First of all, we want to thank both reviewers as well as the editor for their constructive and thorough feedback on the manuscript. After having carefully addressed their valuable comments we are confident to have improved the manuscript substantially. We want to start with outlining the major changes performed:

- As response to valuable comments by both reviewers, we **sharpened how we use the term 'drought'** throughout the manuscript, and are now rigorous with definitions of different drought categories (mild, severe, extreme), as well as use the term 'dry (anomaly)' for dry periods with index values below 0. Besides that, in particular responding to reviewer #2, we also included an explicit **evaluation of the model's ability to detect drought events** (new Table 3, additions to sections 2.4.3 and 3.2)
- In a response to a valid comment by reviewer #1, we **moved the description of the quality-assured observational dataset** from the Results to the Data and methods section (and shortened it; parts are now found in Appendix A). This also means that Figure 2 introducing the study area and observation data is presented earlier in the manuscript, as new Figure 1.
- As response to comments by both reviewers, we r**estructured the Discussion section, where we moved results** to the Results (and Methods and data section): The groundwater sensitivity to summer and winter droughts now is shown in a new section 3.5. The analysis of correlation between geologic variables and drought lag versus the DK-model is now presented as part of section 3.3, and explained in the Methods under section 2.4.2.
- Similar, we added **references and results from comparable studies, both on drought specific and general model evaluation to the Discussion** – see sections 4.2 and 4.3.1.
- We **changed the soil moisture index from SMDI to ESSMI** (Empirical Standardized Soil Moisture Index, https://doi.org/10.1016/j.jag.2015.06.011). Please note that, despite this change obviously necessitating updates throughout the entire manuscript, use of the new soil moisture index did not alter the conclusions drawn from this study. The background for the change to ESSMI is two-fold: Both reviewers asked for results of the normality tests. This made us re-visit the previously performed Kolmogorov-Smirnov tests. Here, we decided to use the Shapiro-Wilk normality test instead, which is considered more rigorous and better suited for small sample sizes (n=30) as in our study. This also more strongly revealed issues with simulated soil moisture distribution (mostly due to these values being partially bound, e.g. to saturation water content). Furthermore, in currently ongoing work by some of the authors, we ourselves stumbled upon issues with the SMDI distribution. As a response to that, we decided to switch the soil moisture index from SMDI to the ESSMI, which handles bound values and resulting distributions due to reliance on empirical distribution.

In the following, we answer the individual comments and outline performed changes in the revision. Formatting below is indicating:

> *Reviewer comments*
>
> Authors' response
> Changes in section x.x / line xxx

(where line numbers refer to the document with tracked changes, and numbers of sections, figures and tables usually to the revised version)

Reviewer #1

Summary:

This manuscript presents a comprehensive and innovative evaluation of the DK-model for drought monitoring across multiple hydrological compartments. The authors compile an extensive observational dataset and analyse drought propagation from meteorological to soil moisture, streamflow, and groundwater droughts. This represents a substantial effort and addresses a highly relevant topic for hydrological drought research. The manuscript fits well within the scope of _Hydrology and Earth System Sciences_ and will be of interest to a broad readership.

The paper tackles an important problem and is based on a unique national-scale modelling and observational dataset. The main strengths are the comprehensive evaluation across hydrological compartments and the explicit focus on drought propagation. The main weaknesses concern (i) manuscript structure, (ii) conceptual clarity regarding drought definitions and thresholds, and (iii) limited quantitative discussion of model limitations and comparison to existing studies.

The study is generally well presented, with clear figures and a comprehensive modelling framework. However, the manuscript is currently quite lengthy and would benefit from a clearer separation of results and discussion, as well as a more in-depth discussion of the findings in comparison to other hydrological modelling systems used for drought assessment. After addressing the comments below, the manuscript would be suitable for publication.

Major Comments:

1 Manuscript structure and focus on drought evaluation
The manuscript would benefit from a clearer structure and stronger focus on the core drought-evaluation results. In particular:
- The selection and quality assurance of observational data should be moved fully into the _Methods_ section as data preprocessing.
- Detailed descriptions of observational datasets and preliminary performance results should be moved to the Appendix.
- This restructuring would shorten the manuscript and sharpen the focus on drought-related findings.

Thank you for this suggestion to make the paper easier readable. This is partly in line with a similar comment by reviewer #2. We have now merged section 2.3 and 3.2 into a pre-processing section (new section 2.3). As suggested, the section 2.3.1 is now shortened, and the detailed description of the selection process for the groundwater levels has been mostly moved to Appendix A. Changes: Re-written section 2.3, removed section 3.2, where parts were moved to Appendix A, and moved Figure 2 from previous section 3.2 to new Figure 1 in section 2.3.

- Model performance results (e.g. Figure 1) are shown for all available stations, whereas drought evaluation is conducted only for a selected subset. For consistency and relevance: Performance results should be shown only for the stations used in the drought analysis. Results for all stations can be provided in the Appendix.

As we feel that the point of the DK-model being a "general-purpose" model is an important part of the story, we did not remove the general/overall performance description in section 3.1 and the related old Figure 1 (now Figure 2), but added the performance across the 53 wells and 153 discharge stations selected for drought analysis as additional ECDF in Figure 2, panels a and b, and c and d, respectively. As also suggested in one of your minor comments. Now, each of these subplots shows two lines: One representing performance across all calibration data, the second one representing performance in the selected wells/stations.
Updated Figure 2 (previously Figure 1) and text changes around lines 431-446 (section 3.1)

3. Definition of drought and drought thresholds
Drought is repeatedly defined as index values below 0, which corresponds to "dry anomalies" rather than

drought. Drought classes are introduced in the Introduction, but only values <0 are analysed later.
- Please clarify the conceptual definition of drought used in the study.
- Why are thresholds for moderate or severe drought (e.g. SPI < −1) not analysed?
- Please ensure consistency between definitions, analysis, and interpretation.

