. The overall tone is constructive, aiming to refine and improve the manuscript's approach and presentation. The review is detailed, providing specific suggestions for improving the clarity and focus of the research presented.
L177: Why do you use 20% as a cutoff, please elaborate.

L179: Data cleaning step 2 in combination of using the model results of PASTAS based on time series of precipitation and evapotranspiration to remove outliers homogenizes the time series. Don’t you lose specific features in the time series especially those related to dry conditions? Is this what you mean by ‘over-filtering’ (L211)?

L196-200: What are the reasons for ‘atypical behaviour’? Could site-specific characteristics, i.e. hydrogeology, play a role? In general, I think it would be good for the reader to see either some of the original groundwater head time series or the model outcome.

L201: Why do you use 60%? It seems to be very low. The distribution (histogram) of the models’ EVP would be interesting. It might also be interesting to compare time series with different EVPs, e.g. 60% and 80%. Where are the differences – extremes/timing etc.?

L254: How do you aggregate the daily data? Do you really need a time series lengths of 30 years in a daily resolution to calculate the SGI values for 2018? Is it not efficient enough to use time series with a coarser temporal resolution and a time series length shorter than 30 years? How many measured data sets out of the 2722 fulfill weaker requirements in terms of measurement frequency and length, e.g. at least weekly data from the past 20 years? How would the SGI values (using normal score transformation) and their spatial distribution look like for these time series. How much extra information do you get of using your approach? This kind of analysis would be interesting in order to show the advantages of this approach.

Secondary Comments

L159: How do you define short-term disturbance? Depending on the aquifer system, human disturbance is able to result or magnify groundwater droughts, for example by increased short-term(?) water consumption as a consequence of high temperatures/evapotranspiration rates, or?

L188: Why do you use a threshold of 4 years missing data?

L211: some errors, overremoved are very vague expressions. Please clarify.

L212: over-filtering? Please elaborate or even better show an example.

L221: What about factors influencing groundwater recharge on different time scales, e.g. long-lasting frost seasons, land-use change, or anthropogenic influences changing the boundary conditions, e.g. pilling, water abstraction?

L255: Is this a typo or do you intentionally use different values than the ones in Bloomfield and Marchant (2013): 1/(2n)?

Table 2: Add abbreviation in section 3.1 or even better, write the words out here. You might consider adding percentage here (in brackets).

L321: Are the terms dry, normal and wet conditions based on the percentiles (L227-229)? How do you define ‘more extreme groundwater levels’?

Figure 3: Why do you show midpoints instead of regions? How is the midpoint defined?

L390-395: 3 cm sounds indeed acceptable but an average absolute SGI error of 0.34...What does this mean for the prediction of droughts using the presented method
also in terms of thresholds etc. for stakeholders? Remember: The 2018 drought was an extreme drought.

L404: Please elaborate. Is depth to groundwater or vadose zone thickness not a site-specific characteristic?