We would like to thank the two anonymous Referees and Jim Freer for their feedbacks and helpful suggestions on this manuscript. Below we give point-by-point responses to all the comments (bold and italic). The new manuscript with tracked changes can be found below the responses.

Editorial comments:

The two core matters seem to be:
1) The tests are rather arbitrary, limited, and they are not comprehensive in nature - I also noted on the submission that they seem rather simplistic and not well justified in terms of any actual expected climatic variation - synthetic or not. I believe this needs to be resolved by improving the experimental base for the analyses, and I believe the authors intend to do this but they still need to be linked to some justification of 'climate change' else the paper should simply be about a sensitivity analyses and I am not sure that is enough for publication. So whilst the authors offer more % changes this is not enough to tie in better and more objectively with potential future scenarios and so the reasoning still has to be improved

We expanded the experimental base of our study by performing more model runs resulting in more robust and reliable findings (see answers to question #1 of reviewer 1 and question #2 of reviewer 2). Additionally, we improved the manuscript to make clear that our approach is complementary and different from the climate projections ensemble approach and starts with testing the general sensitivity of the system first. Subsequently this sensitivity can be used to assess the vulnerability of the system – not limited to but with particular relevance to expected climate change. For putting the results of the stress tests in the context of model projected climate change, of course projections must be considered in addition.

In the revised manuscript, we now consistently use 'stress test' instead of 'scenario' to avoid confusion. We better introduce, argue and explain this approach in the introduction (Section 1), eliminated all obscuring connections to climate change projections from the stress test descriptions (Section 3) and put a reflection regarding climate change in the discussion (Section 5). We think that this reorganization makes the reasoning of the manuscript now clearer and puts the insights for climate change adaptations in the right perspective.

2) That the model evaluation is improved - in the sense that some understanding of caution (or indeed not reported) for regions where the model does not do that well - which I add the authors have in some ways explored within their WRR paper. I believe I also noted this matter of model quality in my initial assessment.

We significantly expanded the manuscript regarding model evaluation specifically for the properties relevant in the stress tests and added our reflections to Section 3, 5 and the Supplement (for details see our response to comment #3 of reviewer 2).

Referee #1:

1) The paper describes a study where a large-scale, high-resolution MODFLOW-groundwater model of Germany has been used to assess a range of potential changes to groundwater and baseflow drought hazard based on

three change scenarios. The scenarios are: i.) a changed recharge regime with wetter winters and drier summers (SSHIFT), ii.) changes to antecedents conditions associated with three major historic episodes of drought in Germany (SEVENT), and iii.) recovery from drought (SRECOV). These scenarios were co-designed in part with the Climate and Water Initiative of southern Germany's federal states (KLIWA) (L67-86) with the aim of stress testing the sensitivity to drought of groundwater and baseflow. Although, the geographical focus of the study is Germany, the paper addresses questions relevant to a wide readership and is clearly in the scope of HESS.

The description of the model setup (Section 2) is adequate given that more details can be found in the paper by Hellwig et al. (2020) who developed the model. However, a critical assessment by the authors of the model's suitability, including method of calibration and appropriateness of it's underlying assumptions, for the current application would be helpful. The description of the scenario design and modelling approach (Section 3) is generally clear and well-reasoned. However, the scenarios appear somewhat arbitrary. In particular, the formulation of the SRECOV scenario is less convincing than the other two scenarios. To assess the maximum duration for groundwater recovery from severe drought, the lowest simulated groundwater heads are taken as an initial condition and groundwater heads are simulated using long-term average monthly recharge as input until an arbitrary recovery has been achieved. Although adequately described, the motivation and justification for the details of this scenario are not given. The results are presented well, both graphically and in their description in Section 4.

The national-scale groundwater model is certainly limited in its local representation. Therefore, we expanded the manuscript regarding model evaluation and added our assessment specifically for the stress tests to Section 3, 5 and the Supplement (for details see our response to comment #3 of reviewer 2).

We agree that the formulation of S_{RECOV} might appear first of all arbitrary, but it is in fact similar to well-established probabilistic real time forecasting practices. We selected this stress test, however, to address the stakeholders request for a better understanding of the drought termination period. To account for the local differences in the groundwater system, we adopted the idea of a 'composite map' and did not select one specific drought year but rather the lowest heads of the simulated period. We agree that the recovery threshold is arbitrary, however, additional analyses with other thresholds (40 and 50-percentile groundwater head) resulted in the same patterns (see section 4.3 of the revised manuscript). To test the influence of the recharge on T_{rec}, **as suggested,** ***we performed two additional model runs for the revised manuscript representing dry (wet) conditions during drought recovery. Accordingly, drought recovery decelerated (accelerated). However, general spatial patterns remained similar as T_{rec} is strongly related to T_{max}, which is an aquifer characteristic independent of meteorology/recharge or initial conditions (see Figure 9 c) + d) of the revised manuscript).***

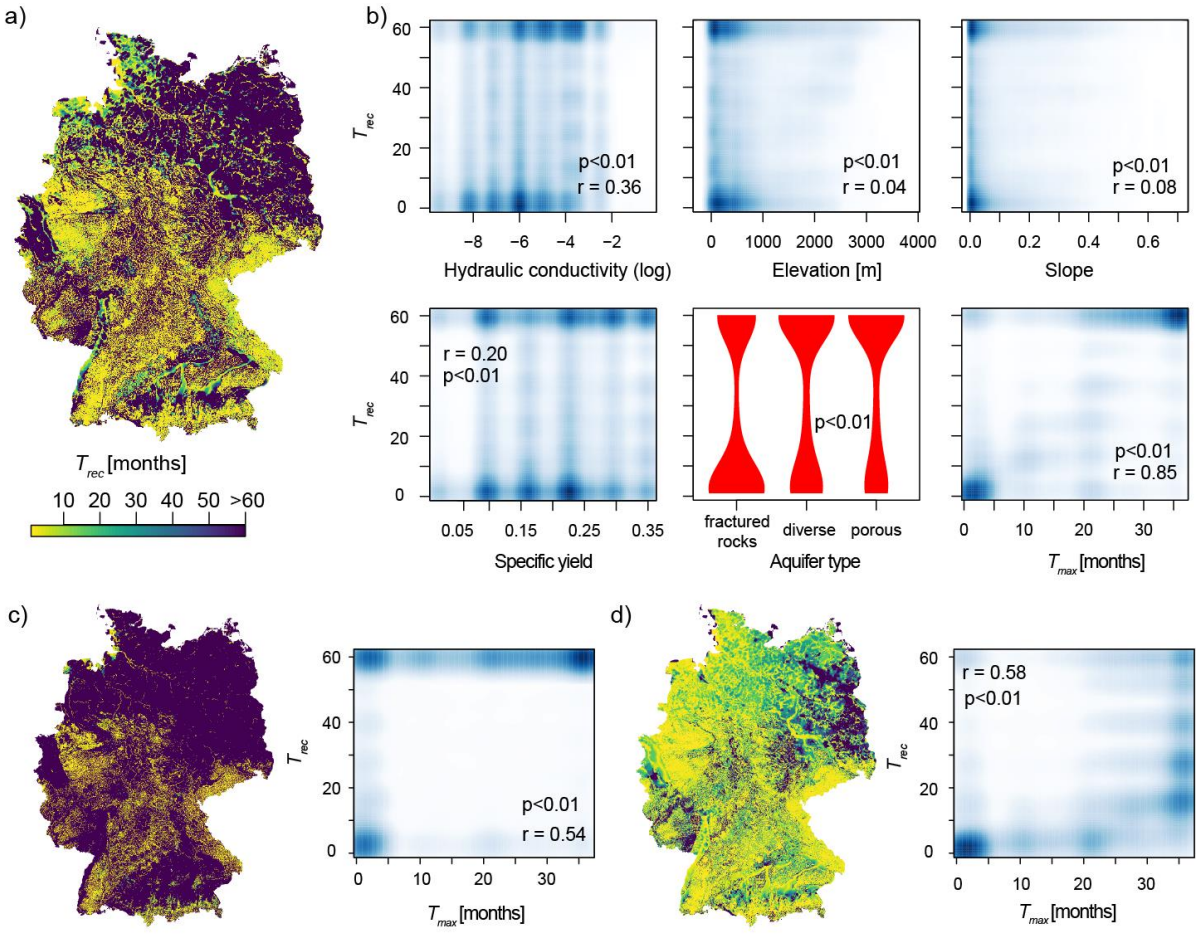


Figure 9: Recovery time T_{rec} for S_{RECOV}. a) spatial distribution of T_{rec} across Germany, b) relationship between T_{rec} and model parameters hydraulic conductivity, elevation, slope and specific yield, aquifer type (taken from HÜK200) and precipitation accumulation time that has the maximum correlation with groundwater T_{max} (taken from Hellwig et al., 2020). c) + d) spatial distribution of T_{rec} and T_{rec} over T_{max} for dry resp. wet conditions during drought recovery. Blue colours indicate the smoothed density derived from all model grid cells. Red violins illustrate the distribution of T_{rec} in three different categories of aquifer type. r is the Pearson correlation coefficient for the variables compared, p is the corresponding p-value.

We added these results to the manuscript and discuss the influences of initial conditions, assumed recharge and recovery threshold on T_{rec} now in more detail (see section 4.3)

2) The Discussion provides a number of interesting insights into the results. For example, the authors make the observation at L287-290 that: “the different responses of baseflow and groundwater are important to consider

for an effective water management in a changing climate. For example, in a climate with higher annual recharge sums but more frequent summer droughts groundwater droughts might become less severe while the baseflow drought hazard becomes more severe with potential impacts on economy and ecology". Given that the scenarios that led to this observation were shaped by stakeholders, it would be interesting to know if and how stakeholders might use such information. More generally, given the nature of the set-up of the paper (e.g. L67-74) it would be interesting to hear the author's views on any specific implications of the results of their study for drought planning and management. These could be described, however briefly, in the Discussion.

We added a brief discussion on potential adaptations that may be considered by stakeholders to our findings.

> ***"The different responses of baseflow and groundwater are important to consider for an effective water management and drought planning in a changing climate. Different stakeholders will face different challenges in future and use the stress tests differently to design adaptation or to plan mitigation measures for emergency plans. For example, in a climate with higher annual recharge sums but more frequent or severe summer droughts groundwater droughts might become less severe while the baseflow drought hazard becomes more severe. Where possible, one option might be to switch or add water use from surface water to groundwater to meet water demands for irrigation, industry, and public water supply. For other purposes relying on a minimal amount of surface water (e.g. navigation, water quality, or ecosystem health) adaptations such as regional water transfers or increased surface water storage capabilities might be more expedient." (ll. 314-321)***

3) Specific comments: L229 I think that the authors meant "relative" not "relevant"?

Agree, changed.

4) Section 4.1. The authors make a number of observations relating to the groundwater and baseflow changes being more pronounced under average conditions than for drought, and this is also highlighted in the Summary at L326. A brief interpretation and discussion of the implications of these observations would be helpful.

Agree. We added the following paragraph to the Discussions:

> ***"Inter- and intra-annual changes in recharge do not only affect the immediate drought hazard in a different way for different hydrogeology but will also cause various changes to the long-term groundwater and baseflow dynamics. A change of recharge variability will not necessarily result in a change of hydrological drought conditions, where response times are long enough or where a change in variability is caused by changes in the mean or the wet climate and recharge extreme. Hence, assessments of potential changes regarding average conditions or variability may have minor or no information for proactive drought planning. Our results suggest that drought assessments directly relevant for specific stakeholders' needs and analysed in the context of the local sensitivity determined by hydrogeological conditions will better allow for adaptation and planning." (ll. 284-290)***

5) Section 4.3 and Figure 9. The main feature of the analysis of recovery time appears to be the essentially bi-modal nature of Trec, this being most evident in the Trec vs. Tmax plot in Fig. 9. It would be interesting to hear what the authors think might be contributing to this result. Does it reflect intrinsic characteristics of the modelled system, is it an artefact of the model structure or calibration, or is it some combination of both? Perhaps such a discussion could be added to Section 4.3?

The bi-modal nature of T_{rec} is strongly related to T_{max}. The large differences in T_{max} with some regions responding within few months and others over several years are a characteristic we can also find in observations (e.g. see Figure 7 in Hellwig et al., 2020), so we don't think it is just an artefact of the modelling. However, in the simulation there is for many grid cells no recovery from the extreme drought within the timeframe of five years. Certainly, this is an important factor influencing the distribution of T_{rec}. In the additional analysis simulating T_{rec} under dry recharge conditions no bi-modal distribution was found since also regions of large T_{max} recover within the timeframe (see Figure 9d above). We added this important point to the Results (Section 4.2).

6) L348-350. The first and only mention of the application of this approach to is in the Conclusions. This seems strange. It may be appropriate to include these observations in the Discussion, but not in the conclusions?

This paragraph was modified with a different focus in the revised manuscript.

Referee #2:

1) This paper tackles an important topic of how groundwater and baseflow will respond to changes in recharge. To test this, the study uses MODFLOW to explore how groundwater and baseflow change in response to three different recharge scenarios across Germany. The recharge scenarios are informed from stakeholder interactions and the combination of the scenarios targets different characteristics of groundwater and baseflow drought responses. The study concludes that a shift in rainfall to wetter winters and drier summers will not cause decreases in groundwater resources in general, but water managers need to consider the potential for more severe groundwater droughts following prolonged dry spells. The figures are well presented and the paper is generally well written.

The results could be of significant interest to the scientific community. However, my overall assessment is that major changes to the paper with additional simulations are required before the paper is suitable for publication. Currently the paper explores a very limited set of scenarios and thus does not robustly "stress test" or truly assess the sensitivity of groundwater and baseflow drought responses to different scenarios. It is difficult to have confidence in the conclusions that are presented in the paper when they are based on a single change for each scenario. This becomes particularly important given the significant non-linearities between changes in groundwater head and baseflow, as highlighted by the authors. A critical assessment of the model's suitability to simulate groundwater and baseflow drought responses is also needed.

These comments are discussed in more detail below, which I hope the authors find useful.

We agree that the confidence in the results will be increased by running more simulations to show that our findings are independent from specific assumptions. However, with our stress tests we specifically aim to meet stakeholders' requests for simple and easily interpretable scenarios that rather give information on possible general directions of change instead of uncertainty ranges depending on specific scenario assumptions. This is also an important difference from scenarios in the sense of ensembles of slightly different pathways. To make this difference of our stress tests more transparent, we adopted the terminology in the revised manuscript and use now consequently "stress test" instead of "scenario". Additionally, we performed additional model runs to make the conclusions more confident and expanded the evaluation (details below).

2) **Scenarios** – The scenarios are very limited. If the aim of the paper is to test and attribute specific sensitivities as noted in the introduction then a larger number of simulations should have been undertaken. Conclusions such as "a seasonal shift of recharge (i.e. less summer recharge and more winter recharge) will therefore have low effects on groundwater and baseflow drought severity" need to be based on more than a single scenario of +/-15% to be robust. Specific comments are:

(1) *SShift* – This scenario applies a 15% increase in recharge for winter months and a 15% decrease in recharge for summer months to the whole time series. Running a single set of percentage changes applied to the whole timeseries provides a very limited view of the question posed of "How will a changed recharge regime with wetter winters and drier summers change the inter-annual variability and water availability during droughts?". The authors should explore this in more depth by running additional scenarios that vary the percentage increases.

(2) *Srecov* – The justification for this scenario is quite weak compared to the other two scenarios and again is very limited in that it only explores the response under the assumption of long term average recharge.

(3) *Comparison between scenarios* – In the discussion and conclusions, comparisons between the scenarios are made. However, it is difficult to be confident in these comparisons as only a single scenario is assessed. For these comparisons to be robust additional simulations need to be performed to assess the sensitivity of the drought response to each scenario.

The intention of our S_{SHIFT} and S_{RECOV} generic stress tests is to identify site-specific sensitivities to certain general hydroclimatic conditions which are of special interest to different. Typically, for the groundwater system these sensitivities are much more driven by the physiographic and hydrogeological conditions compared to the exact climatic forcings tested. However, we agree that the results can become more reliable with a broader range of tested forcings (in our case recharge) and different responses of baseflow and groundwater can be analysed in more detail with more model runs.

To test the influence of the assumed percentual shift on changes of (drought) percentiles in S_{Shift} we performed four additional model runs for the revised manuscript with changes of 5, 10,20 and 30% in winter (increase) and summer (decrease). Patterns of change are for all seasons and percentiles the same as for a 15% change (as it was assumed in the original manuscript). Solely the magnitude of changes differs for the percentages of

change (Figure S3). Also, the characteristic differences in baseflow and groundwater response to the seasonal shift are the same for the different percentage shifts. We added these additional results from the new runs to the revised manuscript (Section 4.1).

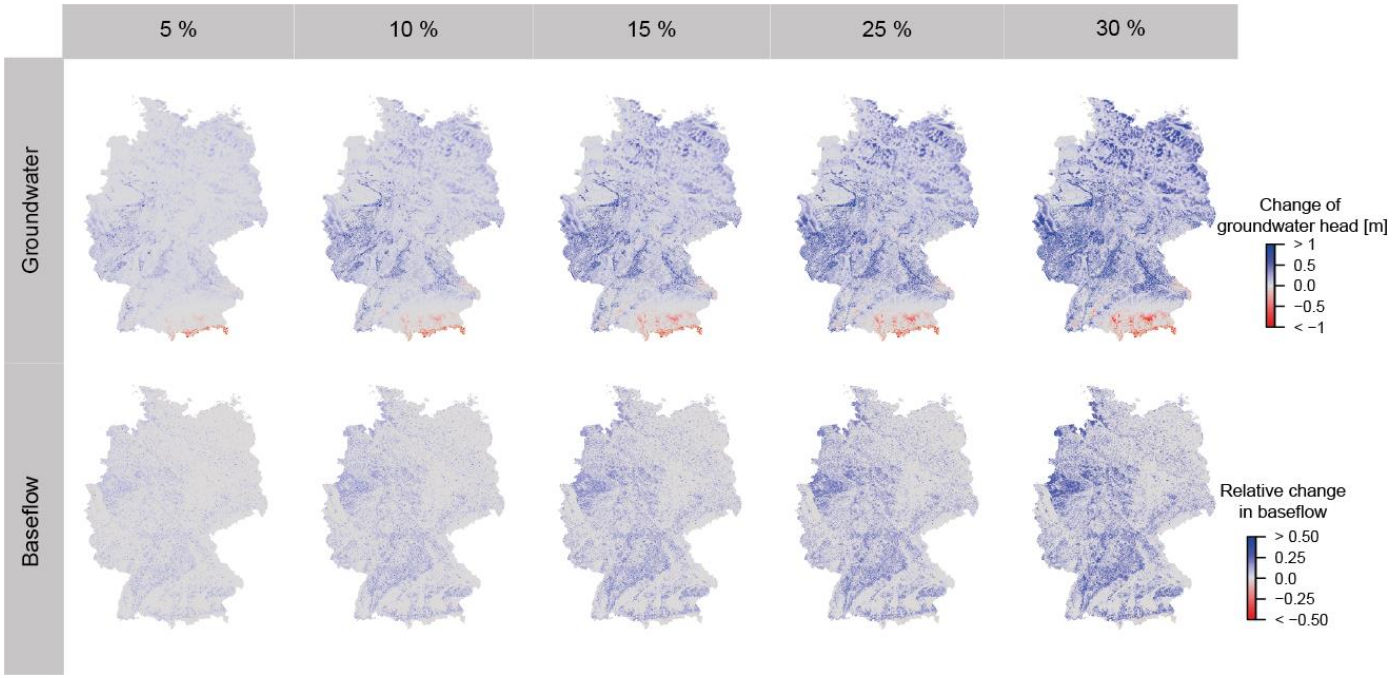


Fig S3: Example for changes of groundwater heads and baseflow (rows) for the different percentage changes in S_{Shift} (columns). Changes are shown for winter and moderately dry conditions (τ = 0.25), all other seasons and percentiles similarly differ most of all in the magnitude of change.

