

This study systematically investigates meteorological-to-groundwater drought propagation at the global scale using hyper-resolution (~1 km) groundwater simulations and classifies groundwater responses into three types to explore the global patterns and drivers of multi-year groundwater droughts. The authors also combine model simulations with observational data, which is an interesting and valuable attempt. Overall, this study provides useful insights into our understanding of meteorological-to-groundwater drought propagation. Before recommending the manuscript for publication, I have several comments and suggestions that I hope the authors will carefully consider and appropriately address.

Main comments:

1.The term “drought propagation” in the title is rather broad, whereas the manuscript primarily examines propagation from meteorological drought to groundwater drought. The authors may consider specifying “meteorological-to-groundwater drought propagation” in the title for greater precision.

Thank you for this suggestion, we agree that ‘meteorological-to-groundwater drought propagation’ is more accurate and will change the title.

2.Please clarify how all reported global percentages were calculated. Are these based on the number of groundwater grid cells or on area-weighted fractions? If the former, please demonstrate that differences in grid-cell area do not bias the reported global percentages. I would also recommend replacing expressions such as “35% of the world” with a more precise description of the denominator.

It is the former and we will make this clearer in the manuscript. These are percentages based on the total number of groundwater grid cells. We found little difference when calculating area-weighted fractions e.g. the 35% of the world which has on average a multi-year groundwater drought remains 35% using the alternative area-weighted method.

We will clarify this choice the first time we reference a global percentage on L255 by adding the following sentence:

Throughout this manuscript where we refer to global percentages we are referencing the percentage of groundwater grid cells that fulfill a criteria, not the area-weighted fraction which we found produced very similar results.

We will also change text references such as ‘35% of the world’ to ‘35% of global groundwater grid cells’.

3.The validation indicates that GLOBGM tends to simulate longer groundwater droughts than observed, particularly against GRACE (mean duration of 9.9 months in the simulations versus 5.5 months in GRACE). This is highly relevant because the central result of the manuscript is the prevalence of multi-year groundwater drought and the classification of systems based on drought duration. Could you quantify how this apparent persistence bias affects the estimated fraction of the globe experiencing multi-year groundwater drought (35%) and the proportion classified as meteo<GW? At minimum, I think this uncertainty needs to be discussed much more explicitly when interpreting the global percentages.

We agree that this uncertainty needs to be discussed in more detail and will add further text to the manuscript.

We chose to include a validation against GRACE data to provide some global, drought specific evaluation of the outputs from GLOBGM, but acknowledge that this decision brings with it significant uncertainty which we should have made clearer. When using the GRACE data alone, only 9% of global groundwater grid cells have an average groundwater drought duration which is multi-year, compared with 35% in the GLOBGM data. Furthermore, when using GRACE data to represent groundwater droughts in our typology, we find that only 7.8% of global data falls into the meteo<GW type, compared to 38% when using GLOBGM. This difference is largely a function of more data falling into the meteo=GW type which now contains 46% of data compared to the previous 19%. There are several reasons as to why this does not come as a surprise. Firstly, the GRACE timeseries is shorter than GLOBGM (it contains 18 years of data compared with 60 years simulated by GLOBGM) meaning that there are fewer drought events, and therefore the uncertainty bounds obtained using our bootstrap sampling are more likely to overlap between meteorology and groundwater. The GRACE data is also provided at a much coarser resolution (~300km vs 1km) and therefore is likely to have a more muted drought signal. While we acknowledge that droughts are a large scale phenomena, local geophysical properties at 1km could result in a different local drought signal that would be smoothed out when moving to spatial scales of 100km or more. The shorter GRACE period may also miss multi-year droughts in many parts of the world, as it was shown by van Mourik et al. (2025) that in some regions MYD can be linked to multi-decadal cycles. The comparison to GRACE data is complicated further by the fact that GRACE measures terrestrial water storage, rather than groundwater table depth, meaning that the GRACE signal can be dominated by changes in surface water and root zone soil moisture rather than changes in groundwater storage (Akl et al. 2025; Hohensinn et al. 2026). GRACE is also known to be less reliable across small islands and archipelagos e.g. Indonesia, the Philippines and the Caribbean due to land-ocean leakage (Pfeffer et al. 2023; van Jaarsveld et al. 2026). We will highlight more clearly in the manuscript that using GRACE data for model evaluation is not a perfect solution and make it clear that this introduces uncertainty.

