The Editor/Reviewer's comments are in *italic* and our response in normal font. The line and the page numbers, we indicate below are provided within the latexdiff version of the manuscript, which can be found at the end of this document.

Response to Editor's (Prof. Gregor Laaha) comments

1. *Thank you for your interesting manuscript. Your paper has been carefully read by three referees. All of them find your paper well suited for HESS bus suggest a number of corrections/modifications which need to be carefully considered before publication. The authors are therefore invited to prepare a author's response which includes a revised version of the manuscript indicating all changes. The revised version shall address all points raised by the referees.*

   We thank the editor Prof. Gregor Laaha for his kind assessment and providing us the opportunity to revise the manuscript.

2. *On concern of the second referee (that was not fully addressed in the responses so far) was why to use an empirical approach to calculate the SPI, when a well-established and widely accepted methodology (based on fitting a theoretical distribution function) exists. Although I think this point is only of subordinate relevance for the scope of your study, I share the concern of the referee. Proposing a new approach without carefully evaluating its merits and possible pitfalls relative to a standard method would be dangerous, but I understand that this is likely not the scope of your paper. Authors should therefore clarify in the MS that the approach will likely not have the full skill of the standard method, but is deemed suitable to get a sufficient approximation on a regional scale.*

   With all due respect to the Editor's comment, we do not agree that our non-parametric approach of fitting distribution function is by any means inferior to a commonly used approach of fitting a theoretical distribution function for estimating SPI. Indeed, several recent studies have pointed out difficulties with the (*a priori*) selection of suitable theoretical distribution functions for the SPI estimation, including other problems such as difficulties when handing the multi-modality of observed datasets using a univariate distribution function. As also detailed in the reviewer #2 response (to this particular comment), that a non-parametric approach to estimate SPI is used here to avoid the problem of assigning a unique distribution function to all datasets (as mentioned above), and to ensure the consistency in the estimation of drought indices for the precipitation and groundwater time series (i.e., both variables use a similar approach so that the resulting drought indices fall within a same range $[0, 1]$).

   We would also like to emphasise that a non-parametric approach used in this study is not new, and certainly not the first time we are proposing (and neither do we claim so). The use of empirical approach to estimate quantile-based drought indices has been the basis for several recent drought related studies (e.g., from the group of D.P. Lettenmaier, J. Sheffield, A. AghaKouchak, J. P. Vidal, J. P. Bloomfield, among others). Furthermore, if deemed necessary the quantile based drought indices can be easily transformed to the unbounded range of the standard normal distribution (commonly used for the SPI detection) based on the mapping of values across the two cumulative density functions. Moreover the empirical approach to estimate SPI is also justified considering that the approach does not influence the ordering of drought events. In this study, we explored the agreement between SPI and SGI for matching drought/no drought events based on the contingency table based skill scores for which the ordering of drought events is an important aspect. We have included these notes in the revised manuscript to avoid any further confusion. See P9, L22-26 and P10, L1-10.

3. *I also share the view of this referee that the role of evaporation is so obvious that one would tend to use the SPEI rather than SPI for assessing hydrological impacts. I acknowledge that the scope of the paper is on a different index, but at least a rough indication should be given (either from additional analyses, or from literature) how far using an index which integrates*

evaporation loss would make atmospheric drought indices a suitable indicator for groundwater drought.

Initially, we were hesitant to comment on the skill of SPEI since our focus in this work was to assess the SPI skill as mentioned explicitly in the title of the paper (and also throughout the manuscript). But based on the Editor and the Reviewer #2 suggestions, we performed additional analyses to assess the skill of SPEI for characterising the groundwater droughts. We, however, wished not to include these results in the present manuscript as it would greatly extend (and modify) the scope of the presented work, and will very likely deviate the readers from the key message of this paper. Therefore we retain the SPEI results in this rebuttal only.

We repeated all our drought analyses using the monthly SPEI estimated at different accumulation periods using the precipitation and potential evapotranspiration (PET) data. The multiscale evaluation was conducted for both point and gridded scales using data of German wells for the illustration purpose. PET is estimated based on the the Hargreaves and Samani (1985) method that uses average, maximum and minimum temperatures.

We find no significant improvement in the skill of SPEI over SPI for characterising groundwater droughts. Indeed both meteorological based drought indices (SPI and SPEI) exhibited similar behaviour when contrasted against the respective SGIs at both point and gridded scales (compare the results shown below in Figures A1 to A5 for the SPEI with those of the SPI in the manuscript - the corresponding manuscript figure number is also mentioned in the caption of the following figures to ease comparison).

Similar to the results of the SPI (presented in the manuscript), we found a large spatial variability in the optimal accumulation period, required to achieve a maximum correlation between SPEI and SGI. Consequently, the application of uniform accumulation period (of SPEI) over the study region would significantly deteriorate the correspondence between SPEI and SGI. On average, the mean absolute errors between SGI and SPEI at different accumulation periods were also high (0.15-0.25) similar to those noticed between SPI and SGI.

The visual inspection of the monthly SPEI timeseries (e.g., for 6 and 12 months of accumulation periods) show almost similar temporal pattern as that of the SPI - and both meteorological drought indices could not adequately capture the higher variability exhibited by the SGI timeseries at both point and gridded scales. Furthermore, the majority of wells and grid cells exhibited low scores of hit rate and high false alarm ratio again demonstrating the low reliability of groundwater drought predictions using the SPEI. In summary, based on our analysis we do not find any substantial improvement in using SPEI over SPI for groundwater drought predictions. Overall our analysis indicated that both atmospheric drought indices (SPI and SPEI) are not suitable indictors for characterising groundwater droughts.

4. *I would also add one comment, that the methodology of assessing the agreement between meteorological and hydrological drought indices is not completely new. It was recently used in a similar context, for assessing the link of SPI and other atmospheric indices to detect low flow events. Because of this similarity a citation to that study would be appropriate: Haslinger, K., Koffler, D., Schöner, W. and Laaha, G.: Exploring the link between meteorological drought and streamflow: Effects of climate-catchment interaction, Water Resour. Res., 50(3), 2468–2487, doi:10.1002/2013WR015051, 2014.*

   Thank you for pointing out this issue. We have included the reference of this study in the revised manuscript. See P12, L22-25.

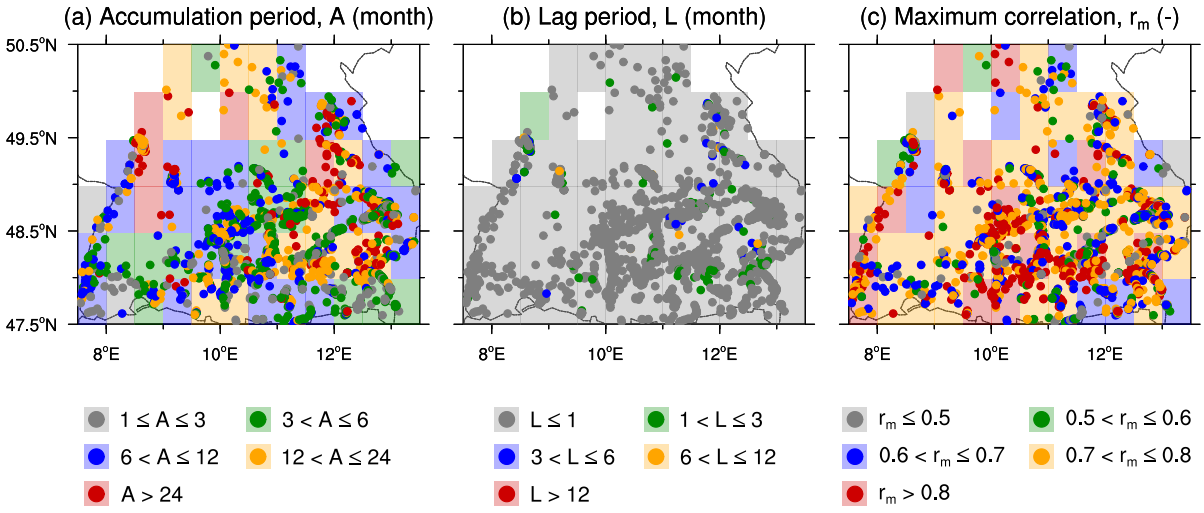

Figure A1: The **(a)** optimal accumulation A (month) and **(b)** lag periods L (month) required to obtain the **(c)** maximum correlation r_{m} (-) between the SGI and SPEI at point and gridded (0.5°) scale. Similar to Figure 2 for the SPI in the manuscript.

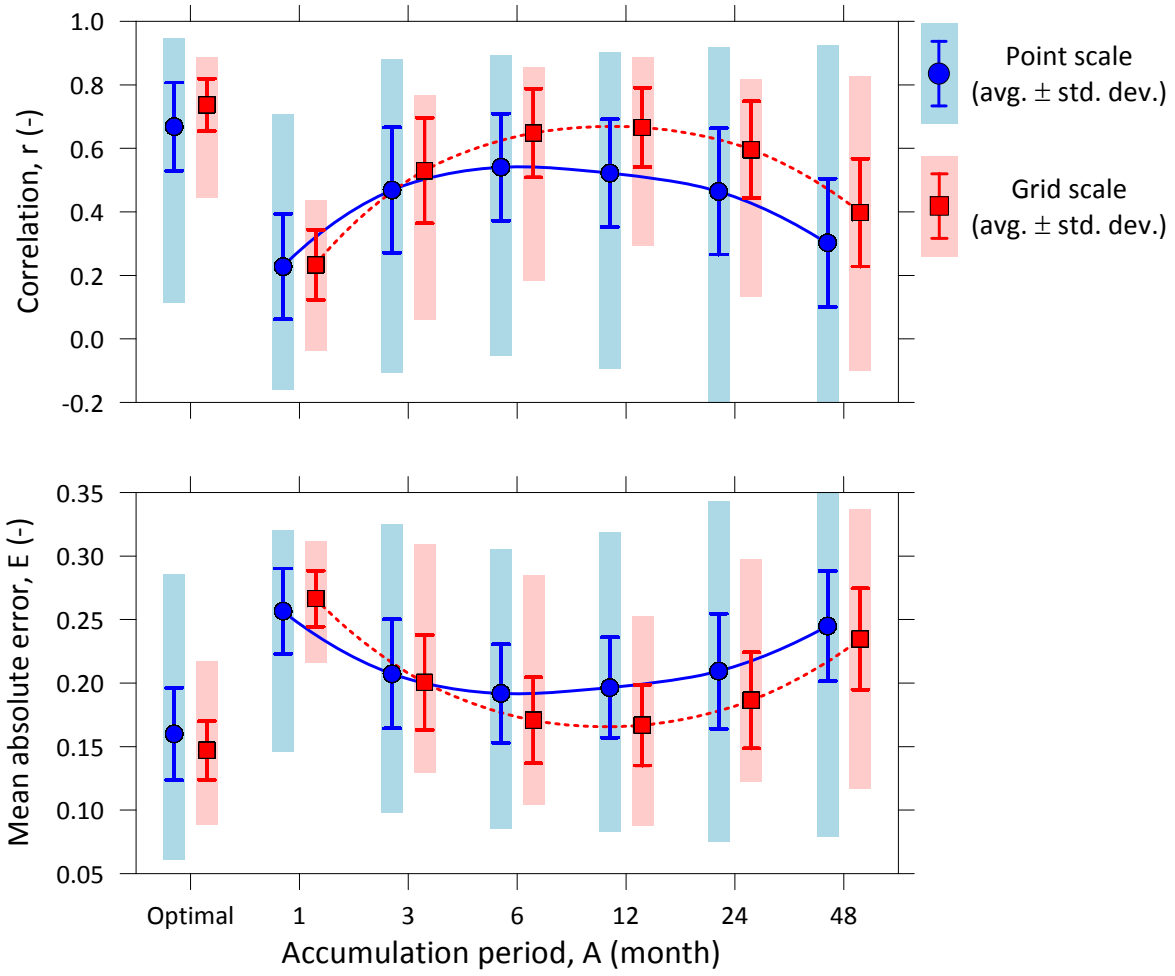

Figure A2: The correlation r (top) and the mean absolute error E (bottom) estimated between the SGI and SPEI of the 1, 3, 6, 12, 24, and 48 months of uniform accumulations for the point and the gridded data sets. Their respective maximum (r_m) and the minimum (E_m) estimates corresponding to the optimal accumulation periods of SPEI are also shown in the leftmost of the panels. Results are summarized here as average $\pm$ one standard deviation. Similar to Figure 4 for the SPI in the manuscript.

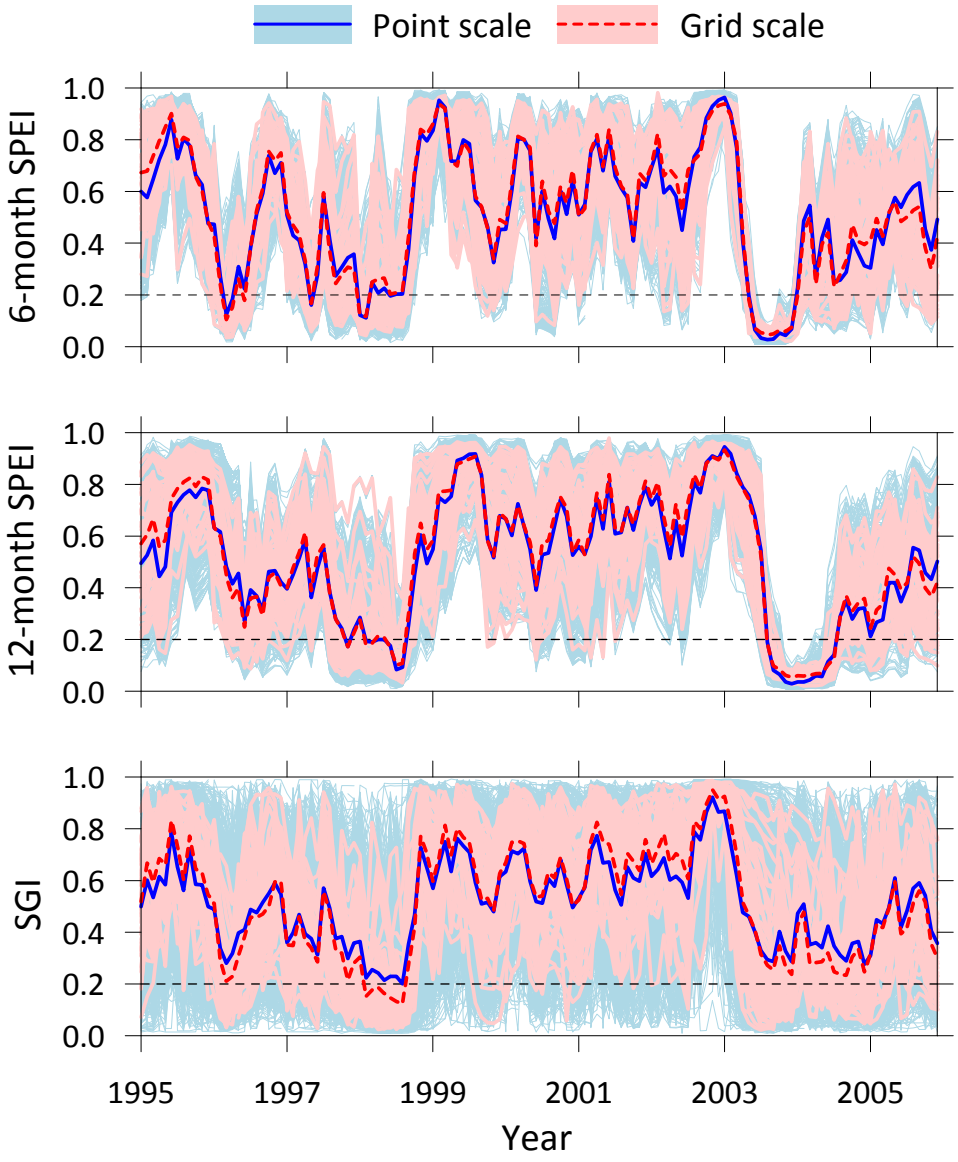

Figure A3: The monthly time series of the 6 and 12-month point (light blue) and gridded (light pink) SPEI and the respective spatial averages (dark blue and dark red). The bottom plots are the corresponding SGI time series and their spatial averages. The black dashed line depicts the drought threshold τ of 0.2. Similar to Figure 5 in the manuscript.