Thanks for catching this partly imprecise use of the term 'drought' by us. In the revised version, we made sure to be consistent with the use of 'drought' for the moderate, severe, or extreme droughts as defined by drought indices below -1, -1.5, or -2, respectively. Index values < 0 are now referred to as 'dry period' or 'dry anomaly'. Moreover, also as a response to a comment by reviewer #2, we added an evaluation of drought detection (moderate, severe, or extreme) of the DK-model – see additions to section 2.4.3, section 3.2, and new Table 3. Changes throughout the manuscript, in particular lines 313-318. Or lines 495, 498, 500, 505, etc

4. Separation of results and discussion
Results are repeatedly introduced and interpreted in the Discussion section.
- All new results should be presented in the Results section.
- The Discussion should focus on interpretation, comparison with previous studies, and implications.
- Several subsections currently labelled as Discussion (e.g. Sections 4.2 and parts of 4.3) read as Results.

Thanks for pointing this out. Former section 4.2 (groundwater sensitivity to summer/winter drought) is now moved to new section 3.5 as parts of the Results. Former section 4.3 (controlling variables on drought propagation) was split up. The method is now explained as part of section 2.4.2 (Drought propagation and lag), whereas the results including former Table 3 are now part of section 3.3 (Accumulation period performance). Only relevant parts of those sections remain in the discussion, under section 4.1.
Section 2.4.2, lines 380-396 now outlines which/how controlling variables were extracted. Results are presented as part of Section 3.3, lines 587-604 with new Table 3. Groundwater sensitivity to winter/summer droughts now in new section 3.5, lines 650-667. Remaining relevant discussion of those results in section 4.1, lines 689-710.

5. Discussion depth and comparison to other models
The discussion is relatively short compared to the breadth of the analysis.
- Please extend the discussion by comparing the DK-model performance to other hydrological models used for drought assessment (e.g. national-scale or continental-scale models).
- Strengths and limitations of the DK-model relative to these systems should be discussed more explicitly.

Thanks for the suggestion, these are valuable additions to the manuscript. We added a new section 4.2 giving context with somewhat comparable studies on hydrological model drought evaluation, and extended section 4.3.1 on the general DK-model performance.
New section 4.2, extended section 4.3.1 (lines 794-797)

Specific Comments:

- "Climate projections indicate more frequent and intense droughts" (p.1, LL10): For Northern Europe, projected drought changes are mixed in the literature. Please nuance this statement (e.g. more summer droughts, fewer winter droughts).

Also in response to a comment by reviewer #2 on shortening that part of the abstract – we removed this statement (line 10-11)

- Abstract (LL21–24): Statements on model performance are very general. Please include quantitative results.

Added median r values for streamflow and groundwater. Line 22-23.

- p.3, LL71–74: You state that fewer studies address groundwater and the entire hydrological cycle, yet cite more studies than for other compartments. Please clarify.

True. However, it is beyond the scope of this article to deliver a review of the entire existing literature – we just wanted to give a few examples. We re-wrote this paragraph, adding two references (Van Lanen et al. 2016, Van Loon et al. 2024) that support the statement that more studies looking at the entire hydrological cycle are needed. Rewritten lines 72-79, with added refernces.

- p.4, LL100: "Weichsel and Saale" – please clarify that these refer to glaciations.

Done. Line 108.

- p.5, LL145: Is there a more recent reference describing developments over the last three decades?

We added Henriksen et al. 2020 as the most recent description, including the addition of the 100m model. More recent documentation doesn't currently exist – some of the authors are currently working on this. Line 153: added reference.

- p.5, LL154–155: How thick is the unsaturated/root zone in the model? This is essential for interpreting soil moisture results

The root zone is a lumped representation across the entire root depth (which is variable in time and space). (this is already discussed as a shortcoming in section 4.3.2). Added clarification, lines 174-175.

- p.6, LL189–190: Please quantify the impact of constant abstraction rates and provide supporting material (Appendix). A map showing trends in water consumption (Appendix) would help identify regions where this assumption is most critical.

We added information on the abstraction amounts, which on national scale are 10 to 25 mm per year, or 3 to 7% of net precipitation. That means that despite temporal variability, the overall magnitude of the abstractions is small, not majorly altering the hydrological cycle. Spatial distribution of abstractions is detailed in the provided reference. (the precise impact of the abstraction variability on the groundwater levels is beyond the scope of this manuscript; and actually something some of the authors currently are working on). Line 197-208: Provided values for abstraction amounts, including reference, added explanation.

- p.7, LL215: Which lag times were tested?

The TFN models allow inherently to detect different lags between input and output, without this having to be defined explicitly/manually. The term 'lag time' may, however, cause confusion in this context. We have therefore changed the wording to 'delays' in line 900 and 913 (this section is now part of Appendix A).

- p.7, LL219–220: Please clarify the criteria used for expert judgement in data evaluation.

Thank you for pointing out this unclear formulation. The criteria referred to in this sentence is the overall objective of the time series being climate-driven and not abstraction affected. We have therefore reformulated the sentence and added information on the particular signs that were the focus during the expert judgement. Lines 904-911: "Two hydrogeologists evaluated them first independently and then jointly to ensure that the final selected datasets are climate-driven, and unaffected by abstractions. This includes visual

screening for typical signs of abstraction influences, e.g., fast drawdowns during irrigation season or decadal trends in mean due to long-term reduction in abstraction. […]" (this section is now part of Appendix A)

- p.9, LL256ff: Only introduce indices actually used in the study and explain why they were chosen over alternatives.

Line 296-306: Removed explanation of not used drought indices.

- p.9, LL275: SPI values between 0 and −1 are not drought. Please correct and add references for drought class definitions.

Thanks for pointing out the inaccuracies surrounding the definition of 'drought' – this is also in line with your major comment 3 and similar comments by reviewer #2. Reformulated lines 313-318 – and more changes throughout the manuscript, where index values < 0 are now referred to 'dry' (period/anomaly) instead of 'drought'.

- p.10, LL285: Why are SMDI and SDI resampled from weekly to monthly?

