Response to Reviewer #1 (J.P. Bloomfield)

The reviewer's comments are in *italic* and our response in normal font.

1. *Kumar et al investigate the suitability of a version of the standard precipitation index (SPI) to characterise groundwater droughts, as defined by standardised groundwater level hydrographs (SGI), at point and regional scales. This is done using monthly groundwater level data from 2000 relatively shallow wells from southern Germany and central Netherlands. The study first characterises the relationships between SPI and SGI for a variety of SPI accumulation periods, including identification of optimal SPI accumulation periods, and then assesses the skill of SPI in predicting groundwater droughts using an assessment of the hit rate and false alarm ratios. The authors find that in the absence of prior information about the hydrogeology of a point or region SPI is a poor indicator of groundwater drought at both scales. The paper is well written, with a clear description of the aims and methods. The results and discussion are combined in a single section. Although generally this is not to be recommended, because the combined results and discussion section is well structured the combined presentation does not detract from the central arguments of the paper.*

   We thank the reviewer Dr. J. P. Bloomfield for the positive summary and the constructive and helpful comments.

2. *The papers main finding is essentially a negative one, i.e. that SPI when used in isolation is a poor indicator of groundwater drought, and consequently the authors should be applauded on reporting this based on a systematic and well argued analysis. Although the authors emphasise that the study focuses on the statistical skill of SPI in predicting groundwater droughts (P7409,L1-4), given the negative findings it would have been interesting to see what effect additional prior information may have had on the correlations between SPI and SGI. For example, information about the geology or aquifer type of each site and some equivalent averaged descriptor for each 0.5 degree grid cell should be available based on even relatively coarse-scale mapping. Would bringing this sort of information into the analysis improve the SPI/SGI correlations, and if so by how much?*

   We appreciate the encouraging words of the reviewer about our analysis. We also understand that performing additional analyses that take into account other geological characteristics might improve the correspondence between SPI and SGI. In fact, prior to putting up this paper, we had a similar discussion amongst ourselves and decided to omit that kind of detailed exploratory analysis from this manuscript for the following reasons:

   i. To keep the focus of the paper simple (easily conveyable) and avoid diverting the readers' attention from the main message. Also noting that the scope and objective of this paper is not to develop a well-functioning drought indicator for two specific regions in Germany and the Netherlands but to assess the statistical skill of commonly used SPI and its feasibility for characterizing groundwater drought.
   ii. Due to the lack of detailed geological data sets at each site to carry out such exploratory analysis. This issue was mentioned earlier in the manuscript (see P. 7421 LL. 25-26).

   Based on the reviewer suggestion, we made an attempt and present here the results of our preliminary analysis investigating the role of geological characteristics based on the large-scale digitized hydro-geological map of Germany (HUEK200; 1:200 000). Specifically we extracted the underlying hydraulic conductivity values of the upper aquifer for every well and grouped them into four dominant classes: High ($>10^{-3}$ m/s), medium (10^{-3}–10^{-5} m/s), moderate (10^{-5}–10^{-7} m/s), and low ($<10^{-7}$ m/s).

   The results of this analysis shown below in Figure A1 indicate that there is no clear trend in the optimal accumulation period (A) between SPI and SGI over these classes. The correspondence between optimal SPI and SGI appears to be relatively weaker at wells located in the low aquifer conductivity class as indicated by a relatively lower value of the maximum correlation r_m (Figure A1b). The optimal accumulation periods (A) appear on average higher for the wells located in the medium to moderate type of aquifer permeability class as compared to that noted for the low conductivity class for which one could have expected the largest smoothing (or attenuation) of precipitation signals. These seemingly contradictory results indicate that the influence of local geological conditions on the propagation of precipitation signals to groundwater flows cannot be assessed by looking single factors (here aquifer conductivity) alone. We note that other geological parameters such as transmissivity and horizontal extent of an aquifer, which are not readily available, would have been more adequate in characterizing the aquifer response time (e.g., Kraijenhoff van de Leur, 1958, Gelhar 1993). Also other local factors such as depth to the groundwater, properties of the unsaturated zone, etc. play an important role and their contribution is neither linear nor independent. It adds to the complexity of this problem that data on local conditions are only available from rather coarse large scale hydro-geological maps with possible large deviations from the actual well-specific conditions. These issues require thus a careful and detailed analyses which are beyond the scope of this study.