We have already added some discussion of the comparability of GRACE data to GLOBGM to the results section (L239) in response to a comment from reviewer 1:

In several of the areas where the anomaly correlations are lower, we find nuances in the GRACE data which can begin to explain these patterns. Since GRACE is a measurement of total water storage, in tropical regions such as the Amazon or central Africa, the GRACE signal can be dominated by changes in surface water and root zone soil moisture rather than changes in groundwater storage (Hohensinn et al. 2026). In more arid regions, the lower anomaly correlations can be partially explained by the time it takes for surface water to reach the groundwater system, meaning that low correlations are not unexpected (Akl et al. 2025).

In response to comments from reviewer 1 we have also added a new section to the Discussion called Model Limitations which we will alter to also address this comment, the new section reads as follows with the new text in red:

The validation of global groundwater models is notoriously difficult due to mismatches in spatial resolution and representation between observed groundwater data and model simulations of groundwater. The relatively low anomaly correlations and mismatches in drought duration calculated between observed and simulated data should influence the confidence we have in the classification of groundwater response types and interpretation of the data in regions where the model does not perform well or cannot be evaluated. In regions where the anomaly correlation is low, the model may not be well suited for regional use and the results must be

interpreted with caution. There are several reasons why the anomaly correlations may be low which are unrelated to GLOBGM. Firstly, because of errors in the observational dataset. As discussed in Results section 3.1, in many places the GRACE data is not suitable for direct comparison to the groundwater data, and the signal should not be compared to groundwater table depth. This has been observed by many other studies in the literature who have identified problems with using this data for drought based evaluation (Akl et al. 2025; van Loon et al. 2017; Long et al. 2013). Although we tried to couple the GRACE evaluation with perhaps more suitable point observations, these are reasonably scarce. Similarly, comparing the length of droughts simulated by GRACE to those simulated by GLOBGM can be misleading. Using GRACE data alone only 9% of global groundwater cells have a multi-year average drought duration, compared with 35% using GLOBGM. This is likely to be influenced by the shorter length of this timeseries (18 years of GRACE data compared with 60 years simulated by GLOBGM), especially since it has already been shown that meteorological multi-year droughts can be clustered in time (van Mourik et al. 2025). As well as the limitations involved in comparing GRACE's total water storage to simulated groundwater table depth, the difference in resolution between the datasets (~ 300km vs 1km) is also likely to influence the number of multi-year groundwater droughts in the record, where a coarser resolution can result in a more muted drought signal.

As a groundwater community, we need to push for more global groundwater data to aid model evaluation, but in its absence, we suggest that there is still value in running models in regions where they cannot be validated. Since GLOBGM has been proven to perform as well as, or better than many alternative global groundwater models (van Jaarsveld et al. 2026), we believe there is value in using the results for global analysis. Currently the model's performance is on par or better than other global dynamic groundwater model simulations. Choosing suitable metrics for model evaluation is also tricky, and in this case, in many areas the anomaly correlation could be over exaggerating a mis-match in simulated and observed data in regions with little water table variability where small trends oppose each other. It is also likely that the meteorological forcing input to GLOBGM could be decreasing the anomaly correlations in some regions such as sub-Saharan Africa and north-eastern Siberian (van Jaarsveld et al. 2026; Liu et al. 2024). Uncertainties remain large for the sub-surface parameterization, however in the absence of better parameterization we have used the state-of-the-art available parameters to perform GLOBGM groundwater simulations. It is possible that the groundwater abstractions implemented in GLOBGM contribute to poor model performance in some regions. However, in general, GLOBGM simulates large scale groundwater depletion signals well, identifying many of the established regions of known depletion e.g. the Arabian Peninsula and California's Central Valley. It also recreates some of the regions where groundwater supplies are recovering e.g. the Guarani Aquifer System in South America (van Jaarsveld et al. 2026).

