We would like to thank Referee #2 for the feedback and helpful suggestions on this manuscript. Below we give point-by-point responses to the comments (bold and italic).

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 acknowledge that the term 'stress-test scenario' might be misleading as 'scenarios' are often linked to ensembles of slightly different pathways. As our scenarios are considerably different from that, we will adopt the terminology in the revised manuscript to make this more transparent. 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. We will put this aim more clearly in the revised manuscript. Additionally, we will expand our model runs and 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 SShift and Srecov generic scenarios is to identify site-specific sensitivities to certain general hydroclimatic conditions which are of special interest to different stakeholders (in the revised manuscript we will phrase this more clearly). Typically, 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 directly recharge) and different responses of baseflow and groundwater can be analysed in more detail with more model runs. Therefore, for the revised manuscript we plan to run additional simulations of S_{Shift} (assuming other percentages of change such as 5%, 10%, 20% and 30%) and S_{Recov} (assuming rather wet and dry conditions during recovery). This will also help to better compare the outcomes of the different scenarios.

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 scenarios in the paper focus on sensitivities during drought. We agree that the model's ability to simulate the dynamics targeted in the scenarios is crucial for the reliability of the results. Hence, we will expand our reflections on the abilities and limits of the groundwater model. Specifically, in the revised manuscript we will define the required model ability for each scenario type (see Table R1) and discuss 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 will state this importance of T_{max} for the interpretation of the results more clearly in the revised manuscript.

Table R1: Required model ability and discussion of model performance for the different scenarios.

	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 R1). 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 scenario are – similar to the other scenarios – largest in regions of weaker model performance regarding T_{max}.

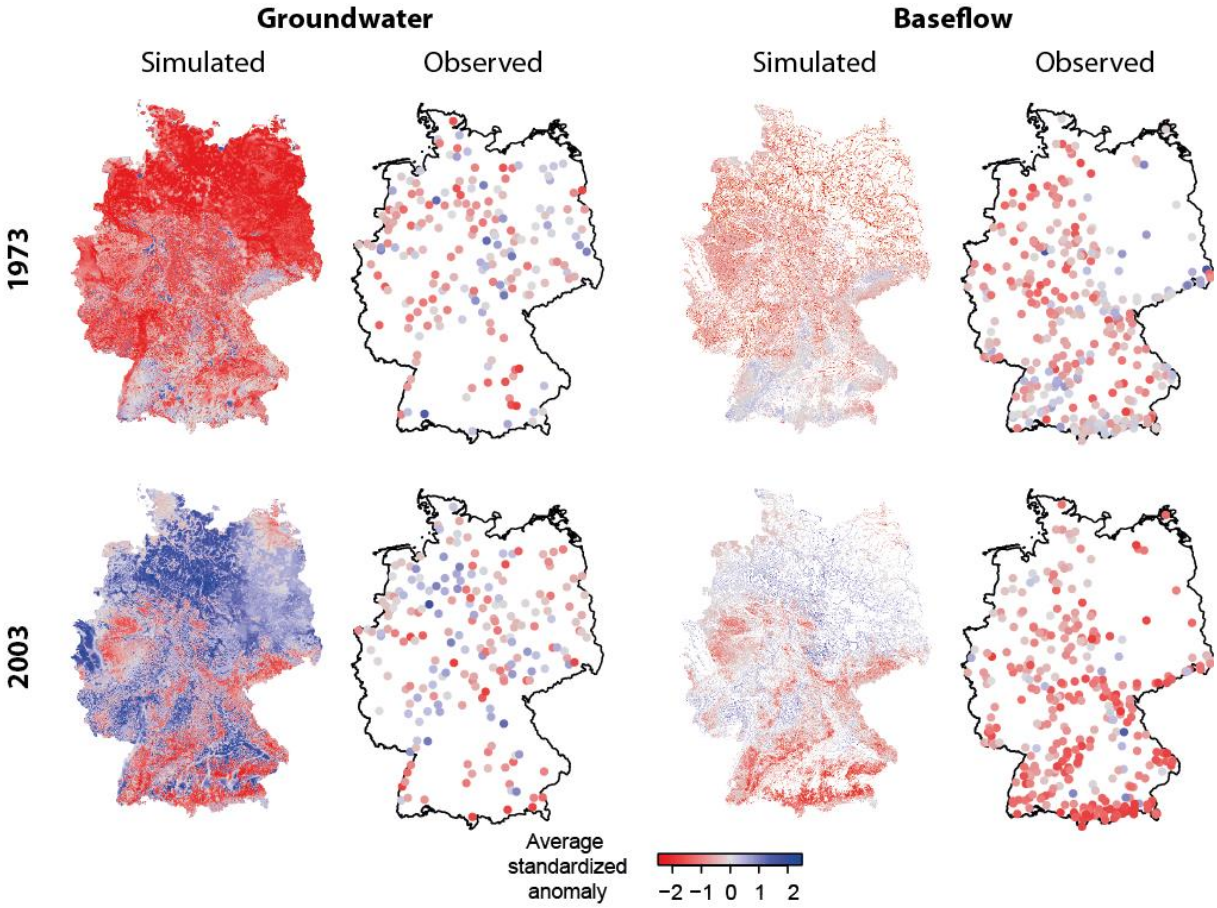

Figure R1: Simulated and observed anomalies averaged for summer months (JJA) of the benchmark drought years 1973 and 2003. Figure based on data taken from Hellwig et al. (2020).

We think with these additional remarks model results will be 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. In light of the commentary published by Gleeson et al. (2020) in the meantime, who discuss the issue of difficult-to-validate groundwater models with local observations, we suggest that we may add a more general discussion and conclusion on the issue.

4) **Minor Comments and Technical Corrections**

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

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.

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

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.

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

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

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

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?

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.

Thanks for pointing out. We will correct for this in the revised version.

References cited in this response:

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., ... & Oshinlaja, N. (2020). HESS Opinions: Improving the evaluation of groundwater representation in continental to global scale models. *Hydrology and Earth System Sciences Discussions*, 1-39.

Hellwig, J., de Graaf, I. E. M., Weiler, M., & Stahl, K. (2020). Large-Scale Assessment of Delayed Groundwater Responses to Drought. *Water Resources Research*, *56*(2), e2019WR025441.