For S_{RECOV} we also performed additional model runs and tested for the impact of different assumptions. We find that exact T_{rec} depends on assumptions such as recovery threshold and recharge conditions during recovery, however, general spatial patterns that are related to T_{max} are valid (for details see our response to comment #1 of reviewer 1). With the additional model runs for S_{SHIFT} and S_{RECOV} the results are now more confident and allow for a more a reliable comparison between the different stress tests.

3) **Model Evaluation** – I agree with reviewer 1 that a critical assessment of the model's suitability for this application is required in Section 2. The authors need to demonstrate that the model can effectively reproduce the metrics that are used in this paper to assess groundwater and baseflow drought responses (e.g. the recovery time T_{rec}, inter-annual variability, percentile thresholds, performance during "benchmark droughts") and how this varies spatially and temporally for Germany. Currently, the discussion in Section 2 centres on model performance for T_{max} which is based on correlations and not focused on the (likely) non-linear drought responses that are being assessed here.

The generic stress tests in the paper focus on sensitivities during drought. We agree that the model's ability to simulate the dynamics targeted in the stress tests is crucial for the reliability of the results. Hence, we

expanded our reflections on the abilities and limits of the groundwater model. Specifically, in the revised manuscript we added Table 2 defining the required model ability for each stress test type and discussing the model evaluation in these specific regards. However, we think that T_{max} is still a very important evaluation metric to understand model behaviour and particularly the non-linearity of baseflow and groundwater head response: Overall T_{max} for baseflow is much shorter than for groundwater directly relating to a larger dependency on intra-annual climate dynamics for baseflow and on inter-annual dynamics for groundwater heads. In Hellwig et al. (2020) it was demonstrated that these differences, which lead to the non-linearities found in our study, are appropriately captured by the model. We stated this importance of T_{max} for the interpretation of the results more clearly in the revised manuscript.

Table 2: Required model ability and discussion of model performance for the different stress tests.

	Required model ability	Evaluation metric	Discussion of model performance
S_{SHIFT}	Reliable propagation of inter- and intra-annual recharge dynamics into groundwater heads and baseflow	T_{max}	Overall, the model depicts both differences of T_{max} across the study area and the systematically shorter T_{max} of baseflow compared to groundwater. However, for baseflow T_{max} was notably overestimated in the North and underestimated in the South while for groundwater it was overestimated in the porous aquifers of the lowlands and underestimated in higher elevations (see Hellwig et al., 2020 for more detailed analyses). Hence, absolute S_{SHIFT} responses may be biased in that same way. The model estimates allow most confidence in the representation of general shift-patterns across the study area.
S_{EVENT}	Reliable model representation of benchmark drought events	Differences between observed and modelled groundwater /baseflow drought severities	Simulations and observations show a considerable variability of groundwater drought severity for different drought years across the study area. Consistent with observations, modelled drought severities were weaker in 2003 compared to 1973 with several regions in the study area not in groundwater drought. These patterns are also consistent with state agency reports (see Hellwig et al., 2020). However, especially in the Northeast the model responds too slowly (corresponding

			with too long T_{max}, see above) leading to deviating groundwater drought severities: the drought severity of 1973 is overestimated in the model while it is underestimated for 2003. For baseflow model performance is similar: while general patterns of drought severity can be depicted, drought severities deviate most in the North (-East) (see also Figure S1). Overall, there are systematic uncertainties arising from the comparison of observational data with model outputs which might relate to some of the differences found (for a more advanced discussion on that see Hellwig et al., 2020, Section 2.3).
S_{RECOV}	Reliable representation of severe drought + propagation of recharge forcing into groundwater	Combination of evaluation metrics of S_{SHIFT} and S_{EVENT}	As both general patterns of drought severities and the propagation of the forcing into groundwater are captured by the model, prerequisites for an appropriate drought termination simulation are given. Uncertainties for this stress test are – similar to the other stress tests – largest in regions of weaker model performance regarding T_{max}.

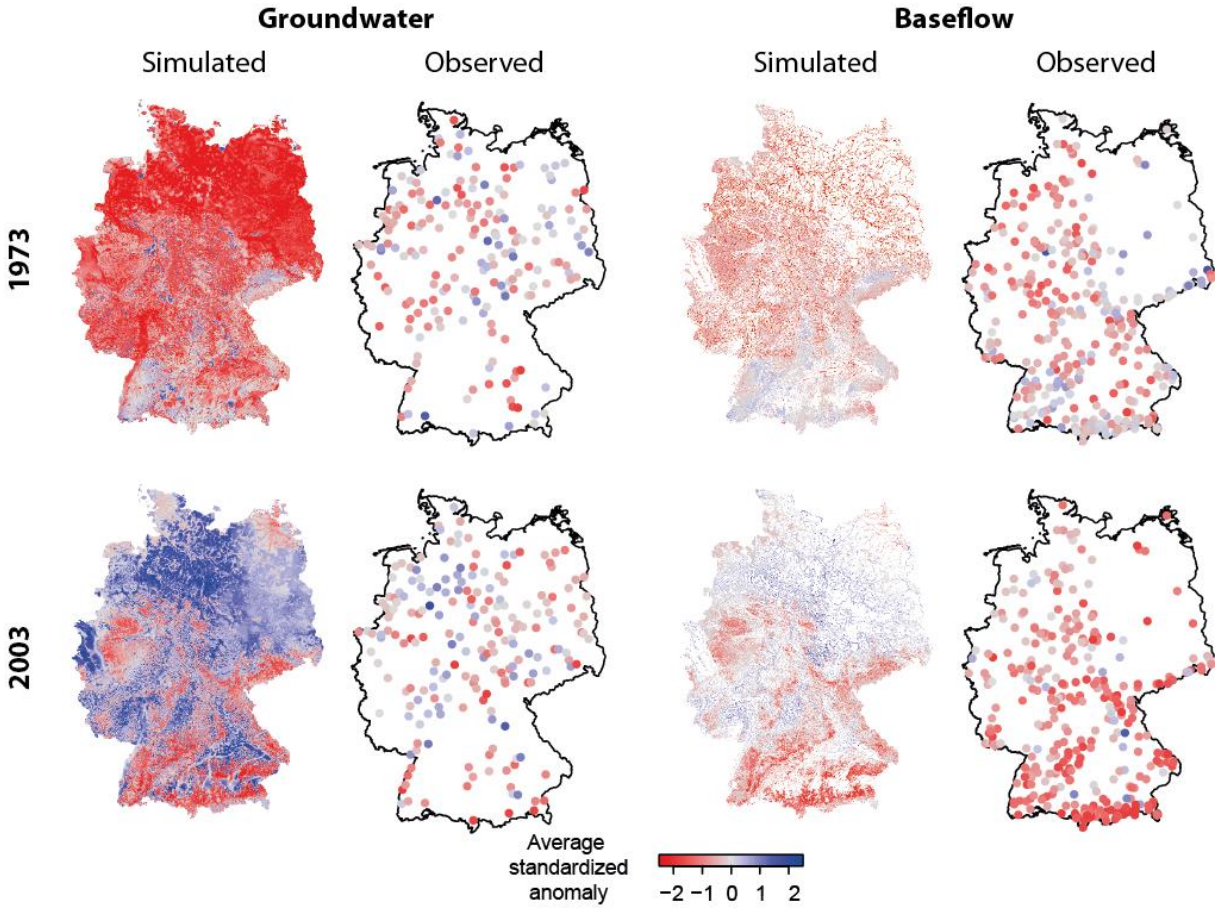


Figure S1: Simulated and observed anomalies averaged for summer months (JJA) of the benchmark drought years 1973 and 2003. Figure based on data taken from Hellwig et al. (2020).

We think with these additional remarks model results are now better interpretable to the reader while maintaining the focus of the study which are rather the different sensitivities found and not the model design and evaluation.

Additionally, we expanded in the revised manuscript the discussion of the model performance and evaluation:

> ***"Moreover, there is uncertainty arising from the aquifer parametrization. Exact model derived T_{max} as well as groundwater and baseflow drought severity must be taken with care and should not be interpreted exactly to the location. In particular, Hellwig et al. (2020) found a decreasing model performance for higher elevation regions with small scale variability of the hydrogeology. Gleeson et al. (2020) conclude in their commentary that profound (observation-based) model evaluations for large-scale groundwater models are currently beyond reach. Groundwater head dynamics measured at boreholes can deviate considerably from grid cell averages due to a large subgrid heterogeneity (e.g. Kumar et al., 2016). Opposingly, baseflow dynamics can be seen as an integrated spatial signal***

but uncertainties arising from the separation of baseflow from streamflow are large (e.g. Stoelzle et al., 2020a). Also, for other observational data there are severe constraints (Gleeson et al., 2020). Even though these uncertainties limit considerations for an effective local water management, they do not affect the general conclusions on regional groundwater sensitivity found." (ll. 326-335)

4) **Minor Comments and Technical Corrections**

Abstract L7. Please change to "depend on the systems' sensitivity"

Done

Introduction L25-28. I would move (or remove) the two sentences starting with "Contrary to surface water, groundwater is hard to…" to L44 where you discuss the absence of observational data and use of groundwater models in more detail.

Done

Introduction L49. Replace "more and more" with "increasingly"

Done

Introduction L55. "Climate models (often) lack alterations in the sequencing of future wet and dry spells". This sentence needs to be supported by some references.

This part of the introduction was reformulated, and the statement is no longer used in the revised manuscript.

Equation 1 L114. What does the 'f' denote?

f denotes the spatially varying depth determining the rapidity of conductivity decrease with depth. It is inversely related to surface slopes (i.e. a faster decrease of conductivity with depth in areas with higher slopes). We added this explanation to the manuscript.

Section 3 L164. It is not entirely clear to me how you calculate inter-annual variability – can you clarify and provide the equation?

We calculate the variability for the different seasons taking all values from that season from all years. We noted that the term 'inter-annual' is insufficient and replaced it with this more detailed description.

Discussion**.** It might be worth adding some sub-section headings to the discussion to break up the text a little for the reader.

Done

Discussion L267-268 "Also, the recovery time Trec from a severe drought varied accordingly (SRECOV)." I am not sure what you mean here – can you clarify?

The spatial patterns found in S_{RECOV} confirm the patterns of the other stress tests used. We reformulated this sentence.

Supplementary Information. Figures S1-S4 are very difficult to interpret and the figure quality is poor (i.e. they are quite blurry). Can you make the maps bigger and ensure the figures are incorporated at high resolution so that they are clear to the reader.

Done.

References cited in this response:

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., ... and Oshinlaja, N.: HESS Opinions: Improving the evaluation of groundwater representation in continental to global scale models. Hydrol. Earth Syst, Sci. Discuss., 1-39, https://doi.org/10.5194/hess-2020-378, 2020

Hellwig, J., de Graaf, I. E. M., Weiler, M., and Stahl, K.: Large scale assessment of delayed groundwater responses to drought, Water Resour Res., 56(2), e2019WR025441, doi: 10.1029/2019WR025441, 2020.

Kumar, R., Musuuza, J. L., Van Loon, A. F., Teuling, A. J., Barthel, R., Ten Broek, J., . . . Attinger, S.: Multiscale evaluation of the Standardized Precipitation Index as a groundwater drought indicator, Hydrol. Earth Syst, Sci., 20(3), 1117-1131, doi: 10.5194/hess-20-1117-2016, 2016.

Stoelzle, M., Schuetz, T., Weiler, M., Stahl, K., and Tallaksen, L. M.: Beyond binary baseflow separation: a delayed-flow index for multiple streamflow contributions. Hydrol. Earth Syst, Sci., 24(2), 849-867, doi: 10.5194/hess-24-849-2020, 2020a.

Most relevant changes to the manuscript:

- We added model runs for the S_{SHIFT} and S_{RECOV} stress tests.
- We added a model evaluation specifically for the properties relevant in the stress tests.
- We reorganized the manuscript (most importantly in Section 1, 3 and 5) to clarify the difference of our approach from climate change ensemble projections and to put our findings (particularly regarding climate change) in the right perspective
- Minor changes to all Sections and most of the Figures.

~~Stress-testing g~~Groundwater and baseflow drought responses to synthetic ~~climate change-informed~~ recharge ~~scenarios~~stress tests

Jost Hellwig[1], Michael Stoelzle[1], Kerstin Stahl[1,2]

[1]Environmental Hydrological Systems, University of Freiburg, Freiburg, 79085, Germany
[2]Freiburg Institute of Advanced Studies (FRIAS), University of Freiburg, Germany

Correspondence to: Jost Hellwig (jost.hellwig@hydrology.uni-freiburg.de)

Abstract. Groundwater is the main source of freshwater and maintains streamflow during drought. Potential future groundwater and baseflow drought hazards depend on the systems' sensitivity to altered recharge conditions. We performed groundwater model experiments using three different generic ~~scenarios~~ stress tests to estimate the groundwater- and baseflow drought sensitivity to changes in recharge. The ~~scenarios~~ stress tests stem from a stakeholder co-design process that specifically followed the idea of altering known drought events from the past, i.e. asking whether altered recharge could have made a particular event worse. Across Germany groundwater responses to the ~~scenarios~~ stress tests are highly heterogeneous with groundwater heads in the North more sensitive to long-term recharge and in the Central German Uplands to short-term recharge variations. Baseflow droughts are generally more sensitive to intra-annual dynamics and baseflow responses to the ~~scenarios~~ stress tests are smaller compared to the groundwater heads. The groundwater drought recovery time is mainly driven by the hydrogeological conditions with slow (fast) recovery in the porous (fractured rock) aquifers. In general, a seasonal shift of recharge (i.e. less summer recharge and more winter recharge) will therefore have low effects on groundwater and baseflow drought severity. A lengthening of dry spells might cause much stronger responses, especially in regions with slow groundwater response to precipitation. ~~As climate models suggest such directional changes for Germany in the future, the results of the stress tests suggest that groundwater resources in Germany may not decrease in general, but w~~Water management may need to consider the spatially different sensitivities of the groundwater system and the potential for more severe groundwater droughts in the large porous aquifers following prolonged meteorological droughts, particularly in the context of climate change projections indicating stronger seasonality and more severe drought events.

1 Introduction

Freshwater is a vital resource for human life and the demand is growing worldwide simultaneously to economic and demographic growth. The largest accessible storage and one of the most important sources for human water demand is groundwater (Gleeson et al., 2016; Wada et al., 2014), especially in case of low surface water availability, and it is expected to become even more important under climate change (Taylor et al., 2013; Kundzewicz and Döll, 2009). ~~Contrary to surface water, groundwater is hard to observe on larger scales in sufficient resolution for assessments of the meteorological,~~

~~hydrogeological and anthropogenic influences on groundwater dynamics. However, these investigations are essential to better understand large-scale groundwater sensitivity to climate variability, changes and other potential threats.~~

Groundwater serves as a buffer against hydroclimatic variations and is a considerable factor influencing the propagation of drought (Eltahir and Yeh, 1999; Peters et al., 2003). Drought is defined as below normal water availability and starts with a meteorological drought that can propagate through all parts of the hydrological cycle (Van Loon, 2015). It can lead to social and economic impacts, especially during seasons with low water availability compared to water demand. As a natural hazard drought affects people worldwide and causes high economic loss (EC, 2007). Hence, the groundwater's potential to attenuate meteorological droughts influences society´s current and future vulnerability to drought events.

The groundwater response to meteorology can be highly diverse both on small and large scales (Stoelzle et al., 2014; Bloomfield et al. 2015; Kumar et al., 2016, Haas and Birk, 2018). Weider and Boutt (2010) showed that groundwater responses to precipitation anomalies are more heterogeneous compared to the responses of streamflow. Accordingly, Bloomfield et al. (2015), Kumar et al. (2016) and Stoelzle et al. (2014) consistently found that typical time scales of drought propagation into groundwater are site-specific, pointing to the importance of hydrogeological characteristics and subsurface storage processes. The sensitivity to changes in the meteorology will hence be site-specific and is often not generalizable, in particular when considering borehole data from specific locations within an aquifer and relative to rivers or recharge areas (Heudorfer and Stahl, 2017). Hellwig and Stahl (2018) found that the differences in the groundwater response to precipitation anomalies also correspond to varying sensitivities of baseflow to precipitation shifts.

To assess the groundwater and baseflow sensitivity to climate change on larger scales, extensive observational data capturing the large diversity of their responses to meteorology would be required. However, unlike surface water, groundwater is hard to observe on larger scales in sufficient resolution for these analyses. As borehole observations are often hardly scalable (Kumar et al., 2016) they are usually not sufficient to investigate groundwater sensitivity to climate variability on larger scales. ~~As borehole observations are often hardly scalable (Kumar et al., 2016), these datasets are rarely available on a larger scale and~~Therefore, groundwater models are often inevitable for detailed investigations. Recently, the use of large-scale groundwater models including gradient driven lateral flows has gained increasing attention (e.g. Maxwell et al., 2015; de Graaf et al., 2015; Reinecke et al., 2019), as large-scale datasets on aquifer parameters become ~~more and more~~increasingly available. Hellwig et al. (2020) demonstrated that these models can depict the differences in propagation time from meteorological water deficits to groundwater (droughts) on ~~a~~ larger ~~-~~scales reasonably well, concluding that they are also suitable to assess the groundwater's and baseflow's sensitivity to ~~climate~~ recharge changes on larger scales.