4. The authors mentioned that long-term groundwater decline violates the stationarity assumption underlying the standardized anomaly approach and may artificially produce very long drought events. However, given that drought duration is the central diagnostic of this study, I do not think that simply recommending interpretation alongside a trend mask is sufficient. Could the authors quantify the sensitivity of the main results after excluding grid cells with strong groundwater trends, or at least stratify the reported percentages by trend magnitude? In particular, it would be important to know whether the reported 35% global multi-year fraction and the meteo<GW category are disproportionately associated with strongly declining groundwater systems.

This is an interesting point and has highlighted that we should have spent more time discussing our approach to dealing with data with trends in the manuscript. We have now evaluated how the locations influenced by trends impact the statistics quoted in our manuscript and will elaborate on the implications of this in the discussion. The new text carries on from the existing L425 and will read:

Another challenge for groundwater drought definition is the presence of trends in groundwater levels. Globally, many groundwater aquifers are experiencing significant decline both from unsustainable abstractions and from climate change (Jasechko et al., 2024; Nazari et al., 2025; Berghuijs et al., 2024). These trends remove the stationarity assumption that is often used in drought characterisation and can artificially inflate or deflate drought duration. Solutions such as de-trending the data have not been successful in other global groundwater drought studies (e.g. Herbert and Döll, 2025; Wanders et al., 2015), so instead, here we implement a simple screening to remove drought events likely to be a function of trends. By removing droughts which have already begun on the first timestep or not ended on the final timestep of our groundwater simulations, we see that the 54% of global groundwater cells which have on average a multi-year groundwater drought duration is reduced to 35%. To supplement this initial screening we also redistributed the global groundwater data into the response types after excluding places where a significant trend is observed. To calculate which locations are impacted by trends we fit a linear regression to the standardized monthly groundwater anomaly timeseries and calculated the total change in standardized anomaly over the time period. Where this is more than 3 standard deviations of change across the period and the p value is < 0.01 we conclude there is a significant trend. Masking out this regions makes only a very small change to the global distribution of response types. The meteo>GW category goes from containing 42% of data to 38% when we mask out locations with trends, meteo=GW stays at 20% and meteo<GW changes from 38 to 42%. This minimal change assures us that our initial screening procedure is removing most of the significant trend-influenced drought events, but it also highlights a potential artifact of the methodology where in a small number of locations, removing one particularly long event can instead skew the average groundwater drought duration down, leading to a slightly higher proportion of data in the meteo<GW category.

Other comments:

1.Line 123: Please briefly discuss the sensitivity of the results to the 12-month threshold used to distinguish ND and MYD. For example, would the major spatial patterns remain similar using 18 or 24 months?

The major spatial patterns are similar no matter the chosen threshold, but as the threshold is increased we observe fewer MYD. With a 12 month threshold we observe that 35% of global groundwater cells have an average drought duration which is multi-year, and this number falls to 26% with an 18 month threshold and 21% with a 24 month threshold. Figure 1 displays the global distribution of average groundwater drought duration below, this time with the colour bar extended to show the spatial patterns amongst locations where the average duration is more than 12 months.