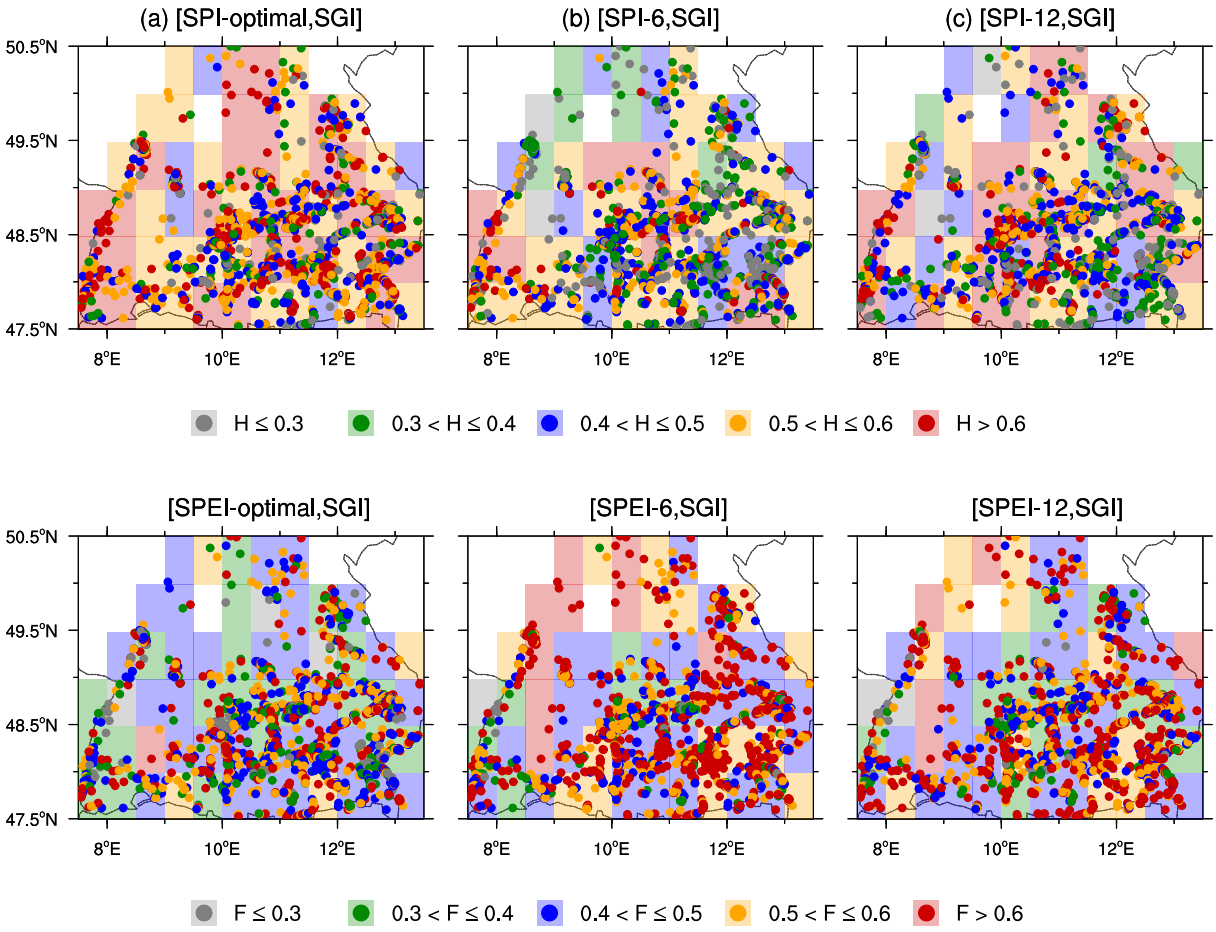

Figure A4: The (top) hit rate H and the (bottom) false alarm ratio F to detect SGI based groundwater droughts using the SPEI with the (a) optimal accumulation period and (b and c) 6 and 12 months of uniform accumulation periods at the point and gridded scales. A threshold value τ of 0.2 is used to identify drought events. Similar to Figures 6 and 7 for the SPI in the manuscript.

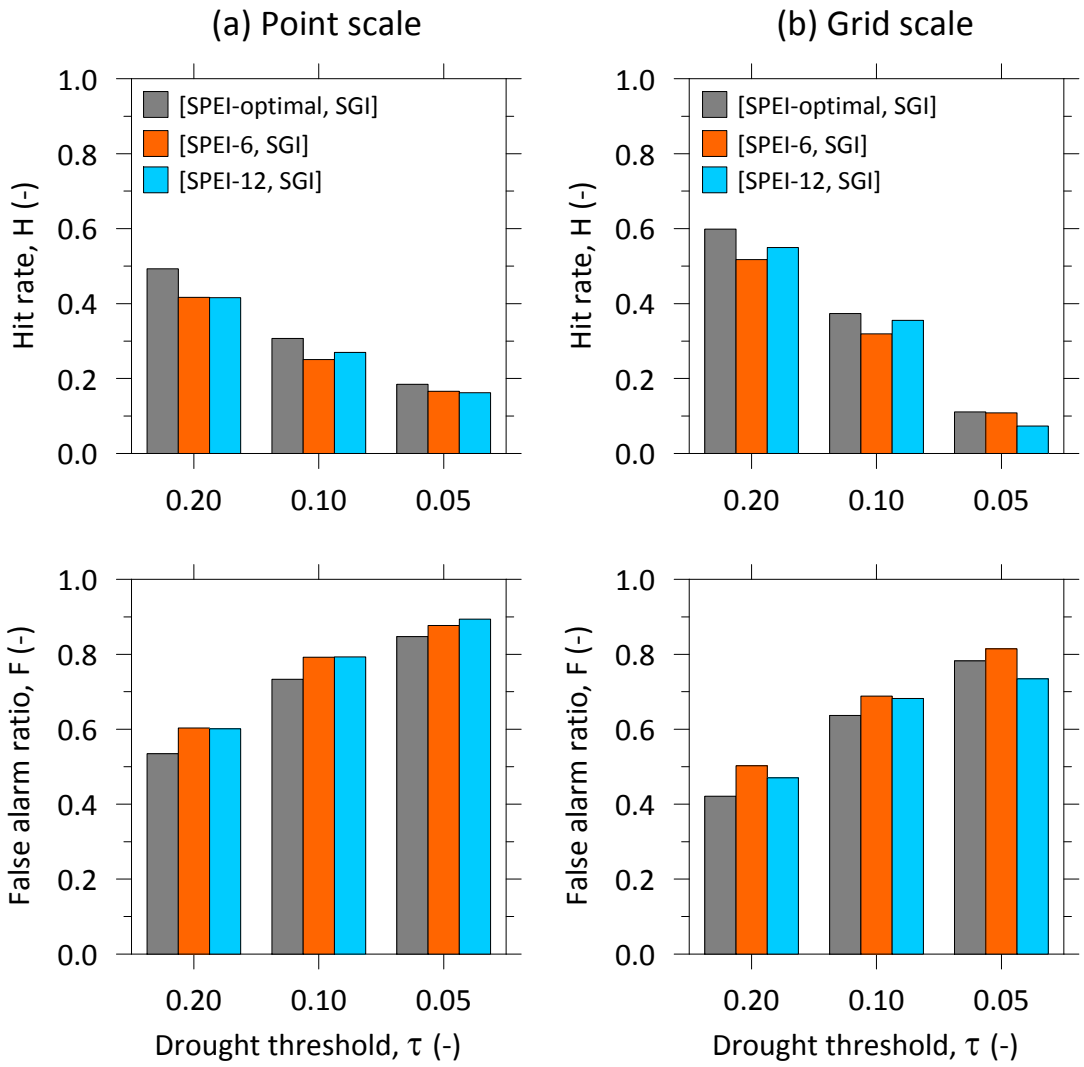

Figure A5: The hit rate (H) and the false alarm ratio (F) averaged over all investigated (a) wells and (b) grid cells to detect SGI based groundwater droughts using the SPEI with the optimal accumulation and 6 and 12 months of uniform accumulation periods for varying levels of threshold value τ (0.2, 0.1, and 0.05) used to identify drought events. Similar to Figure 8 for the SPI in the manuscript.

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

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.

   Nevertheless based on the reviewer suggestion, we made an attempt and present here the results of our preliminary analysis investigating the role of geological characteristics on the spatial variability of optimal accumulation period (A) and maximum correlation (r_{m}) between the SPI and SGI. For this purpose, the underlying hydraulic conductivity values of the uppermost aquifer were extracted from the available large-scale hydro-geological map of Germany (HUEK200; available at a scale of 1:200 000). The wells were grouped into four dominant conductivity classes: High ($>10^{-3}$ m/s), medium (10^{-3}–10^{-5} m/s), low (10^{-5}–10^{-7} m/s), and very low ($<10^{-7}$ m/s). Results of this analysis indicate that there is no clear trend in the optimal accumulation period (A) between SPI and SGI over these classes (Figure A6). The correspondence between optimal SPI and SGI appears to be relatively weaker at wells located in aquifers with lower conductivity as indicated by a relatively lower value of the maximum correlation (r_m).

The optimal accumulation periods (A) appear on average higher for the wells located in the medium to low type of aquifer permeability class as compared to that noted for the very 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 (e.g., HUEK200 map) with possible large deviations from the actual well-specific conditions. These issues thus require careful and detailed analyses which are beyond the scope of this study. We have now included these aforementioned results in the revised manuscript. See P14, L4-29, and Figure 3 (middle column).

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 Dr. 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.

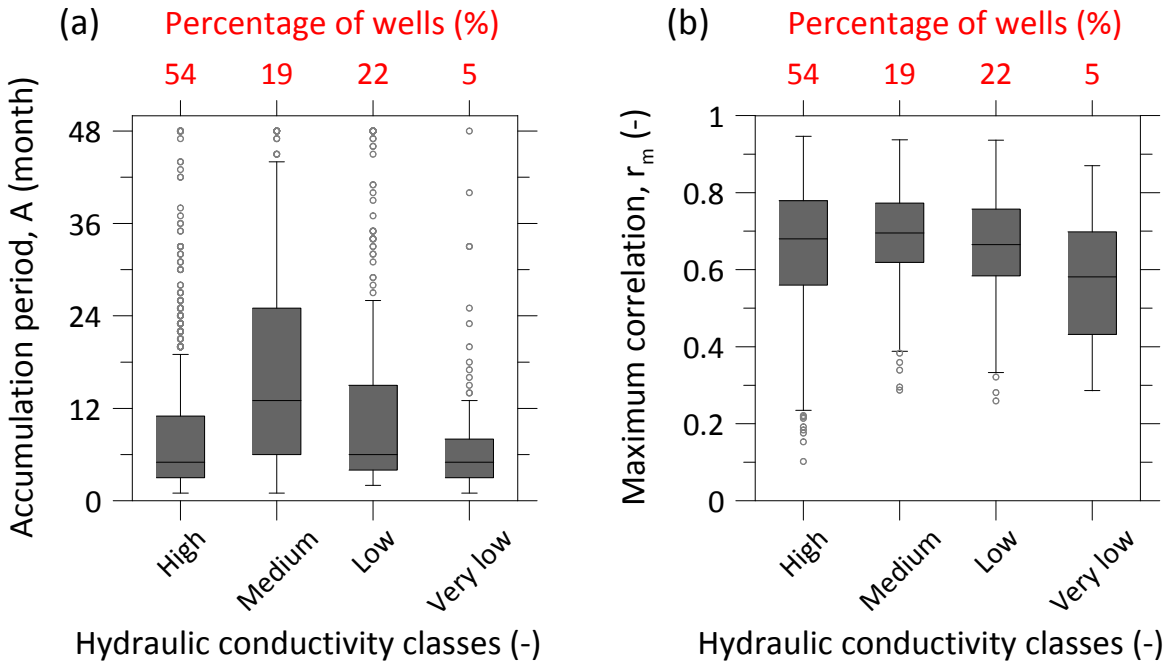

Figure A6: Box-and-whisker plots of the optimal accumulation period (A) and the maximum correlation (r_m) estimated for a group of wells with varying aquifer hydraulic conductivity classes: High ($>10^{-3}$ m/s), medium (10^{-3}–10^{-5} m/s), low (10^{-5}–10^{-7} m/s), and very low ($<10^{-7}$ m/s).

[1] 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.

[2] 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 for pointing this out. We have re-ordered this list in the revised manuscript (see P4, L24).

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 the analysis (i.e., left missing). Finally, we consider only those wells which have at least 10 years of valid monthly records (i.e. without missing value). We have elaborated more clearly on these steps in the revised manuscript (see P7, L14-24).

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 P11, L5-7). 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 have added a note on this in the revised manscript (see P11, L7-13).

   Based on the reviewers concern, we also 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 A7 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. 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 A7. For completeness, we have included these results in the revised manuscript. See P20, L26-28 and P21, L1-5.

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. We have included a sentence in the revised manuscript to better explain this experimental set-up (see P11, L17-19). 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.

   To avoid any further confusion, we revised the text to clearly state that our analyses in addition to looking at the entire range of drought indices (SPI and SGI), focus on analysing the skill of SPI to predict groundwater drought events at various severity levels (see P23, L17-22).

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 have revised the text to reflect this comment. See P19, L2-4.

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

   Thank you. We revised the text accordingly. See P22, L7.

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 for this comment. We made our best possible effort to check all references for missing information and updates.