A systematic assessment of sensitivities is often realised based on a scenario-neutral ensemble approach, ~~The most common approach to estimate the impact of climatic changes on hydrological systems are model chains starting from emission pathways and global climate models and leading to regional impact models (Keller et al., 2019). In general, climate change scenarios allow the assessment of system changes but quantitative predictions of future changes are subject to large uncertainties (e.g., Lehner et al, 2020). Climate models (often) lack alterations in the sequencing of future wet and dry spells. Both, time~~

~~sequencing and small magnitudes of change, however, matter strongly to low flow and drought responses (Vormoor et al., 2017).~~

~~Alternative approaches such as scenario neutral ensembles testing systems' sensitivities have therefore been proposed~~ for ~~example~~example, to inform planning processes for floods (Prudhomme et al., 2010). Other than commonly used scenarios based on climate change projections, scenario-neutral approaches aim to provide robust information on potential change directions based on the system's characteristics and independent from specific emission scenarios and climate change uncertainties. Designing similar approaches for drought, a slowly developing phenomenon with time lagged signal in streamflow and groundwater, requires the consideration of longer lead times and resulting depletion of catchment storage ~~(e.g. climate change informed seasonal wetting or drying)~~. For example, Staudinger et al. (2015) used ~~scenarios~~ model experiments of progressive drying to assess the streamflow sensitivity to drought for catchments across Switzerland. Stoelzle et al. (2014) developed a model-based ~~scenario~~ stress test approach to study the sensitivity of streamflow to changes in climate based on modifications of the recharge. More applied synthetic stress-testing approaches often use worst-case scenarios to estimate the consequences of specific events (Stoelzle et al., 2020b). Stress-testing sensitivity to drought will help to better understand the degree of resilience of various hydrological systems (Hall and Leng, 2019).

As part of the Climate and Water Initiative of southern Germany's federal states (KLIWA) different types of ~~"~~stress test~~s scenarios"~~ or "what-if experiments~~scenarios~~" were explored as means to better understand and more easily communicate potential future changes to low flow (Stoelzle et al., 2018; 2020b). ~~Scenario~~ Stress test designing included for example a progressive recharge reduction before the 2003 summer drought, as this event is often used as planning benchmark or to assess follow-up costs: the ~~scenarios~~ stress tests ask whether the effect may even have been worse, e.g. with different antecedent conditions. The co-design process of KLIWA revealed different preferences, including rather arbitrary repetitions of sequences of past (known) dry years, very straightforward 'wetter-drier' modifications of past periods or specific drought events, and more systematic approaches with larger model ensembles of modified conditions. In this study we employ~~ed~~ three of the approaches from this co-design process that also allow for a systematic analysis of stress responses (e.g. drought recovery).

Specifically, the ~~scenarios~~ stress tests focus on pre-drought recharge reduction effects on the hydrological drought sensitivity simulated in the groundwater-baseflow domain. Directly modifying groundwater recharge allows to focus the research question to the storage-outflow processes relevant to the hydrology in dry periods. In this study ~~It~~ this modification ~~is justified by the~~ aims ~~of~~ at testing and attributing specific system sensitivities rather than an ~~general~~ overall system response to ~~specific~~ climatic change projections ~~in this study~~. As groundwater has a recharge memory, antecedent recharge conditions are a key factor for groundwater drought severity and the effect of perturbed recharge on drought severity can provide information on the site-specific groundwater and baseflow drought sensitivity. The approach by Stoelzle et al. (2014) illustrated an assessment of the sensitivity to altered recharge in reservoir or box-type hydrological models and was limited to the investigation of baseflow sensitivity.

In this study, we used similar recharge scenariosstress tests, as well as the stress-test ideas of KLIWA, for entire Germany in a large-scale high-resolution MODFLOW-groundwater model to assess a range of potential changes to groundwater and baseflow drought hazard. Specifically, this study aims to

(1) assess potential changes inthe sensitivity of groundwater and baseflow availability during drought due to a climate-change informed a seasonal wetting and drying shift,

(2) identify large-scale sensitivity patterns of groundwater and baseflow drought events to extreme recharge drought conditions with particular return periods, and

(3) quantify characteristic groundwater drought recovery times.

2 Study area and groundwater model setup

The study area of this work is the state of Germany. Germany consists of four main natural regions with different groundwater characteristics (Figure 1): the lowlands in the North with slow responding groundwater in porous aquifers, the uplands in Central Germany with faster responses and mixed aquifer types including fractured rocks and karst aquifers, the Alpine foothills in southern Germany with porous aquifers and the high elevation Alps in the far South with mostly fractured rocks aquifers. Germany's temperate humid climate is characterized by evenly distributed precipitation throughout the year and an annual temperature cycle that results in climatic water deficits due to higher evapotranspiration rates. As a result, groundwater recharge largely takes place during the winter months (Jacob et al., 2012;, Kopp et al.,, 2018). Future climate projections indicate – despite all uncertainties emerging from different models and emission pathways – as a general pattern that precipitation will increase during winter and decrease during summer (e.g. Jacob et al., 2012; Paparrizos et al., 2018; Herrmann et al., 2016). Combined with increasing temperatures over the whole year, recharge will most likely increase in winter and decrease in summer (Eckhardt and Ulbrich, 2003; Stoll et al., 2011; Dams et al., 2012; Hunkeler et al., 2014; Chen et al., 2018). The magnitude of change is highly uncertain with low model agreement compared to other regions in the world (e.g. Reinecke et al., 2019) and depends on the choice of recharge model (e.g. Moeck et al., 2016) as well as the choice of compared reference and future periods.

To assess the groundwater response to recharge scenarios stress tests we apply applied a large-scale groundwater model covering Germany. The model consists of one MODFLOW layer (Harbaugh et al., 2000), simulating groundwater heads, baseflow (i.e. groundwater discharge to surface water) and lateral flows in weekly time steps. It covers all basins intersecting Germany (i.e. river Rhine in the West, river Danube in the South, river Elbe and river Oder in the East) with a spatial resolution of approximately 1 km (latitudinal: 1/22°, longitudinal: 1/14°). Hellwig et al. (2020) developed and evaluated the model, demonstrating its ability to depict the heterogeneous groundwater response to precipitation anomalies even though model performance markedly declined in the mountainous regions due to the larger topographic variability. In the following the model structure and input data are briefly described, for detailed information refer to Hellwig et al. (2020).

Specific yield values were taken from the porosity values in the GLobal HYdrogeology MaPS (GLHYMPS: Gleeson et al., 2014). Initial hydraulic conductivity values k_0 for Germany were derived from the "Hydrogeologische Übersichtskarte" (hydrogeological map HÜK200: BGR and SGD, 2016), for the rest of the model domain k_0 was based on GLHYMPS' permeability values. Consistent with other groundwater models based on a single layer (e.g. Fan et al., 2007; Miguez-Macho et al., 2008), hydraulic conductivity was assumed to decrease exponentially with depth. The characteristic decrease is described by an exponential ~~the *e*-folding~~ spatially varying depth function f which inversely relates hydraulic conductivity to the slopes~~.~~ of surface terrain (i.e. a faster decrease of conductivity with depth in areas with steeper slopes). Then, transmissivity T depends on k_0, f and the current groundwater table depth d_{gw}:

$$T = \int_{d_{gw}}^{100} k_0 e^{\frac{-z'}{f}} dz' \quad (1)$$

where z' is the depth below surface and~~.~~ T is updated every time step.

Interactions between surface water and groundwater were implemented using the RIV-package, simulating flow dependent on the difference of groundwater and surface water heads. Each cell contains either a large river (width > 10 m) with strong interactions with the aquifer or a small stream (width < 10 m) with less interactions. Channel depth, riverbed conductivities and river head over riverbed were derived from long-term average routed baseflow of previous model runs (Hellwig et al., 2020). Baseflow and infiltration was assumed to be proportional to the difference of groundwater heads and surface water heads as well as riverbed conductivity. Hence, with decreasing water tables baseflow reduces and stops when groundwater heads fall below surface water heads.

Groundwater recharge was calculated using a conceptual recharge model consisting of a soil storage and a snow storage. Rainfall, snow and evaporation (following Hargreaves and Samani, 1985) were derived from the European Climate Assessment & Dataset (Haylock et al., 2008), version 16. The soil storage was parameterized with data from the 'Hydrologischer Atlas Deutschlands' (HAD, hydrological atlas of Germany; BMU, 2003). To ensure realistic recharge rates, recharge was rescaled using long-term average recharge estimates from the HAD.

~~The groundwater model was evaluated by Hellwig et al. (2020) using 202 groundwater borehole time series and 338 streamflow observations. Their results suggested that the model can reproduce the standardized time series as well as precipitation accumulation times that have the maximum correlation (T_{max}) with groundwater and baseflow, even though the model is still too coarse for the small-scale variability in the mountainous regions of Germany. The T_{max} measuring the time needed to propagate anomalies from precipitation to groundwater were found to be a good measure for the patterns of groundwater drought following different meteorological drought events. Moreover, T_{max} were found to be related to the model parameters conductivity, specific yield and elevation.~~

This study uses time series of water table and baseflow dynamics from 1970 to 2016 (reference run). For different ~~scenarios~~stress tests, recharge and boundary conditions in the model are altered and resulting water table and baseflow time series are compared to the reference run.

3 ~~Scenario~~ Stress test design and modelling approach

Three types of generic recharge ~~scenarios~~ stress tests addressing different questions for drought management were applied to the groundwater model (Table 1). To do this the ~~scenarios~~ stress tests have different boundary conditions and different recharge modifications. All ~~scenarios~~ stress tests apply relative changes over entire Germany, thus allowing the results to be analysed as composite maps of the same relative change but with respect to the specific local conditions. This sensitivity analysis approach should not be confused with the more common climate change model chain experiments that would apply locally varying changes stemming from the combination of climate model output and hydrology or soil water balance models with particular assumptions and parametrizations of vegetation and soils. The composite maps therefore represent response differences to the designed ~~scenario~~ stress test inputs due to hydrogeology.

The first ~~scenario~~ stress test S_{SHIFT} assumes a ~~potential future~~ change in drought hazard due to an increased seasonality of precipitation and temperature. This stress test aims to answer practitioners' questions how an intra-annual climatic shift in Germany can affect inter-annual variability as well as extreme events such as droughts in groundwater and baseflow (Table 1). Therefore, for S_{SHIFT} precipitation is assumed to increase in winter and decrease in summer whereas temperature increases over the whole year. The stress test experiment consequently increase~~d~~s (decrease~~d~~s) recharge during winter (summer), directly amplifying recharge~~'s~~ seasonality. ~~increase during winter and decrease during summer (e.g. Jacob et al., 2012, Paparrizos et al, 2018, Herrmann et al., 2016). Climate projections for Germany include large uncertainties emerging from different models and emission pathways. In general, projected climatic changes and resulting estimates such as groundwater recharge are small and with low model agreement compared to other regions in the world (e.g. Reinecke et al., 2019). As a general pattern, precipitation - which so far has shown little seasonality in Germany - is projected to increase during winter and decrease during summer (e.g. Jacob et al., 2012, Paparrizos et al, 2018, Herrmann et al., 2016). Combined with increasing temperatures over the whole year, recharge also can be assumed to increase (decrease) in winter (summer) (Eckhardt and Ulbrich, 2003; Stoll et al., 2011; Dams et al., 2012; Hunkeler et al., 2014; Chen et al., 2018). The magnitude of change is highly uncertain (e.g. Moeck et al., 2016)with low model agreement compared to other regions in the world (e.g. Reinecke et al., 2019) and depends on the choice of (e.g. Moeck et al., 2016) reference and future period. Due to this uncertainty in the magnitude, the more general question, whether and where the expected contrasting seasonal change has a general potential to influence low flow and groundwater drought is what water management at the moment is most interested in (Table 1). In S_{SHIFT}, we therefore run T~~t~~he model is run from 1970 to 2016 with different assumptions of ~~shift~~the magnitude (5, 10, 15, 20, 30%) of ~~decreased (increased)~~ recharge shift ~~during~~ from a decrease during summer months (JJA) to an increase during winter months (DJF) (Figure 2). ~~This scenario lies in the range of potential precipitation changes for winter and projected evapotranspiration changes in summer predicted by regional climate and water balance models for Germany until the end of the 21st century (Jacob et al., 2012, Herrmann et al., 2016, Paparrizos et al., 2018). Therefore, S_{SHIFT} should be seen as a generic scenario to gain a composite insight into groundwater and baseflow responses under the assumed seasonal recharge shifts in Central Europe (Table 1).~~

For the assessment of the response to S_{SHIFT} we compare ~~inter-annual~~the variability for different seasons (i.e. variability is calculated for water table/baseflow of selected months taken from all simulated years) and percentile thresholds for water table/baseflow during drought from the ~~scenario~~ stress test run with the reference run forced by original recharge. As a spatially and temporally varying threshold τ we use 0.10, 0.25 and 0.50 representing an exceedance probability of 90, 75 and 50% within the specific season (Van Loon and Van Lanen, 2012; Heudorfer and Stahl, 2017). An increase (decrease) of the water table/baseflow under S_{SHIFT} indicates a higher (lower) water availability for the selected drought severity.

The second ~~scenario~~ stress test type S_{EVENT} focuses directly on the scale of selected drought events and is designed to assess the groundwater's drought sensitivity to systematic changes in the antecedent recharge conditions (Table 1). ~~Changes related to single events might also become relevant in the future (Taylor et al., 2013), but in particular regarding dry spells are generally not well represented in downscaled and bias-corrected climate model derived input and difficult to analyse regarding changes in hydrological drought (Vormoor et al., 2017, Kohn et al., 2019). One potential future hazard is the occurrence of more severe and prolonged meteorological drought events.~~ Practitioners often use past events for the design of drought management plans and ask whether there might be ~~scenarios~~ conditions that had the potential to make ~~these~~ similar events even worse (Table 1). For this study the events of 1973, 2003 and 2015 are selected for the analysis of a range of different but well-known severe benchmark drought years. These drought years have received attention in previous publications, and although they all had large precipitation deficits also differences were noted (e.g. Tallaksen and Stahl, 2014; Laaha et al., 2017; Hellwig, 2019~~; and~~ Hellwig et al., 2020). Due to differences in the recharge conditions before the droughts, the groundwater situation was very different in each case (Hellwig, 2019). While the 1973 event can be characterized as a long-term water deficit leading to depleted water tables across Germany (Figure 3a), the events in 2003 and 2015 were rather severe short-term summer drought events. As the winter 2002/03 was exceptionally wet, ~~in general~~most water tables were not depleted in summer 2003 (Figure 3b). The 2015 event followed ~~on~~ a winter of average recharge and led to a severe groundwater drought in the following summer in the fast responding aquifers in the South whereas the slower responding aquifers in the North did not ~~show~~ develop anomalies corresponding to a groundwater drought (Figure 3c). ~~For~~ With S_{EVENT} these real antecedent recharge conditions for every modelled grid cell ~~were~~ awere further stressed by ~~altered~~ altering recharge for three different durations (3, 9 and 24 months) to investigate ~~groundwater responses on~~ different time scales. The month of the groundwater drought's start is set in May. For the 3-month (9-month, 24-month) ~~scenarios~~ stress tests we modify~~ied~~ recharge backwards from the drought's start for 3 (9, 24) months starting in February (August of the year before, May two years before) and compare~~d~~ the resulting groundwater situation from May to November in the drought year to the reference simulation (Figure 2).

The amount of ~~Antecedent~~ antecedent recharge is modified to represent a "recharge deficit event" with a return period T_{RP} of 50 and 100 years based on the modelled 57 years of reference recharge series for each grid cell (1960-2016). The use of return periods allows a consistent spatial comparison of the same ~~scenario~~ stress test intensity. First, for all three durations the corresponding 57 recharge sums are used to fit a generalized extreme value distribution with Weibull plotting positions. Then, fitted distributions are used to estimate the recharge sums of drought events with $T_{RP} = 50$ and $T_{RP} = 100$ years representing

different drought severities. Finally, the reference recharge time series is rescaled to match these recharge sums while conserving the original variability of the recharge time series (Stoelzle et al., 2014). The reduced recharge is then used as an input for the groundwater model. Altogether, this ~~scenario~~ stress test type consists of 18 model runs: for 3 drought years (1973, 2003, 2015) antecedent recharge ~~was~~ is modified on three time-scales (3, 9, 24 months) to match that of a drought event of two return periods (50y, 100y).

For the assessment of the response to S_{EVENT} we analyse changes in water table/baseflow for all different benchmark droughts, time scales and return periods. Effects of S_{EVENT} are related to potential explanatory variables from the groundwater model: hydraulic conductivity, specific yield, elevation, slope, aquifer type and precipitation accumulation times that have the maximum correlation (T_{max}) with groundwater and baseflow~~T_{max}~~. T_{max} can be understood as the time scale of anomaly propagation from climate to the groundwater system and ranges between one month~~s~~ and several years.

The third ~~scenario~~ stress test type ~~(~~S_{RECOV}) is strictly speaking not a test that applies additional stress but a test of system recovery. It focuses on the recovery of the worst drought events in the historical record and aims to answer practitioners' question how long the drought will last if the following months are normal, dry or wet ~~)~~(Table 1). As groundwater dynamics are often more damped than climate anomalies, groundwater droughts usually last longer than meteorological droughts. To assess the maximum duration the groundwater system needs to recover from severe drought conditions, the lowest groundwater heads simulated between 1970 and 2016 are taken as the initial condition for each grid cell in this ~~scenario~~simulation experiment. Then, starting in October (in general, the beginning of the main recharge period in Germany), groundwater heads are simulated using ~~the~~ three assumed recharge tests~~long-term average monthly recharge~~ as input: average monthly recharge, continuously dry (25-percentile monthly recharge) and wet (75-percentile monthly recharge) recharge conditions, derived from the long-term historical recharge record (Figure 2). Drought termination is set to when the simulation exceeds the recovery threshold for the first time. As a recovery threshold we also ~~use~~ test three options: the monthly variable 25-percentile groundwater head (i.e. the groundwater head that is exceeded 75% of the time in that calendar month considering all simulated years), and the 40- and 50-percentile groundwater head. The time between ~~the~~ each ~~scenario stress test~~simulation start~~ing to apply the long-term average recharge~~ and the drought termination is the groundwater recovery time T_{rec}, i.e. the time needed to recover from ~~initial~~ the worst drought conditions. Like for the interpretation of the results from S_{EVENT} we relate T_{rec} to potential explanatory variables. ~~To test the impact the of stress- test assumptions on the results two additional model runs with dry (long-term 25-percentile monthly recharge) and wet (long-term 75-percentile monthly recharge) conditions during recovery and two other recovery thresholds (40- and 50-percentile groundwater head) were used.~~

The groundwater model used for these experiments was evaluated by Hellwig et al. (2020) using 202 groundwater borehole time series and 338 streamflow observations. Their results suggested that the model can reproduce the standardized time series as well as T_{max}, even though the model is still too coarse for the small-scale variability in ~~the~~ mountainous regions of Germany. However, for the different stress tests, specific model abilities will be required (Table 2). While for S_{SHIFT} the appropriate

simulation of T_{max} measuring the time needed to propagate anomalies from precipitation to groundwater is most relevant, for S_{EVENT} it is more the depiction of drought severity during the selected benchmark drought events. These two model abilities are also essential for S_{RECOV}. In general, overall patterns of the stress test- results can be expected to be reliable for both groundwater heads and baseflow with largest uncertainties of the actual groundwater levels and the magnitude of their fluctuations in the porous aquifers in North-East and the mountainous South.