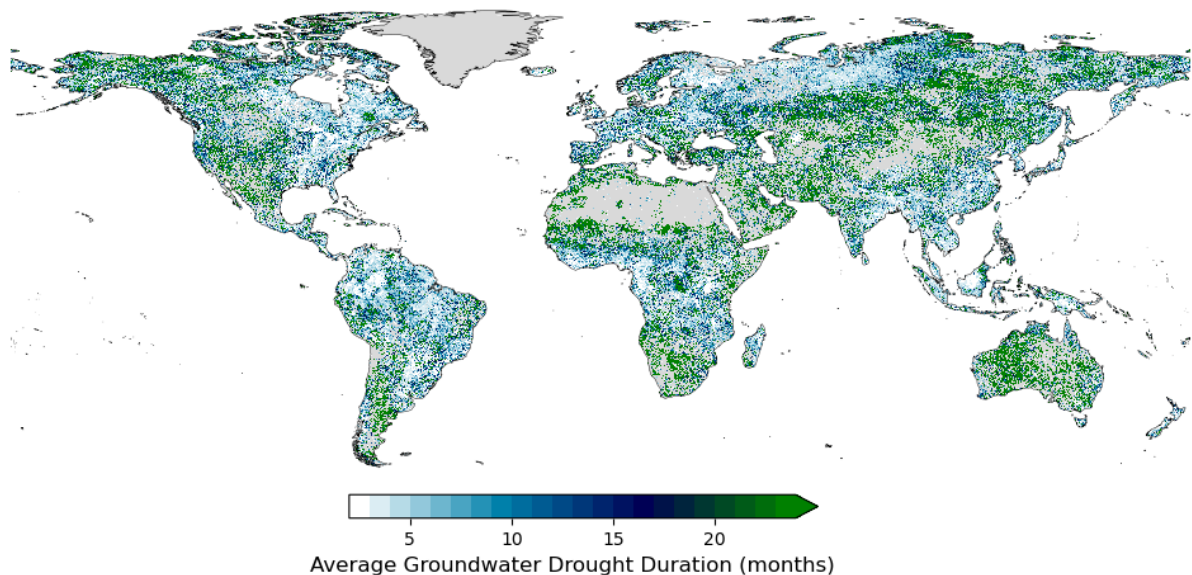


Figure 1. Global distribution of Mean Groundwater Drought Duration.

We have also analyzed how a change in this threshold would impact the distribution of data across our typology which can be seen below. As anticipated, a longer definition for a MYD results in fewer MYD's overall, but their distribution across the response types remains similar.

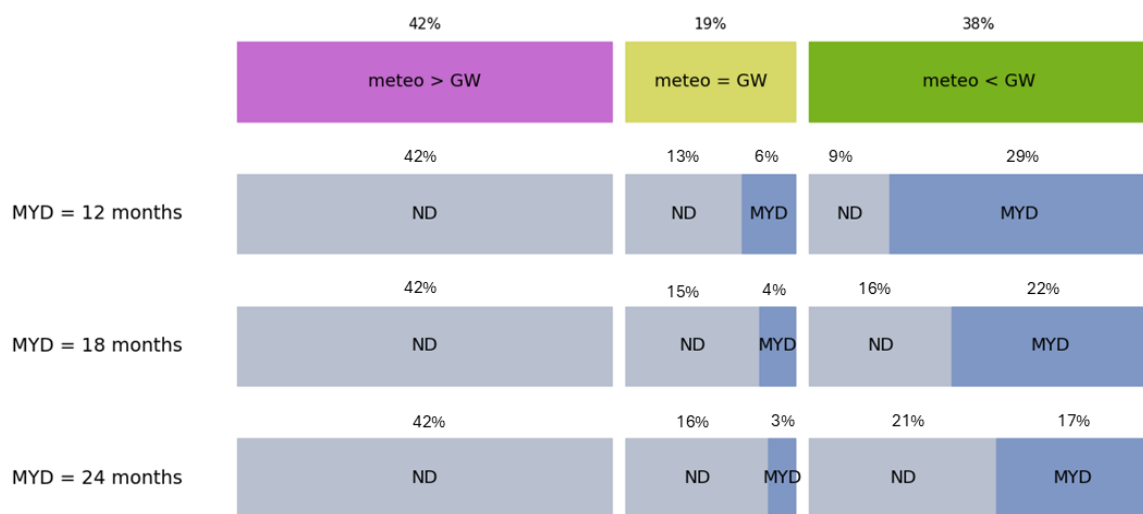


Figure 2. Global groundwater response types (top row). Rows 2-4 divide data in each type based on whether the average groundwater drought duration is multi-year (MYD) or normal (ND) using three MYD thresholds.

2.Line 134: Has the sensitivity of the results to the drought threshold been tested (e.g., -0.5, -1.0, -1.5)? Since event duration can be highly threshold-dependent, a brief sensitivity analysis would strengthen the robustness of the results.

Yes, we have tested this, and find that changing the threshold does change the frequency and duration of drought events (see Figure 4). To test this, we selected approximately 11,000 random groundwater cells from across the globe (see Figure 3 for their spatial distribution) and looked at

how the sample of drought statistics changed across three drought thresholds (-0.5, -1 and -1.5). We used a sample of points to reduce the computational burden of this analysis.