Those values are calculated originally on weekly basis to allow for a better representation of the seasonal development. For the faster reacting variables streamflow and soil moisture, climatologies can change quickly. That means that the use of weekly climatologies will allow better representation of e.g. current anomalies for example for real-time monitoring, as in our case with the DK-model. Line 325-328: Added a short justification.

- p.10, LL289–290: Are results of the normality test shown?

Also in response to a similar request by Reviewer #2, we now present results of a Shapiro-Wilk test (more rigorous/suited to small samples like in our case) in lines 332-346.

- p.11, LL305: Please specify the length of soil moisture time series.

The observed soil moisture time series cover 10 years. Clarified in line 360.

- p.12, LL342: Interpretation ("little bias") should be moved to the Discussion and compared with literature.

Removed those statements (line 436, 437) and added a statement to section 4.3.1 (lines 794-798).

- p.12, LL344–345: A mean absolute error of 0.65 m relative to an average amplitude of 1.06 m appears large. Please discuss.

Yes, that is a valid objection. We both removed interpretative statements from the results section, and now provide some context: Amplitude mean errors are skewed by large values/outliers; the median absolute error, for example, is 0.41m. Adapted lines 438-440.

- Figure 1 (p.13): Show performance only for stations used in drought analysis.

This is indeed a good suggestion. See also the response to your major comment 2 – we ended up adding the performance across the selected drought stations to the Figure. Changed figure with performance across drought observation dataset (now Figure 2).

- p.13, LL356: Consider renaming to "Quality-assured observational dataset for drought evaluation".

As part of the restructuring of the manuscript, the entire section 3.2 which is referred to here was removed (and merged into sections 2.3 and Appendix A).

- p.13, LL358: Figure numbering and narrative would be clearer if the study area and data were introduced before calibration results.

That is true. As part of the restructuring of the manuscript, we now moved the previous Figure 2 (which shows the study area) into section 2.3, which makes it the new Figure 1.

- p.17, LL400–401: Statements regarding reduced correlation during drought periods should be quantitatively tested and better explained.

Correlations calculated for the full time series are generally higher than those obtained when restricting the analysis to dry conditions (index < 0). This behaviour is expected because the restriction truncates the joint distribution of the variables, reducing variance and covariance within the sample and thereby lowering Pearson's correlation coefficient. This is often referred to as truncation and range restriction. Lines 502-505: Added more explicit explanation of truncating effect on correlation coefficient.

- p.17, LL406: Drought index <0 does not equal drought.

Changed to "dry" (lines 495, 496, 500, 505, etc) – as part of the overall clarification of "drought" definition (major comments by both reviewers).

- p.19, LL435: The colour coding of wells is already explained in the text; repetition is unnecessary.

In our understanding, figures together with their caption should ideally be self-explanatory. That's why we opted to keep it as is.

- p.20, LL440: Testing accumulation periods up to 60 months for soil moisture with only ~10-year records is not meaningful. Consider limiting this to ~36 months.

That is correct. In the discussion, we added clarification along the lines that the long correlations to aggregation periods of 30 months and longer might be spurious (lines 689-690).

- p.25, LL507: Why was May 2020 chosen? Please provide context.

(also as response to a similar comment by reviewer #2) We chose May 2020 because it is a good example of different accumulation periods (of SPI and SPEI, i.e. meteorological condition) controlling different hydrological compartments: Fast reacting soil moisture (ESSMI) is in drought in May 2020, as the previous 2 months – shown as SPI2 – have been dry. When looking back an entire year – shown by SPI12 – that period has been wet. Which is reflected in groundwater (SGDI) and also streamflow (SDI) showing clear wet anomalies. In the revised version, we made this more clear (lines 639-646)

- p.26, LL518–520: The discussion starts very generally; consider linking more directly to your results

Lines 669-671: We removed the first two sentences of the Discussion – they were not strictly necessary and are basically a summary of statements from the Introduction.

- p.26, LL523: The key research question should already be clearly stated in the Introduction.

Lines 118-120: We added a summary of the key research question to the Introduction

- p.26, LL530: Statements about future soil moisture observations are vague—please provide references or concrete examples.

Currently, some of the co-authors and colleagues are working on establishing a Denmark-wide soil moisture monitoring network. It will include 10+ CRN sensors, situated to cover different land use and soil types across Denmark. Added "additional CRN sensors throughout Denmark as part of an upcoming soil moisture network (10+ stations across different land use and soil types)" to line 817.

- p.29, LL611–612: Please specify the thickness of the root zone. Given that the modelled soil moisture represents the entire unsaturated zone (i.e. a larger volume than observations), one might expect the opposite behaviour. Please explain this discrepancy more clearly.

The root zone is variable in time and space, and mostly varies between ~0.5m and ~2m. Unfortunately, we fail to understand what "opposite behaviour" is being referred to. Added root zone thickness in lines 811-812.

- p.31, LL656–658: The model does not appear to capture soil moisture lag times well. Please adjust this statement accordingly.

We adapted these statements to make clear that it is mostly streamflow and groundwater which are modelled well. Lines 858-860

Reviewer #2

Summary:

This manuscript addresses the need to consider different aspects of drought phenomena when evaluating hydrological models' ability to model drought, a topic relevant within the scope of HESS, and of importance to the hydrological modelling community. Specifically, they assess the general-purpose DK-model's ability to simulate the temporal variability of normalized soil moisture, streamflow and groundwater, and its ability to capture lag-times between normalized precipitation and the abovementioned hydrological components. By using an operational model, they bridge research and operational hydrology, making the results relevant for both spheres.

A major weakness of the paper is the lack of drought-specific quantitative evaluations of the model, despite that being a main aim of the study. The assessments of drought dynamics are limited to general evaluation metric of the dry 50% of the time series. Assessments that address drought in particular, is needed for the paper to meet its own objective and to be accepted for publication. Other weaknesses that need to be addressed include a consistent lack of benchmarking or justification of what to accept at good metric values, the issue of including the calibration period in the evaluation period, and a need of clearer organization of the paper.