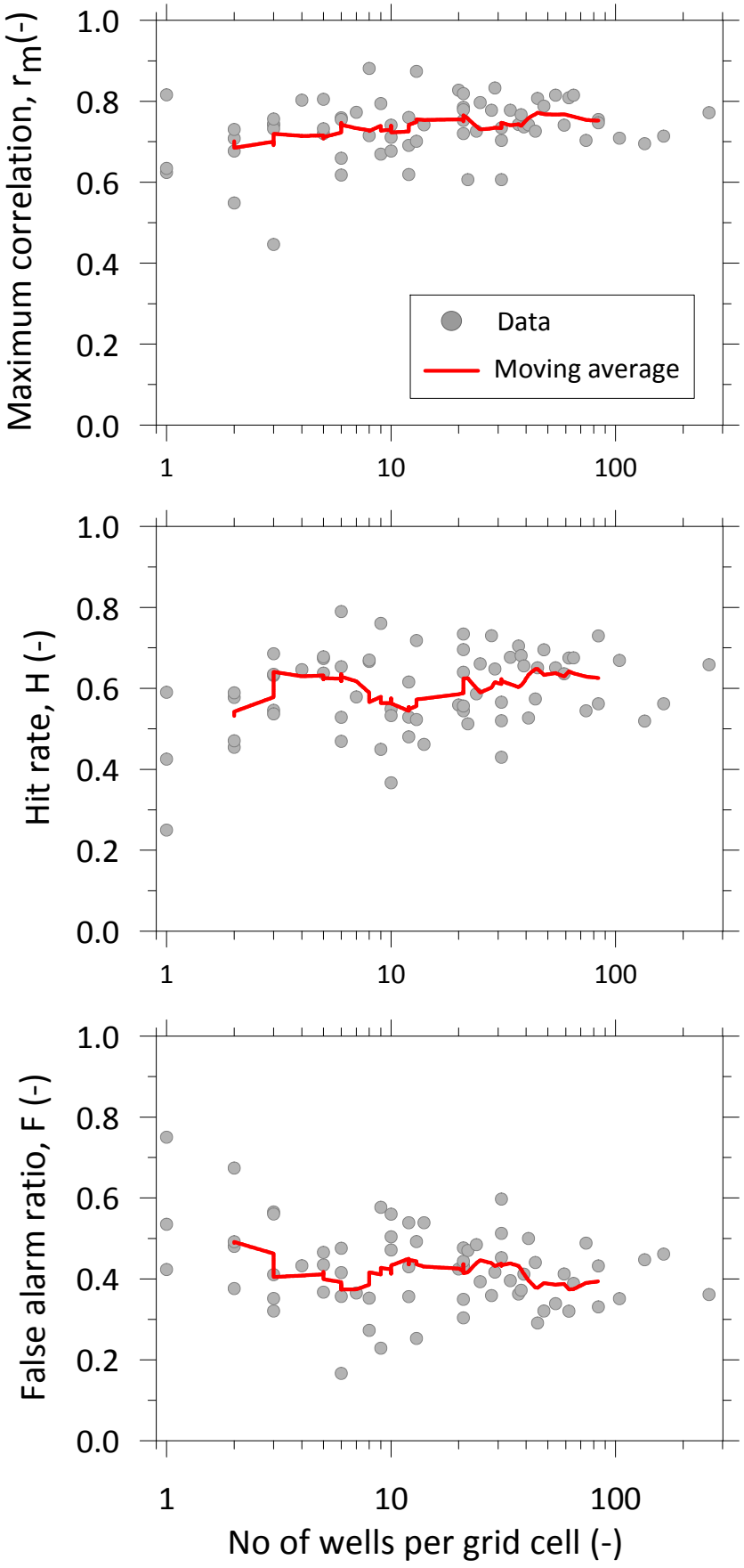

Figure A7: Variation in the cell specific maximum correlation (r_m), hit rate (H), and false alarm ratio (F) with number of underlying wells used to create 0.5° gridded estimates of SPI and SGI. All SPI estimates correspond to the grid specific optimal accumulation period. The moving averages with a window size of eleven wells data are shown in the red line.

Response to Reviewer #2 (Dr. S. M. Vicente Serrano)

1. *Review of "Multiple evaluation of the standardized precipitation index as a groundwater drought indicator". The manuscript analyses the relationship between precipitation and groundwater droughts in South Germany and the Netherlands. The manuscript focusses in knowing on the capacity of the Standardized Precipitation Index as a drought monitoring metric to determine groundwater drought. The article is well-written and structured. The research topic is suitable for HESS and it has great potential given the current interest of moving from the use of climate drought indices (easy to calculate) to drought impacts (difficult to estimate). I would recommend the acceptance of the manuscript in HESS but I would like to draw attention to different issues that would be interesting that the authors consider or at least mention in the discussion of the results.*

   We thank the reviewer Dr. S. M. Vicente Serrano for his encouraging words and helpful comments.

2. *Page 7407. Lines 11-13. I agree that drought monitoring based on precipitation data may have advantages regarding data availability. Nevertheless, this approach may have also deficiencies since it does not consider other key variables that affect drought severity, mainly the atmospheric evaporative demand (AED). Although the AED effect could be considered negligible for ground water recharge, we cannot forget that approximately 3/5 parts of the precipitation returns to the atmosphere via evapotranspiration processes. Probably in the Netherlands and Germany the AED is not a relevant stressing variable given high precipitation amounts (although not negligible for drought impacts, e.g., the year 2003) but from sub-humid to arid regions AED is a determining factor that affects water resources availability in a determining manner. Thus, it is expected that AED does not only affect soil moisture and runoff but also water infiltration and ground water since AED is affecting the vegetation respiration and the water exchange between plants and the atmosphere. A comment or discussion about this issue would be welcome. Page 7407. Line 22. Also the role of AED should be mentioned.*

   We fully understand and acknowledge the role of atmospheric evaporative demand (AED) in a hydrologic drought propagation. Indeed we explicitly cited the work by Teuling et al., 2013 (in the same lines: P3, L22) that highlighted the role of evapotranspiration during the 2003 European drought event. Considering that the presented work focuses mainly on assessing the skill of SPI for the groundwater drought, in our earlier manuscript we did not explicitly consider putting more weight on the AED role which itself would be an interesting work, but certainly beyond the scope of the present study. We have now amended the text in the revised manuscript to reflect this issue. See e.g., P3, L22-28, P4, L1-7, P19, L9-29, and P20, L1-4. Furthermore, based on yours and the Editor's comments, we made a preliminary investigation on the assessment of the skill the Standardised Precipitation-Evapotranspiration Index (SPEI) for characterising groundwater droughts (similar to those performed for the SPI in the manuscript). Please refer above for more details on results of this analysis in our responses to the Editor's comments.

3. *Page 7408. Line 11 Some other references dealing directly with this topic: Climate Research. 58, 117-131; Journal of Hydrology. 477: 175-188; Earth Interactions 16, 1-27. Page 7408. Line 16. There are previous studies analyzing the relationship between drought indices and groundwater (e.g., Natural Hazards and Earth System Sciences 15: 1381-1397; Hydrology and Earth System Sciences 19: 2353-2375; Water Resources Management 24: pp. 1867-1884). These studies should be cited here.*

   Thank you for pointing out these studies. We have assimilated much of these references in the revised manuscript wherever appropriate (see P4, L25-26, P5, L1-3).

4. *Page 7409, Line 21. Is there any aquifer exploitation like pumping for water supply and irrigation?, please detail.*

In this study, we screened and selected only those well records that did not exhibit obvious signs of anthropogenic influences through visual inspection and some basic data analysis. We would like to however note that the observational German wells are located in quite densely populated regions (approx. 15 million population) and groundwater forms the main source of drinking water. Irrigation is not widely applied. There are two reasons why this doesn't have large impact on the presented analysis:

i. The observation wells used in this analysis are typically relatively far away from pumping wells and thus the influence of pumping is in most cases negligible. It is estimated that only about 3% of the potentially available water resources (Precipitations-Evapotranspiration) in the region are used (Nickel et al. 2005, doi:10.1016/j.pce.2005.06.004). The regional consequences of groundwater extraction are thus very low.

ii. The groundwater withdrawal in the region is relatively constant all year round as it is mainly domestic and industrial use (no irrigation) without peak loads in specific seasons. Thus, fluctuations in groundwater levels can be mainly attributed to weather/climate and not to fluctuations of groundwater use.

We included these notes in the revised manuscript (see P6, L11-28).

5. *Page 7410. Lines 19-22. More details on the filtering analysis are needed. If only the months with available groundwater are used to select precipitation months, what about previous months needed to obtain longer time-scales?*

   We recognize that we had not been clear on this point (as also reflected in the Reviewer #3 comment). By filtering we mean, the months with missing groundwater values are also set to *missing* in the precipitation time series. We however applied the filtering after the accumulation of precipitation (for any selected time periods e.g., 3, 6, 12 months) had been performed. This way, we ensured that the consistency of longer time scale SPI estimates was maintained and not affected by the filtering procedure. We emphasise that the filtering step was necessary to ensure the comparability between the (accumulated) precipitation and groundwater time series so that both had the same sample size for the estimation of SPI and SGI. We have revised the text detailing more clearly about the filtering procedure to avoid any further misunderstanding (see P8, L7-13).

6. *Page 7411. Lines 4-6. The correct references to support this statement should be McKee et al. 1993 and Guttman 1999.*

   Thank you for pointing this. We have revise the text accordingly (P8, L24).

7. *Page 7411. Line 8. Guttman (1999) suggested the Pearson III distribution based on large study in USA. In any case, the uncertainty associated to the selected distribution should be minimal and there is a standard methodology to calculate the SPI by the World Meteorological Organization* `http://www.wamis.org/agm/pubs/SPI/WMO_1090_EN.pdf`. *For this reason, I do not find suitable to use an empirical approach to calculate the SPI when a well-established and widely accepted methodology exists. Empirical approximations to obtain cumulative distribution functions are much more depending of the available sample than the use of pdfs. I understand that groundwater data availability prevents of fitting a a pdf given low data availability in some wells, but given high density of groundwater stations (which are expected to be highly correlated among them), the regional analysis (Hosking, J.R.M., Wallis, J.R., 1997. Regional Frequency Analysis, An Approach Based on L-Moments. Cambridge University Press, Cambridge, UK) could be a better approach to obtain the groundwater drought index. In any case, since the statistical analysis are based on rank correlations, in which the magnitude of the series is not taken into account, the procedure used to standardize of the precipitation and groundwater is secondary. Thus, the authors could have used directly the raw series of ground water and the series of precipitation accumulated on different time-scales for the analysis.*

In our opinion, both the empirical approximation or the fitting of a theoretical distribution to obtain pdfs (and corresponding cdfs) suffer from the problem of sample size, and there is no unique solution to this problem. To maintain consistency we used a non-parametric kernel density estimator to compute the cdfs of the precipitation and groundwater data. As was mentioned in the text, the kernel also removes the problems related to multi-modality of the data and the subjectivity in the *a priori* selection of the analytic pdf. Please aslo refer to our response (above) to the Editor's comment on a similar issue. We have also now included a note on this issue in the revised manuscript to avoid any further misunderstandings (see P9, L22-26 and P10, L1-10).

8. *Page 7413. Line 16 and following. I think you could have used better approaches to compare the agreement between groundwater and precipitation drought events (e.g., comparing the duration, maximum intensity, total magnitude and spatial extent of droughts). Really a categorical contingency table is useful but I think that more information could be extracted from the available data, at least for the longest groundwater series in which individual drought episodes can be identified.*

   We appreciate the reviewer advice. In this paper, we were mainly concerned with assessing the statistical skill of the SPI for the groundwater drought predictions for which we used the skill scores based on the categorical contingency table. Clearly the next step would be to look more deeply into differences between drought characteristics such as magnitude, severity, duration, intensity, etc. derived based on SPI and SGI time series. However, such analysis would certainly be beyond the scope of the current study. In our opinion, analyzing drought characteristics in detail would divert the reader from the main message which comes with sufficient clarity from the skill analysis based on a categorical contingency table. We therefore left such investigations for future studies and pointed out this in the concluding part of the revised manuscript (see P23, L11-16).

9. *Section 3.2 This stresses the diversity of relationships that are usually recorded between drought time-scales and impacts, and the need of testing initially the best time scale of a drought index to determine possible impacts. This is quite relevant and not specific for groundwater but also for several hydrological and ecological systems (e.g., PNAS 110: 52-57; Climate Research. 58, 117-131; Journal of Hydrology, 386: 13-26; Agricultural and Forest Meteorology. 151: 1800-1811; Journal of Hydrology. 477: 175-188, among others). I think this should be stressed and discussed in more depth (see further discussion about this issue in Journal of Geophysical Research-Atmosphere. 116, D19112, doi:10.1029/2011JD016410).*

   Thank you for outlining this issue which is now detailed in the revised manuscript (see P17, L21-29, P18, L1-6).

Response to Reviewer #3

1. *The authors investigate the connection between groundwater levels expressed through the standardized groundwater level index (SGI) and the standardized precipitation index (SPI) in Southern Germany and the Netherlands. The study aims to characterize the relationship on different accumulation time scales of both the SPI and the SGI as well as via evaluating the skill of the SPI in predicting groundwater droughts using hit rates and false alarm rates. The general message of the manuscript is that the SPI is a poor indicator for groundwater droughts in the given areas. The manuscript is well written, clearly structured and the figures are suitable. The science questions are clearly stated and I recommend publication in HESS after considering some recommendations listed below. I also appreciate the publication of mainly "negative" results which is unfortunately often avoided, although there may be potential to learn even more from negative results than from positive ones.*

   We thank the reviewer for his/her encouraging words and helpful comments.

2. *P7407, L22-23: Aren't there other reasons than "non-linearity" of the transformation of a precipitation signal to a groundwater drought? What about the role of evapotranspiration as a key process in the terrestrial water cycle? Some comments and references on that issue should be added here.*

   There are several reasons for that non-linear transformation of the precipitation signal to groundwater. Apart from the underlying (sub)-surface properties like terrain, soil, vegetation and geological properties, climatic properties such as precipitation seasonality, snowmelt timing, availability of atmospheric water supply and demand as evapotranspiration plays a crucial role. Specifically for the role of evapotranspiration in drought propagation, we acknowledge the study by Teuling et al., 2013 (see P3, L22) which looked into this issue. We have now amended the text in the revised manuscript to reflect this issue (see P3, L22-28, P4, L1-7).

3. *P7408, L12: years in the reference of Peters et al. should be 2003, 2005, 2006.*

   Thank you for pointing this out. We have re-ordered this list in the revised manuscript (see P4, L24).

4. *P7409, L15: The climate of Southern Germany is not ?continental?. It is not as close to the sea as the Netherlands and therefore less maritime, but the wording continental is not appropriate in this respect. Suggestion: ?...a region with hilly to mountainous terrain, less oceanic influence on climate and a wide range...?*

   We have revised the text as suggested (see P6, L5).

5. *P7410, L19-21: These lines are not clear to me. What do you mean by filtering the precipitation time series? Are missing groundwater dates set missing in the precipitation time series? If yes, how does this affect the accumulation on different time periods? Please clarify.*

   We recognize that we had not been clear on this point (as also reflected in the Reviewer #2 comment). By filtering we mean, the months with missing groundwater values are also set to *missing* in the precipitation time series. We however applied the filtering after the accumulation of precipitation (for any selected time periods e.g., 3, 6, 12 months) had been performed. This way, we ensured that the consistency of longer time scale SPI estimates was maintained and not affected by the filtering procedure. We emphasise that the filtering step was necessary to ensure the comparability between the (accumulated) precipitation and groundwater time series so that both had the same sample size for the estimation of SPI and SGI. We have revised the text detailing more clearly about the filtering procedure to avoid any further misunderstanding (see P8, L7-13).

6. *Section 4: I think the manuscript would benefit if there is a more in depth discussion on why the SPI is not a proper groundwater drought identifier. Particularly the role of the demand-side of the water balance seams a critical point, as well as the underlying geology. Please add some*

discussion and reference dealing with these issues. See for example: Natural Hazards and Earth System Sciences, 15, 1381?1397; Journal of Hydrology 477, 175-188, Water Resources Research 50, doi:10.1002/2013WR015051

Based on your and other two reviewers' suggestions, we have amend the text in the revised manuscript to reflect the possible role of geological properties and evapotranspiration in groundwater drought evolution. We have also include much of the suggested references wherever appropriate. See e.g., P14, L3-29, P17, L21-29, P18, L1-6, P19, L9-29, P20, L1-4. We would also like to draw the attention of the reviewer to our responses (above) to the Editor's comment where we show results of our preliminary investigation on the assessment of the skill the Standardised Precipitation-Evapotranspiration Index (SPEI) for characterising groundwater droughts (similar to those performed for the SPI in the manuscript). Please refer above for more details on results of this analysis in our responses to the Editor's comments.