4 Results

4.1 Groundwater drought under a seasonal recharge shift

The assumed S_{SHIFT} affects groundwater heads and baseflow throughout the year. As recharge increases (decreases) during winter (summer) recharge variability increases (decreases) correspondingly (Figure 4). Most recharge in Germany (outside the Alps) occurs during winter, therefore, the seasonal differences are amplified by S_{SHIFT} and inter-annual variability for recharge as well as groundwater tables and baseflow is increased. While in general, the changes in seasonal baseflow variability correspond to the changes in recharge variability, alterations of groundwater head variability are much more heterogeneous. Not only in winter but also during spring and autumn there is an increase in variability across large parts of Germany and even in summer variability increases in the Northeast.

Under S_{SHIFT} groundwater heads increase due to the higher winter recharge except in the alpine South, where groundwater recharge mostly occurs during summer (Figure 5). Changes of groundwater heads are smaller during drought than for median conditions, with negligible differences between the seasons. Absolute head changes are stronger in aquifers of large head variability (i.e. the fractured rock aquifers). On the contrary, relative head changes standardized by the mean and standard deviation of natural variability are most pronounced in the large porous aquifers in the North (Figure S21) where changes of variability are strongest as well (Figure 4). The general pattern of head changes is similar for all different assumed shift magnitudes (Figure S3).

Baseflow also increases under S_{SHIFT} in most parts of Germany (Figure 6). However, there are relevant differences between the seasons: during winter there is a large increase of baseflow, particularly under average conditions. In spring and autumn there are only small increases in the north of Germany (not shown). Baseflow changes during summer are bidirectional with increases in the North and decreases in the South, again more pronounced for average conditions than for drought. On an annual scale changes in baseflow are rather small following the same pattern of increases in the North and decreases in the South. Changes of baseflow relative to its variability are in general much smaller compared to changes of groundwater heads (Figure S42). As for groundwater patterns of baseflow changes are independent from the assumed shift magnitude with stronger responses for larger relative recharge shifts (Figure S3).

4.2 The groundwater drought sensitivity to antecedent recharge

All S_{EVENT} ~~scenarios~~ stress tests exacerbate the selected benchmark groundwater droughts ~~chosen for this stress test~~ (Figure 7). However, the magnitude of declines in groundwater head and baseflow vary for different drought events and durations. In comparison, the effect of the chosen return period ~~of the recharge scenario~~ is low. The differences between S_{EVENT} with T_{RP} = 50y and T_{RP} = 100y are about one order of magnitude smaller than the differences ~~between~~ among the different T_{RP} = 50y ~~scenarios~~ recharge reduction durations. ~~and~~ The median ~~D~~deviations ~~to~~ the reference simulation ~~(median~~ rang~~es~~ing~~~~ between 4 % and 21 % for the ~~different generic~~ different~~scenarios)~~ S_{EVENT}.

Differences between the drought events are similar for water table and baseflow changes (Figure 7). For the 1973 drought event declines are most pronounced for a reduced recharge over 3-months ~~scenario~~ whereas for the short-term summer droughts in 2003 and 2015 longer ~~scenarios~~ durations of recharge reductions cause~~d~~ more severe declines. However, the magnitude of ~~scenario~~stress test~~-~~ caused decreases is different for water tables and baseflow. Water table declines are largest for ~~scenarios~~ stress tests of the 2003 drought and smallest for the 1973 drought (Figure 7a) whereas relative baseflow decreases are similar for all events (Figure 7b). The differences between the ~~scenarios~~ stress tests as well as water tables and baseflow also show distinct spatial patterns (Figures S~~3~~5-S6~~4~~). For example, for the 3-months duration only specific regions in the Central German Uplands are affected with most pronounced head declines for the 1973 event.

The ~~scenario~~ effects of S_{EVENT} are related to different parameters (examples in Figures S~~5~~7-S8~~6~~), most significantly to the anomaly propagation time T_{max}. In general, longer T_{max} are related to stronger head decreases ~~in the scenarios~~ whereas baseflow reductions are larger for shorter T_{max} (Figure 8). However, the exact relationship between T_{max} and ~~scenario~~ stress test depends on the event year and ~~scenario length~~duration of the recharge reduction.

4.3 Recovery times of groundwater drought

Consistent with the results from S_{SHIFT} and S_{EVENT}, there is a large heterogeneity of T_{rec} across Germany (Figure 9~~a~~). For average recharge conditions and a 25-percentile recovery threshold T_{rec} is shorter than 10 months in large parts of Germany, particularly in the Central German Uplands with its fractured rock aquifers (Figure 9a). In these regions, a single average recharge season can be enough to terminate a severe groundwater drought. In the north-eastern part of Germany, which is characterized by large porous aquifers, groundwater heads will still not recover to the 25-percentile recovery threshold after up to 60 months of average recharge. In these regions, average recharge is not enough to terminate a severe groundwater drought. Accordingly, a bi-modal distribution of T_{rec} is found for regions with fast recovery and for regions with no recovery at all in the timeframe. For dry recharge conditions most of Germany will not recover within 60 months apart from some fast responding regions in the Central German Uplands (Figure 9b). On the contrary there are only few regions (most of them in the northeast of Germany) that do not recover within a year given continuously wet recharge conditions (Figure 9c). The larger recovery thresholds lead to increased T_{rec} but the general spatial pattern of regions with slower and faster recovery remains the same (not shown).

-T_{rec} increases with hydraulic conductivity and specific yield used in the model grid cell and is significantly higher in porous aquifers compared to aquifers in fractured rocks (Figure 9b). However, the strongest relationship is found between T_{rec} and propagation time T_{max}. The strong relationship between T_{max} and T_{rec} is found for all S_{RECOV} independent from the choice of recharge conditions and recovery threshold.

5 Discussion

5.1 Groundwater and baseflow sensitivity to altered recharge

All ~~scenarios~~ stress tests revealed a spatially highly heterogeneous groundwater response due to changes in recharge. In the northeast of Germany where large porous aquifers are prevalent, groundwater heads respond to long-term recharge characteristics. Accordingly, in this region changes on the 24-months duration (S_{EVENT}) or changes in the annual average recharge sum (S_{SHIFT}) cause the strongest responses. Contrasting, in the fractured aquifers of the Central German Uplands intra-annual recharge dynamics are much more relevant, demonstrated by the stronger responses to 3-months ~~scenarios~~ stress tests (S_{EVENT}). Also, the recovery time T_{rec} from a severe drought ~~varied~~ showed the same patterns with faster recovery in the uplands and slower recovery in the large porous aquifers~~accordingly~~ (S_{RECOV}). These results highlight the importance of the hydrogeological ~~conditions~~ characteristics for assessing the groundwaters' sensitivity to drought and for drought propagation, supporting the findings of Stoelzle et al. (2014).

Inter- and intra-annual changes in recharge do not only affect the immediate drought hazard in a different way for different hydrogeology but will also cause various changes to the long-term groundwater and baseflow dynamics. A change of recharge variability will not necessarily result in a change of hydrological drought conditions, where response times are long enough or where a change in variability is caused by changes in the mean or the wet climate and recharge extreme. Hence, assessments of potential changes regarding average conditions or variability may have minor or no information for proactive drought planning. Our results suggest that drought assessments directly relevant for specific stakeholders' needs and analysed in the context of the local sensitivity determined by hydrogeological conditions will better allow for adaptation and planning.

The hydrogeological conditions are also linked to the locally specific precipitation accumulation time that has the maximum correlation with water table variation T_{max}. Hellwig et al. (2020) analysed the T_{max} ranging from few months to several years across Germany. Their results suggested that T_{max} can be a good proxy for heterogeneous reactions of the groundwater to droughts. The patterns of T_{max} were similar to those found here for the groundwater's response to the more specific ~~scenarios~~stress tests, hence the propagation time from meteorological to groundwater anomalies also has the potential to be a predictor of the general groundwater drought sensitivity to recharge ~~scenarios~~stress tests.

The drought-specific stress test ~~scenarios~~modelling, however, ~~do~~ provides a more nuanced insight into the hazard. The results for both S_{SHIFT} and S_{EVENT} revealed systematic differences for groundwater heads and baseflow. The main reason here is the non-linear relationship between the two variables: the baseflow dynamics are mainly driven by groundwater fluctuations in the wet range, when groundwater heads are closer to the surface and more groundwater discharge is possible through the

dynamic drainage network (Godsey and Kirchner, 2014). For low groundwater heads, the drainage system shrinks and less baseflow results in a lower sensitivity to changes in groundwater heads. In the model this is represented by the variable number of grid cells in a catchment that contribute to baseflow with less cells in case of low groundwater heads. Changes in groundwater heads due to the event ~~scenarios~~ stress tests are most pronounced in regions with long propagation times T_{max} (taken from Hellwig et al., 2020) where the antecedent recharge has more influence. However, aquifers with long propagation times are usually characterized by large dynamic storages leading to a smaller baseflow variability (i.e. more stable flow regimes). Correspondingly, large changes of baseflow occur predominantly in regions with short T_{max} opposite to the regions of large groundwater head change.

For Germany, climate change is expected to increase the seasonality of the water cycle with higher water availability during winter and lower water availability during summer. The assumed changes of S_{SHIFT} lie in the range of potential precipitation changes for winter and evapotranspiration changes in summer predicted by regional climate and water balance models for Germany until the end of the 21st century (Jacob et al., 2012, Herrmann et al., 2016, Paparrizos et al., 2018). As the magnitude of change is uncertain, the general sensitivity of a system as investigated in this study can help to assess, whether and where the expected contrasting seasonal change has a general potential to influence baseflow and groundwater drought.

The different responses of baseflow and groundwater are important to consider for an effective water management and drought planning in a changing climate. Different stakeholders will face different challenges in future and use the stress tests differently to design adaptation or to plan mitigation measures for emergency plans. For example, in a climate with higher annual recharge sums but more frequent or severe summer droughts groundwater droughts might become less severe while the baseflow drought hazard becomes more severe. Where possible, one option might be to switch or add water use from surface water to groundwater to meet water demands for irrigation, industry, and public water supply. For other purposes relying on a minimal amount of surface water (e.g. navigation, water quality, or ecosystem health) adaptations such as regional water transfers or increased surface water storage capabilities might be more expedient. ~~with potential impacts on economy and ecology.~~

5.2 Uncertainties of large-scale groundwater simulations under climate stress

The model used in this study is limited in that it simulates groundwater head and baseflow dynamics under natural conditions only. The usual anthropogenic response to drought is an increased groundwater pumping, which causes a positive feedback which accelerated drying (Famiglietti, 2014). Therefore, anthropogenic influences also need to be considered as significant contributors to real changes in groundwater heads (Kløve et al., 2014). Moreover, there is uncertainty arising from the aquifer parametrization. ~~and~~ Eexact model derived T_{max} as well as groundwater and baseflow drought severity must be taken with care and should not be interpreted exactly to the location. In particular, Hellwig et al. (2020) found a decreasing model performance for higher elevation regions with small scale variability of the hydrogeology. Gleeson et al. (2020) conclude in their commentary that profound (observation-based) model evaluations for large-scale groundwater models are currently beyond reach. Groundwater head dynamics measured at boreholes can deviate considerably from grid cell averages due to a

large subgrid heterogeneity (e.g. Kumar et al., 2016). Opposingly, baseflow dynamics can be seen as an integrated spatial signal but uncertainties arising from the separation of baseflow from streamflow are large (e.g. Stoelzle et al., 2020a). Also, for other observational data there are severe constraints (Gleeson et al., 2020). Even though these uncertainties limit considerations for an effective local water management, they do not affect the general conclusions on regional groundwater sensitivity ~~reported above~~found.

~~There are~~Climate change projections ~~also~~ contain considerable uncertainties about future precipitation and predictions for recharge are even more uncertain as it might change even more strongly (Ng et al., 2010; Taylor et al., 2012; Jing et al., 2020). ~~Previous s~~Studies on recharge changes in Central Europe consistently predicted increases during winter and decreases during summer (Eckhardt and Ulbrich, 2003; Stoll et al., 2011; Dams et al., 2012; Hunkeler et al., 2014; Chen et al., 2018), however, recharge is variable with potentially large year-to-year variations (Kopp et al., 2018). The spatially different groundwater sensitivities identified in this study allow to assess the general potential of changes of groundwater and baseflow drought in a changing climate. Key findings of the stress tests using S_{SHIFT} (a general increase of head variability, increase of average water table) are also in line with recent findings of Jing et al. (2020) who use climate change projections to study impacts on the groundwater system in a small catchment in Central Germany.

~~intra-annual shift used for S_{SHIFT} is based on a relatively simple assumption that only represents a climate change informed consensus estimate of recharge changes but is supported by recent findings of Jing et al. (2020) reporting increases in recharge and groundwater heads under different climate change scenarios for a small catchment in Central Germany. Additionally, t~~There is evidence that different hydro~~other~~ meteorological characteristics that might change in future are relevant for groundwater and baseflow drought. Bloomfield et al. (2019) demonstrated an influence from changes in evapotranspiration due to increasing temperatures on changes in groundwater drought. Longobardi and Van Loon (2018) showed that changes in dry spell length can alter groundwater contributions to streamflow. Applying recharge frequency analysis to derive a 50-year or 100-year recharge drought event extrapolating beyond the range of the observational time period is a pragmatic hydrological design concept. As always, it comes with uncertainty and may be questioned due to climate-change induced non-stationarity. But as a sensitivity testing framework, it is found useful and suitable for communication to practitioners used to dealing for example with flood frequency terminology. The S_{EVENT} for the first-time provides country-scale composite estimates of groundwater and baseflow sensitivity to such assumed more severe recharge droughts and should also be considered for future water management plans.

5.3 Benefits of complementary stress testing for sensitivity assessments

The different ~~scenarios~~ stress tests are complementary to modelling chains from climate change scenarios to hydrogeology as they target the groundwater's sensitivity against different characteristics that are important to consider ~~under climate change~~for water management. S_{SHIFT} focusses on systematic intra-annual changes in the recharge regime and its consequences for droughts. S_{EVENT} assesses the specific response to prolonged dry spells whereas S_{RECOV} investigates the groundwater's ability to recover after a severe drought. With the combination of these different ~~scenarios~~ stress tests different aspects of the

groundwaters' sensitivity can be assessed and the following main points regarding the baseflow and groundwater drought sensitivity emerge:

1. Changes in the annual average recharge sum alter the groundwater heads in regions with slow groundwater response over the entire year, mitigating (or exacerbating if annual recharge is ~~reducing~~reduced) the groundwater drought hazard here for all seasons. In regions with fast groundwater responses, intra-annual recharge trends are more relevant than changes of the annual recharge sum.
2. An intra-annual shift of the recharge ~~like~~ as it was assumed in S_{SHIFT} has larger effects on baseflow and groundwater ~~heads~~ under average conditions than on ~~groundwater heads~~water availability during drought. The general increase in baseflow and groundwater ~~head~~ variability following ~~higher~~ stronger recharge ~~variability~~ seasonality does not necessarily result in a change of hydrological drought conditions~~is rather a result of wetter average conditions than drier drought conditions~~.
3. Baseflow and g~~G~~roundwater ~~heads~~ respond to recharge on characteristic time scales. Hence, reduced antecedent recharge over a longer duration which could be a result of a changed climate with prolonged dry spells can lead to much more severe ~~groundwater~~ droughts in aquifers and surface waters reacting on the corresponding time scales.
4. Groundwater recovery times for a severe drought are mainly related to the hydrogeology. This finding supports recent approaches for predictions on groundwater drought development several months ahead based on the site-specific characteristics of groundwater dynamics (e.g. Prudhomme et al., 2017; Parry et al., 2018).

6 Conclusions

Future changes of recharge are relevant for the groundwater drought hazard and groundwater~~'~~'s potential to mitigate drought impacts. In this study a stress ~~-~~test approach was employed ~~as an alternative to climate change model chains~~to test the groundwater's system sensitivity to changes in recharge: three generic recharge ~~scenarios~~ stress tests were used in a country-scale German groundwater model simulating groundwater heads and baseflow. Different ~~Despite uncertainties of future recharge~~from climate change scenarios, the ~~scenarios~~ stress tests systematically apply different types of recharge change (e.g. proportional shifts or extreme events of a given return period) allowing for general conclusions on the diversity of groundwaters' sensitivity ~~to projected directions of climate change~~. While the assumed intra-annual recharge shifts can be expected to weaken the groundwater drought hazard, prolonged dry spells may aggravate droughts, particularly in regions with slow responding aquifers. Baseflow is not linearly related to changes of groundwater heads and is more prone to intensified drought event conditions on a shorter time scale, especially in regions with fast responding aquifers. The groundwaters' drought recovery time is strongly related to the aquifers' characteristic response time scale. Hence, ~~it is not~~spatial patterns of recovery times are only secondarily depending on the meteorological drought characteristics but rather an inherent property of the aquifer with large regional differences.

The ~~scenario~~stress test~~-~~ approach applied in this study allows for a detailed composite assessment of a controlled environmental change. ~~-~~Regional sensitivities to changes in recharge differ considerably. Hence, key regions most vulnerable to recharge changes can be identified and may enable proactive adaptations for different stakeholders independent of specific climate projections ~~are possible. Simple recharge scenarios (e.g. below-average, average, above-average recharge) in a country-scale groundwater model~~Different regional sensitivities could also be used for probabilistic real-time groundwater drought forecasting as an informative tool for water supply and other stakeholders. ~~The application of~~While recently developed country-to-global scale transient and gradient-based groundwater models ~~could allow for forecasts of groundwater heads with long lead time~~can guide decision-making on these scales. ~~. F~~For local management decisions it will be important to consider local hydrogeological conditions and include also anthropogenic feedbacks such as increased pumping during drought (e.g. due to higher irrigation demand). Such feedback could be also implemented as generic stress tests. Therefore, future work evaluating the groundwater response to scenarios of human water use during drought will be needed to complement the findings of this study.