The manuscript presents a comprehensive and interesting work in terms of modelling, data selection and lag-time analyses. By properly addressing the abovementioned gaps, as well at the comments below, the manuscript could be suitable for publication.

General comments:

1. *Despite the study's aims to evaluate drought dynamics, it lacks a clear definition of drought (e.g. occurrences, deficit volumes, durations based on thresholds) and quantitative evaluations based on those definitions. The time series of the lowest 50% values (presented in Fig 4 and Table 5) cannot be considered drought.*

Thanks for catching this partly imprecise use of the term 'drought' by us. In the revised version, we made sure to be consistent with the use of 'drought' for the moderate, severe, or extreme droughts as defined by drought indices below -1, -1.5, or -2, respectively. Index values < 0 are now referred to as 'dry period' or 'dry anomaly'. Moreover, we added an evaluation of drought detection (moderate, severe, or extreme) of the DK-model – see additions to section 2.4.3, section 3.2, and new Table 3. Changes throughout the manuscript, most notably lines 313-318 with definition of moderate/severe/extreme drought, or lines 495, 498, 500, 505, etc. New Table 3 presenting results of drought detection evaluation.

2. *The interpretation of the results as satisfactory lacks a benchmarking or justification. The manuscript explicitly states that the model performs "good" (L413), "very well" (e.g. L21) and "very good" (L532), for correlations between 0.56 and 0.63 (Table 2). Correlations of 0.56-0.63 are not generally accepted as very good, and references or justifications are needed for such statements. This comment also applies for other evaluation metrics.*

Thanks for your critical review of our text. In the revisions, we carefully examined the text to remove those potentially subjective evaluative statements. Moreover, we added some literature on comparable large-scale modelling to the Discussion, justifying our confidence in our results. Changes e.g. to line 21, 436-437, 683-684

3. *The DK-model is calibrated for 2000-2010 and evaluated for 1990-2023. Hence, the model has seen and been adjusted to about 30% the evaluated data during calibration. In general, models should be evaluated for data not seen during the calibration, to avoid too optimistic results. The authors need to justify their choice of not doing so, and preferably provide evaluation metrics for the unseen part of the evaluation period in appendix to ensure readers that the results are robust.*

The choice of overlapping calibration (2000-2010) and evaluation periods (1990-2023) is due to the need for an as long evaluation period as possible to determine robust drought indices. We are limited to starting our simulations with the DK-model in 1990 due to input data constraints. That leaves us only with a historic period of ~35 years to today, which we deem just enough for a rigorous evaluation of drought dynamics and propagation throughout the hydrological cycle, including deep groundwater. So, inevitably, there will be an overlap between calibration and evaluation period; a full split-sample experiment would likely deteriorate the validity of the drought index standardization. In the revised version of the manuscript, however, we provide a version of Figure 1 in the supplement, showing validation period (1990 – 1999 and 2011 – 2019) performance – with little difference to calibration period performance. Changes: Added a sentence about validation period performance to lines 444-446, that hint at the new Fig. B 1 in the supplement.

4. *The manuscript needs a clear organization, including a clear separation between results and discussions, and a methods section that prepares the reader for the results. The structure would benefit from moving the presentation of the evaluation datasets from results to methods, moving new results from discussion to results and introduce methods in methods section. Comparable studies/literature should be discussed in the Discussion (currently lacking).*

Thanks for pointing this out. In line with similar comments by reviewer #1, we restructured relevant parts of the manuscript (see also our 'major changes' outlined in the beginning of this document)
Separation of Results and Discussion: The groundwater sensitivity to summer and winter droughts now is shown in a new section 3.5. The analysis of correlation between geologic variables and drought lag versus the DK-model is now presented as part of section 3.3, and explained in the Methods under section 2.4.2.
Discussion of comparable studies: New section 4.2 in Discussion
Observation dataset part of Methods: We have now merged section 2.3 and 3.2 into a pre-processing section (new section 2.3, i.e. parts of Methods). The detailed description of the selection process for the groundwater levels has been moved to Appendix A. Changes: Re-written section 2.3, removed section 3.2, where parts were moved to Appendix A, and moved Figure 2 from previous section 3.2 to new Figure 1 in section 2.3.

Specific comments:

1. *Include quantitative results backing your conclusions in the abstract*

Added median r values for streamflow and groundwater. Line 23.

2. *Nuance or rephrase the statement of "Denmark's dense hydrological monitoring network" (does not apply for soil moisture), and/or note reader about the sparse soil moisture data when providing results in Abstract.*

Changed to "dense monitoring network for streamflow and groundwater". Line 19.

3. *Suggest to shorten first 15 lines of the Abstact before the study is presented.*

Thanks for this suggestion. We removed "Although historical drought trends in northern Europe are uncertain, climate projections indicate more frequent and intense droughts." and "…despite their importance for baseflow and water supply.", lines 10-11 and 13-14.

4. *Other studies /evaluation results using DK-model should be presented in the introduction or methods.*

We added some studies with evaluation of the DK-model; more detail follows in the Methods. Line 108-110: new references.

5. *Methods: add the currently lacking descriptions of the temporal resolution of the model and data.*

We added the clarification that the DK-model is ran with a maximum timestep of 24 hours, and that validation data is daily. Meteorological forcing is daily, as already described. Lines 166-167, 191.

6. *L158-160: Briefly describe the meteorological data used (how it is produced)*

Added clarification: "which are interpolated from daily or sub-daily data from in situ stations across Denmark", lines 169-170. (more information can be found in the provided references)

7. *L190: Define reference period first time it is mentioned*

Added '1991 to 2020', line 205.

8. *L176-192: Clarify why you use of all data for optimization (including human influence) when you use fixed abstraction rates and fixed wastewater outflow. I.e. explain why the choices described in sect 2.2.4 does not affect the choice of data used in the optimization described in 2.2.3?*

We added a clarification here, expanding on the existing text: For standard model calibration and validation, we always use the best available data on groundwater abstractions, which are the annually varying data from the Jupiter database. For drought analysis, however, this would mean that in parts of the country where groundwater abstraction rates have been variable, the drought signal especially in the groundwater would be dominated by those changes of abstractions – and not meteorological variability which is our interest here. Hence, we opted to run the model with constant abstractions, at mean levels over the reference period. Lines 197-208: added clarification.