7. *P7421, L12: should be: "...on the basis of this data-based..."*

   Thank you for pointing this, we have revised the text (see P22, L7).

Manuscript prepared for Hydrol. Earth Syst. Sci. Discuss.
with version 2015/09/17 7.94 Copernicus papers of the LaTeX class copernicus.cls.
Date: 3 February 2016

Multiscale evaluation of the standardized precipitation index as a groundwater drought indicator

R. Kumar[1], J. L. Musuuza[1,2], A. F. Van Loon[3,4], A. J. Teuling[3], R. Barthel[5], J. Ten Broek[3], J. Mai[1], L. Samaniego[1], and S. Attinger[1,6]

[1]UFZ-Helmholtz Centre for Environmental Research, Leipzig, Germany
[2]Now at the Department of Civil Engineering, University of Bristol, Bristol, UK
[3]Hydrology and Quantitative Water Management Group, Wageningen University, Wageningen, the Netherlands
[4]Now at the School of Geography, University of Birmingham, Birmingham, UK
[5]Department of Earth Sciences, University of Gothenburg, Gothenburg, Sweden
[6]Institute of Geosciences, University of Jena, Jena, Germany

Correspondence to: R. Kumar (rohini.kumar@ufz.de)

Abstract

The lack of comprehensive groundwater observations at regional and global scales has promoted the use of alternative proxies and indices to quantify and predict groundwater droughts. Among them, the Standardized Precipitation Index (SPI) is commonly used to characterize droughts in different compartments of the hydro-meteorological system. In this study, we explore the suitability of the SPI to characterize local and regional scale groundwater droughts using observations at more than 2000 groundwater wells in geologically different areas in Germany and the Netherlands. A multiscale evaluation of the SPI is performed using the station data and their corresponding 0.5° gridded estimates to analyze the local and regional behavior of groundwater droughts, respectively. The standardized anomalies in the groundwater heads (SGI) were correlated against SPIs obtained using different accumulation periods. The accumulation periods to achieve maximum correlation exhibited high spatial variability (ranges 3 to 36 months) at both scales, leading to the conclusion that an a priori selection of the accumulation period (for computing the SPI) would result in inadequate characterization of groundwater droughts. The application of the uniform accumulation periods over the entire domain significantly reduced the correlation between SPI and SGI ($\approx$ 21–66 %) indicating the limited applicability of SPI as a proxy for groundwater droughts even at long accumulation times. Furthermore, the low scores of the hit rate (0.3–0.6) and high false alarm ratio (0.4–0.7) at the majority of the wells and grid cells demonstrated the low reliability of groundwater drought predictions using the SPI. The findings of this study highlight the pitfalls of using the SPI as a groundwater drought indicator at both local and regional scales, and stress the need for more groundwater observations and accounting for regional hydrogeological characteristics in groundwater drought monitoring.

1 Introduction

Drought as a natural hazard is often associated with high socio-economic losses and damage to ecosystems (Wilhite, 2000). Many of these drought effects are not directly caused by rainfall deficits, but are related to below-average storage conditions in surface water, reservoirs, and groundwater that are the consequences of the propagation of a meteorological drought into the hydrological system (Tallaksen and Van Lanen, 2004; Mishra and Singh, 2010; Sheffield and Wood, 2011; Seneviratne et al., 2012). Due to a lack of large-scale groundwater and surface water observations, most scientists and water resources managers interested in drought predictions have to rely on proxy data to quantify storage conditions.

One widely-used approach is to use drought indices based solely on precipitation (e.g. Standardized Precipitation Index; SPI), because precipitation records generally have good spatial coverage and long observation periods required for drought analysis. It is then assumed that by computing the SPI over longer time scales (e.g., 3, 6, 12 or more months), it mimics the filtering effect of catchment storage conditions and hence captures the smooth precipitation deficits typical for hydrological (groundwater) droughts (Seneviratne et al., 2012; Joetzjer et al., 2013; Li and Rodell, 2015). Although the SPI is recognized as an effective meteorological drought index (Hayes et al., 2010) due to its relative ease of computation and comparability across climates, some studies have questioned its application for groundwater drought monitoring because the translation of precipitation deficits into hydrologic (groundwater) droughts is non-linear (Bloomfield and Marchant, 2013; Teuling et al., 2013; Van Loon et al., 2014). Both catchment and climate characteristics such as the differences in underlying soil, terrain, vegetation and geological properties, precipitation seasonality, snowmelt timing, and the availability of atmospheric water supply and demand (evapotranspiration) control the development hydrologic droughts and the resulting drought characteristics (Bloomfield and Marchant, 2013; Haslinger et al., 2014; Van Loon et al., 2014; Stoelzle et al., 2014; Van Loon, 2015). Recently, (Vicente-Serrano et al., 2010) introduced the Standardised Precipitation-

Evapotranspiration Index (SPEI) having a similar multitemporal characteristic of the SPI, but accounting for both the atmospheric water supply (precipitation) and evaporative demand (potential evapotranspiration). SPEI can account for the influence of temperature variability and thus it is better suited than SPI for drought studies under global warming conditions. In regions with high precipitation variability (e.g., humid areas), both the SPI and SPEI are expected to generally exhibit a similar behaviour, albeit having slight differences among each other during a specific calendar month and time period (Vicente-Serrano et al., 2012).

Another approach to quantify drought is based on the use of large-scale gridded data products, e.g., from hydrologic models or satellites (e.g., Sheffield et al., 2004; Andreadis et al., 2005; Vidal et al., 2010; Samaniego et al., 2013; van Huijgevoort et al., 2013; Prudhomme et al., 2014; Mo and Lettenmaier, 2013; Nijssen et al., 2014; Hao et al., 2014; Li and Rodell, 2015; Damberg and AghaKouchak, 2014; Wanders et al., 2015; AghaKouchak et al., 2015). An extensive multi-model study (Prudhomme et al., 2014) for example, projected increases in hydrological drought severity in many areas around the world. This approach also has limitations in its application at local to regional scale hydrological drought monitoring because of scale mismatches and some issues in the correct representation of storage in models (Gudmundsson et al., 2012; Van Loon et al., 2012; Tallaksen and Stahl, 2014). The importance of spatial variation in groundwater drought conditions resulting from complexity in subsurface conditions is increasingly recognized (Peters et al., 2006; Bloomfield and Marchant, 2013; Stoelzle et al., 2014).

While there are many studies that have focused on analyzing the propagation of meteorological droughts through the hydrologic systems for improved process understanding of the evolution of ~~groundwater droughts~~ hydrologic (groundwater) droughts (e.g., Eltahir and Yeh, 1999; Peters et al., 2003, 2005, 2006; Tallaksen et al., 2006, 2009; Weider and Boutt, 2010; Vicente-Serrano et al., 2012; Bloomfield and Marchant, 2013; Haslinger et al., 2014; López-Moreno et al., 2013; Van Loon et al., 2014), there is still lack of comprehensive observational-based studies to verify whether hydrological drought proxies, like precipitation-based indices (SPI) and gridded data products, are suitable for groundwater drought monitoring at regional to local scales relevant for water management.

~~Recently,~~ In recent years there had been some efforts on analyzing the relationship between meteorological and groundwater based drought indices (Bloomfield and Marchant, 2013; Folland et al., 2015; Bachmair et al., 2015). Bloomfield and Marchant (2013), for example, introduced the Standardized Groundwater level Index (SGI), similar to the SPI, and found a site specific relationship between the two indices. Their study was, however, limited to the analysis of local scale behavior of groundwater droughts at 14 sites across the UK.

In this data-based exploratory study, we test the suitability of the SPI to characterize groundwater droughts using observations at more than 2000 groundwater wells located in Germany and the Netherlands. We used this large set of groundwater wells to comprehensively analyze the local to regional behavior of groundwater droughts and investigate the scale mismatch between local and regional scale estimates. A focus on groundwater was preferred to other hydrological variables because of the immense multi-sectoral importance of the resource (Famiglietti, 2014). Given the wide-spread availability and usage of precipitation-based drought indices, we hypothesize that if adequate accumulation periods and lead times are applied to the precipitation signal, the observation-based SPI can predict groundwater droughts. In this maiden attempt we carry out a quantitative evaluation of the performance of the widely-used SPI for groundwater drought monitoring on a local to regional scale using a large collection of groundwater well records, focusing on the statistical skill and not on the causing factors. The results of this study will provide insight to the water-sector practitioners and managers on the precautions demanded if they are to use local precipitation data or large-scale gridded estimates to characterize groundwater droughts.

2 Method

2.1 Study area and data

The study was performed using monthly groundwater observations from two hydro-geologically different regions located in southern Germany and central Netherlands (Dutch

province of Gelderland) with 1991 and 49 groundwater wells, respectively (Fig. 1). The Dutch region is characterized by ~~the~~ maritime climate and the wells are located on a relatively low terrain, but with large spatial differences in unsaturated zone and groundwater conditions. The German wells are located in a region with hilly to mountainous terrain, ~~with continental~~ less oceanic influence on climate and a wide range of unconsolidated and consolidated geological formations. The monthly groundwater data for the German wells were collected from the Bavarian Environment Agency (LfU Bayern) and the State Institute for Environment, Measurements and Nature Conservation Baden-Württemberg (LUBW). The data for the Dutch wells were acquired from the Dutch institute TNO (www.dinoloket.nl/).

~~The selected wells ' records did not exhibit significant~~ To be able to attribute groundwater level changes to climatic causes, it is necessary to exclude the possibility that these changes are a consequence of anthropogenic influences such as pumping or hydraulic structures. Therefore, those wells which exhibit obvious signs of anthropogenic influences ~~and could therefore allow the understanding of the natural response of the groundwater to the precipitation signal.~~ were excluded from the analysis. In general, in both regions it can be expected that the effects of groundwater withdrawals are only local as water consumptions constitutes only a very minor portion of the potentially available water resources (precipitation - evapotranspiration). In the German study area it is estimated that only about 3% of the potentially available water are used (Nickel et al., 2005). Irrigation is not widely applied. Moreover, the groundwater withdrawal in the region is relatively constant all year round as it is mainly domestic and industrial use without peak loads in specific seasons. Notably the German wells are located in quite densely populated regions (approx. 15 million population) and groundwater forms the main source of drinking water. This, however, does not have large impact on the presented analysis since the observation wells used in this analysis are typically located far away from pumping wells. Thus, fluctuations in groundwater levels can be mainly attributed to weather/climate and not to fluctuations of groundwater use.

The majority of the selected wells (around 90 %) are located in shallow aquifers with an average depth to the water table within 20 m below the ground surface (see Fig. 3a for the well distribution). The length of records varied from well to well with a minimum of 10 years (Fig. 1) starting from the year 1951 for the German wells and 1988 for the Dutch wells. It should be noted that the 10-year criterion does not meet the recommended minimum 30 years (McKee et al., 1993; Guttman, 1999) for estimating drought indices (i.e., SPI or SGI). However, a longer cutoff of 30 years would lower the number of qualifying wells significantly. For example, all the Dutch wells would have been excluded under this criterion (Fig. 1). The limited availability of in situ groundwater data records as well as the variable record lengths are inevitable problems in performing groundwater drought studies over a large domain (see e.g., Peters et al., 2006; Weider and Boutt, 2010; Li and Rodell, 2015). We nevertheless have tested the reliability of our results to this data availability issue (as discussed later in Sect. 3.1).

~~The daily~~ The sampling time interval of groundwater observations 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 for German wells were measured at least on weekly time interval until about 1990, from then on a steadily increasing number of observations switched to daily measurements. Roughly from 2000 onwards, all stations provide data at a daily time interval. To harmonize these disrate data sets at a common time scale, we performed the analysis at a monthly time scale wherein shorter time scale data sets were averaged to produce the monthly groundwater time series. The missing groundwater observations were left out from the analysis (i.e., left missing). Finally, we consider only those wells which have at least 10 years of valid monthly records (i.e., without missing values).

The daily precipitation time series at every well were extracted from their gridded estimates computed based on the available ~~rain-gage~~ raingauge network (Samaniego et al., 2013; ten Broek et al., 2014). The underlying point measurement data from about 5600 rain gauges for Germany and 51 rain gauges for the Netherland were acquired from the German Meteorological Service (DWD) and the Royal Netherlands Meteorological Institute (KNMI),

respectively. Interested readers may refer to Samaniego et al. (2013) and ten Broek et al. (2014) for more details on processing with ~~the~~ precipitation data sets for the German and the Dutch regions, respectively. The monthly total precipitation was then computed from their respective daily estimates to match the temporal resolution of groundwater records. Additionally, prior to the SPI calculations, the precipitation time series was filtered based on the temporal availability of groundwater records to ensure the comparability between the two time series. In other words, the months with missing groundwater records are also set to missing in the precipitation time series. This filtering step is however applied after the accumulation of precipitation (for any selected time periods e.g., 3, 6, 12 months) for longer time scales had been performed. In this way, we ensured the consistency of the longer time scales (accumulated) SPI estimates, as well as their compatibility with the availability of groundwater records such that both variables had same sample size for the estimation of the corresponding drought indices (i.e., SPI and SGI).

2.2 Drought indices

The Standardized Precipitation Index (SPI) was developed by McKee et al. (1993) to characterize the wetness and dryness conditions of a region based on the departure of the monthly precipitation estimate from their (average) normal value. The SPI can be estimated for different time scales by accumulating the monthly precipitation over different periods, typically at 3, 6, 12, 24, or 36 months (see McKee et al., 1993 for a detailed treatment). In most applications of the SPI, an analytic distribution function (e.g., the gamma) is fitted to the long-term precipitation record for a given accumulation period, and then the corresponding cumulative probability distribution is computed. Finally, the cumulative probability distribution is transformed to the standard normal distribution to estimate the SPI (McKee et al., 1993; Guttman, 1999). Any month with an SPI value below (above) zero is assumed to reflect dry (wet) conditions. Fitting theoretical distribution functions to data is potentially problematic because it is difficult to determine the structural form of the distribution function in advance. For example, Guttman (1999) found for the SPI that the ~~gamma~~ Person Type III distribution was the best universal model based on large set of US

datasets, whereas Lana et al. (2001) found that data from Catalonia in Spain could best be modeled with the Poisson-gamma distribution. Additional problems may arise if the data exhibits multi-modality.

