Reviewer 1

R: The manuscript is very well written, the topic is perfectly suited for HESS and the methodology and scientific results are very good (very good is missing in the rating). Figures' quality has to be improved.

I suggest therefore revision.

A: We thank the reviewer for the thorough and constructive assessment of our manuscript. The comments have been highly valuable in highlighting aspects of the model setup, process representation and parameter choice that require further clarification or refinement. Several points raised directly help us improve the transparency and robustness of our methodological approach.

Due to significant concerns regarding the methodological framework and the model setup, we critically re-examined the model structure, including discussions with the developers of the LWFBrook90 model and implemented adjustments to several flux representations. These modifications led to partly substantial changes in the results, which have been fully incorporated into the revised version of the manuscript.

In the following responses, we address each comment in detail and indicate where revisions or additional adjustments to the model setup or the manuscript will be made. We appreciate the reviewer's critical input and believe that the resulting revisions will substantially strengthen the scientific quality and clarity of the study.

Following additional comments could be provided:

R: More information about the SVAT model is necessary to better understand the underlying processes and assumptions for quickflow fraction, multiplier to partially activate drainage, root water uptake, interception, are transpiration and evaporation uptake handled in a different way?, run-off,

A: We agree that a more detailed description of the underlying processes and assumptions will improve the transparency of our methodology. We will also clarify how surface runoff and other relevant model processes are represented. We revised the manuscript accordingly and added an expanded model description in the supplement to address these points and enhance the overall clarity of the paper.

R: The 100,000 random combinations are set without considering correlation between parameters, which is not really true for soils of similar texture. Please discuss.

A: It is correct that the random generation of 100,000 parameter combinations did not explicitly account for known correlations between the parameters, which are typically constrained by soil texture and structure.

In our study, however, the sampling was performed on the hydraulic model parameters (θ_s, θ_r, α, n, K_s), which represent empirical degrees of freedom in the van Genuchten–Mualem formulation rather than directly measured texture fractions. The well-known pedogenic correlations between texture variables therefore do not translate directly into the hydraulic parameter space, where such relationships are typically weaker, nonlinear or site-specific.

For this reason, we used independent random sampling to explore the parameter space. This approach is widely applied in hydrological uncertainty analysis and is considered appropriate when the objective is to characterise the range of possible system responses rather than to generate parameter sets that strictly represent natural soil populations (e.g., Beven, 2006a; Vrugt et al., 2008). Moreover, parameter selection in our study did not rely solely on this sampling step but was additionally informed by published BROOK90 documentation, site-specific measurements and expert judgement, ensuring that key processes were represented realistically.

We acknowledge that correlation-preserving sampling strategies (e.g., correlated Latin Hypercube sampling or copula-based approaches) could further constrain the parameter space and may be valuable for future studies aiming to generate parameter sets that more closely reflect natural soil variability. We have added a corresponding note to the manuscript.

R: How was the warm-up period of three months estimated? (may depend on climatic conditions).

A: Due to the relatively short data series of less than two years, we decided to use only the period from the beginning of 2023 until the start of data collection in March 2023 as a warm-up period. This is admitteldy a too short; normally, the warm-up period should be at least six months long. For the revision of the paper, we performed the calibration with extended warm-up period of 15 months. An even longer period would be desirable to capture a broader range of hydrological conditions, including interannual variability in soil moisture, evapotranspiration and recharge dynamics to reduce more the influence of initial conditions and improve the robustness of parameter estimation, particularly in systems with strong seasonal memory.

R: Figure 5 clearly shows that there are some problems with the soil moisture measurements. These outliers may impact the KGE (everything is squared that gives a higher weight to these singularities). Did you remove these ‘errors’?

A: We assume that by "problems with the soil moisture measurements" or "outliers" you're referring to the repeated, sudden spikes in soil moisture in some profiles. These are very high soil moisture readings due to waterlogging following heavy rainfall events caused by poorly permeable soil layers. These conditions are consistent with field observations, where we were also able to observe this phenomenon of stagnant water in some soil profiles. Therefore, we believe these are not erroneous measurements.