9. *L202: Link or reference to Jupyter. Is it open or restricted? Information can be provided in the data availability section.*

Line 178, 220: added reference and clarification that access to the Jupiter database is public.

10. *L205-206: Please clarify. Requiring minimum 20 yrs of data allowing for 20% gaps can be interpreted as requiring minimum 16 yrs of data. Is only 20% of the series allowed to have bi-monthly data, or do the 20% refer to gaps larger than two months.*

What we meant was that the first and last observation have to be at least 20 years apart. Then, only wells were considered that had at least one 20 year window with a maximum of 20% of missing data. So 20% refers to gaps larger than two months. Lines 224-226: clarified the just mentioned.

11. *L208 (and elsewhere): Consider changing "climate" to "meteorological", "atmospheric" or similar.*

Where appropriate, we changed to "meteorological" throughout the manuscript.

12. *L220: Please specify "criteria", and in particular specify applied criteria important for your drought analyses.*

(in line with a similar comment by reviewer #1) Thank you for pointing out this unclear formulation. The criteria referred to in this sentence is the overall objective of the time series being climate-driven and not abstraction affected. We have therefore reformulated the sentence and added information on the particular signs that were the focus during the expert judgement. Please note, however, that this is now part of Appendix A. Line 906-913: "Two hydrogeologists evaluated them first independently and then jointly to ensure that the final selected datasets are climate-driven, and unaffected by abstractions. This includes visual screening for typical signs of abstraction influences, e.g., fast drawdowns during irrigation season or decadal trends in mean due to long-term reduction in abstraction. […]"

13. *L230: Why exclude catchments smaller than 15 km2. The model has a 500m spatial resolution.*

Both measurement and model error increases for very small discharges / catchments. Also, the smallest typically used catchment size in Denmark is ~15km^2, as reflected in national "ID15" dataset of catchments (the 15 referring to catchment size of 15km^2), see e.g. https://essd.copernicus.org/articles/17/1551/2025/essd-17-1551-2025.html
Line 257: Added "as both measurement and model error increases for very small catchments"

14. *L234-235: Why does this argument not apply for groundwater wells? Please give general information about recharge area or similar to justify why point (i.e. well) observations are ok for evaluation of model at 500m resolution in sect 2.3.1.*

Lines 264-265: We added the clarification that soil moisture values can vary dramatically within such small areas (including references); this heterogeneity cannot be modelled by our models. (Groundwater levels from wells are generally less variable in space across small scales)

15. *Sect 2.3.1-2.3.3 Please consider moving number of stations and Fig. 2 to here.*

Thanks for this suggestion. Along with the larger restructuring of the manuscript, old Figure 2 was now moved to section 2.3 as new Figure 1, and the resulting dataset is described here. Whereas details surrounding the quality assurance of the groundwater level dataset were moved to Appendix A.

16. *L244: You state "hundreds" based on reference that states "more than 100". Please consider rephrasing as hundreds may be interpreted as much more.*

Line 279: Changed to "A multitude of different drought indices exist"

17. *L275 and L277: Please add reference to categories, and consider removing "mild drought", as this implies drought half of the time.*

We removed "mild drought" (along with the general clarification of the terms "drought" and "dry anomaly" etc. throughout the manuscript). Furthermore, we added the reference that established these categories that are commonly used ever since. Lines 312-318: Added reference to McKee et al. (1993)

18. *L275-277: Please consider using (some of) these thresholds as a basis for drought definition (ref major comment a).*

Yes. Amongst others, as also mentioned above, we now added an evaluation of drought detection based on these drought categories. See e.g. the new Table 3.

19. *L275-277: Thresholds and names does not align with what is used in Fig 9. Please fix.*

Thanks for noticing that. Fixed in an updated Fig. 9

20. *Table 1:*
    - *Please justify choice of Makkink evapotranspiration method in main text.*
    - *Please define "q-points" in table legend*
    - *Please be consistent in abbreviation for groundwater (SGI or SGDI)*
    - *Please justify the transformation ln(Q) in main text.*

A modified Makkink formula is the basis of the meteorological data from the Danish Meteorological Institute. Line 169: We added an explanation of this and a reference to the used formula.
We removed "q-points" from the Table, as it isn't strictly necessary here; q-point then is defined later in the text (line 355). Thanks for noting the mistake – it is corrected to "$SGDI_{deep}$" now in Table 1. The ln-transformation of streamflow follows the SDI definition of Nalbantis and Tsakiris (2009) to account for typically skewed streamflow distributions; we added this to the text (line 305)

21. *L280: Using 1-month accumulation period for SPI and SPEI is not "most commonly practiced" (3-, 6- and 12-months are generally more common). Please justify your choice, e.g. by your knowledge of drought development in Denmark, or other studies.*

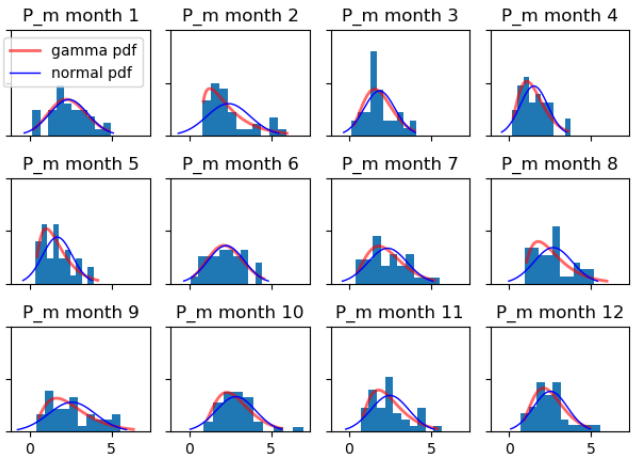
We are aware that other accumulation periods are often used. We also use them ourselves in this study; section 2.4.2. However, SPI/SPEI across all accumulation periods are calculated based on monthly values – in the sense that they account for, e.g., precipitation across the last 3 months, and are calculated from monthly values/climatology, updated every month with the new 3-month aggregate (not only 3-monthly).