Data availability

~~All~~ The model outputs from the reference run and ~~scenario~~ stress test runs can be downloaded from FreiDok (https://doi.org/10.6094/UNIFR/167379~~https://freidok.uni-freiburg.de/~~, will be activated ~~complete link~~ upon acceptance~~in the final paper~~).

Author contribution

JH developed the main ideas, the design of the ~~scenarios~~ stress ~~-~~tests was jointly developed by all authors. JH performed the analyses and prepared the manuscript which was reviewed by the co-authors.

Acknowledgements

We acknowledge the comments by Editor Jim Freer and two anonymous reviewers ~~for their~~who provided valuable reviews which significantly helped to improve the quality of the paper. JH was funded by the DFG project TrenDHy STA632/4-1 and KS by the DFG Heisenberg programme STA632/3-1, MS by the LUBW grant "Low Flow Stress Tests".

References

BGR and SGD: Bundesanstalt für Geowissenschaften und Rohstoffe and Staatliche Geologische Dienste: Hydrogeologische Übersichtskarte von Deutschland 1:200.000, Oberer Grundwasserleiter (HÜK200 OGWL), Digitaler Datenbestand, Version 3.0. – Hannover, 2016.

Bloomfield, J. P., Marchant, B. P., Bricker, S. H., and Morgan, R. B.: Regional analysis of groundwater droughts using hydrograph classification, Hydrol. Earth Syst, Sci., 19(10), 4327-4344, doi: 10.5194/hess-19-4327-2015, 2015.

Bloomfield, J. P., Marchant, B. P., and McKenzie, A. A. (2019). Changes in groundwater drought associated with anthropogenic warming, Hydrol. Earth Syst, Sci., 23(3), 1393-1408, doi: 10.5194/hess-23-1393-2019, 2019.

BMU: Federal Ministry for the Environment, Nature Conservation and Nuclear Safety: Hydrologischer Atlas von Deutschland. Lieferung 1-3 mit 51 Kartentafeln. Bonn/Berlin, 2003

Chen, Z., Hartmann, A., Wagener, T., and Goldscheider, N.: Dynamics of water fluxes and storages in an Alpine karst catchment under current and potential future climate conditions, Hydrol. Earth Syst, Sci., 22(7), 3807-3823, doi: 10.5194/hess-22-3807-2018, 2018.

Dams, J., Salvadore, E., Van Daele, T., Ntegeka, V., Willems, P., and Batelaan, O.: Spatio-temporal impact of climate change on the groundwater system, Hydrol. Earth Syst, Sci., 16(5), 1517-1531, doi: 10.5194/hess-16-1517-2012, 2012.

de Graaf, I. E. M., Sutanudjaja, E. H., van Beek, L. P. H., and Bierkens, M. F. P.: A high-resolution global-scale groundwater model, Hydrol. Earth Syst, Sci., 19(2), 823-837, doi: 10.5194/hess-19-823-2015, 2015.

EC: Communication from the Commission to the European Parliament and the Council addressing the challenge of water scarcity and droughts in the European Union, Commission of the European Communities, COM(2007), 414 final, Brussels, 2007.

Eckhardt, K., and Ulbrich, U.: Potential impacts of climate change on groundwater recharge and streamflow in a central European low mountain range, J. Hydrol., 284(1-4), 244-252, doi: 10.1016/j.jhydrol.2003.08.005, 2003.

Eltahir, E. A. B., and Yeh, P.: On the asymmetric response of aquifer water level to floods and droughts in Illinois, Water Resour. Res., 35(4), 1199-1217, doi: 10.1029/1998wr900071, 1999.

Famiglietti, J. S.: The global groundwater crisis, Nat Clim Change, 4(11), 945, doi: 10.1038/nclimate2425, 2014.

Fan, Y., Miguez-Macho, G., Weaver, C. P., Walko, R., and Robock, A.: Incorporating water table dynamics in climate ~~modeling~~29odelling: 1. Water table observations and equilibrium water table simulations, J. Geophys. Res.-Atmos., 112(D10), 17. doi: 10.1029/2006jd008111, 2007.

Gleeson, T., Moosdorf, N., Hartmann, J., and Beek, L. P. H.: A glimpse beneath earth-'s surface: Global Hydrogeology MaPS (GLHYMPS) of permeability and porosity, Geophys Res Lett, 41(11), 3891-3898, doi: 10.1002/2014GL059856, 2014.

Gleeson, T., Befus, K. M., Jasechko, S., Luijendijk, E., and Cardenas, M. B.: The global volume and distribution of modern groundwater, Nat Geosci, 9(2), 161, doi: 10.1038/ngeo2590, 2016.

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., ... and Oshinlaja, N.: HESS Opinions: Improving the evaluation of groundwater representation in continental to global scale models. Hydrol. Earth Syst, Sci. Discuss., 1-39, https://doi.org/10.5194/hess-2020-378, 2020

Godsey, S. E., and Kirchner, J. W.: Dynamic, discontinuous stream networks: hydrologically driven variations in active drainage density, flowing channels and stream order, Hydrol. Process., 28, 5791– 5803, doi: 10.1002/hyp.10310, 2014.

Haas, J. C., and Birk, S.: Characterizing the spatiotemporal variability of groundwater levels of alluvial aquifers in different settings using drought indices, Hydrol. Earth Syst, Sci., 21(5), 2421-2448, doi: 10.5194/hess-21-2421-2017, 2017.

Hall, JW, Leng, G. Can we calculate drought risk… and do we need to? WIREs Water., 6:e1349. https://doi.org/10.1002/wat2.1349, 2019.

Harbaugh, A. W., Banta, E. R., Hill, M. C., and McDonald, M. G.: MODFLOW-2000, The U. S. Geological Survey Modular Ground-Water Model-User Guide to Modularization Concepts and the Ground-Water Flow Process. Open-file Report. U. S. Geological Survey(92), 134, 2000.

Hargreaves, G. H., and Samani, Z. A.: Reference crop evapotranspiration from temperature, Appl. eng. agric., 1(2), 96-99, 1985.

Haylock, M. R., Hofstra, N., Tank, A., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006, J Geophys. Res.-Atmos., 113(D20), 12, doi: 10.1029/2008jd010201, 2008.

Hellwig, J.: Grundwasserdürren in Deutschland von 1970 bis 2018, Korrespondenz Wasserwirtschaft, 12(10), 567-572, doi: 10.3243/kwe2019.10.001, 2020.

Hellwig, J., and Stahl, K.: An assessment of trends and potential future changes in groundwater-baseflow drought based on catchment response times, Hydrol. Earth Syst, Sci., 22(12), 6209-6224, doi: 10.5194/hess-22-6209-2018, 2018.

Hellwig, J., de Graaf, I. E. M., Weiler, M., and Stahl, K.: Large scale assessment of delayed groundwater responses to drought, Water Resour Res., 56(2), e2019WR025441, doi: 10.1029/2019WR025441, 2020.

Herrmann, F., Kunkel, R., Ostermann, U. Vereecken, H., Wendland, F..: Projected impact of climate change on irrigation needs and groundwater resources in the metropolitan area of Hamburg (Germany). Environ Earth Sci 75, 1104, https://doi.org/10.1007/s12665-016-5904-y, 2016.

Heudorfer, B., and Stahl, K.: Comparison of different threshold level methods for drought propagation analysis in Germany, Hydrol Res, 48(5), 1311-1326, doi: 10.2166/nh.2016.258, 2017.

Hunkeler, D., Möck, C., Käser, D., and Brunner, P.: Klimaeinflüsse auf Grundwassermengen, Aqua & Gas, 11, 43-49, 2014.

Jacob, D., Bülow, K., Kotova, L., Moseley, C., Petersen, J., and Rechid, D.: Regionale Klimaprojektionen für Europa und Deutschland: Ensemble Simulationen für die Klimafolgenforschung. MPI für Meteorologie, Climate Service Center, 2012.

Jing, M., Kumar, R., Heße, F., Thober, S., Rakovec, O., Samaniego, L., and Attinger, S.: Assessing the response of groundwater quantity and travel time distribution to 1.5, 2 and 3°C global warming in a mesoscale central German basin. Hydrol. Earth Syst, Sci., 24, 1511-1526. doi: 10.5194/hess-24-1511-2020, 2020.

Keller, L., Rössler, O., Martius, O., and Weingartner, R.: Comparison of scenario-neutral approaches for estimation of climate change impacts on flood characteristics, Hydrol. Process., 33(4), 535-550, doi: 10.1002/hyp.13341, 2019.

Kløve, B., Ala-Aho, P., Bertrand, G., Gurdak, J. J., Kupfersberger, H., Kværner, J., . . . Pulido-Velazquez, M.: Climate change impacts on groundwater and dependent ecosystems, J. Hydrol,, 518, 250-266, doi: 10.1016/j.jhydrol.2013.06.037, 2014.

Kohn I., Stahl K., and Stoelzle M.: Low Flow Events – a Review in the Context of Climate Change in Switzerland. Comissioned by the Federal Office for the Environment (FOEN), Bern, Switzerland, 75 pp, doi: 10.6094/UNIFR/150448, 2019.

Kopp, B., Baumeister, C., Gudera, T., Hergesell, M., Kampf, J., Morhard, A., . . . J.: Entwicklung von Bodenwasserhaushalt und Grundwasserneubildung in Baden-Württemberg, Bayern, Rheinland-Pfalz und Hessen von 1951 bis 2015, Hydrol. Wasserbewirts., 62(2), 62-76, doi: 10.5675/HyWa_2018,2_1, 2018.

Kumar, R., Musuuza, J. L., Van Loon, A. F., Teuling, A. J., Barthel, R., Ten Broek, J., . . . Attinger, S.: Multiscale evaluation of the Standardized Precipitation Index as a groundwater drought indicator, Hydrol. Earth Syst, Sci., 20(3), 1117-1131, doi: 10.5194/hess-20-1117-2016, 2016.

Kundzewicz, Z. W., and Döll, P.: Will groundwater ease freshwater stress under climate change? Hydrolog. Sci. J., 54(4), 665-675, doi: 10.1623/hysj.54.4.665, 2009.

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., Knutti, R., and Hawkins, E.: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, Earth Syst. Dynam., 11, 491-508, https://doi.org/10.5194/esd-11-491-2020, 2020.

Longobardi, A, and Van Loon, A. F.: Assessing baseflow index vulnerability to variation in dry spell length for a range of catchment and climate properties, Hydrol. Process., 32(16), 2496–2509, doi: 10.1002/hyp.13147, 2018.

Maxwell, R. M., Condon, L. E., and Kollet, S. J.: A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3, Geosci. Model Dev., 8(3), 923-937, doi: 10.5194/gmd-8-923-2015, 2015.

Miguez-Macho, G., Li, H., and Fan, Y.: Simulated water table and soil moisture climatology over North America, B. Am. Meteorol. Soc., 89(5), 663-672, doi: 10.1175/BAMS-89-5-663, 2008.

Moeck, C., Brunner, P., and Hunkeler, D.: The influence of model structure on groundwater recharge rates in climate-change impact studies, Hydrogeol. J., 24(5), 1171-1184. doi: 10.1007/s10040-016-1367-1, 2016.

Ng, G.-H. C., McLaughlin, D., Entekhabi, D., and Scanlon, B. R.: Probabilistic analysis of the effects of climate change on groundwater recharge, Water Resour. Res., 46(7), doi: 10.1029/2009WR007904, 2010.

Paparrizos, S., Schindler, D., Potouridis, S., ~~&~~ and Matzarakis, A.: Spatio-temporal analysis of present and future precipitation responses over South Germany. Journal of Water and Climate Change, 9(3), 490-499, doi: 10.2166/wcc.2017.009
, 2018.

Parry, S., Wilby, R., Prudhomme, C., Wood, P., and McKenzie, A.: Demonstrating the utility of a drought termination framework: prospects for groundwater level recovery in England and Wales in 2018 or beyond, Environ. Res. Lett., 13(6), 064040, doi: 10.1088/1748-9326/aac78c, 2018.

Peters, E., Torfs, P., van Lanen, H. A. J., and Bier, G.: Propagation of drought through groundwater ~~–~~ a new approach using linear reservoir theory, Hydrol. Process., 17(15), 3023-3040, doi: 10.1002/hyp.1274, 2003.

Prudhomme, C., Wilby, R. L., Crooks, S., Kay, A. L., and Reynard, N. S.: Scenario-neutral approach to climate change impact studies: application to flood risk, J. Hydrol., 390(3-4), 198-209, doi: 10.1016/j.jhydrol.2010.06.043, 2010.

Prudhomme, C., Hannaford, J., Harrigan, S., Boorman, D., Knight, J., Bell, V., . . . Jenkins, A.: Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales, Hydrolog. Sci. J., 62(16), 2753-2768,. doi: 10.1080/02626667.2017.1395032, 2017.

Reinecke, R., Foglia, L., Mehl, S., Trautmann, T., Cáceres, D., and Döll, P.: Challenges in developing a global gradient-based groundwater model (G3M v1.0) for the integration into a global hydrological model, Geosci. Model Dev., 12(6), 2401-2418, doi: 10.5194/gmd-12-2401-2019, 2019.

Staudinger, M., Weiler, M., and Seibert, J.: Quantifying sensitivity to droughts – an experimental modeling approach, Hydrol. Earth Syst, Sci., 19(3), 1371-1384, doi: 10.5194/hess-19-1371-2015, 2015.

Stoelzle, M., Stahl, K., Morhard, A., and Weiler, M.: Streamflow sensitivity to drought scenarios in catchments with different geology, Geophys. Res. Lett., 41(17), 6174-6183, doi: 10.1002/2014gl061344, 2014.

Stoelzle, M., Blauhut, V., Kohn, I., Krumm, J., Weiler, M. and~~,~~ Stahl, K.: Niedrigwasser in Süddeutschland. Analysen, Szenarien

und Handlungsempfehlungen, KLIWA Heft 23, Arbeitskreis KLIWA, www.kliwa.de, 2018 (in German).

Stoelzle, M., Schuetz, T., Weiler, M., Stahl, K., and Tallaksen, L. M.: Beyond binary baseflow separation: a delayed-flow index for multiple streamflow contributions. Hydrol. Earth Syst, Sci., 24(2), 849-867, doi: 10.5194/hess-24-849-2020, 2020a.

Stoelzle, M., Staudinger, M., Stahl, K., and Weiler, M.: Stress testing as complement to climate scenarios: recharge scenarios to quantify streamflow drought sensitivity, Proc. IAHS, 383, 43-50, 2020b

Stoll, S., Hendricks Franssen, H. J., Butts, M., and Kinzelbach, W. K.: Analysis of the impact of climate change on groundwater related hydrological fluxes: a multi-model approach including different downscaling methods, Hydrol. Earth Syst, Sci., 15(1), 21-38, doi: 10.5194/hess-15-21-2011, 2011.

Taylor, R. G., Scanlon, B., Döll, P., Rodell, M., van Beek, R., Wada, Y., . . . Treidel, H.: Ground water and climate change, Nat. Clim. Change, 3, 322, doi: 10.1038/nclimate1744, 2013.

Van Loon, A. F.: Hydrological drought explained, WIRES Water, 2(4), 359-392, doi: 10.1002/wat2.1085, 2015.

Van Loon, A. F., and Van Lanen, H. A.: A process-based typology of hydrological drought, Hydrol. Earth Syst, Sci., 16(7), 1915-1946, doi: 10.5194/hess-16-1915-2012, 2012.

~~Vormoor, K., Rössler, O., Bürger, G. et al. When timing matters considering changing temporal structures in runoff response surfaces. Climatic Change 142, 213–226, doi: 10.1007/s10584-017-1940-1, 2017.~~

Wada, Y., Wisser, D., and Bierkens, M. F. P.: Global ~~modeling~~32odelling of withdrawal, allocation and consumptive use of surface water and groundwater resources, Earth Syst. Dynam., 5(1), 15-40, doi: 10.5194/esd-5-15-2014, 2014.

Weider, K., and Boutt, D. F.: Heterogeneous water table response to climate revealed by 60 years of ground water data, Geophys. Res. Lett., 37(24), doi: 10.1029/2010GL045561, 2010.

Table 1: Overview of the three generic ~~scenarios~~ stress tests used in this study and the related question to be answered by the ~~scenario~~stress test.

	Question to be answered	Time frame	Boundary conditions	Recharge modifications
S_{SHIFT}	How will a changed recharge regime with wetter winters and drier summers change the inter-annual variability and water availability during droughts?	Corresponding to reference simulation (57 years)	Apart from recharge same as for the reference simulation	Winter decrease, summer increase of different strength (-± 5, 10, 15, 20, 30 % relative to reference simulation~~)~~, ~~summer increase (+15 %)~~
S_{EVENT}	Could ~~it~~ the effect have been worse? How sensitive are hydrological droughts to antecedent recharge conditions on different durations?	Historical events	Taken from the historical event in the reference simulation	Recharge from reference simulation rescaled to match a drought event with a return period of 50 (100) years for three different durations
S_{RECOV}	What is the recovery time needed to terminate a severe drought event?	Hypothetical event	Most severe drought modelled in the reference simulation for every grid cell taken as initial conditions	Long-term monthly average/ 25-percentile/ 75-percentile recharge from the reference simulation

Table 2: Required model ability and discussion of model performance for the different stress tests.

	Required model ability	Evaluation metric	~~Discussion of m~~Model performance assessment
S_{SHIFT}	Reliable propagation of inter- and intra-annual recharge dynamics into groundwater heads and baseflow	T_{max}	Overall, the model depicts both, differences of T_{max} across the study area and the systematically shorter T_{max} of baseflow compared to groundwater. However, for baseflow T_{max} was notably overestimated in the North and underestimated in the South while for groundwater it was overestimated in the porous aquifers of the lowlands and underestimated in higher elevations (see Hellwig et al., 2020 for more detailed analyses). Hence, absolute S_{SHIFT} responses may be biased in that same way. The model estimates allow for ~~most~~highest confidence in the representation of general shift-patterns across the study area.
S_{EVENT}	Reliable model representation of benchmark drought events	Differences between observed and modelled groundwater /baseflow drought severities	Simulations and observations show a considerable variability of groundwater drought severity for different drought years across the study area. Consistent with observations, modelled drought severities were weaker in 2003 compared to 1973 with several regions in the study area not in groundwater drought. These patterns are also consistent with state agency reports (see Hellwig et al., 2020). However, especially in the Northeast the model responds too slowly (corresponding with too long T_{max}, see above) leading to deviating groundwater drought severities: the drought severity of 1973 is overestimated in the model while it is underestimated for 2003. For baseflow model performance is similar: while general patterns of drought severity can be depicted, drought severities deviate most in the North (-East) (see also Figure S1). Overall, there are systematic uncertainties arising from the comparison of observational data with model outputs which might relate to some of the differences found (for a more advanced discussion on that see Hellwig et al., 2020, Section 2.3).
S_{RECOV}	Reliable representation of severe drought + propagation of recharge forcing into groundwater	Combination of evaluation metrics of S_{SHIFT} and S_{EVENT}	As both general patterns of drought severities and the propagation of the forcing into groundwater are captured by the model, prerequisites for an appropriate drought termination simulation are given. Uncertainties for this ~~scenariostress~~ test are – similar to the other ~~scenarios~~stress tests – largest in regions of weaker model performance regarding T_{max}.