As expected a lower drought threshold is associated with fewer drought events and on average shorter drought durations. The mean groundwater drought duration from this sample of points goes from 11 to 19 to 28 months as the threshold changes from -1.5 to -1.0 to -0.5. The median groundwater drought duration is much less sensitive to the threshold, starting at 3.5 months with a threshold of -1.5, rising to 4 months when using the -1.0 threshold, but remaining at 4 months with the -0.5 threshold, suggesting that the thresholds are more influential on the distribution of extreme drought events. With a -1.0 drought threshold 23% of the selected points have an average groundwater drought duration which is multi-year, this number rises to 28% with a -0.5 threshold and decreases to 14% with a threshold of -1.5. The distribution of data across the global response types is conditional on the chosen drought thresholds both for groundwater and meteorology and we will make this clearer in the manuscript. We will add the following text to L429:

The results of this study are conditional on the drought definition methodology and we acknowledge that were we to use alternative thresholds or definitions we would see different results.

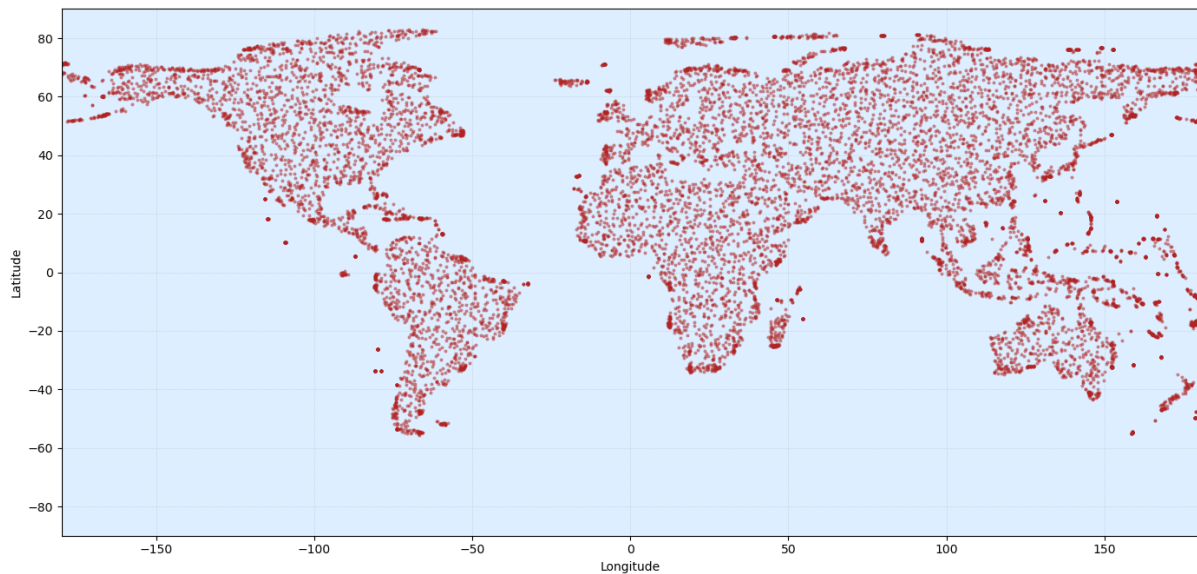


Figure 3. Global distribution of approximately 11,000 points used in drought threshold analysis.