Nevertheless, we will include the aforementioned results regarding the role of geological properties in characterizing groundwater droughts with available data sets, while also recognizing their limitations in the revised manuscript. We note that the results presented for this analysis are for the (well specific) point-scale data sets only because:

i. This would better fit to the current analysis and the storyline of the presented study (similar in line of results shown in Figure 3),

ii. We wished to avoid the complications associated with aggregation of hydraulic conductivity fields (given here as categorical values). We are convinced that J.P. Bloomfield is well aware that the aggregation of extremely heterogeneous hydrogeological information to a representative 0.5° cell value is in itself a research topic for which no standard solution exists. To do and to discuss this would have resulted in many more pages, if not in several more papers. Not only this, it would also have distracted to the clear message we wanted to convey in this paper.

D.A. Kraijenhoff van de Leur. A study of non-steady groundwater flow with special reference to a reservoir coefficient. Ingenieur, 70 (1958), pp. B87–B94.

Gelhar, L.W., 1993. Stochastic Subsurface Hydrology. Prentice-Hall, NJ, USA, 390 pp

3. *P7408, L12: re-order Peters et al references to 2003, 2005, 2006*

   Thank you. We will re-ordered this list in the revised manuscript.

4. *Section 2.1 states that the study was performed using monthly groundwater observations. Are these averaged from more dense observations or are all observations on the same day of each month? Is there any missing data in the time series, if so how has this been handled, e.g. left missing, or infilled and if so how? If there is missing data how much is acceptable?*

   The sampling time interval of groundwater available varied from well to well and also within a single well from one time period to another, at daily, weekly, and monthly time intervals. For example, the original data from German was measured at at least weekly time intervals until about 1990, from then on a steadily increasing number of observations switched to daily measurements. Roughly from 2000 on, all stations provide daily data. To harmonize these disrate data sets at a common time scale, we performed our analysis at monthly time scale by averaging shorter time scale data set. In any case, we would like to emphasis that the aggregation method hardly plays a role in the vast majority of cases, since groundwater is slowly evolving process and much of a (seasonal) signal is well captured by monthly observations. There were missing values and they were left out from further analysis (i.e., left missing), and only the data when they are available were used. Finally, we consider only those wells which have at least 10 years of valid monthly records (i.e. without missing value). We will elaborate more clearly on these steps in the revised manuscript.