To avoid these problems and minimize the uncertainty associated with the selection and estimation of parametric distribution functions, we used a non-parametric kernel density estimator to compute the cumulative probability distributions of the precipitation and groundwater data. The kernel density $\hat{f}(x)$ is given as

$$\hat{f}(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h}\right) \tag{1}$$

where h represents the bandwidth, $K(x)$ the kernel smoothing function, $x_1, \ldots, x_n$ the set of variable of interest (i.e., precipitation or groundwater level), and n the sample size. We used the Gaussian kernel in this study because of its unlimited support and estimated the bandwidth h by an optimization against a cross-validation error estimate (see Samaniego et al., 2013 for details). The distribution functions and the corresponding bandwidths were estimated for each well and calendar month separately. The resulting quantiles, bounded on $[0, 1]$, are denoted hereafter as SPI and SGI for precipitation and groundwater, respectively. The quantile-based index has been used in several recent drought studies (Sheffield et al., 2004; Andreadis et al., 2005; Vidal et al., 2010; Samaniego et al., 2013), and can be easily transformed to the unbounded range of the standard normal distribution (Vidal et al., 2010). The SPI and SGI values below (above) 0.5 denote dry (wet) conditions. Compared to the absolute values of groundwater heads (precipitation estimates), the transformed SGI (SPI) values facilitate better the comparison across space and season (Sheffield et al., 2004). We note that our approach of estimating the SPI time series differs from a more conventional approach of fitting a defined distribution function to the precipitation time series and then estimating the corresponding SPI estimates (Guttman, 1999; Hayes et al., 2010). A non-parametric approach is used here to avoid the problem of assigning a unique distribution function to all datasets (as mentioned above), and to ensure the consistency in

the estimation of drought indices for the precipitation and groundwater time series (i.e., both variables use a similar approach so that the resulting drought indices fall within the same range [0, 1]). We note that many recent drought studies have adopted a non-parameteric approach for the estimation of drought indices (see, e.g., Andreadis et al., 2005; Vidal et al., 2010; Bloomfield and Marchant, 2013; Samaniego et al., 2013; Hao et al., 2014). Bloomfield and Marchant (2013), for example, had difficulties in identifying an unique best distribution function that fits to all groundwater records at various locations, and even at a given location a fitted distribution function varied from one calendar month to another. Here we adopt their approach and estimate precipitation and groundwater drought indices (SPI and SGI) through a non-parameteric method.

2.3 Experimental setup and evaluation criteria

In this study, we explore the ability of SPI to characterise the local and regional scale behavior of groundwater droughts. Therefore, we carried out our analysis at two disparate scales denoted hereafter as the point and the grid scales. The point scale analysis was performed on well-by-well basis using their available SPI and SGI time series. We based this analysis on the assumption that the zone of influence for changes in groundwater levels is limited to the area directly surrounding the well as most of the wells are located within shallow aquifers (see, Fig. 3 for the distribution of average depth to the water table across the investigated wells). We note that the approach chosen here is consistent with that commonly used in previous groundwater studies (e.g., Bloomfield and Marchant, 2013; Li and Rodell, 2015). On the other hand, the grid scale analysis was carried out using the monthly estimates of drought indices gridded at a 0.5° spatial resolution – the scale which is commonly used in regional and global scale drought studies (e.g., Sheffield et al., 2004; Andreadis et al., 2005; Gudmundsson et al., 2012; Seneviratne et al., 2012; Van Loon et al., 2012; Tallaksen and Stahl, 2014; Wanders et al., 2015). The gridded fields of drought indices were estimated using a procedure similar to the ones employed in creating multi-model drought indices (see, e.g., Mo and Lettenmaier, 2013; Nijssen et al., 2014). Following this procedure, the individual estimates of well-specific drought index were combined into

a single grid representative estimate by averaging the drought index from those wells which lie within the selected grid cell. The resulting monthly estimates at each grid were then converted into a percentile based drought index following the (non-parametric kernel) density estimator approach illustrated in Sect. 2.2 for the well-specific data sets. The number of qualifying wells with at least 10 years of records per grid cell varied across the study domain between (1 and 261) for Germany and (1 and 14) for the Netherlands with a median value of around 21 and 5 wells, respectively. We note that as a preliminary investigation towards the regional assessment of groundwater droughts, we used a well adopted ensemble mean approach to estimate 0.5° gridded fields of SPI and SGI. In this simple approach, we used data of all available wells that fall within a particular grid cell, without accounting for the differences in sample size (i.e., the number of wells within a grid cell), to create the gridded estimates of SPI and SGI. We have analyzed the differences in sample size on the gridded SPI and SGI skill scores (reported in Sect. 3.4).

To provide a qualitative skill of the SPI to characterize the SGI, we first examine the spatio-temporal relationship between the two indices based on the cross-correlation analysis. Here we consider the entire ~~spectrum~~ spectra $[0, 1]$ of the SPI and the SGI, without distinguishing between dry or wet regimes. In other words, this part of the analysis was conducted using the entire time series of SPI and SGI that covers the whole spectrum of hydro-meterological conditions spanning from the extremely dry to very wet conditions. The analysis was performed separately for both point and grid scale data sets, with different accumulations and lags of SPI ranging from 1 to 48 months. In this analysis, we used the Spearman rank correlation coefficient (r) as a non-parametric measure to quantify the strength of a monotonic relationship between the SPI and SGI. Since the propagation of precipitation signals to the groundwater is highly nonlinear, the rank correlation was preferred over the traditional Pearson (linear) correlation coefficient in this analysis. The goal here was to identify what accumulations and lags of SPI are required to align the signals of precipitation with the groundwater heads; and how they vary in space for both point and the gridded data sets.

In the subsequent analysis, we focus on assessing the ability of SPI to detect groundwater droughts based on SGI. A drought is defined when the indices (i.e., SPI or SGI) fall below a certain threshold (τ), taken here as 0.2 following previous studies (Sheffield et al., 2004; Andreadis et al., 2005; Vidal et al., 2010; Samaniego et al., 2013). According to the drought classification scheme used by the U.S. Drought Monitor (USDM; http://droughtmonitor.unl.edu), more severe and extreme drought conditions appear when the indices fall below the τ value of 0.1 and 0.05, respectively. The reliability of groundwater drought predictions made by the SPI for different drought classes can be assessed using probabilistic scores based on the probability of detection or hit rate (H) and the false alarm ratio (F). Following the (2×2) contingency table, the hit rate (H) is given by

$$H = \frac{a}{a+c} \tag{2}$$

and the false alarm ratio (F) represented as

$$F = \frac{b}{a+b} \tag{3}$$

where a, b, and c are the hits, false alarms and misses, respectively (Wilks, 2011). In our case, the hit rate H is the fraction of all groundwater drought events correctly predicted by the SPI (i.e., the ratio of the number of times the SPI predicts a groundwater drought when the SGI indicates the occurrence of one, to the total number of times the SGI indicates drought conditions). The false alarm ratio F represents the fraction of forecasted drought events that were false alarms (i.e., the ratio of the number of times the SPI predicts a groundwater drought when the SGI does not indicate one, to the total number of times the SPI predicts droughts). The best scores for H and F are 1 and 0, respectively; and the worst values being 0 and 1, respectively. More recently, Haslinger et al. (2014) used a similar approach to assess the link between the SPI and other atmospheric indices (e.g., SPEI and PDSI) to detect low flow events in the Austrian catchments based on hit rates (H).

3 Results and discussion

3.1 Cross-correlation analysis between SPI and SGI

The results of the cross-correlation analysis between SGI and SPI at different accumulations and lags reveal a large degree of spatial variability in the accumulation period A required to achieve maximum correlation r_{m} at both point and grid scales (Fig. 2). The A value corresponding to the r_{m} is referred hereafter as an "optimal" accumulation period. The estimates of A across the majority of wells and grid cells ($> 90\,\%$) in both study regions varied broadly between 3 and 36 months with an overall median value of around 6 to 12 months. The relatively large variation of A values across the investigated wells signifies the importance of the underlying climate, soil, vegetation, and aquifer properties in modulating the precipitation signals for groundwater flows.

Our preliminary analysis indicated that the wells located in comparatively very thick unsaturated zones or deeper groundwater tables exhibited on average higher accumulation periods, and vice-versa (Fig. 3a). For example, the higher accumulation values (> 24 months) in the middle of the (Gelderland) Dutch region are due to the presence of a relatively thicker unsaturated zone going up to 30 m deep (Fig. 2). Consistent with the theoretical expectation, a similar relationship between the accumulation periods and the depth to the water table was reported recently by Li and Rodell (2015) when analyzing groundwater droughts at wells located in the Mississippi river basin and ~~near-by~~ nearby regions. In general, deeper groundwater tables (or thicker unsaturated zone) cause more attenuation of the high frequency precipitation signals, and require longer accumulations of precipitation to properly align with the smoothed variability of groundwater signals (Barthel, 2011). On the other hand, the shallower groundwater table responds more quickly to high frequency precipitation events and the variability of the groundwater anomalies is better explained by the shorter timescale of the SPI. There are, however, exceptions to this general behavior and the temporal dynamics of groundwater indices (SGI) at some wells in shallower aquifers exhibited better correlation with a longer time scale SPI, going up to 48 months (Fig. 3a). This highlights the need to take into account other hydrogeological and

well-specific information like aquifer release and storage characteristics, perforation type, borehole location, among others (Bloomfield and Marchant, 2013; Stoelzle et al., 2014).

~~The lack of complete information on these characteristics at every well hinders further investigation of this issue.~~We also examined the role of geological characteristics on the spatial variability of optimal accumulation period (A) and maximum correlation (r_{m}) between the SPI and SGI. For this purpose, the underlying hydraulic conductivity values of the uppermost aquifer were extracted from the available large-scale hydro-geological map of Germany (HUEK200; available at a scale of 1:200 000). The wells were grouped into four dominant conductivity classes: High ($> 10^{-3}$ m/s), medium (10^{-3}–10^{-5} m/s), low (10^{-5}–10^{-7} m/s), and very low ($< 10^{-7}$ m/s). Results of this analysis indicate that there is no clear trend in the optimal accumulation period (A) between SPI and SGI over these classes (Fig. 3b). The correspondence between optimal SPI and SGI appears to be relatively weaker at wells located in aquifers with lower conductivity as indicated by a relatively lower value of the maximum correlation (r_m). The optimal accumulation periods (A) appear on average higher for the wells located in the medium to low type of aquifer permeability class as compared to that noted for the very 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 (e.g., HUEK200 map) with possible large deviations from the actual well-specific conditions. These issues thus require careful and detailed analyses which are beyond the scope of this study. We note that the focus of this study is not on identifying

potential factors or relationships explaining the spatial variability of accumulation periods. Nevertheless, we emphasize ~~here~~ that the results of ~~this (limited) exploratory~~ our above presented analysis (Fig. 3a) show the opportunity for establishing a first-order regional relationship between the accumulation period and the average depth to water table, for which global estimates are now becoming available (Fan et al., 2013).

The lag times (L) leading to maximum correlation (r_{m}) between the SPI and SGI show a limited spatial variability across the majority of wells and grid cells with values generally close to zero (Fig. 2b). This implies that the temporal anomalies of the groundwater heads (SGI) at those locations are aligned to those of the (accumulated) precipitation (SPI). Results of our analysis did indicate a substantial variation in the maximum correlation (r_{m}) across the investigated wells, pointing out the lack of a uniform strong relationship between SPI and SGI (Fig. 2c). The r_{m} values ranged between 0.40 and 0.87 for the majority of German wells, and between 0.47 and 0.87 for the Dutch wells with the overall median r_{m} value of around 0.68 and 0.70, respectively. A relatively weaker correlation between SPI and SGI was found for wells located in a shallower aquifer, where the average depth to the water table is less than 5 m (Fig. 3b). The r_{m} value estimated across these shallower wells was on average around 0.64, whereas for wells located in a relatively deeper aquifer (with water table depth > 5 m) the average r_{m} was 0.72. This trend of the correlation (r_{m}) with the average water table depth is, however, not so strongly pronounced as in a case of the accumulation period (Fig. 3a and b).

We also tested the reliability of the above results against the data availability issue. The A and r_{m} obtained across all wells were grouped into three categories according to their available record lengths (i.e., in 10–20, $>$20–30, and $>$30 years). Both the spread and the average ~~behavior~~ behaviour of the optimal accumulation period (A) and the maximum correlation (r_{m}) were comparable across the group of wells with different record lengths (Fig. 3c and d). This shows that the above presented results are reliable and are not contingent on the selection of wells with either short or long record lengths. We also emphasize here that our results are not biased to the selected statistical criteria (i.e., rank correlation). Similar results (not shown here) were obtained using other criteria such as the

Pearson correlation coefficient and the mean absolute error; both exhibited substantially large (small) variations in the accumulation (lag) periods across the analyzed wells and grid cells.

3.2 SPI with spatially uniform accumulation periods

The spatial variation in the optimum accumulation periods shown in Fig. 2 demonstrates that there exists no single representative value that is applicable over the entire domain. A noticeable reduction in the correlation values (r) between SGI and SPI was observed when a uniform accumulation period was applied to all wells or grid cells (Fig. 4). For instance, the correlation estimated across the investigated wells on average reduced from the r_{m} value of 0.67 to 0.23, 0.46, 0.53, 0.50, 0.44, and 0.27 for the 1, 3, 6, 12, 24, and 48 months of uniform accumulations, respectively. Around 10 to 65 % of the wells exhibited a notably low correlation with r values less than 0.3. The gridded data sets exhibited slightly better correspondence between SPI and SGI than the point ones, with the maximum correlation being dropped from 0.73 to 0.26, 0.54, 0.62, 0.64, 0.56, and 0.35 for the 1, 3, 6, 12, 24, and 48 months of uniform accumulations, respectively. In this case, nearly 5–60 % of the grid cells exhibited notably low correlation values below 0.3. Among the different uniform accumulation periods, the strongest correlation between SGI and SPI was observed for 6–12 months of accumulations, while the weakest link was found for the one month precipitation accumulation (Fig. 4). These results suggest that the changes in the monthly groundwater levels can not be explained by the month-to-month precipitation variability, rather the smoothed response of groundwater requires the contribution from seasonal to annual precipitation. These results are consistent with findings of other recent studies performed in different regions (e.g., Bloomfield and Marchant, 2013; Li and Rodell, 2015), and the findings here assert to the general notion of the groundwater system acting as a low pass filter, responding to moderate climate forcings (e.g., Eltahir and Yeh, 1999; Weider and Boutt, 2010).

The discrepancy between SGI and SPI was further quantified using the mean absolute error (E) criterion to provide a quantitative estimate of the error E in the units of the SPI or

SGI (i.e., between 0 and 1). The resulting E value for both point and gridded data sets on average ranged between 0.17 and 0.26 for different accumulation periods of the SPI (Fig. 4). These are quite substantial errors considering that the threshold used to classify between a drought and a no-drought event is usually taken as 0.2 for the quantile based drought indices (Sheffield et al., 2004; Andreadis et al., 2005; Vidal et al., 2010; Samaniego et al., 2013). In this case, even the minimum mean absolute error (E) estimates corresponding to the spatially varying optimal accumulation periods were fairly large with an average estimate of around 0.15–0.16 for the point and the gridded data sets. These high degrees of discrepancies between the SGI and the SPI clearly indicate the inability of the precipitation based drought index to adequately characterize groundwater drought events.

Results of our analysis also show a relatively larger spread in both statistical criteria (i.e., r and E) estimated for the point data sets as compared to the gridded ones (Fig. 4). This once again emphasizes the importance of local scale heterogeneities in propagating the precipitation signals to groundwater. Clearly, the exhibited high variability of precipitation and groundwater anomalies at a point scale are smoothed out at a grid scale due to the spatial averaging that resulted in a better correspondence between the gridded indices at a regional scale (Fig. 4). Despite the better agreement, the error between the gridded SGI and SPI at any of the uniform or the optimal accumulation periods remained substantially high with an average value of at least 0.15 – the error level that is comparable to a threshold value ($\tau = 0.2$) used to classify droughts.