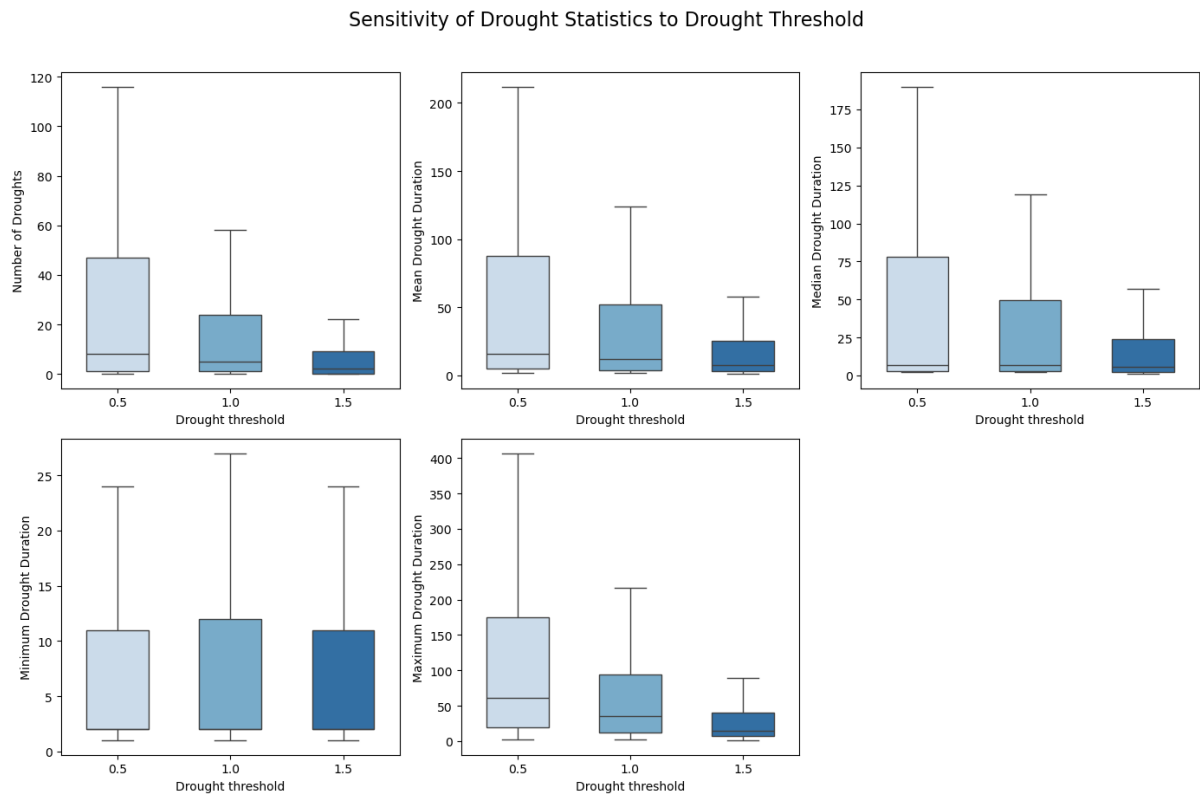


Figure 4. Box plots showing the sensitivity of several key groundwater drought statistics to changes in the drought threshold. Data comes from the points displayed in Figure 3.

3.Line 179: “DRR” should presumably be “DDR”.

Yes, we will correct this.

4.Line 191: I would consider replacing “confidence bounds” with “sensitivity bounds” or another term unless a formal confidence level can be associated with these intervals.

We will change this to uncertainty bounds.

5.Line 217: The term “groundwater drought predictability” appears not suitable for the ROC-AUC analysis. As currently implemented, the ROC-AUC quantifies the ability of contemporaneous SPEI-12 values to discriminate groundwater drought from non-drought months, rather than predictive skill in a forecasting sense. Unless an out-of-sample or lead-time prediction framework is introduced, I suggest using terms such as “discriminatory ability” or “meteorological indicator skill” and moderating the forecasting interpretation.

We will change our language here to use the ‘discriminatory ability’ phrase as we agree this is a better way to describe the metric.

6.The groundwater simulations are at ~1 km, whereas recharge is ~10 km and SPEI is 0.5°. Please discuss more explicitly how this substantial scale mismatch affects interpretation of the apparent fine-scale spatial variability in drought propagation and response types.

We fully agree that there is a scale difference with regard to the different fluxes. However, we believe this doesn’t have a major impact on our findings, as the original meteorological forcing is statistically downscaled to 10km before being passed onto the hydrological model. The meteorological signal in general is rather homogeneous over larger spatial extents and thus we

used the SPEI at 50km, as the statistical downscaling doesn't change any of the temporal dynamics of the signal. Most of the heterogeneity is actually added by the sub-surface properties which is resolved at the 1km resolution. Due to computational constraints it was not possible to fully resolve the groundwater recharge at 1km, but the surface water interactions are resolved at 1km, rather than 10km. We have acknowledged the impact of this by adding the following in the revised manuscript:

While the meteorological forcing comes at roughly 50km spatial resolution, the heterogeneity in groundwater response time mostly comes from the sub-surface geophysical properties of the groundwater system.

7. Please explain how the three “representative” time series in Fig. 5 were selected. Were they selected using objective criteria or chosen visually? Providing their geographic locations would also be helpful.