But we agree that outliers in soil moisture measurements affect the Kling-Gupta Efficiency (KGE), due to its squared error component. High overall KGE scores do not necessarily imply that all aspects of soil moisture dynamics are represented equally well. In particular, we acknowledge that the model shows limitations in reproducing sharp soil moisture peaks following intense precipitation events. So we added a discussion to the single components of KGE in terms of the interpretation of model performance and its impact on groundwater recharge, the values can be found in the supplement Table S6 to S8 and S10 to S12. We revised the manuscript to clarify this point and to explicitly acknowledge that the inability to reproduce preferential flow events may lead to an underestimation of recharge during such conditions.

R: The CV values of Tharandt plots is higher than the CV values of Kienhorst. I agree with the interpretation. However, because the soils are very different, Richards model and MVG models may be more suited for one site than for the other, leading to more variability of the parameter's values.

A: We agree that differences in soil characteristics between the two sites may influence how well the Richards equation and the van Genuchten–Mualem parameterisation represent local hydraulic behaviour. In structured, heterogeneous or macroporous soils such as those at Tharandt, the MVG formulation is known to provide a less constrained description of hydraulic properties, which can lead to a wider range of parameter combinations yielding acceptable model performance. This effect may contribute to the higher coefficient of variation observed for the Tharandt profiles, in addition to the genuine physical heterogeneity of the site.

At the same time, we note that the increased variability in Tharandt is consistent with independent field evidence of stronger spatial heterogeneity in soil structure and hydraulic behaviour. We therefore interpret the higher parameter variability as the combined result of true physical heterogeneity and reduced parameter identifiability under conditions where the MVG model is less representative. We have added a corresponding clarification to the manuscript.

R: The discussion about model calibrated parameters could be improved. It could be interesting to test if each parameter is statistically different from one soil profile to the other for the same site, and if some correlation can be highlighted (usually, tetaS and tetaR are correlated, due to MVG model formulation where the 'driving' parameter is TetaS-TetaR).

A: We agree that a deeper analysis of the calibrated parameter sets can provide valuable insights into soil profile heterogeneity and parameter interdependence. As suggested, we can assess whether calibrated parameter values differ significantly between the soil profiles at each site by using statistical tests and visualization through correlation matrices and parameter scatter plots. These additions would strengthen the interpretation of the calibrated parameters and help to better contextualize differences between soil profiles and the structural links between hydraulic parameters.

However, conducting this analysis in a statistically robust manner would require a substantial methodological extension beyond the scope of the present study. Our primary objective was to evaluate groundwater-recharge dynamics and uncertainty at the plot scale, rather than to perform an exhaustive examination of parameter identifiability and inter-parameter correlations. We therefore consider this a valuable direction for future work, particularly in studies explicitly focused on parameter behaviour, model structure evaluation or pedotransfer-based parameter constraints.

R: The discussion about the groundwater recharge is interesting. Of course, doing some statistics with only 11 profiles can be discussed but, considering the huge amount of work, the results can be considered has a first order quantification (6 profiles minimum among 11). It also questioned the definition of mean values for a highly non-linear model as Richards model. In my opinion, the scaling infiltration at the size of a plot is very difficult and may depend on forcing terms (raining period or dry period). This is nicely shown in the manuscript.

A: We agree that statistical inference based on 11 soil profiles is inherently limited, yet we consider the resulting estimates to provide a meaningful first-order quantification of spatial variability, particularly given the substantial effort required to obtain profile-resolved measurements.

We also acknowledge the non-linearity of the Richards model, which complicates the interpretation of mean values. In highly heterogeneous soils, averaging parameters or outputs may obscure critical dynamics, especially during episodic forcing events such as heavy rainfall or drought. Our intention was not to derive statistically robust spatial averages, but to provide a first-order quantification of recharge variability across representative soil profiles.

You are also right, that scaling infiltration and recharge processes to the plot scale is challenging and strongly dependent on the prevailing hydrometeorological conditions, such as the occurrence of wetting events or extended dry periods. This sensitivity is reflected in our results and further underscores the importance of profile-resolved modelling for understanding the range of possible system responses. We have added a short discussion of transferibility to large-scale in chapter “4.5 Transferability and Generalizability of the Findings”.

R: The unit of ksat in figure S3 is unclear.