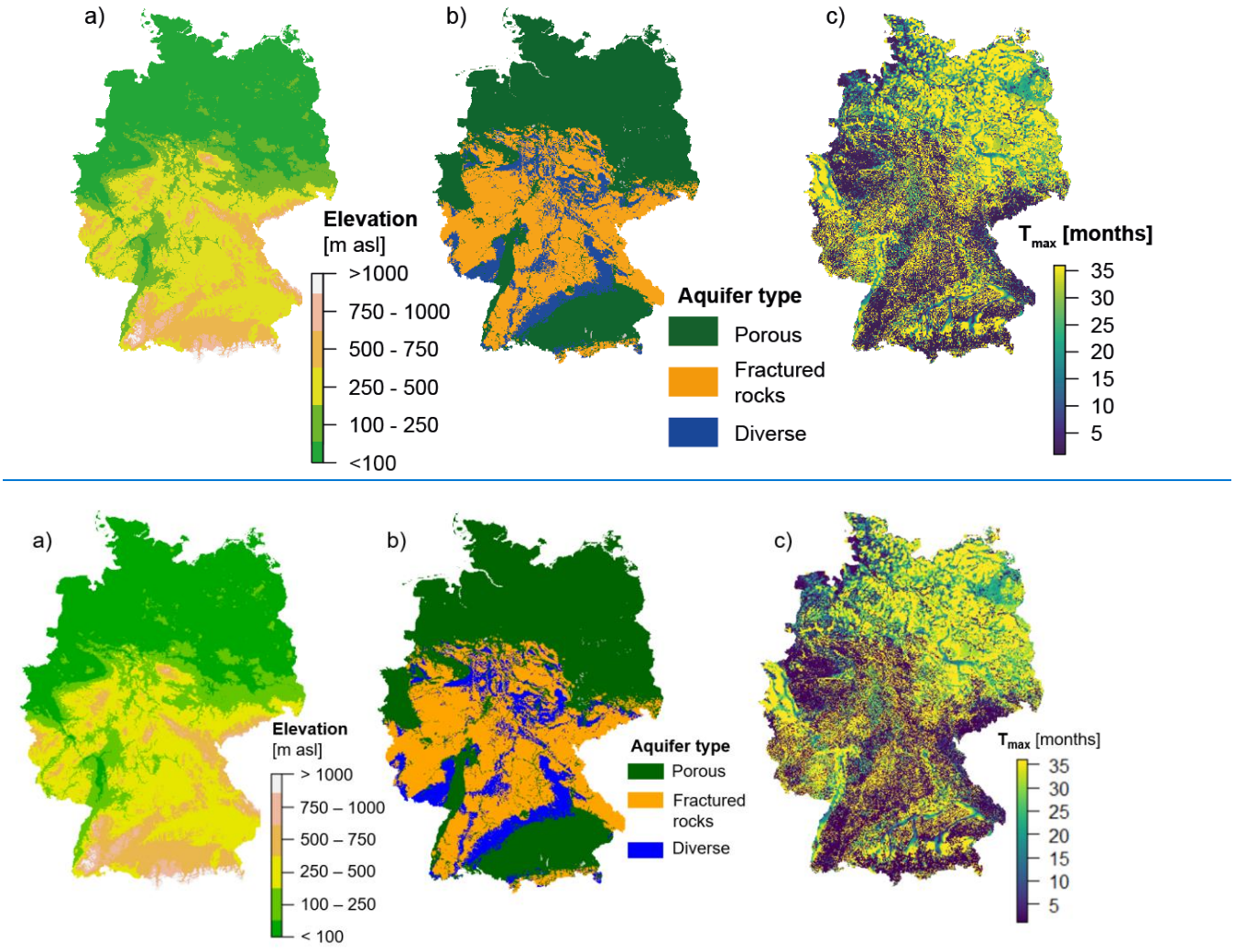


Fig. 1: Study area. a) Topographic map, b) main aquifer types (taken from BGR and SGD, 2016) and c) precipitation accumulation times that have the maximum correlation with groundwater T_{max} (taken from Hellwig et al., 2020).

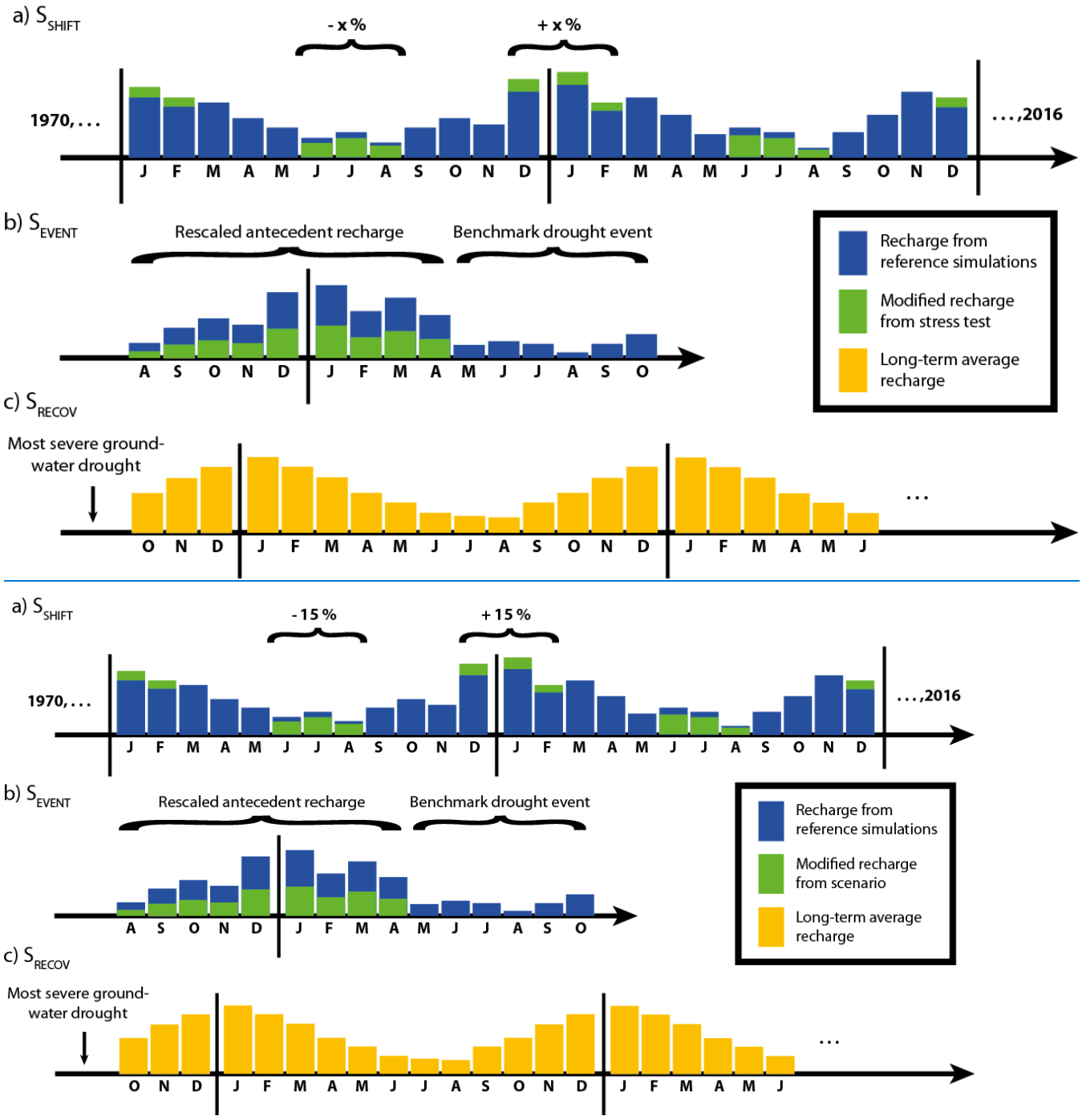


Fig. 2: Recharge modifications used for the three different ~~scenario~~ stress test types in this study.

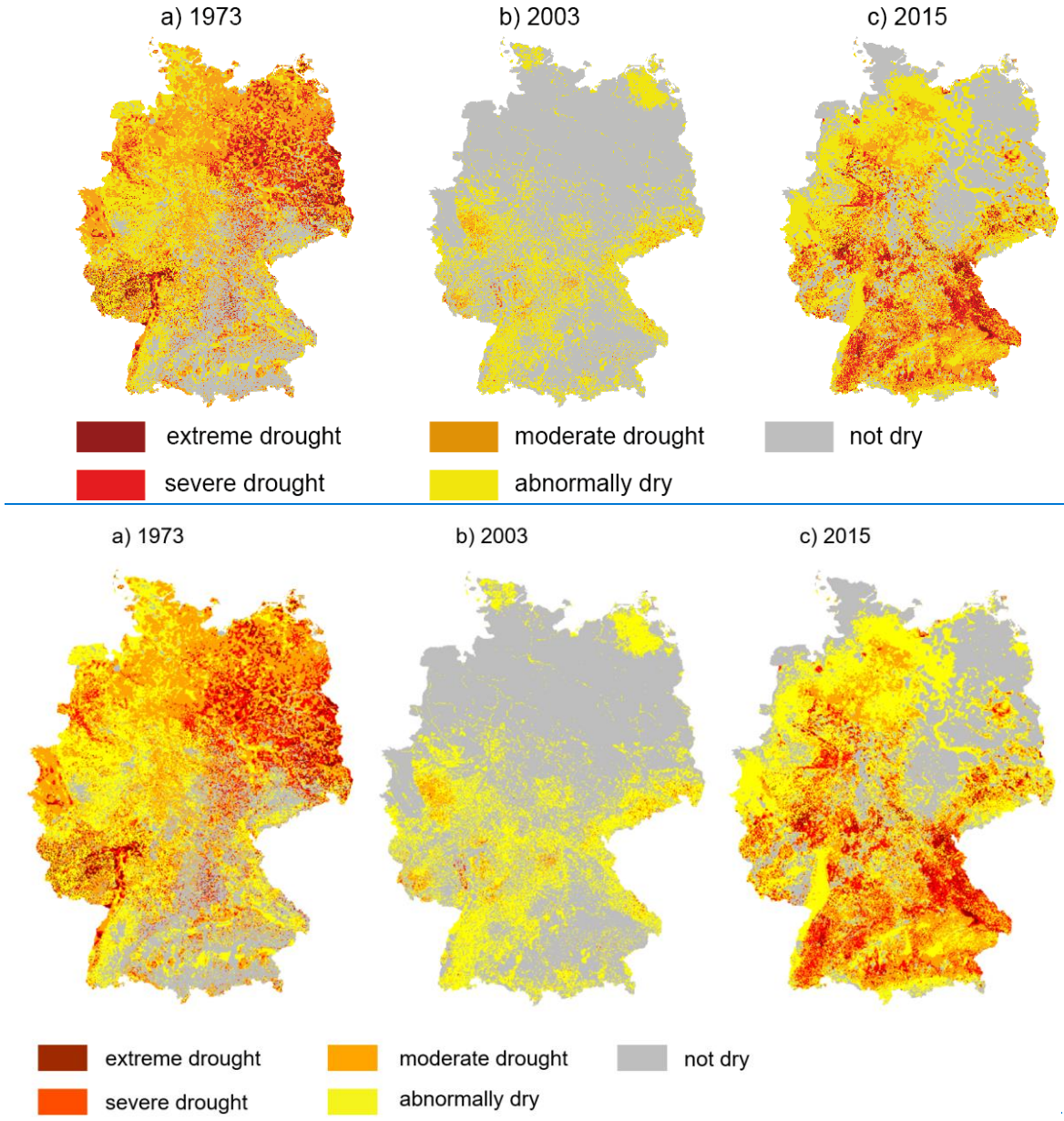
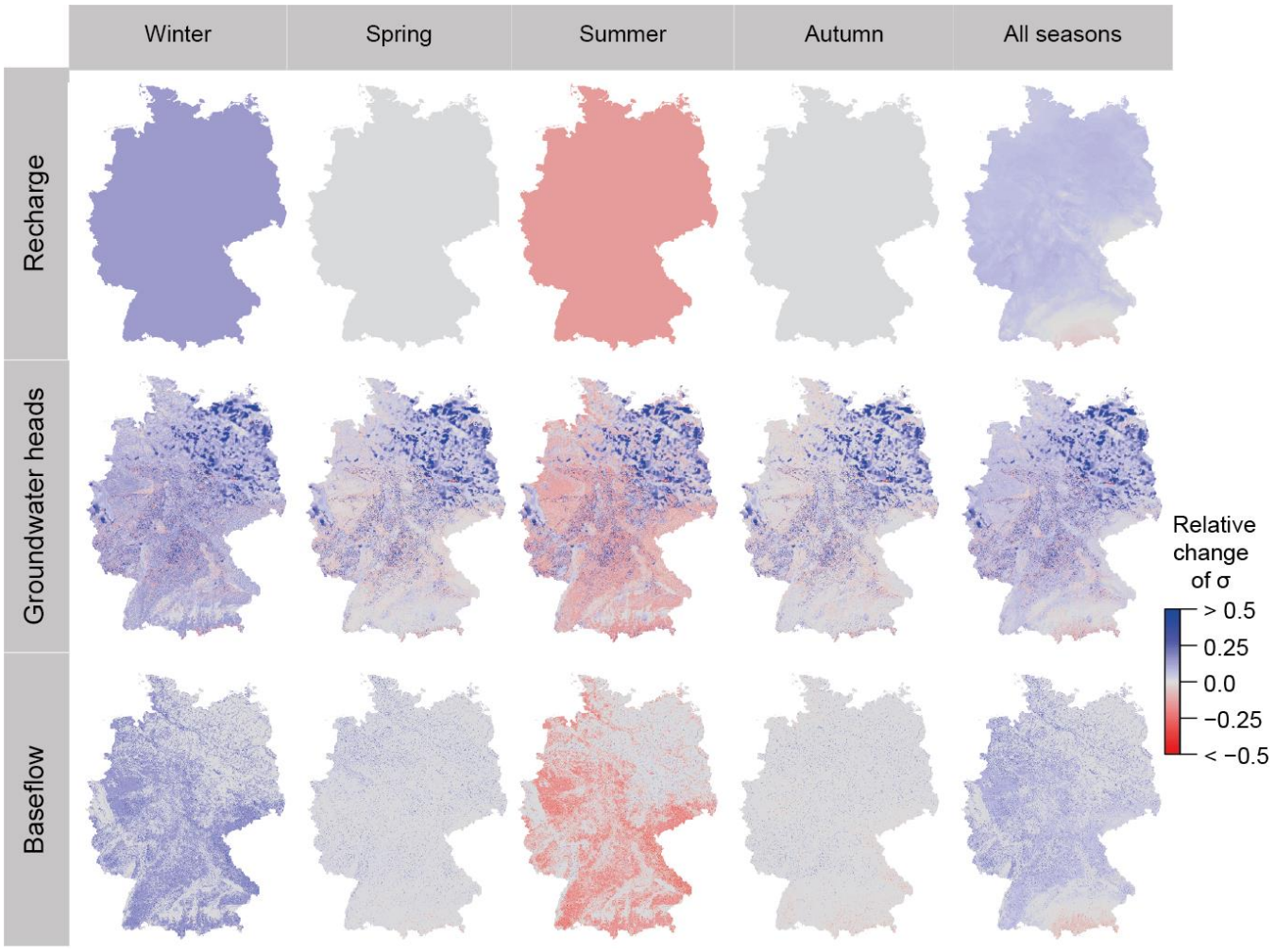


Fig. 3: Modelled groundwater drought situation during summer months (JJA) for benchmark drought events. Drought classes are derived from average standardized water table referring to the thresholds -2 (2.3 % of time: extreme drought), -1.5 (6.7 % of time: severe drought), -1 (15.9 % of time: moderate drought) and 0 (50 % of time: abnormally dry).