They were chosen using broad objective criteria and within that chosen visually. For example, for the pooling and lag plot we subset the data to locations where there was a lag of more than 6 months and a pooling metric of more than 2. We then selected one that was most representative of the process we were trying to demonstrate. We will add their geographic locations into the figure.

8. Line 403: The use of “tipping points” seems speculative based on the current analysis. Unless a threshold/nonlinear transition is explicitly demonstrated, I suggest describing this as a hypothesis or potential mechanism more cautiously.

We will update this text to read:

We hypothesize that these regimes have a mechanism whereby the groundwater system can buffer shorter, or less intense meteorological droughts, but when recharge is reduced beyond a certain threshold the groundwater becomes disconnected from the surface water, descending into an intense, multi-year drought event.

References

- Akl, M., Thomas, B. F., & Clarke, P. J. (2025). Global groundwater drought assessment revisited: A holistic re-evaluation of the GRACE-groundwater drought index across major aquifers. *Water Resources Research*, 61(12), e2025WR040389.
- Berghuijs, W. R., Collenteur, R. A., Jasechko, S., Jaramillo, F., Luijendijk, E., Moeck, C., ... & Allen, S. T. (2024). Groundwater recharge is sensitive to changing long-term aridity. *Nature Climate Change*, 14(4), 357-363.
- Herbert, C., & Döll, P. (2025). Assessing groundwater drought hazard in groundwater depletion regions: Recommendations for large-scale drought early warning systems. *Water resources research*, 61(6), e2024WR038684.
- Hohensinn, R., Gou, J., Meyer, U., Humphrey, V., Dorigo, W., Boergens, E., ... & Güntner, A. (2026). Water storage trends derived from the GRACE/-FO global gravity-based groundwater product (G3P). *EGU sphere*, 2026, 1-42.

Jasechko, S., Seybold, H., Perrone, D., Fan, Y., Shamsudduha, M., Taylor, R.G., Fallatah, O. and Kirchner, J.W., 2024. Rapid groundwater decline and some cases of recovery in aquifers globally. *Nature*, 625(7996), pp.715–721.

Liu, R., Zhang, X., Wang, W., Wang, Y., Liu, H., Ma, M., and Tang, G.: Global-Scale ERA5 Product Precipitation and Temperature Evaluation, *Ecol. Indic.*, 166, 112481, <https://doi.org/10.1016/j.ecolind.2024.112481>, 2024.

Long, D., Scanlon, B.R., Longuevergne, L., Sun, A.Y., Fernando, D.N. and Save, H., 2013. GRACE satellite monitoring of large depletion in water storage in response to the 2011 drought in Texas. *Geophysical Research Letters*, 40(13), pp.3395–3401.

Nazari, S., Reinecke, R. and Moosdorf, N., 2025. Global estimates of groundwater withdrawal trends and uncertainties. *Environmental Research Letters*.

Pfeffer, J., Cazenave, A., Blazquez, A., Decharme, B., Munier, S., and Barnoud, A.: Assessment of pluri-annual and decadal changes in terrestrial water storage predicted by global hydrological models in comparison with the GRACE satellite gravity mission, *Hydrol. Earth Syst. Sci.*, 27, 3743–3768, <https://doi.org/10.5194/hess-27-3743-2023>, 2023.

van Jaarsveld, B., Wanders, N., Otoo, N. G., Sutanudjaja, E. H., Verkaik, J., Zamrsky, D., & Bierkens, M. F. (2026). Global hyper-resolution modeling of historical and future groundwater dynamics. *Earth System Dynamics*, 17(5), 1201-1236.

van Loon, A.F., Kumar, R. and Mishra, V., 2017. Testing the use of standardised indices and GRACE satellite data to estimate the European 2015 groundwater drought in near-real time. *Hydrology and Earth System Sciences*, 21(4), pp.1947–1971.

van Mourik, J., Ruijsch, D., van der Wiel, K., Hazeleger, W., & Wanders, N. (2025). Regional drivers and characteristics of multi-year droughts. *Weather and Climate Extremes*, 48, Article 100748.

Wanders, N., Wada, Y., & Van Lanen, H. A. J. (2015). Global hydrological droughts in the 21st century under a changing hydrological regime. *Earth System Dynamics*, 6(1), 1-15.