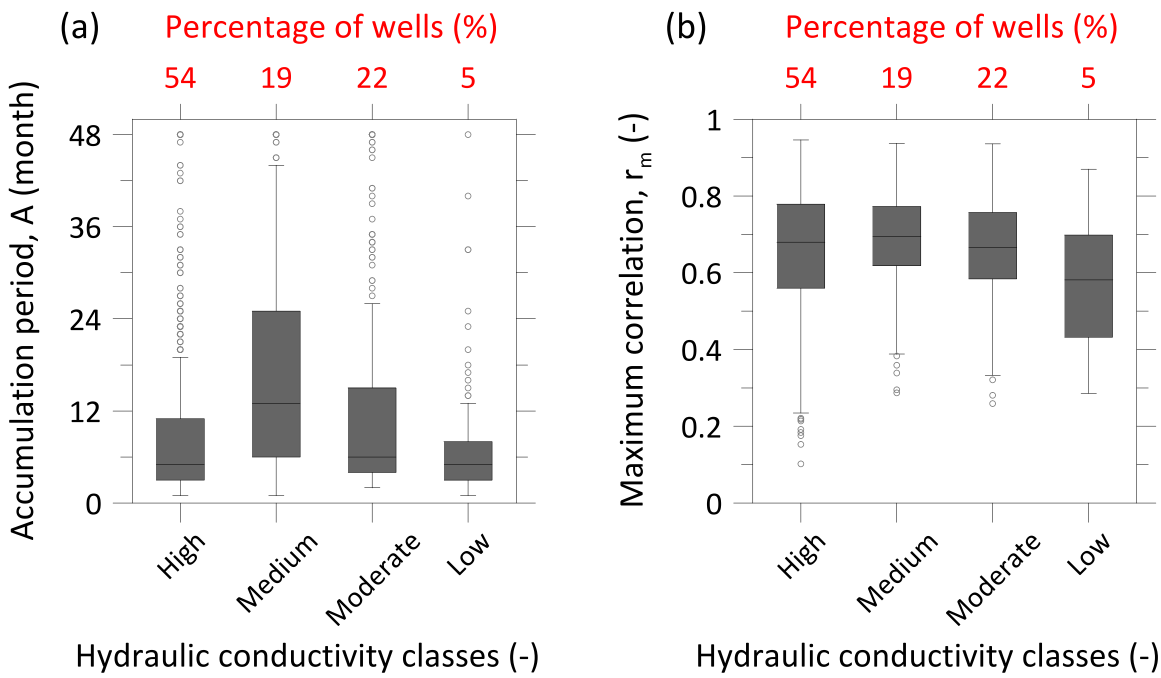

Figure A1: Box-and-whisker plots of the optimal accumulation period (A) and the maximum correlation (r_m) estimated for a group of wells with varying aquifer hydraulic conductivity classes: High ($> 10^{-3}$ m/s), medium (10^{-3}–10^{-5} m/s), moderate (10^{-5}–10^{-7} m/s), and low ($< 10^{-7}$ m/s). The number of wells in four classes are 1080, 382, 433, and 96, respectively.

5. *P7412, L15-25 describes the method used to produce monthly estimates of SPI and SGI at 0.5 degree grid scales. What analysis has been undertaken to investigate the effect of sample size on the relative confidence of estimated mean gridded SPI and SGI values and the consequent implications for calculated hit rates and false alarms (Figs. 6 and 7)? For example, some of the grid cells for the Dutch study area contain only 2 or 3 sites, whereas some grids cells in Germany appear to have many 10s of sites. Also, it appears that the better hit rates for Optimal SPI in Fig. 6 are associated with grid cells with the most sites. Is this correct? If so, what are the implications for the analysis?*

   As a preliminary investigation towards the regional assessment of groundwater droughts, we used a well adopted approach to estimate the ensemble mean of 0.5° gridded SPI and SGI values based on their corresponding point estimates (i.e., well specific SPI and SGI). In this approach we simply used data of all available wells that fall within a particular grid cell to create the gridded estimates at regular intervals of 0.5°. As a consequence the number of underlying wells varied from cell to cell, as rightly pointed out by the reviewer - we have also mentioned this in the manuscript (see P. 7412, LL.22-25). Since this is a simple approach, we do not account for the differences in sample size (i.e., the number of wells falling in a given cell) when estimating the ensemble mean. We, however, fully recognize that there may be several other approaches to optimally upscale point data to gridded estimates, that take into account the differences in sample size among grid cells. We will amend the text in the revised manuscript to highlight these statements.

   Based on the reviewers concern, we conducted *a posteriori* investigation to analyze the effect of sample size on the gridded SPI and SGI relationships. Specifically, we analyze the variation in the gridded estimates of the Spearman rank correlation, Hit rate, and False alarms of the optimal SPI and SGI with the number of wells in every grid cell (see Figure A2 below). The results indicate a slight deterioration in correspondence between SPI and SGI for grid cells with very few underlying wells (< 3). After this threshold (where the majority of grid cells fall), there is no clear improvement in the correspondence of SPI and SGI with the increasing number of wells present in a cell. Consequently, it can be safely concluded that our results are not very much affected by changes in sample size, beyond a certain threshold level as also seen from the moving average estimate plotted in Figure A2. For completeness, we will include these results in the revised manuscript.

6. *P7413, L5 it is stated that "we consider the entire spectrum [0,1] of the SPI and the SGI, without distinguishing between dry or wet regimes". It would be helpful to add a brief discussion of the implications of this statement. Also note a slight contradiction with the statement at P7422, L24-25 that "here we specifically aimed at analyzing the ability of the SPI to predict groundwater drought conditions at different levels". Consider a short clarification to reconcile these statements.*

   We think there has been some misunderstanding here about these two statements. There is no contradiction. We performed two sets of experiments to analyze the feasibility of the SPI to characterize the SGI. In the first set, we took the entire range of quantile based drought indices, SPI and SGI, varying between [0,1] for performing the cross-correlation analysis. Here we investigated about the optimal accumulation and lag periods required to achieve a maximum correlation between the monthly SPI and SGI time series. In the second set of analysis, we analyze the skill of SPI to predict groundwater droughts, i.e., when the SGI $\leq \tau$, and we tested this for the τ value of 0.2, 0.1, and 0.05 indicating different severity levels. Here, we used the scores based on the hit rate and the false alarm ratio to assess the reliability of groundwater drought predictions made by the SPI. In the revised manuscript, we will amend the text to clearly state that our analyses looked into both the entire range of drought indices (SPI and SGI) and the parts focusing on drought periods.

7. *P7419, L9-12. This is describing the well known phenomenon of drought attenuation in the groundwater compartment of the terrestrial water cycle. It may be helpful to explicitly acknowledge this here with a suitable reference.*

   We will amend the text in the revised manuscript with a suitable reference.