Overall, the above presented results signify the importance of identifying an appropriate drought time scale i.e., the optimal accumulation period of precipitation based drought indices which is best correlated to impact variables (e.g., streamflow or groundwater levels indicating hydrological or groundwater drought indices). The application of a single accumulation period over the entire domain or among different impact variables could induce large errors and therefore is not recommended. The diversity of relationships that are usually recorded between drought indices and impact variables stresses the need of testing initially the best time-scales of a drought index to determine possible impacts. It is however noted that such analysis would require a good quality of impact variable datasets and

for many regions for which we need reliable and accurate datasets (e.g., on groundwater drought information) these observations are often not readily available. Nevertheless, the issue of analysing an appropriate drought time scale is not only specific for the groundwater system but also relevant for several other hydrological and ecological systems (e.g., Pasho et al., 2010; Vicente-Serrano et al., 2011, 2012; López-Moreno et al., 2013; Vicente-Serrano et al., 2013; Haslinger et al., 2014; Bachmair et al., 2015; Van Loon, 2015).

3.3 Temporal evolution of SPI and SGI

Figure 5 shows the exemplary time series of the SGI and SPI at 6 and 12 months of the accumulation periods for all wells and grid cells and their respective spatial averages for an overlapping period of 1995–2006. The SGI estimates for both point and gridded data sets exhibited higher spatial variability that cannot be adequately represented by their respective SPIs regardless of the accumulation periods used. This points out the enhanced role of soil, vegetation, and hydro-geological properties in propagating the precipitation signal through the subsurface. These observations are consistent with the findings of Weider and Boutt (2010) who also found that groundwater levels in New England have higher (spatial) variability in their responses than other hydro-meteorological variables including precipitation and streamflows.

The SPI and the SGI for large-scale drought events like that of 1996 and 2003 show a remarkable regional difference between German and Dutch wells (Fig. 5). A drought is defined when the indices (e.g., SPI) fall below a threshold (τ) value of 0.2. For instance, the regionally averaged SPI estimates indicate the most severe and extended (prolonged) droughts during the 1996 event for Dutch wells, which was not so strongly pronounced at the German wells (Fig. 5). The opposite behavior was, however, noticed for the 2003 drought event, where the SPI pointed towards more severe drought situations at the German wells than at the Dutch wells. The regional differences were also apparent in the anomalies of groundwater heads (SGI) with German wells showing on average a relatively smoother groundwater response compared to the highly fluctuating and variable groundwater anomalies at the Dutch wells (Fig. 5). In comparison to the SPI, the regionally

averaged SGI exhibited far less severe drought condtions, although there were some wells at which the drought severity based on SGI and SPI were comparable (Fig. 5). This is in accordance with a well known phenomena of the drought attenuation while propagating through sub-suface media and the groundwater compartment of the terrestrial water cycle (H. Hisdal and L. M. Tallaksen, 2000; Van Loon, 2015). Notably, the 1996 and the 2003 drought events that appeared in the averaged SPI at the Dutch and the German wells, respectively, were not strongly pronounced in their respective SGI estimates to characterize these events as severe large-scale groundwater droughts.

~~This underpins~~ The above presented results underpin the inability of the SPI to satisfactorily ~~characterize droughts~~ track the drought events in the groundwater compartment even when applied at longer time scales. The propagation of precipitation signals to groundwater droughts is largely controlled by both catchment and climatic characteristics such as terrain, soil and geological properties, and precipitation seasonality, snowmelt timing, and atmospheric water supply and evaporative demand. This results in a pronounced spatial variation of hydrologic (groundwater) drought characteristics (see e.g., Peters et al., 2006; Bloomfield and Marchant, 2013; Haslinger et al., 2014; Vicente-Serrano et al., 2012; Teuling et al., 2013; Van Loon et al., 2014; Stoelzle et al., 2014; Van Loon, 2015). On the aspect of climatic variables, the SPI that fully accounts for the atmospheric water supply side does not include the effects of the evaporative water demand which could be a determining factor in a hydrologic (groundwater) drought analysis (Vicente-Serrano et al., 2010; Teuling et al., 2013). Another meterological index such as the SPEI Vicente-Serrano et al. (2010) which accounts for both the atmospheric water supply and evaporative demand is expected to be better suited for characterizing hydrologic (groundwater) droughts. However, results of our preliminary investigation (not shown here) indicated that there is not much benefit in using the SPEI over the SPI for the groundwater drought analysis in the study regions. This could be because of the fact that these regions are characterized by a high precipitation variability which dominates over the influence of temperature variability (expressed in the evaporation term of the SPEI). We however recognise that both meterological based drought indices may exhibit a slight difference

during some specific (summer) months and time periods. We note that our study mainly focuses on assessing the skill of SPI and the evaluation of other drought indices (like SPEI or some model based indicators) which in itself would be an interesting research work is beyond the scope of current study.

3.4 Skill of the SPI to predict groundwater droughts

The skill of the SPI to predict groundwater droughts is assessed using the probabilistic scores based on the hit rate (H) and the false alarm ratio (F) (see Sect. 2.3 for a description of their estimation). The results shown in Fig. 6 for H indicate that for a drought threshold τ of 0.2, the SPI was only able to correctly predict three out of five ($H \geq 0.6$) SGI-based groundwater droughts at less than 12 % of the German and 16 % of the Dutch wells for any of the two uniform accumulation periods (6 and 12 months) of SPI. Even in the case of the SPI corresponding to the spatially varying optimal accumulation period (Fig. 6a), only 21 and 18 % of the German and the Dutch wells exhibited an H score greater than 0.6, respectively. The low reliability of the groundwater drought predictions using the SPI was also confirmed from the F scores, shown in Fig. 7, for which at least three in every five events ($F > 0.6$) were wrongly predicted at around 50 % of the wells for both uniform accumulation periods (6 and 12 months) of SPI. In this case, around 30 % of the wells in both regions exhibited a high false alarm ratio ($F > 0.6$) for groundwater drought predictions using the SPI with the optimal accumulation periods (Fig. 7a).

Although the skill of the SPI for the gridded data sets was better than that of the well-specific ones, both the H and the F scores for the gridded data were far from their best scores (Figs. 6 and 7). Overall, the H score on average ranged between 0.52 and 0.58 for the optimal and uniform accumulation periods of SPI, and the corresponding F score varied between 0.44 and 0.50. These results clearly highlight the limited skill of the gridded SPI to capture regional scale groundwater droughts with either optimal or uniform accumulation periods of SPI. Furthermore, the grid cells for which the SPI and SGI is constructed based on the point scale data of very few underlying wells (< 3) exhibited a slightly lower H (and higher F) scores compared to the others. This lower correspondence between SPI (optimal)

and SGI was noticed in a few grid cells (7 out of a total 69 cells). For the remaining grid cells, there was no systematic pattern of improvement or deterioration in the skill scores with the increasing number of underlying wells, which indicated that the difference in number of underlying wells among the grid cells had a relatively minor to no effect on the results presented here for the regional scale groundwater drought analysis.

Results of the further analysis for predicting more severe and extreme groundwater drought conditions also revealed a significantly poor skill of the SPI at both point and grid scales (Fig. 8). For example, the 6 and 12 months of uniform accumulation period based SPI predictions for the severe groundwater drought conditions ($\tau = 0.1$) exhibited an average hit rate H of around 0.26 (i.e., only one in every four events is correctly predicted) for the point data sets, and around 0.33 (i.e., only one in every three events is correctly predicted) for the gridded data sets. The corresponding average F score was quite high around 0.79 (i.e., nearly four in every five events predicted are false alarms) and 0.67 (i.e., two in every three events predicted are false alarms), respectively. Even with the spatially varying optimal accumulation period, the overall skill of the SPI was poor, with an average H score of 0.30 and 0.39 for the point and the gridded data sets, respectively. The corresponding F score was 0.72 and 0.60, respectively.

The performance of the SPI further deteriorated drastically for the predictions of the extreme groundwater drought conditions ($\tau = 0.05$), regardless of the accumulation periods and spatial resolution of the data sets (Fig. 8). These results highlighted the limited reliability of the SPI for predicting groundwater droughts at different severity levels. Among other things, these levels are used for watching (or tracking) the onset, development, and termination of drought events – essential elements to any effective drought monitoring system (e.g., USDM). The skillful predictions of these drought conditions are of critical importance because planners and water managers need to know for example the onset of droughts to take appropriate drought mitigative actions to reduce damages (Hayes et al., 2010).

4 Conclusions

In this study we assessed the ability of the precipitation based drought index (SPI) to characterize groundwater droughts at more than 2000 wells located in two regions in Germany and the Netherlands. These two groundwater networks consisting of a large number of wells and available records allowed us to quantitatively evaluate the skill of the SPI for groundwater drought monitoring at a local and regional scale using the well-specific and the 0.5° gridded data sets, respectively. On the basis ~~on~~ of this data-based exploratory analysis, we found that the precipitation needs to be accumulated over several (3–24) months to (temporally) align the SPI with the SGI time series at both local and regional scales, reflecting the significantly smoothed response of groundwater to precipitation signals. Despite this alignment and with a relatively fair degree of correlation, the SPI lacked the skill to predict groundwater droughts based on SGI. The necessary accumulation periods varied considerably in space however and were not known beforehand. We found that the thickness of the unsaturated zone (expressed here as the average depth to the water table) partly but not entirely controlled the spatial variation of the accumulation period. The groundwater levels at the wells located in relatively deeper aquifers exhibited on average stronger correlation with longer time scale SPI, and vice-versa. There was, however, a considerable noise in this relationship and further studies are required to investigate the possible role of other land-surface and hydro-geological properties including aquifer storage and transmission characteristics.

The application of the uniform accumulation periods over the entire domain significantly reduces the correlation between SPI and SGI indicating the limited applicability of SPI as a proxy for groundwater droughts even at long accumulation times. The differences between the SGI and SPI at both point and gridded data sets were substantially high and generally comparable to the often used threshold value ($\tau = 0.2$) to classify droughts. Based on the results of this multiscale analyses, the assumption of an average smoothing of precipitation to mimic groundwater response during droughts is highly unrealistic.

Depending on the region, the severity of SPI-based drought events differed greatly from those based on the SGI. In some cases the SPI-based extreme droughts (e.g., 1996 or 2003) only showed up in some groundwater wells but not in the spatially-averaged SGI, indicating the enhanced role of the subsurface medium in modulating the precipitation signal. Future studies may look into disentangling the roles of the individual subsurface medium attributes and climatic factors. The predictions of groundwater droughts at different severity levels are crucial for water utilities and regulators for planning (e.g., management of abstraction rates and tariffs, etc) and decision making (e.g., restricting water usage and rationing). The results of the probabilistic scores based on the hit rate and the false alarm ratio clearly indicated the inability of the SPI to capture these aspects of drought conditions, and would therefore be inadequate for monitoring and planning purposes. While these categorical contingency table based skill scores clearly outlined the limitations of the SPI to detect the SGI based groundwater drought events, more insights could be gained by analyzing the differences among different drought characteristics (e.g., duration, severity and maximum intensity) derived based on the SPI and SGI time series. Future studies may therefore look into these aspects.

~~In contrast to recent studies that have mainly focused on analyzing the correlation between precipitation-based index and groundwater drought index~~ In addition to the analysis focusing on the correspondence between SPI and SGI over their entire ranges $[0, 1]$, representing both dry and wet conditions, ~~here we specifically aimed at analyzing the ability~~ we put specific emphasis on assessing the skill of the SPI to predict groundwater drought conditions at different severity levels ~~.~~ (i.e., SGI $\leq$ 0.2 or 0.1 or 0.05). These analyses have allowed us to gain more insights into the limitations of a precipitation-based drought index to properly identify the groundwater droughts. Based on the results obtained in this study, the hypothesis that the observation-based SPI can adequately predict groundwater droughts could not be supported for the analyzed point and gridded data sets.

The evidence presented in this study regarding the inability of the SPI to characterize groundwater drought events at both local and regional scales calls for a different observation-based indicator like the SGI. If for data availability reasons, the precipitation-

based drought indicator is used for groundwater drought studies, the aforementioned limitations should be borne in mind. ~~Finally we~~ We stress the need for putting more efforts in the collection and collation of groundwater data, so that groundwater observations become available on global scale to characterize groundwater drought and the availability of subsurface water resources during drought, at spatial scales small enough to be relevant for water resources management. Finally, in this study we screened our observational wells to keep minimal human influence on groundwater levels, but we note that anthropogenic changes in land use and water use in most of the world today are contributing to the discrepancy between SPI and SGI. This human influence can and should not be disregarded by only using SPI to characterise hydrological drought because it creates a false image of the drought situation on the ground and its impact on people (Van Loon et al., 2016).

Acknowledgements. We are grateful to the Bavarian Environment Agency (LfU Bayern) and the State Institute for Environment, Measurements and Nature Conservation Baden-Württemberg (LUBW) who provided the southern Germany groundwater data sets for use in GLOWA-Danube (www.glowa-danube.de), a project funded by the German Ministry of Education and Research (BMBF); and the German Weather Service (DWD) for providing the precipitation data for Germany. We also wish to thank the Royal Netherlands Meteorological Institute (KNMI) in De Bilt, the Netherlands (www.knmi.nl/index_en.html) for providing the precipitation data and the Dutch institute TNO (www.dinoloket.nl/) for the Dutch groundwater level data. The Dutch authors were partly funded by the EU-FP7 project DROUGHT-R&SPI (contract no. 282769). This paper supports the work of the UNESCO-IHP VIII FRIEND programme and the Helmholtz Climate Initiative REKLIM project.

The article processing charges for this open-access publication were covered by a Research Centre of the Helmholtz Association.

References

AghaKouchak, A., Farahmand, A., Melton, F. S., Teixeira, J., Anderson, M. C., Wardlow, B. D., and Hain, C. R.: Remote sensing of drought: Progress, challenges and opportunities, Reviews of Geophysics, 53, 452–480, 2015.

Andreadis, K. M., Clark, E. A., Wood, A. W., Hamlet, A. F., and Lettenmaier, D. P.: Twentieth-Century Drought in the Conterminous United States, J. Hydrometeorol., 6, 985–1001, 2005.

Bachmair, S., Kohn, I., and Stahl, K.: Exploring the link between drought indicators and impacts, Natural Hazards and Earth System Sciences, 15, 1381–1397, 2015.

Barthel, R.: An indicator approach to assessing and predicting the quantitative state of groundwater bodies on the regional scale with a special focus on the impacts of climate change, Hydrogeology Journal, 19, 525–546, 2011.

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, 2013.

Damberg, L. and AghaKouchak, A.: Global trends and patterns of drought from space, Theoretical and Applied Climatology, 117, 441–448, 2014.

Eltahir, E. A. B. and Yeh, P. J.-F.: On the asymmetric response of aquifer water level to floods and droughts in Illinois, Water Resources Research, 35, 1199–1217, 1999.

Famiglietti, J. S.: The global groundwater crisis, Nature Clim. Change, 4, 945–948, 2014.

Fan, Y., Li, H., and Miguez-Macho, G.: Global Patterns of Groundwater Table Depth, Science, 339, 940–943, 2013.

Folland, C. K., Hannaford, J., Bloomfield, J. P., Kendon, M., Svensson, C., Marchant, B. P., Prior, J., and Wallace, E.: Multi-annual droughts in the English Lowlands: a review of their characteristics and climate drivers in the winter half-year, Hydrology and Earth System Sciences, 19, 2353–2375, 2015.

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

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., , and Koirala, S.: Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe, J. Hydrometeor., 13, 604–620, 2012.

Guttman, N. B.: Accepting the standardized precipitation index: A calculation algorithm, J. Am. Water Resour. As., 35, 311–322, 1999.