Winter
Spring
Summer
Autumn
All seasons
Recharge
Groundwater heads
Baseflow
Relative change of σ
> 0.5
0.25
0.0
−0.25
< −0.5

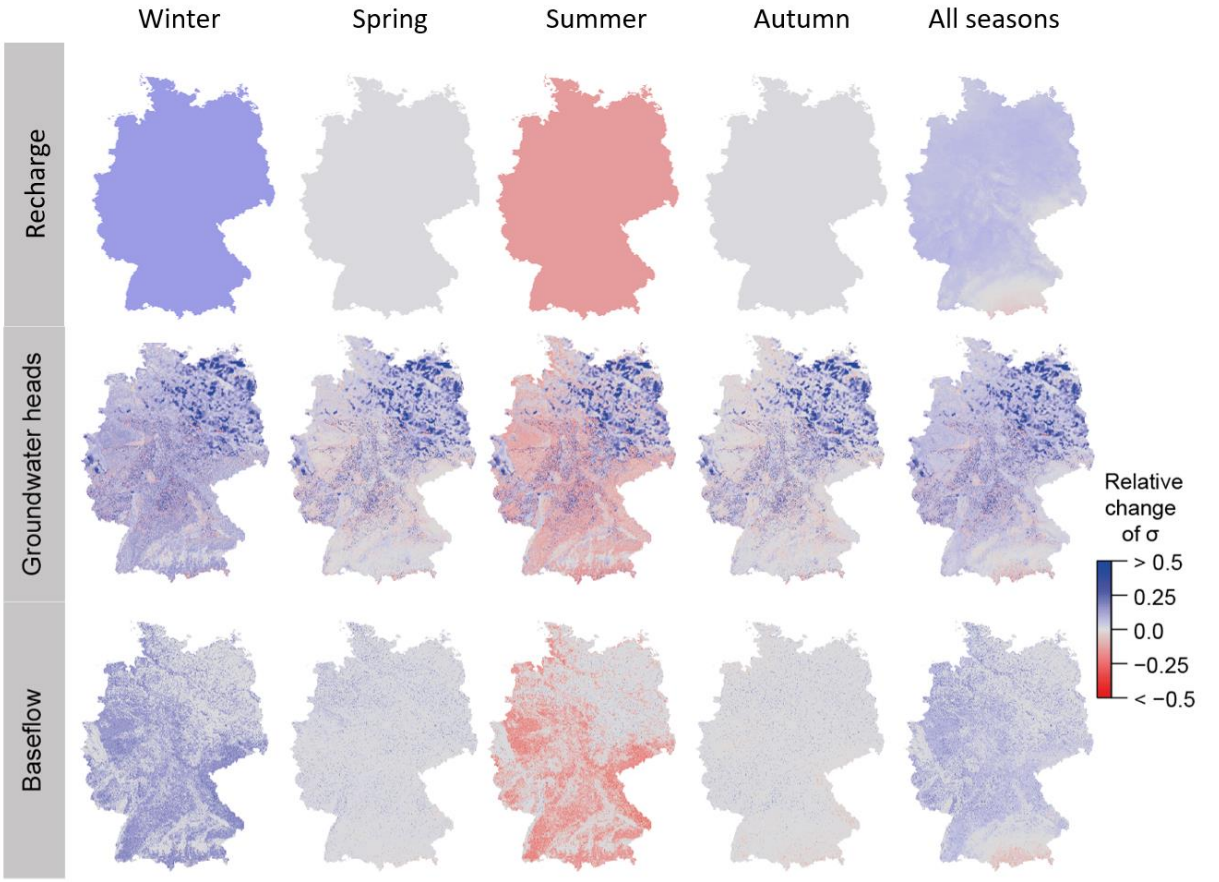
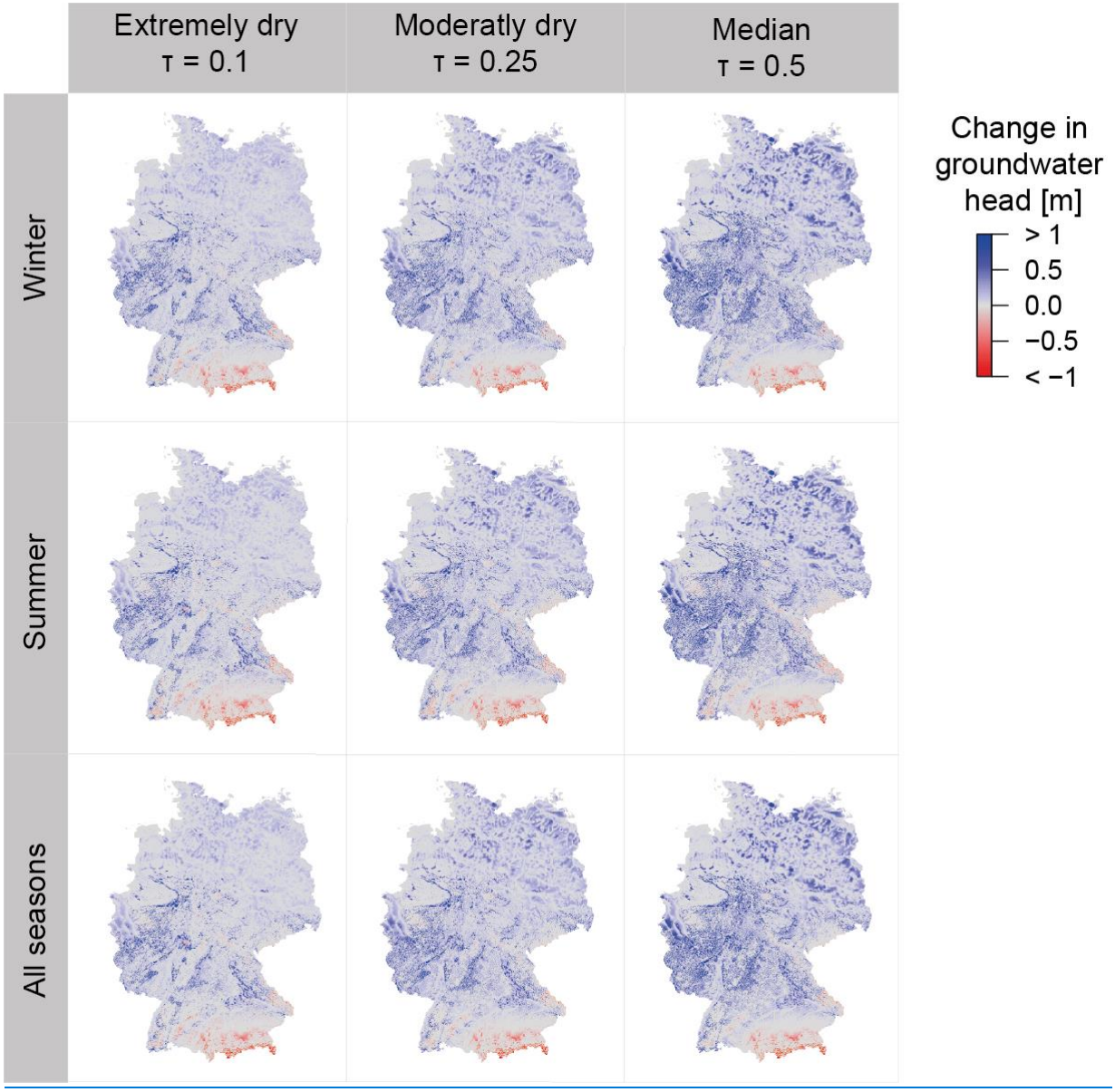


Fig 4: Relative changes in the inter-annual variability of recharge, groundwater head and baseflow for different seasons (with winter: DJF, spring: MAM, summer: JJA, autumn: SON) and a seasonal shift of 15%.


Extremely dry
τ = 0.1
Moderatly dry
τ = 0.25
Median
τ = 0.5
Winter
Summer
All seasons
Change in groundwater head [m]
> 1
0.5
0.0
−0.5
< −1

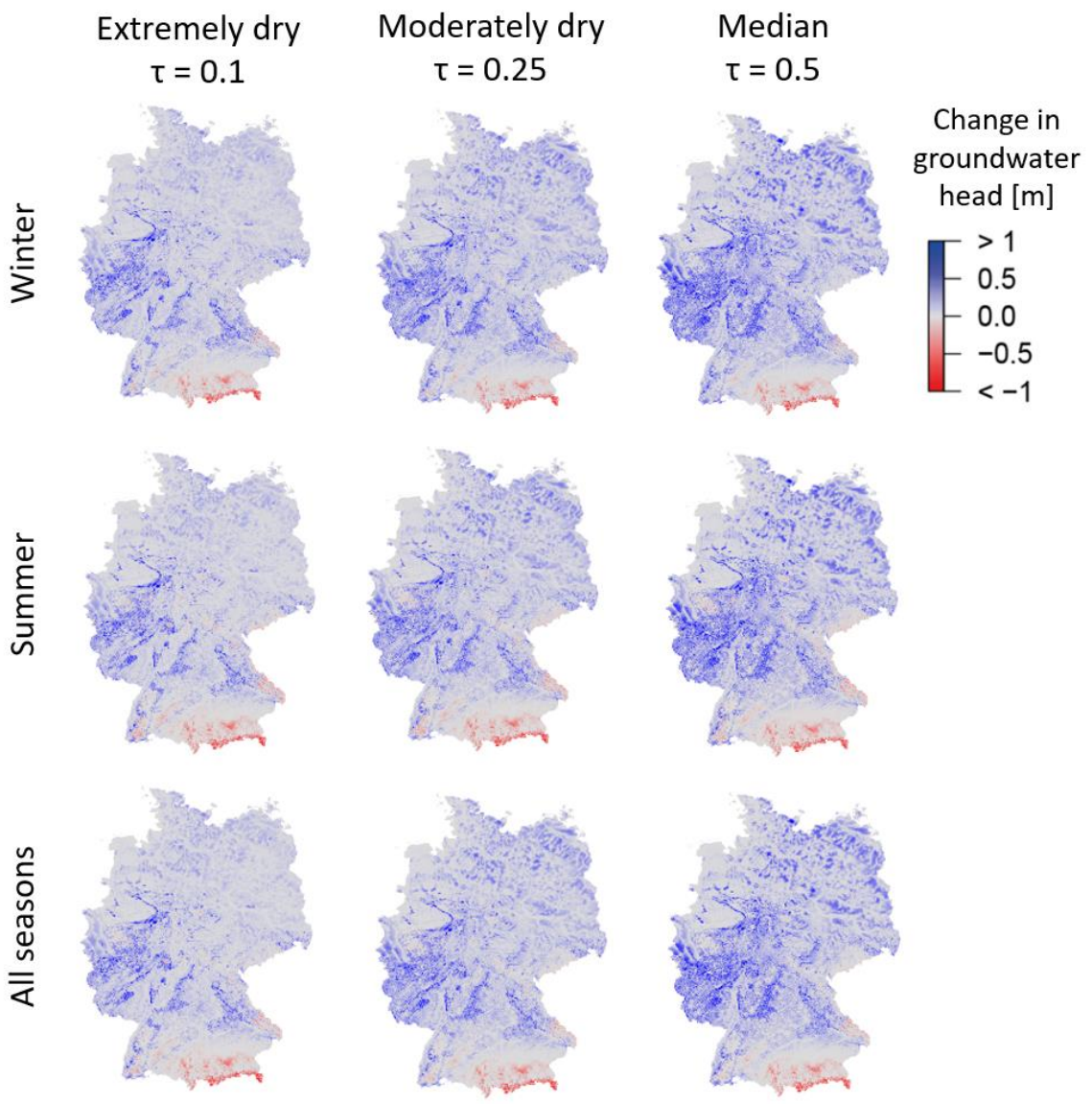
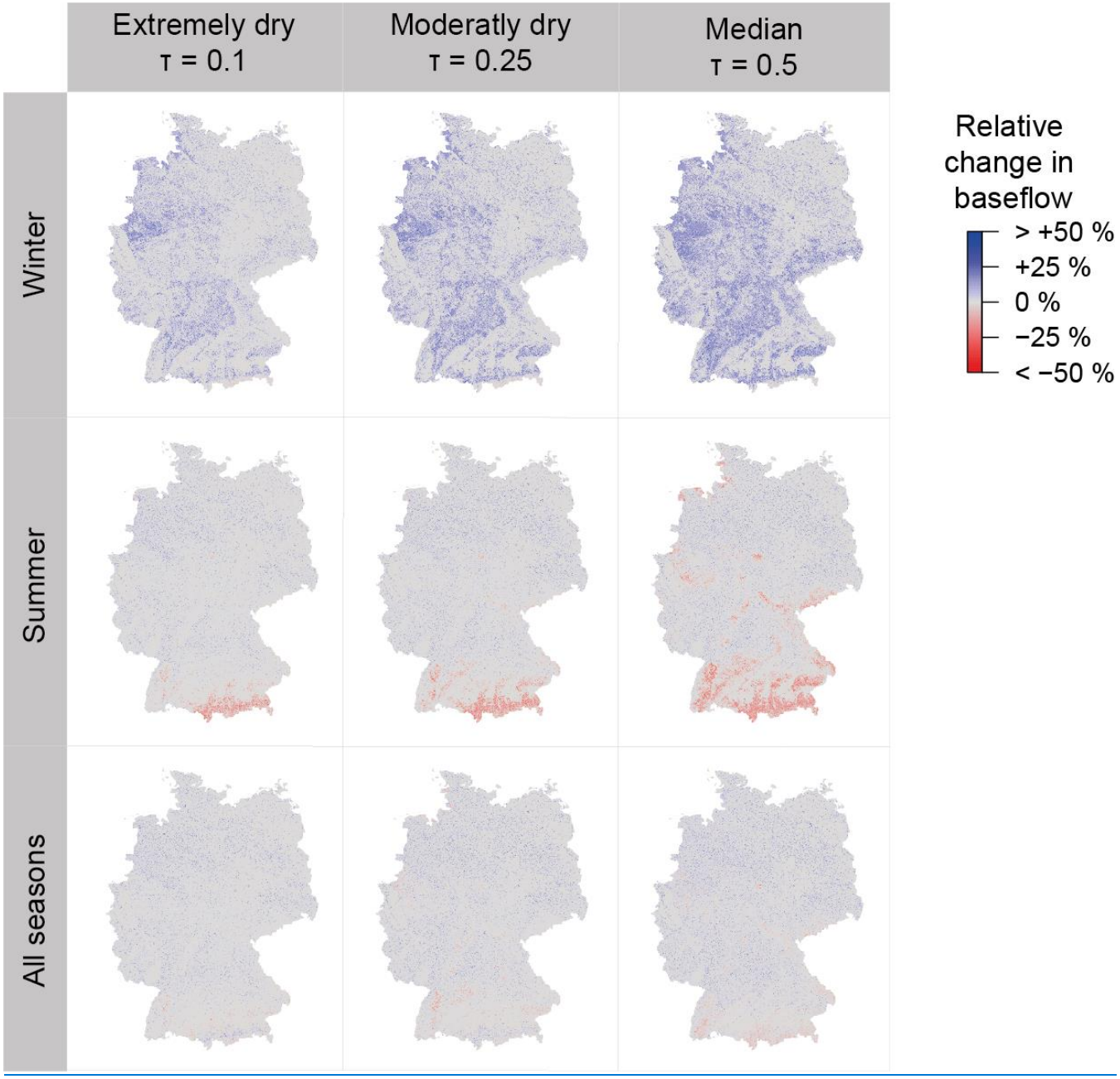


Fig. 5: Groundwater head changes for S_{SHIFT} in Germany for selected drought thresholds (columns) for different seasons (rows) for a shift of 15%.


Extremely dry
τ = 0.1
Moderatly dry
τ = 0.25
Median
τ = 0.5
Winter
Summer
All seasons
Relative change in baseflow
> +50 %
+25 %
0 %
−25 %
< −50 %

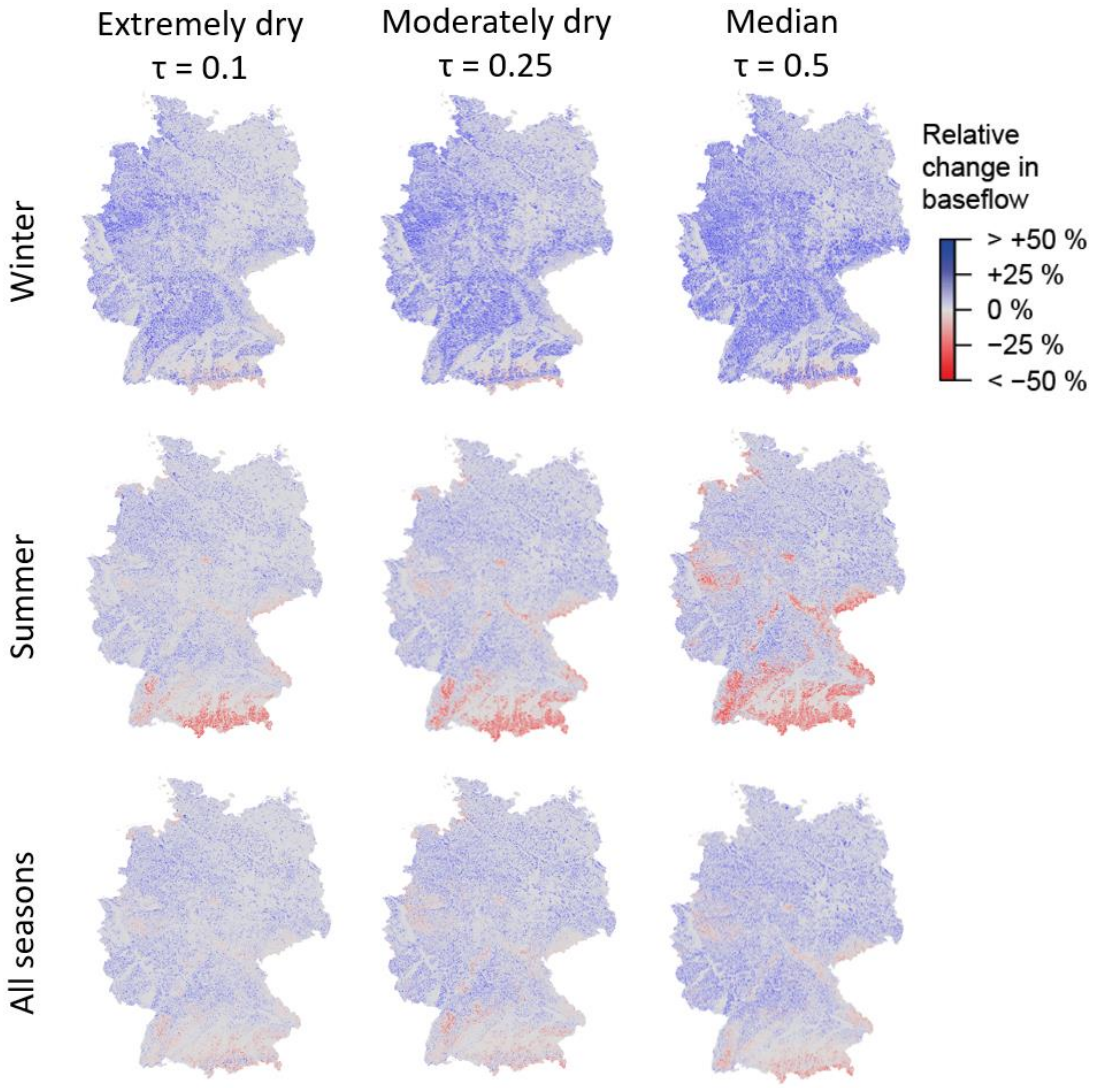


Fig. 6: Same as Figure 5 for relative changes of baseflow.