8. *P7421, L12 should read "on the basis of this data-based exploratory analysis"*

   Thank you. We will consider your suggestion in the revised manuscript.

9. *Please check all references. For example, a number have missing volume or page numbers, e.g. AghaKouchak et al.; Hao et al; Li et al; Samaniego et al; Teuling et al; and Weider and Boutt.*

   Thank you, the references will be taken care of.

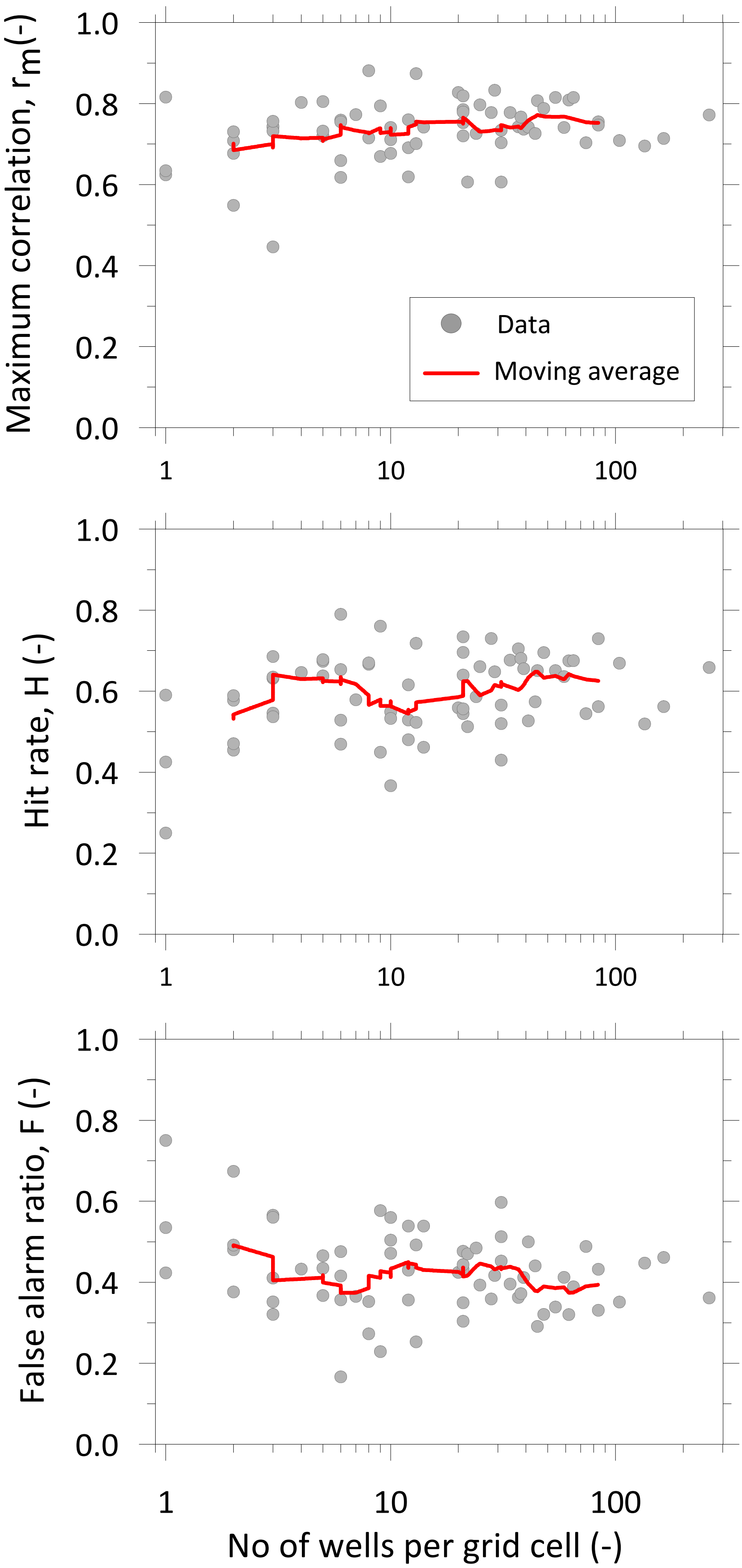

Figure A2: Variation in the cell specific maximum correlation (r_m), hit rate (H), and false alarm ratio (F) with number of underlying wells used to create 0.5° gridded estimates of SPI and SGI. All SPI estimates correspond to the grid specific optimal accumulation period. The moving averages with a window size of eleven wells data are shown in the red line.