H. Hisdal and L. M. Tallaksen: Drought event definition, Tech. Rep. 6, ARIDE, University of Oslo, Norway, 2000.

Hao, Z., AghaKouchak, A., Nakhjiri, N., and Farahmand, A.: Global integrated drought monitoring and prediction system., Scientific data, 1, 1–10, 2014.

Haslinger, K., Koffler, D., Schoener, W., and Laaha, G.: Exploring the link between meteorological drought and streamflow: Effects of climate- catchment interaction, Water Resour. Res., 50, 2468–2487, 2014.

Hayes, M., Svoboda, M., Wall, N., and Widhalm, M.: The Lincoln Declaration on Drought Indices: Universal Meteorological Drought Index Recommended, Bull. Amer. Meteor. Soc., 92, 485–488, 2010.

Joetzjer, E., Douville, H., Delire, C., Ciais, P., Decharme, B., and Tyteca, S.: Hydrologic benchmarking of meteorological drought indices at interannual to climate change timescales: a case study over the Amazon and Mississippi river basins, Hydrol. Earth Syst. Sci., 17, 4885–4895, 2013.

Kraijenhoff-van de Leur, D.: A study of non-steady groundwater flow with special reference to a reservoir coefficient, Ingenieur, 70, 87–94, 1958.

Lana, X., Serra, C., and Burgueño, A.: Patterns of monthly rainfall shortage and excess in terms of the standardized precipitation index for Catalonia (Spain), Int. J. Climatol., 21, 1669–1691, 2001.

Li, B. and Rodell, M.: Evaluation of a model-based groundwater drought indicator in the conterminous U.S., J. Hydrol., 526, 78–88, 2015.

López-Moreno, J. I., Vicente-Serrano, S. M., Zabalza, J., Beguería, S., Lorenzo-Lacruz, J., Azorin-Molina, C., and Moran-Tejeda, E.: Hydrological response to climate variability at different time scales: A study in the Ebro basin, Journal of Hydrology, 477, 175–188, 2013.

McKee, T. B., Doesken, N. J., and Kleist, J.: The relationship of drought frequency and duration to time scales, Eighth Conference on Applied Climatology, Anaheim, California, 17-22 January 1993, 1993.

Mishra, A. K. and Singh, V. P.: A review of drought concepts, J. Hydrol., 391, 202–216, 2010.

Mo, K. C. and Lettenmaier, D. P.: Objective Drought Classification Using Multiple Land Surface Models, J. Hydrometeor, 15, 990–1010, 2013.

Nickel, D., Barthel, R., and Braun, J.: Large-scale water resources management within the framework of GLOWA-Danube - The water supply model, Physics and Chemistry of the Earth, Parts A/B/C, 30, 383–388, 2005.

Nijssen, B., Shukla, S., Lin, C., Gao, H., Zhou, T., Ishottama, Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: A Prototype Global Drought Information System Based on Multiple Land Surface Models, J. Hydrometeor., 15, 1661–1676, 2014.

Pasho, E., Camarero, J. J., de Luis, M., and Vicente-Serrano, S. M.: Impacts of drought at different time scales on forest growth across a wide climatic gradient in north-eastern Spain, Agricultural and Forest Meteorology, 151, 1800–1811, 2010.

Peters, E., Torfs, P. J. J. F., van Lanen, H. A. J., and Bier, G.: Propagation of drought through groundwater - a new approach using linear reservoir theory, Hydrol. Process., 17, 3023–3040, 2003.

Peters, E., van Lanen, H. A. J., Torfs, P. J. J. F., and Bier, G.: Drought in groundwater-drought distribution and performance indicators, J. Hydrol., 306, 302–317, 2005.

Peters, E., Bier, G., van Lanen, H. A. J., and Torfs, P. J. J. F.: Propagation and spatial distribution of drought in a groundwater catchment, J. Hydrol., 321, 257–275, 2006.

Prudhomme, C., Giuntoli, I., Robinson, E. L., Clark, D. B., Arnell, N. W., Dankers, R., Fekete, B. M., Franssen, W., Gerten, D., Gosling, S. N., Hagemann, S., Hannah, D. M., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., and Wisser, D.: Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment, Proceedings of the National Academy of Sciences, 111, 3262–3267, 2014.

Samaniego, L., Kumar, R., and Zink, M.: Implications of Parameter Uncertainty on Soil Moisture Drought Analysis in Germany, J. Hydrometeor., 14, 47–68, 2013.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., , and Zhang, X.: Changes in climate extremes and their impacts on the natural physical environment, in: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change, edited by Field, C. B., Barros, V., Stocker, T., Qin, D., Dokken, D., Ebi, K., Mastrandrea, M., Mach, K., Plattner, G.-K., S.K. Allen, M. T., , and Midgley, P., pp. 167–172, Cambridge University Press, Cambridge, UK, and New York, NY, USA, 582pp., 2012.

Sheffield, J. and Wood, E. F.: Drought: Past problems and future scenarios, Earthscan, 2011.

Sheffield, J., Goteti, G., Wen, F., and Wood, E. F.: A simulated soil moisture based drought analysis for the United States, J. Geophys. Res., 109, D24 108, 2004.

Stoelzle, M., Stahl, K., Morhard, A., and Weiler, M.: Streamflow sensitivity to drought scenarios in catchments with different geology, Geophys. Res. Lett., 41, 6174–6183, 2014.

Tallaksen, L., Hisdal, H., and Lanen, H. V.: Space-time modelling of catchment scale drought characteristics, J. Hydrol., 375, 363–372, 2009.

Tallaksen, L. M. and Stahl, K.: Spatial and temporal patterns of large-scale droughts in Europe: Model dispersion and performance, Geophys. Res. Lett., 41, 429–434, 2014.

Tallaksen, L. M. and Van Lanen, H. A. J.: Hydrological Drought: Processes and Estimation Methods for Streamflow and Groundwater, Developments in Water Sciences, Elsevier BV, the Netherlands, 2004.

Tallaksen, L. M., Hisdal, H., and van Lanen, H. A. J.: Propagation of drought in a groundwater fed catchment, the Pang in the UK, in: Climate variability and change: Hydrological Impacts, edited by Demuth, S., Gustard, A., Planos, E., Scatena, F., and Servat, E., vol. 308, pp. 128–133, International Association of Hydrological Sciences (IAHS), 5th FRIEND World Conference Havana, Cuba, November 2006, Wallingford, UK, IAHS Publication, 2006.

ten Broek, J., Teuling, A. J., and Van Loon, A. F.: Comparison of drought indices for the province of Gelderland, the Netherlands, Tech. Rep. 16, DROUGHT-R and SPI, 2014.

Teuling, A. J., Van Loon, A. F., Seneviratne, S. I., Lehner, I., Aubinet, M., Heinesch, B., Bernhofer, C., Grünwald, T., Prasse, H., and Spank, U.: Evapotranspiration amplifies European summer drought, Geophys. Res. Lett., 40, 2071–2075, 2013.

van Huijgevoort, M. H. J., Hazenberg, P., van Lanen, H. A. J., Teuling, A. J., Clark, D. B., Folwell, S., Gosling, S. N., Hanasaki, N., Heinke, J., Koirala, S., Stacke, T., Voss, F., Sheffield, J., and Uijlenhoet, R.: Global Multimodel Analysis of Drought in Runoff for the Second Half of the Twentieth Century, J. Hydrometeor., 14, 1535–1552, 2013.

Van Loon, A. F.: Hydrological drought explained, Wiley Interdisciplinary Reviews: Water, 2, 359–392, 2015.

Van Loon, A. F., van Huijgevoort, M. H. J., and Van Lanen, H. A. J.: Evaluation of drought propagation in an ensemble mean of large-scale hydrological models, Hydrol. Earth Syst. Sci., 16, 4057–4078, 2012.

Van Loon, A. F., Tijdeman, E., Wanders, N., Van Lanen, H., Teuling, A. J., and Uijlenhoet, R.: How climate seasonality modifies drought duration and deficit, Journal of Geophysical Research: Atmospheres, 119, 4640–4656, 2014.

Van Loon, A. F., Gleeson, T., Clark, J., van Dijk, A. I. J. M., Stahl, K., Hannaford, J., Di Baldassarre, G., Teuling, A. J., Tallaksen, L. M., Uijlenhoet, R., Hannah, D. M., Sheffield, J., Svoboda, M., Verbeiren, B., Wagener, T., Rangecroft, S., Wanders, N., and Van Lanen, H. A. J.: Drought in the Anthropocene, Nature Geoscience, 9, 89–91, 2016.

Vicente-Serrano, S. M., Begueria, S., and López-Moreno, J. I.: A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index, Journal of Climate, 23, 1696–1718, 2010.

Vicente-Serrano, S. M., Begueria, S., and López-Moreno, J. I.: Comment on "Characteristics and trends in various forms of the Palmer Drought Severity Index (PDSI) during 1900–2008" by Aiguo Dai, Journal of Geophysical Research, 116, D19 112, 2011.

Vicente-Serrano, S. M., Begueria, S., Lorenzo-Lacruz, J., Camarero, J. J., López-Moreno, J. I., Azorin-Molina, C., Revuelto, J., Morán-Tejeda, E., and Sanchez-Lorenzo, A.: Performance of Drought Indices for Ecological, Agricultural, and Hydrological Applications, Earth Interactions, 16, 1–27, 2012.

Vicente-Serrano, S. M., Gouveia, C., Camarero, J. J., Begueria, S., Trigo, R., López-Moreno, J. I., Azorin-Molina, C., Pasho, E., Lorenzo-Lacruz, J., Revuelto, J., Morán-Tejeda, E., and Sanchez-Lorenzo, A.: Response of vegetation to drought time-scales across global land biomes., Proceedings of the National Academy of Sciences, 110, 52–57, 2013.

Vidal, J. P., Martin, E., Franchistéguy, L., Habets, F., Subeyroux, J. M., Blanchard, M., and Baillon, M.: Multilevel and multiscale drought reanalysis over France with the Safran-Isba-Modcu hydrometeorological suite, Hydrol. Earth Syst. Sci., 14, 459–478, 2010.

Wanders, N., Wada, Y., and Van Lanen, H. A. J.: Global hydrological droughts in the 21st century under a changing hydrological regime, Earth System Dynamics, 6, 1–15, 2015.

Weider, K. and Boutt, D. F.: Heterogeneous water table response to climate revealed by 60 years of ground water data, Geophysical Research Letters, 37, 1–6, 2010.

Wilhite, D. A.: Drought as a natural hazard: Concepts and definitions, in: Drought: A Global Assessment, edited by Wilhite, D. A., pp. 3–18, Routledge Hazards and Disasters Series Vol. 2, 2000.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, 2011.

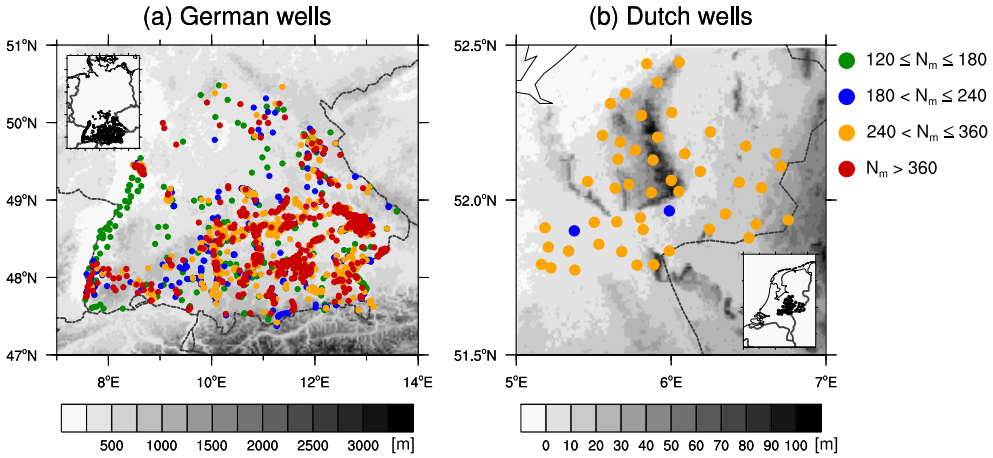

Figure 1. The locations of **(a)** German and **(b)** Dutch wells overlaid on the respective terrains. The marker colors show the number of months N_{m} with available records during the periods 1950–2013 and 1988–2013 for German and Dutch wells, respectively.

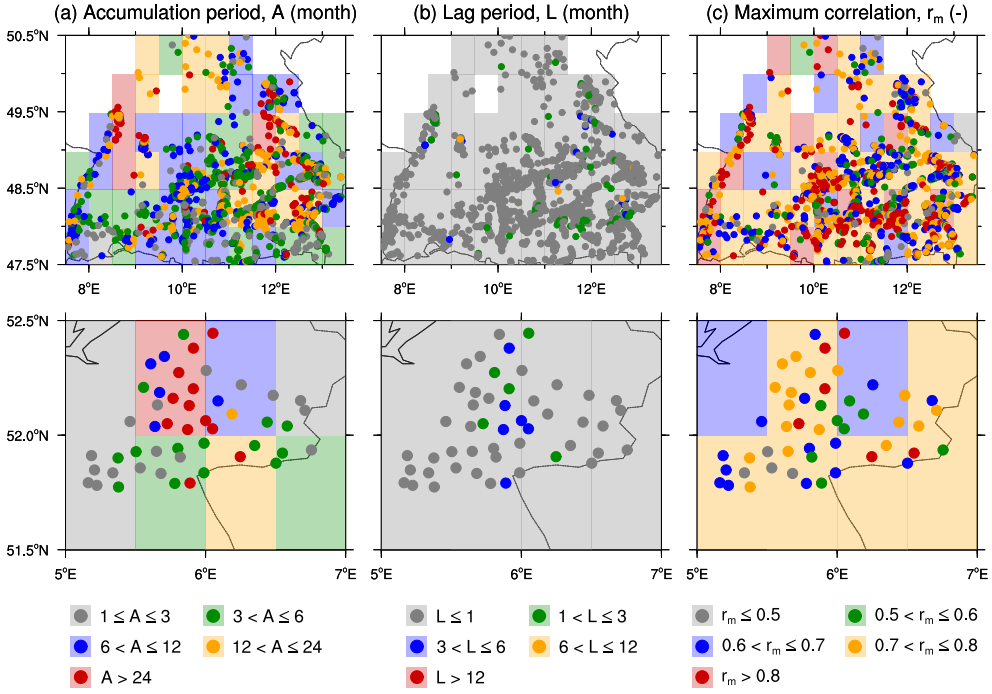

Figure 2. The **(a)** optimal accumulation A (month) and **(b)** lag periods L (month) required to obtain the **(c)** maximum correlation r_{m} (–) between the SGI and SPI at point and gridded (0.5°) scales for German (top) and Dutch (bottom) data sets.