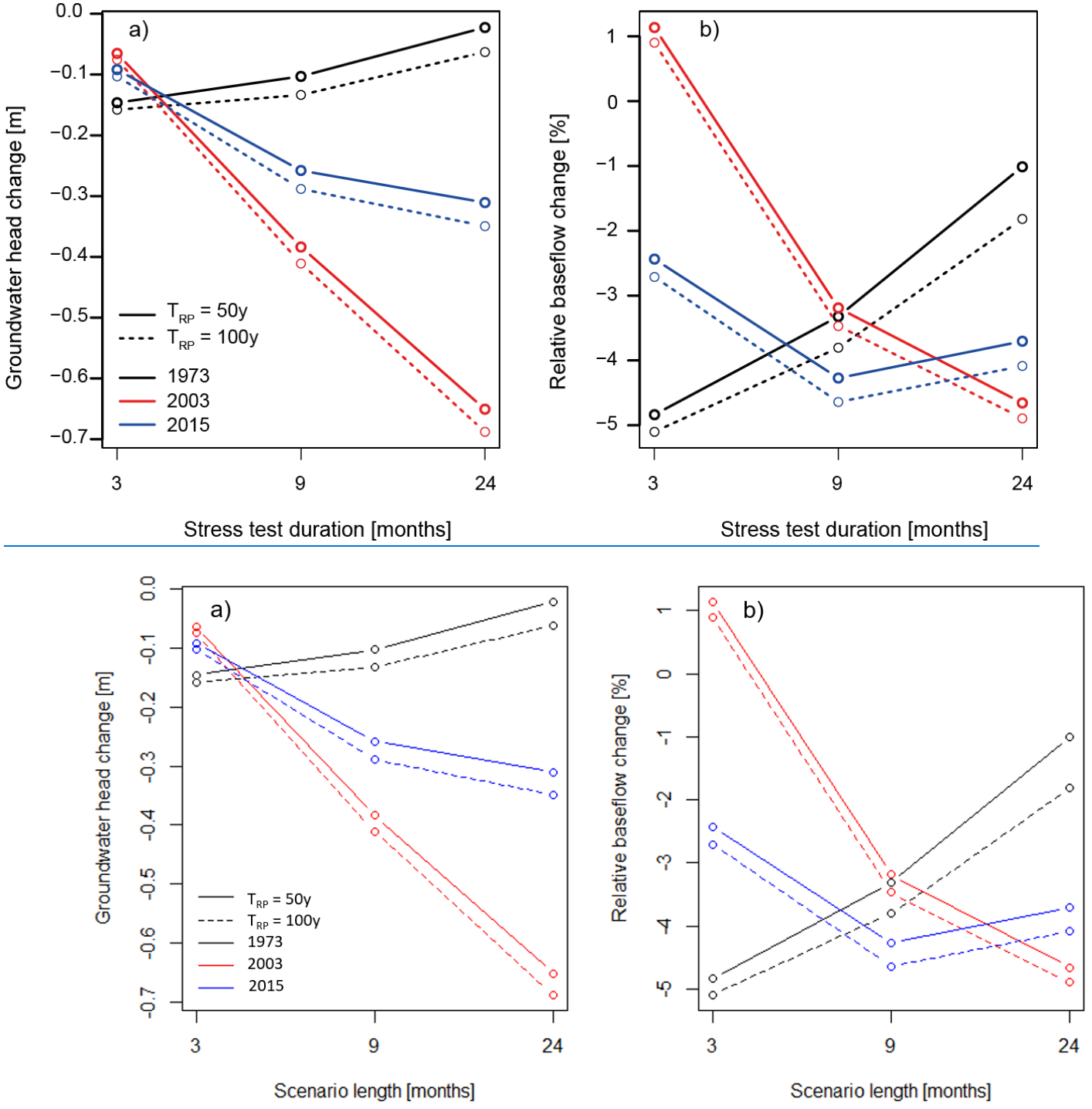
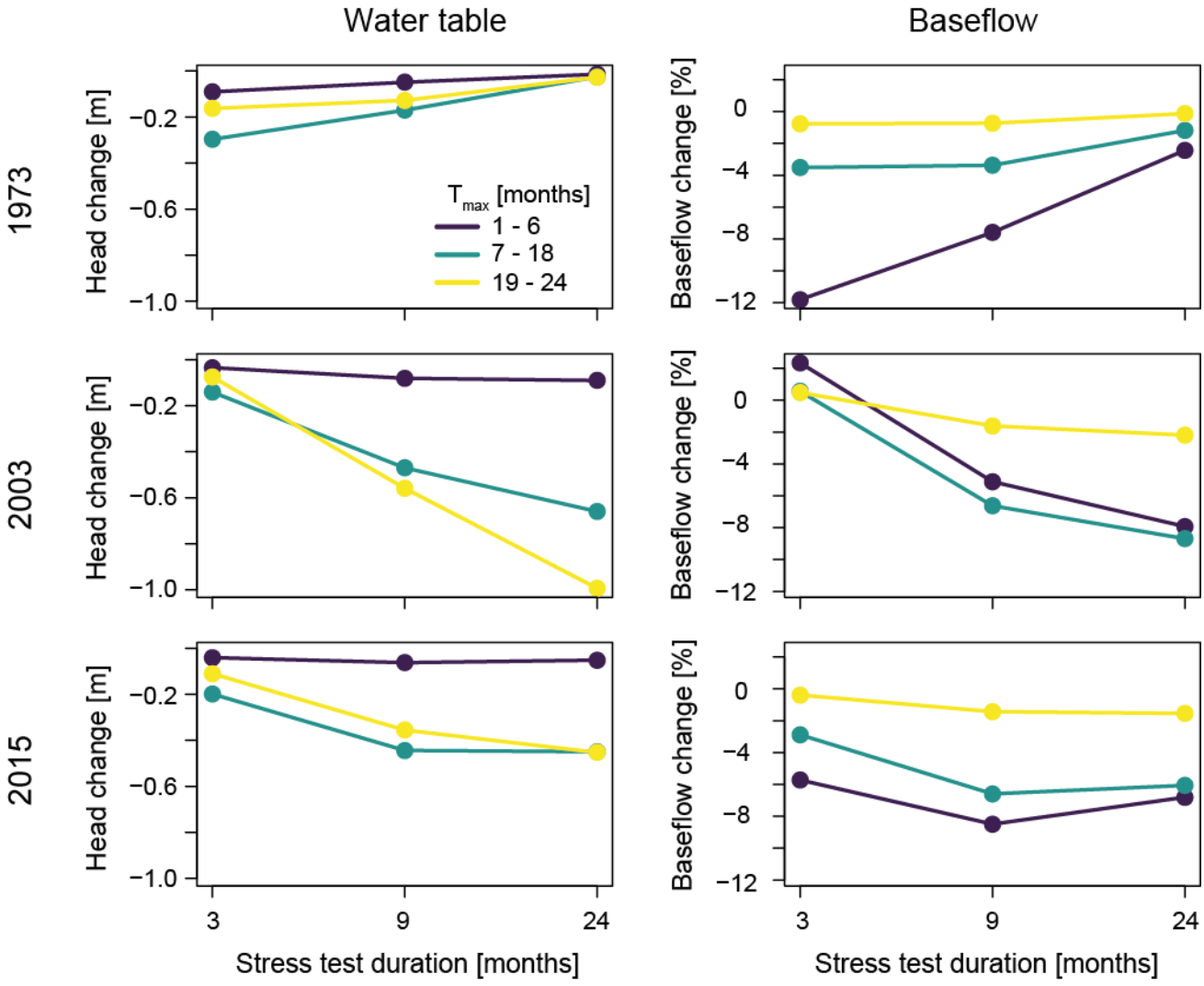


Fig. 7: Changes during drought averaged over Germany for all different SEVENT ~~scenarios~~stress tests: response of different events (1973, 2003, 2015), different antecedent recharge reduction time scales (3, 9, 24 months) and two return periods (T_{RP} = 50 and T_{RP} = 100 years). a) Groundwater head changes, b) relative baseflow changes.


Water table
Baseflow
1973
2003
2015
Head change [m]
Baseflow change [%]
−0.2
−0.6
−1.0
0
−4
−8
−12
T_{max} [months]
1 - 6
7 - 18
19 - 24
3
9
24
Stress test duration [months]

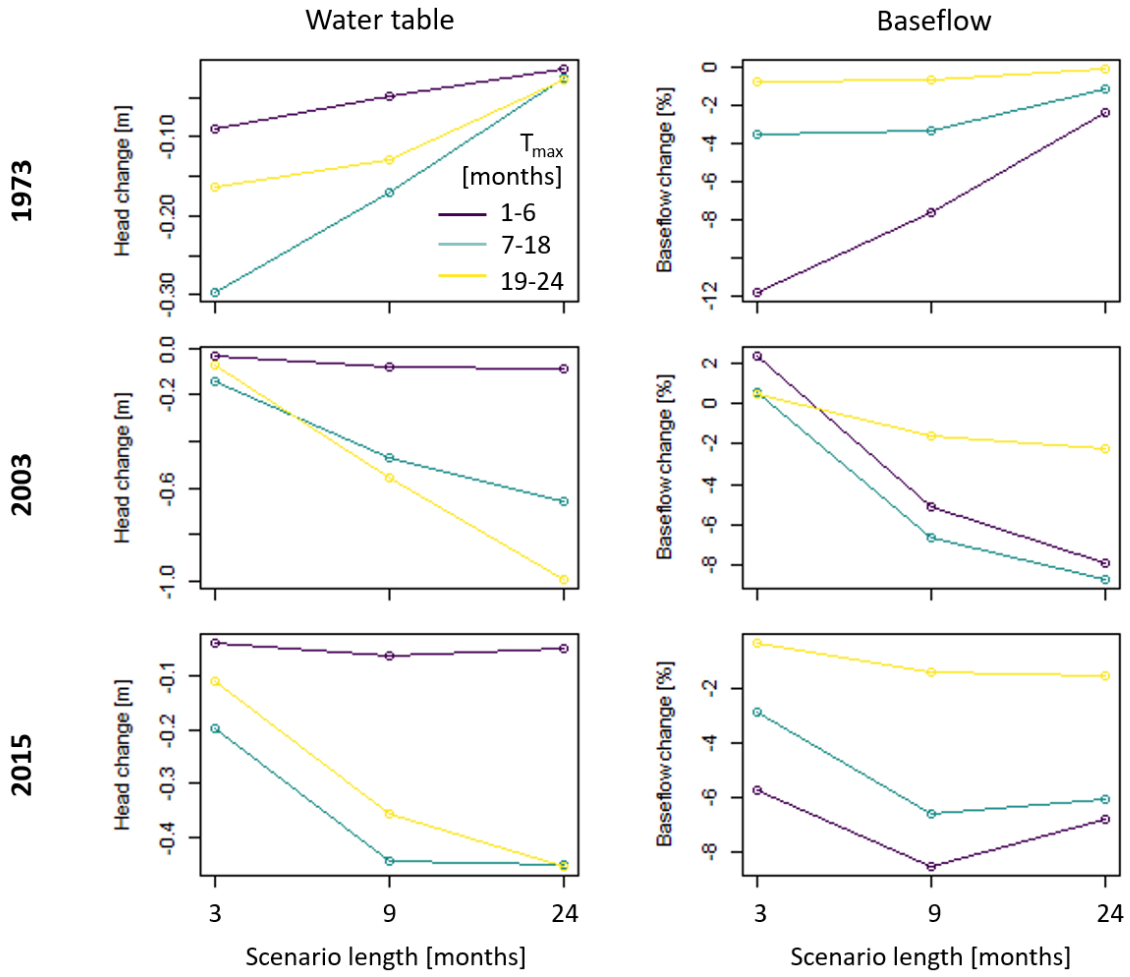
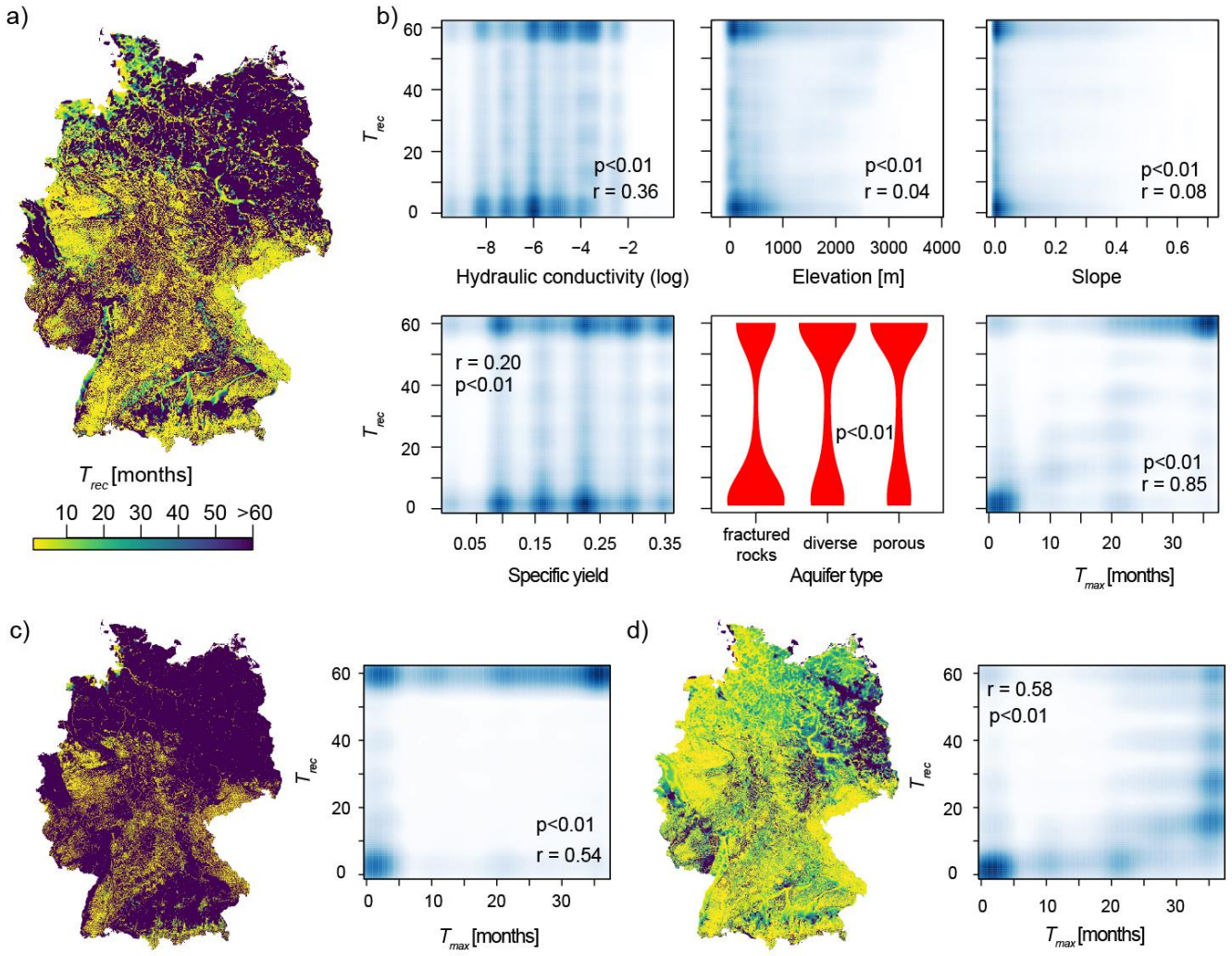


Fig. 8: Effects of SEVENT with T_{RP} = 50y for three different classes of T_{max} averaged over Germany. Note the different scales for the y-axes.


a)
T_{rec} [months]
10 20 30 40 50 >60
b)
T_{rec}
p<0.01
r = 0.36
Hydraulic conductivity (log)
p<0.01
r = 0.04
Elevation [m]
p<0.01
r = 0.08
Slope
r = 0.20
p<0.01
Specific yield
p<0.01
fractured rocks
diverse
porous
Aquifer type
p<0.01
r = 0.85
T_{max} [months]
c)
T_{rec}
p<0.01
r = 0.54
T_{max} [months]
d)
r = 0.58
p<0.01
T_{rec}
T_{max} [months]

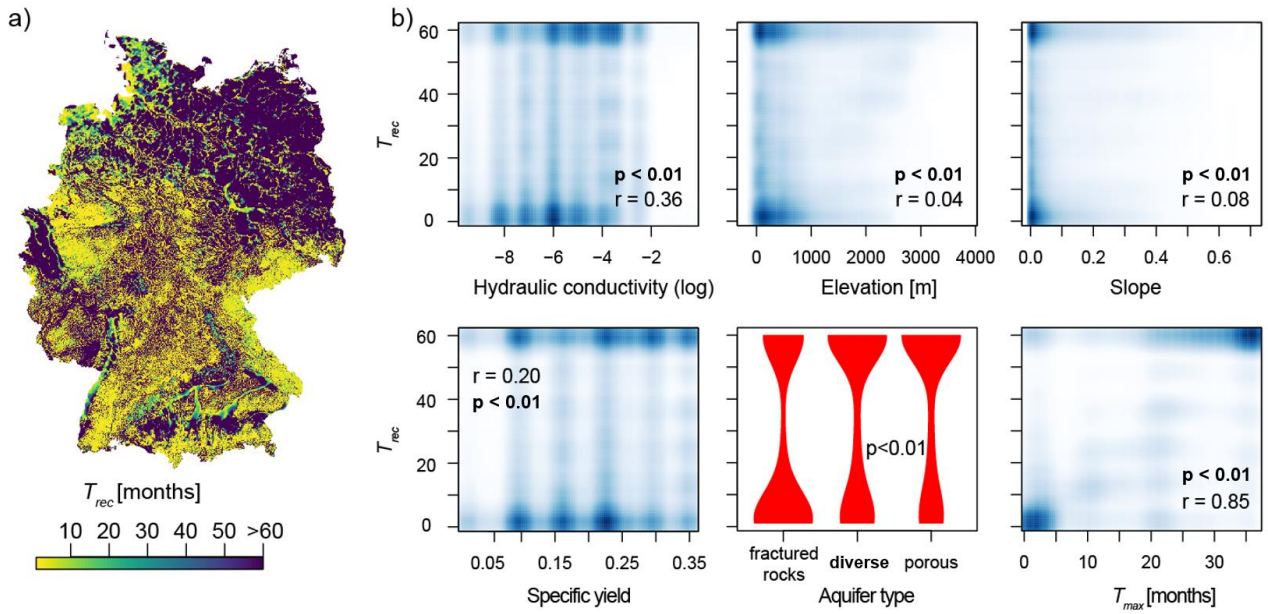


Fig. 9: Recovery time T_{rec} for ~~the reference simulation~~ S_{RECOV}. a) spatial distribution of T_{rec} across Germany, b) relationship between T_{rec} and model parameters hydraulic conductivity, elevation, slope and specific yield, aquifer type (taken from HÜK200) and precipitation accumulation time that has the maximum correlation with groundwater T_{max} (taken from Hellwig et al., 2020). c) + d) spatial distribution of T_{rec} and T_{rec} over T_{max} for dry resp. wet conditions during drought recovery. Blue colours indicate the smoothed density derived from all model grid cells. Red violins illustrate the distribution of T_{rec} in three different categories of aquifer type. r is the Pearson correlation coefficient for the variables compared, p is the corresponding p-value.