22. *L285-286: Please clarify advantage of computing on weekly and resample to monthly instead of computing directly on monthly data.*

Lines 325-328: We explain why weekly climatologies were chosen for ESSMI and SDI: To allow for these fast-reacting variables to better follow their seasonal variability.

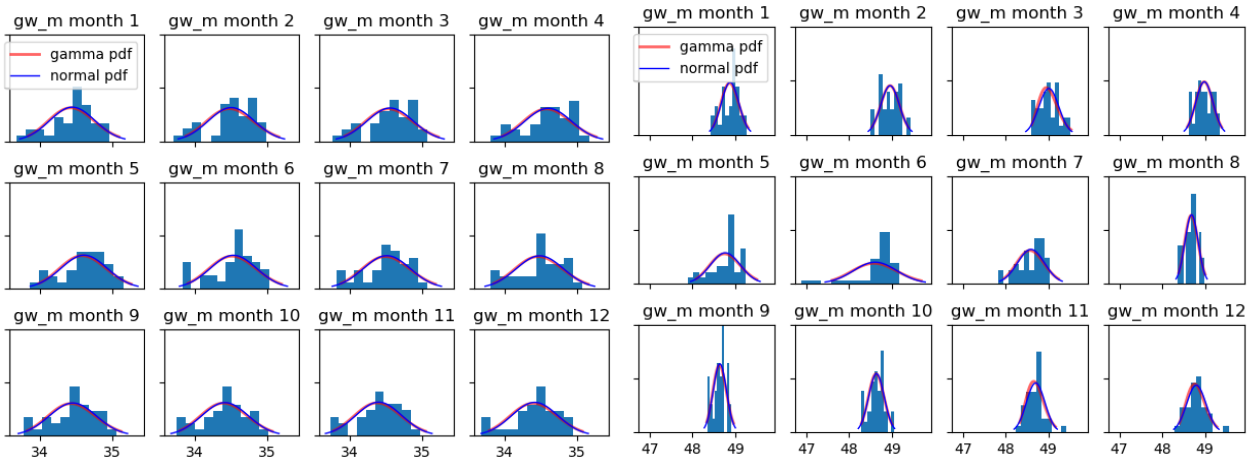
23. *L290: Fitting precipitation to a normal distribution is rare. Can histograms or similar be visualized to back up the choice?*
24. *L290-294: Please quantitively back up the conclusion from the Kolmogorov-Smirnov normality tests. And add the reasoning behind the choices for SDI, which is currently lacking.*

Normal distribution for precipitation: Yes, that is rare. However, we find that this works for the Danish case, probably due to its relatively stable temperate climate. We performed visual checks of the data before the decision (histograms etc) and decided that even though a gamma distribution sometimes delivers a better fit, it also more prone to overfitting in these small sample sizes (n=30). One grid example for monthly precipitation is shown below:

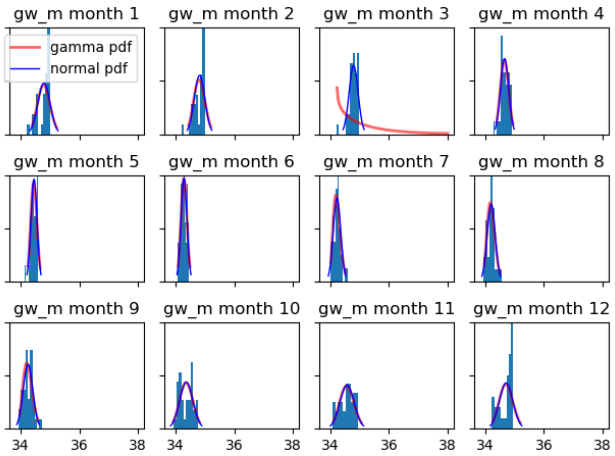


For a row of months (Jan, Mar, Jun, Jul, Dec) the normal and gamma distribution fit are similar. For some months (especially Feb) it seems plausible that the gamma distribution delivers the better fit. For some months, Aug in this example, however, we are not certain whether the gamma distribution simply is fitting to an anomaly mostly caused by the relatively small sample size.

Similar applies to **groundwater**, where it can be seen in two examples below that (i) normal and gamma distribution produce very similar fits, (ii) the small sample size makes distribution fitting difficult, as the seemingly discrete values groundwater level partially assumes must be considered random artefacts.



Whereas yet another example below shows how gamma distribution fitting can fail (Mar) for groundwater levels:



Given this, and the wish for a simple stringent method across indices, including the possibility of extrapolation (to unseen values e.g. in future climate), we opted for the normal distribution for the SPI, SPEI, SDI, and SGDI.
We now also performed **Shapiro-Wilk normality tests** (more rigorous and better suited to small sample sizes as here with n=30) and quantify the results in the updated version of the manuscript.

The only **exception to that is soil moisture**, as simulated soil moisture is bound, and often takes on discrete values (in Denmark e.g. saturation water content during parts of winter). Here, a closer look revealed that even though the previously used SMDI does not directly rely on fitting a normal (or other) distribution, it still can be problematic when used with such partially bound inputs. Hence, we decided to **change the soil moisture index to the ESSMI** which is designed to handle such bound values by relying on the empirical distribution function, and using a kernel density estimate to obtain a smooth approximation that permits limited extrapolation beyond the observed data.

Re-written section 2.4.1 (lines 332-351) to include results of Shapiro-Wilk normality test and outline reasons for choice of normal distribution.