A: We changed this in the revised manuscript

Reviewer 2

R: Review of “Effects of spatial soil moisture variability in forest plots on simulated groundwater recharge estimates”

The manuscript investigates how the number of soil moisture measurement plots affects the parameterisation, calibration, and estimation of soil moisture and subsequently groundwater recharge (GWRCH) using the LWF-BROOK90 model, applied to two pine and spruce forest sites in Germany. The manuscript is generally well written and clearly structured. The topic itself is relevant and fits well within the scope of HESS.

However, I have substantial concerns regarding the methodological setup and the scientific and practical relevance of the presented results. Nevertheless, the study has the potential to make a meaningful contribution to both the BROOK90 modelling community and forest hydrology research. Therefore, I recommend major revisions.

A: We would like to thank for the reviewer’s detailed, constructive, and scientifically insightful comments. The feedback has been extremely valuable in helping us to identify aspects of the manuscript where clarification, additional analyses, or methodological refinement are warranted.

Due to significant concerns regarding the methodological framework and the model setup, we critically re-examined the model structure, including discussions with the developers of the LWFBrook90 model and implemented adjustments to several flux representations. These modifications led to partly substantial changes in the results, which have been fully incorporated into the revised version of the manuscript.

In particular, the reviewer’s remarks concerning model parameterization, preferential flow processes, groundwater recharge dynamics and the representation of site-specific hydrological conditions have provided important guidance for strengthening both the conceptual framing and the technical robustness of the study.

In the following reply, we address each comment systematically and outline the revisions and the additional steps we will undertake to improve the model setup and the interpretation of the results.

Where the reviewer has pointed to methodological limitations, we acknowledge these explicitly and describe how they will be addressed either through revised analyses or through a clearer discussion of uncertainties and assumptions.

General comments:

Model setup and description

R: The model setup section needs further elaboration. Please provide more information about the LWF-BROOK90 model itself, not only its soil moisture component.

A: We agree that a more detailed description of the underlying processes and assumptions will improve the transparency of our methodology. We revised the manuscript accordingly, expanded the model description and added an expanded model description in the supplement to address these points and enhance the overall clarity of the paper.

R: It is unclear which model output was used to represent groundwater recharge (GWRCH). I assume it is the drainage from the lowest soil layer rather than outflow from the conceptual subsoil bucket with delay, but this should be explicitly stated and justified.

A: In our study, groundwater recharge was defined as the drainage flux from the lowest soil layer, without applying an additional conceptual delay or storage bucket. Hence, it is rather potential or latent recharge. Analysing this potential recharge allows for a direct comparison across soil profiles with varying hydraulic properties and the study focuses on relative recharge dynamics rather than absolute aquifer recharge timing. We explain this point in the added expanded model description.

R: It should be discussed that the model can only be used to estimate GWRCH under several restrictive assumptions: it is a 1D model with no explicit groundwater module, no lateral flow/routing, and no upward capillary flux representation.

A: We fully acknowledge that LWF-BROOK90 is a one-dimensional SVAT model without an explicit groundwater module. As such, it does not simulate groundwater table dynamics, lateral flow or routing. Potential groundwater recharge in this context is defined as vertical drainage beyond the root zone, and should be interpreted as a first-order estimate of potential recharge, rather than actual aquifer replenishment.

We discussed this in the added expanded model description as well as in the manuscript when discussing calibration results.

R: Please provide sources for all parameter values and parameter ranges used in the calibration. How the soil hydraulic parameters were derived for each layer (which PTF, etc.)?

A: We added corresponding literature sources or data bases for all fixed and calibrated parameters to table S1 and S2 in the supplement.

Soil hydraulic parameters (, θs, θr, α, n and Ks) were derived using the pedotransfer function of Wösten et al. (1999) based on measured texture fractions and bulk density for each layer. These PTF-derived values served as initial estimates for the reference simulation.

R: I was surprised not to find the maximum LAI parameter or its inclusion in the sensitivity analysis. Together with glmax, this is one of the most sensitive parameters in BROOK90. Further, the method used to simulate seasonal LAI dynamics (phenological scheme and parameters) is not described, although it strongly affects evapotranspiration and thus water availability for percolation.