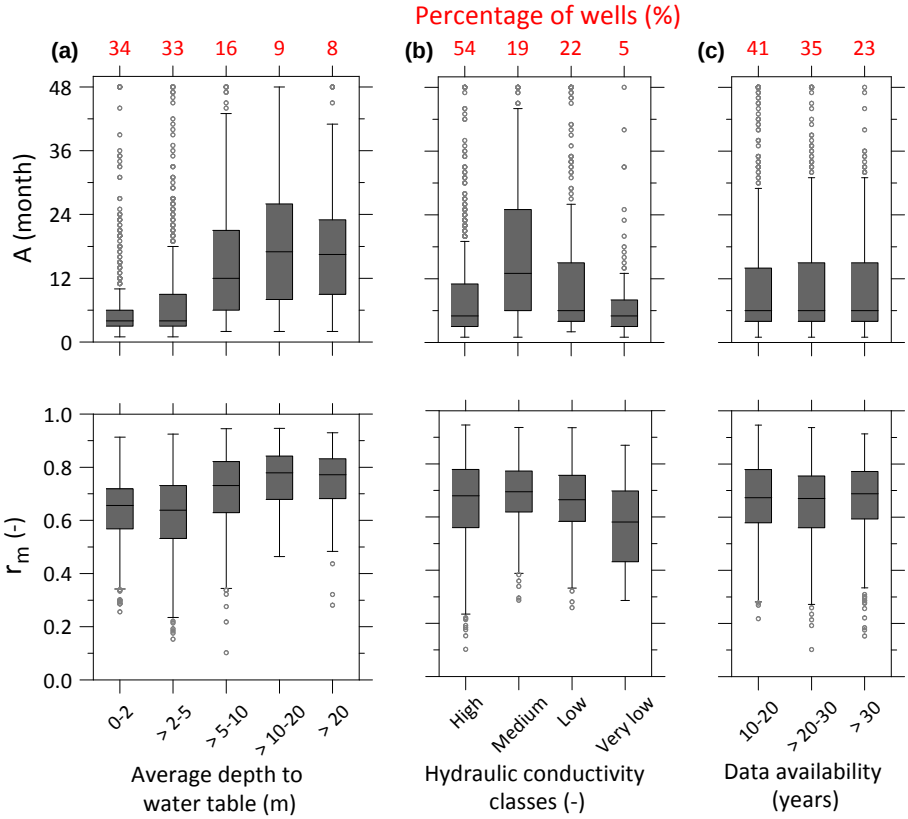
Percentage of wells (%)
(a) 34 33 16 9 8
(b) 54 19 22 5
(c) 41 35 23
A (month)
48
36
24
12
0
r_m (-)
1.0
0.8
0.6
0.4
0.2
0.0
0-2
> 2-5
> 5-10
> 10-20
> 20
Average depth to water table (m)
High
Medium
Low
Very low
Hydraulic conductivity classes (-)
10-20
> 20-30
> 30
Data availability (years)

Figure 3. Box-and-whisker plots of the optimal accumulation period ~~(~~A (top) and the maximum correlation ~~(~~r_{m} (bottom) estimated for a group of wells with varying depth to water table (~~top~~left: **a**~~and~~), aquifer hydraulic conductivity classes (middle: **b**), and record lengths (~~bottom~~right: **c**). Results shown for the aquifer hydraulic conductivity correspond to the German wells are grouped into four distinct classes: high ($> 10^{-3}$ m/s), medium (10^{-3}–10^{-5} m/s), low (10^{-5}–10^{-7} m/s), and ~~d~~very low ($< 10^{-7}$ m/s). The percentage of wells falling within each group is indicated at the top of every plot.

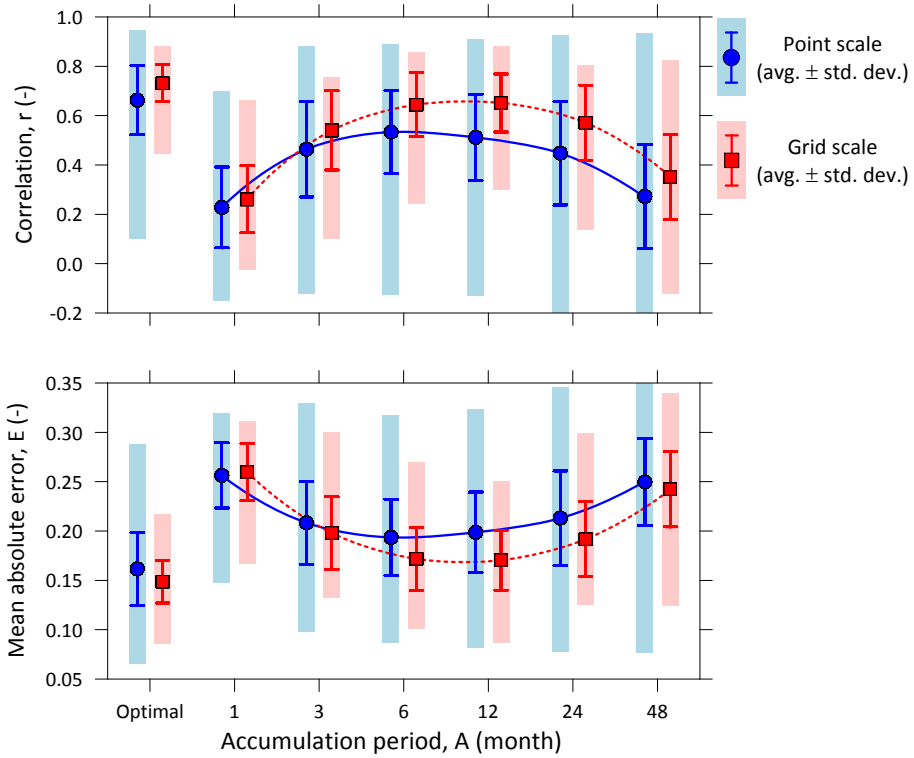

Figure 4. The correlation r (top) and the mean absolute error E (bottom) estimated between the SGI and SPI of the 1, 3, 6, 12, 24, and 48 months of uniform accumulations for the point and the gridded data sets. Their respective maximum (r_{m}) and the minimum (E_{m}) estimates corresponding to the optimal accumulation periods of SPI are also shown in the leftmost of the panels. ~~Results~~ Summary statistics are ~~summarized here~~ provided as an average $\pm$ one standard deviation, and the entire range is depicted as filled bars in the background.

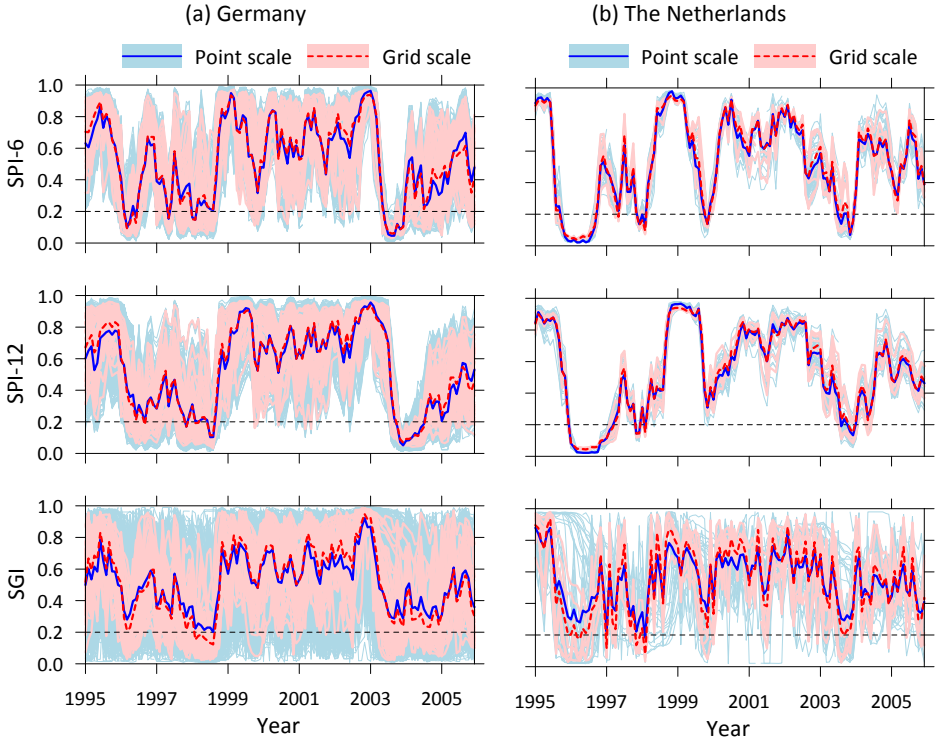

Figure 5. The monthly time series of the 6 and 12 month point (light blue) and gridded (light pink) SPI and the respective spatial averages (dark blue and dark red) for **(a)** German and **(b)** Dutch data sets. The bottom plots are the corresponding SGI time series and their spatial averages. The black dashed line depicts the drought threshold τ of 0.2.

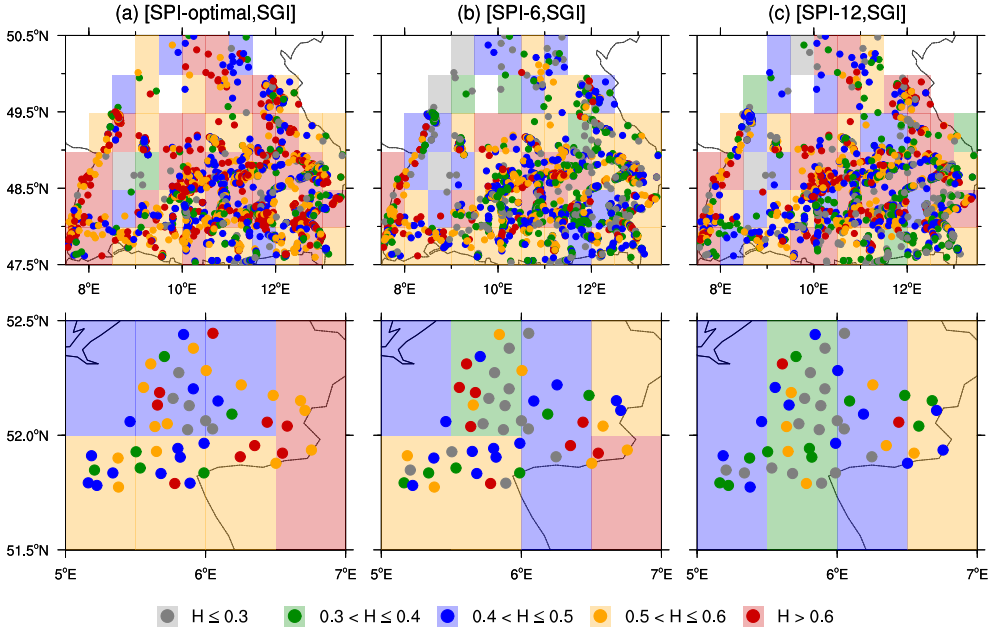

Figure 6. The hit rate (H) to detect SGI based groundwater droughts using the SPI with the **(a)** optimal accumulation period and **(b, c)** 6 and 12 months of uniform accumulation periods at the point and gridded scales for German (top) and Dutch (bottom) data sets. A threshold value τ of 0.2 is used to identify drought events.

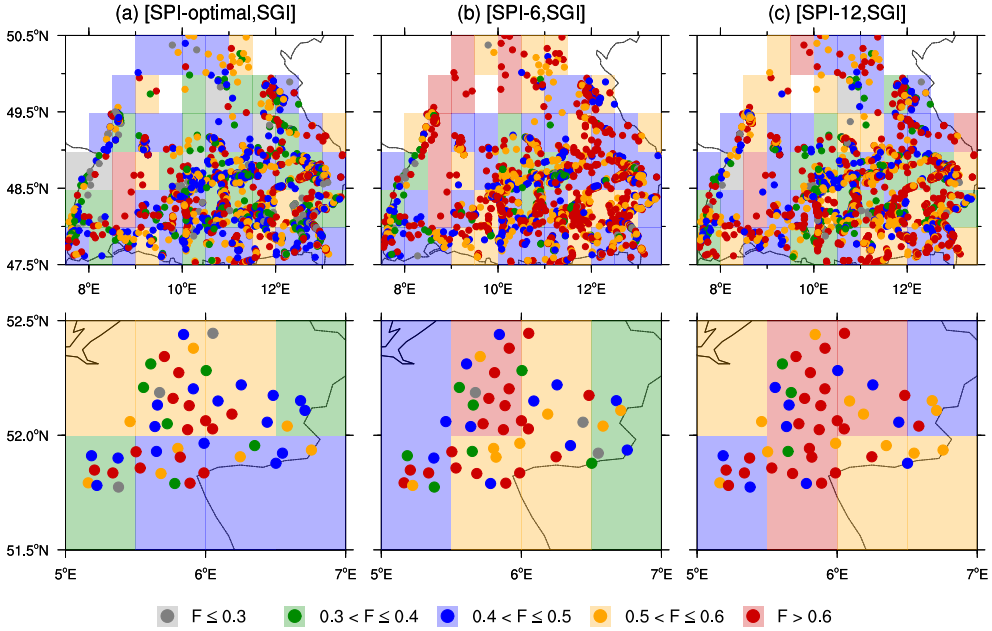

Figure 7. The false alarm ratio (F) to detect SGI based groundwater droughts using the SPI with the **(a)** optimal accumulation period and **(b, c)** 6 and 12 months of uniform accumulation periods at the point and gridded scales for German (top) and Dutch (bottom) data sets. A threshold value τ of 0.2 is used to identify drought events.

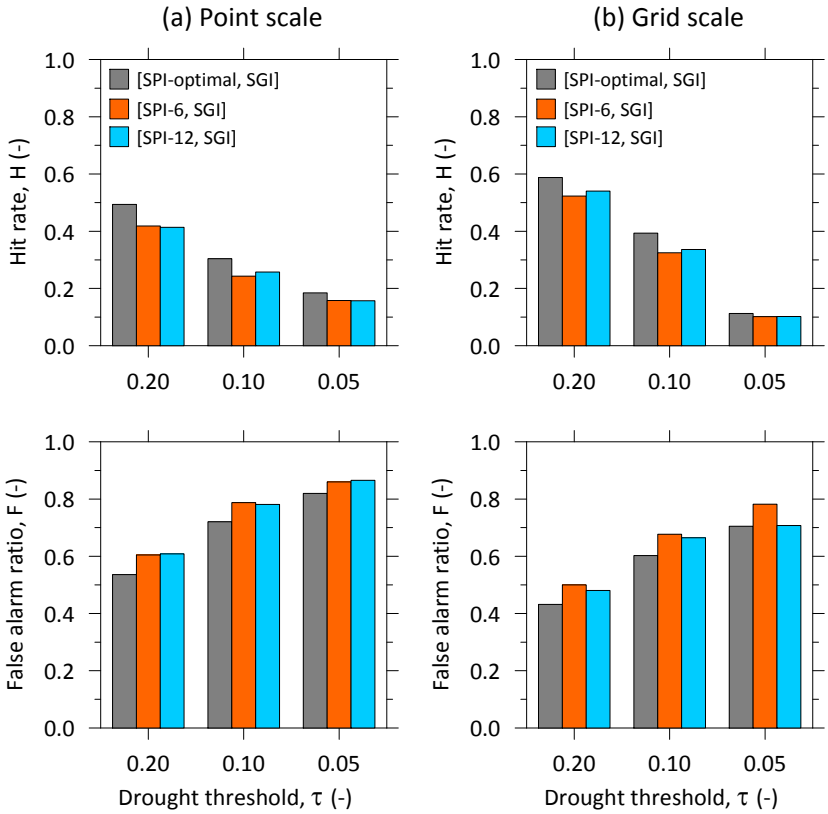

Figure 8. The hit rate (H) and the false alarm ratio (F) averaged over all investigated **(a)** wells and **(b)** grid cells to detect SGI based groundwater droughts using the SPI with the optimal accumulation and 6 and 12 months of uniform accumulation periods for varying levels of threshold value τ (0.2, 0.1, and 0.05) used to identify drought events.