25. *L301: Briefly define what a "streamflow calculation point" is, either here, or under the description of the model.*

Line 355: Added explanation of 'q-point'

26. *L333-335: The evaluation of representativeness of the observation points of the entirety of Denmark does not take into account that you would want that stations to represent the variability of streamflow and groundwater behavior across Denmark, which you yourself highlight as an important aspect e.g. by separating into two different regions in Fig 8. Please consider to discuss the dataset's representativeness in terms of representing the different hydrological regimes/behaviors across Denmark in the discussion (or under the presentation of the evaluation datasets).*

This is part of the discussion now (added an explanation lines 691-693; also partly in lines 693-703) where we now in more detail explain that the differences in observed drought propagation can be reproduced to some degree by the DK-model. Futhermore, the analysis of drought propagation lag to groundwater together with the controlling variables on that (now explained as part of the methods, section 2.4.2, lines 380-396, and results in section 3.3, lines 587-604, and Table 4) underline the same "variability of behaviour".

27. *Sect 2.4.3: Please align with results (including results in the discussion section) – ref major comment d). When presenting metrics (e.g. correlations and RMSE), it would be beneficial if you stated (and justified) values that indicate good performance and not, as this is unclear in the results section when there are no alternatives to compare the performance with.*

(also in line with other similar comments, such as your major comment 2) We generally removed potentially subjective evaluative statements from the Results section. Besides that, we added some references to somewhat comparable large-scale model evaluation studies (line 718), as well as examples of explicit model evaluations against drought (line 718-732), which generally support that our model performance is decent (e.g. TAI on groundwater droughts from Hellwig et al. 2020 of 0.25 – our TAI is 0.46 (see line 723-725).

28. *L331: Please state how you aggregated.*

Line 409: Made clear that we mean average across Denmark.

29. *Sect 3.1 is missing – maybe a numbering issue.*

Thanks for noting that – the missing heading "3.1 General DK-model performance" (line 425) now is added

30. *L338: Please include in discussion whether a median KGE of 0.67 is satisfactory, with reference to other modelling studies.*

We included references to some large-scale model evaluation studies, underlining that our DK-model performance indeed is satisfactory. Line 713-731: Added references to other large-scale model evaluation studies in Discussion

31. *Figure 1 and L337: State which period is used for the overall performance metric scores.*

Figure 2 (previous Figure 1) (and related lines 444-446 hinting to validation performance in Fig. B1) now makes it clear that we refer to the calibration period 2000 to 2010.
Line 398-400: Now more clear that for drought metrics evaluation, full available time series (overlap between observed and simulated data) is used.

32. *L340: Please introduce equation in methods, not results.*

Fbal is only used here, in section 3.1 and Fig. 1. Defining it further up would in our eyes not help the readers. As it is a fairly trivial formula, we moved it into the text itself, instead of a dedicated equation (lines 428-430).

33. *L344-345: Are mean absolute error of 0.65 for amplitudes of 1.06 "reasonably well" reproduced? Please justify, or rephrase.*

(Similar to a comment by reviewer #1 and your comment above). We both removed interpretative statements from the results section, and now provide some context: Amplitude mean errors are skewed by large values/outliers; the median absolute error, for example, is 0.41m. Adapted lines 438-440.

34. *Sect 3.2: Suggest to move to relevant methods sections, ref previous comments, and delete repetitions.*

Again, thanks for the suggestion. As part of the larger restructuring of the manuscript, the previous section 3.2 was removed, and most content from that moved to Appendix A. Only most relevant content was added to section 2.3, along with moving old Figure 2 to section 2.3; now Figure 1.

35. *Figure 2: Move to methods, and please consider more contrast for points in both a and c to clearly separate Q from SM and shallow from deep. Please provide threshold between shallow and deep in figure caption*

As suggested, we increased the colour contrasts between Q and SM stations as well as shallow and deep wells. Also, the old Figure 2 referred to here was moved to section 2.3 as part of the larger restructuring of the manuscript and now is Figure 1. Modified Figure 1 as suggested.

36. *L389 and Fig 3: Example time series: please state the reasons for your choice of these stations, and whether these are near best-cases, medium cases, or other. Can there locations be shown on a map? Please state why you have included this example figure, as this is currently unclear.*

We wanted to give the readers an impression of some actual time series. The examples were chosen to be somewhat representative of overall performance, which can also be seen by comparing the time series r values stated in the plot to the distribution in Fig. 4. Line 484-487: Added justification for Fig. 3.

37. *Fig 3: figure caption colour description opposite of figure, and "vertical" should be replaced with "horizontal"*

Thanks for noting this mistake. Corrected Fig. 3 caption accordingly.

38. *Fig 4: If the correlation of the entire period and the dry periods are not comparable, why combine them in the figure?*

It is still the same metric, just calculated across a different part of the time series. Hence, we consider it justifiable to remain in the same figure (together with the fact that we now better explain the effect of truncating on Pearson correlation coefficients in lines 502-505, as response to a comment by reviewer #1).

39. *Sect 3.3 would benefit from a geographical visualization (i.e. map) or regional summary of the results, to see if the performance values depend on region, to better understand potential reasons for the better and worse results. In Fig 7 (and related text), you underline the different processes and drought signal in different regions in Denmark, however, you currently have no evaluation metrics*

shown either on a map or for regional averages. Knowing where the model can be more or less trusted is beneficial when using the model operationally, and helpful for further model development.

This is a valid point, thanks. We modified Figure 4 to not only show the cumulative distribution of correlation coefficients, but in a new panel also show the respective correlation coefficients on a map. Changed Figure 4: Included map showing performance (r) at observation points across Denmark.

40. *L413: move interpretations of results as "good" etc. to discussion.*

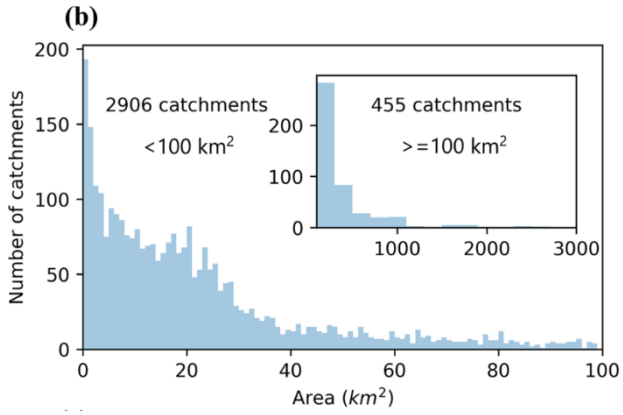
As in line with similar comments, we removed these statements from the Results section (here: line 515, but changes throughout the manuscript)

41. *L426: define "hydrological drought index", preferably in methods (this is the first time this term is used)*

We define them now in section 2.4. Line 295-296: "The other indices, covering different compartments of the hydrological cycle, are referred to as hydrological indices."

42. *Sect 3.4: Is the matching of SPI and streamflow done for the overlapping grid cell or for precipitation over the entire catchment of the streamflow station? If the latter, please state in the methods. If the former, please justify in the methods, as one would expect the entire catchment precipitation (deficit) to affect streamflow. If not using catchment precipitation as the basis, this choice should also be discussed in the discussion section, and how this choice may have affected the results.*

We use an overlap of streamflow point and grid cell. This is justifiable as (i) precipitation and evapotranspiration are fairly homogeneous across Denmark, especially at monthly timescale; (ii) precipitation and evapotranspiration data anyway is coarse in resolution with each grid covering 100km^2 or 400km^2, respectively; (iii) catchment sizes in Denmark are typically small – see below (from Figure 1 in https://essd.copernicus.org/articles/17/1551/2025/)



Line 375-377: We added an explanation about the simple overlap of SPI grid and q-point, but want to refrain from a more lengthy justification there, as justified above.

43. *Fig 5:*
    - *Please clarify in caption that also deep wells have gray colour for non-signifcant correlations in (g) and (h).*
    - *Not clear if deep or shallow wells when there are blue outlines on points representing more than one well in (i).*

- *Please comment on the outliers (>20 months for SDI and >50months for SGDI) in the main text. What can explain them?*

Figure 5 caption is changed for clarification, as well as marking of multiple (max. number per point is two) SGDIdeep wells per point is implemented.
Line 687-688: Added a sentence that some of the outliers potentially can be explained by the peak accumulation correlation being weakly defined.

44. *L443: please consider replacing "variability" with "accumulation period" or similar.*

Thanks for spotting that mistake. Line 557: Replaced with "accumulation periods"

45. *L465-466 vs L493-494: How does similar results for SPI and SPEI align with the statement that "SPI and SPEI values are largely uncorrelated in time"?*

This was a misunderstanding. In the latter statement, we meant rather "little auto-correlation in time". Line 625-626: Clarified.

46. *L493-494: provide the correlation value underlying this statement ("largely uncorrelated")*

This statement was changed to "show little auto-correlation" – see comment above. The AR1 for both monthly SPI and SPEI is 0.02; however, we want to avoid stating this number for the sake of brevity.

47. *Fig 9 and related text: Please state why you chose to show a case study and why you chose this event (and not e.g. a more severe event from the events seen in fig7-8). The choice of SPI2 is also unclear.*

(also responding to a similar comment by reviewer #1) We chose May 2020 because it is a good example of different accumulation periods (of SPI and SPEI, i.e. meteorological condition) controlling different hydrological compartments: Fast reacting soil moisture (ESSMI) is in drought in May 2020, as the previous 2 months – shown as SPI2, which is the dominating accumulation period for ESSMI; compare Fig. 6 – have been dry. When looking back an entire year – shown by SPI12 – that period has been wet. Which is reflected in groundwater (SGDI) and also streamflow (SDI) showing clear wet anomalies. In the revised version, we made this more clear (lines 639-647)

48. *L532: "very good results" – I advice to be more modest in the interpretation of the performance results unless you have reference numbers or benchmarking to compare with. This applies throughout the manuscript.*

Line 683: removed "very" etc (and more changes throughout the manuscript)

49. *L532-533: Please state/or refer to place where number underlying this statement can be found.*

Line 683-684: We removed the statement, as this indeed was not the focus of the manuscript.

50. *L553: Please add reference or justify your statement that the groundwater sensitivity to winter drought is often overlooked*

We wrote this coming across statements "drought is a summer phenomenon" or similar. However, in literature surrounding groundwater and recharge, the picture is more diverse. Hence, in the revised version, we removed the term 'often overlooked' (line 708)

51. *L562-563: Underline your statement that there is significant spatial variability in acc.periods for streamflow with a p-value or similar. According to L446-447, the acc.periods are not well defined and hence the spatial variability is not necessarily significant. Please also discuss here that the spatial variability acc.periods for observed streamflow is much lower (according to fig 5(f)).*

This was a misuse / imprecise use of the term 'significant'. We removed it, and clarified that the spatial variability is "represented to some degree by the DK-model". Lines 692-693.

52. *L606: Ref major comment a), "extreme conditions" are not evaluated in this study. Please rephrase to be in line with what is actually evaluated or can be implied based on that.*

"Extreme conditions" is in our understanding a generic term that can refer to droughts or floods (or other conditions) – but not necessarily to extreme droughts with index value < -2. Besides, we now added a drought detection performance evaluation, including on extreme droughts (Table 3). Hence, we left the statement unchanged.

53. *L648: Does conventional hydrological model refer to a specific model (e.g. the DK-model)? Please clarify.*

Line 849: Replaced 'conventional' with 'physically-based'

54. *L615: please specify the plans for CRN sensors, e.g. approx. number, locations etc.*

Currently, some of the co-authors and colleagues are working on establishing a Denmark-wide soil moisture monitoring network. It will include 10+ CRN sensors, situated to cover different land use and soil types across Denmark. Added "additional CRN sensors throughout Denmark as part of an upcoming soil moisture network (10+ stations across different land use and soil types)" to line 817-818.

55. *L672: "drought occurrence" not evaluated.*

In the revised version of the manuscript, also drought occurrence is evaluated. Statement unchanged.

56. *Data availability section should include meteorological data, as well as the soil moisture, streamflow and groundwater data used for calibration and evaluation of the model.*

Lines 964-970: We extended the data availability section accordingly, with references to the model setup, meteorological data and observational data.