A: Seasonal LAI dynamics were derived from site-specific phenological observations provided by the ICOS managers, ensuring that the temporal evolution of canopy development was represented realistically for each plot. The seasonal LAI dynamics are contained in the input file "Data-NEW_meteoVeg.csv", where the daily relative LAI value in percent is specified. For the calculation of the relative LAI values, we used a maximum LAI of 4 for the Kienhorst (pine) location and 5 for the Tharandt (spruce) location Table S1 and S2 in the supplement

R: Information on meteorological forcing data (source, temporal resolution, spatial representativeness) is missing. If daily data were used together with standard DURATN values, this may strongly affect infiltration and percolation processes.

A: We will provided the missing data in the revised manuscript,.

We have generated daily data from 30 minutes of data; this data is contained in the input file "Data-New_meteoVeg.csv".

The average number of hours per precipitation event for all 12 months is defined in the file „Data-NEW_meteo_storm_durations.csv“, we have used the value 4 here, which is described as a „satisfactory approximation“ according to the Brook90 documentation (Federer 2002).

Soil moisture data and calibration

R: I appreciate the substantial effort required to collect such a large soil moisture dataset. However, its application within the current modelling setup—given the research questions—raises concerns.

You calibrated the model using the entire soil moisture dataset (20 months), leaving no data for validation. While validation may not be the core focus, it is good (and standard) scientific practice and would strengthen the credibility of your modelling results.

A: You are right that independent validation is important to check the generalizability of the model. For the first version of the manuscript, the entire data set was used for calibration because the amount of data was limited (data series of less than two years, march 2023 to november 2024). As our measurement compaign contibued during the revisions of this paper we were able to collect more data. So we had an extended observation period available and we performed a split-sample validation (use of 70% of data for calibration and 30% for validation). We will also extend the warm-up period from 3 to 15 months to make sure the model is not affected by estimated initial conditions.

R: Given the large number of calibration runs, I am concerned about potential model overfitting, which may reduce model performance outside the calibration period.

A: In our opinion, any risk of overfitting is primarily related to the number of calibration parameters rather than the number of calibration runs, which are intended to explore the parameter space. The large number of runs can only reveal the problem.

To mitigate this risk, we substantially reduced the number of calibration parameters in the revised version and constrained their plausible ranges, thereby limiting the potential for parameter compensation and overfitting.

The remaining uncertainty in model predictions was assessed using a split-sample validation test. These approache allowed us to quantify the robustness of the model and the influence of the remaining degrees of freedom, ensuring that predictions remain reliable beyond the calibration dataset.

We added a discussion of these steps in the revised manuscript to clarify how overfitting is minimized and how residual uncertainty is addressed.

R: Groundwater recharge is mostly generated during winter, yet only one winter season (20232024) is included. This raises questions regarding the robustness of GWRCH estimates. If sensors are still operating, I strongly recommend including additional data and performing a split-sample calibration/validation.

A: It is correct that the dataset of the first version of the manuscript includes only one full winter season (2023/2024), which limits the temporal representativeness of the groundwater recharge estimates.

As mentioned before, we included additional available measurement data from 2025 during the revision of the paper. So we are able to determine the groundwater recharge for an additional winter season. This will help ensure that the recharge estimates are not biased by single-year anomalies and reflect broader climatic variability.

R: Please discuss how soil moisture measurement uncertainty (±3%) affects calibration results, especially during dry periods where simulated SM values often fall below 5%—resulting in possible relative errors of up to 100%.

A: This is indeed an important aspect for the discussion. The uncertainty becomes particularly relevant under very dry conditions (< 5 vol.-%), where the relative influence of sensor noise is high and the numerical dominance of dry states in the calibration dataset may bias parameter optimisation. Under such circumstances, there is a risk that a calibration based on the full time series favours parameter sets that reproduce the prolonged dry periods well, while underrepresenting the wetter conditions that are more relevant for groundwater-recharge generation.

We have addressed this issue in the revised version of the manuscript. To evaluate the potential impact of this bias more systematically, future work should compare calibrations based on the full time series with calibrations restricted to periods with θ > 10 vol.-%, thereby allowing an assessment of how strongly the representation of dry conditions influences recharge estimates. While we acknowledge the value of the proposed approach, we did not implement it in the present study, as it would have required substantial methodological adjustments beyond the scope of the current revision.

Missing reference simulation

R: I recommend including uncalibrated/reference simulations (using measured vegetation parameters and Mualem–van Genuchten parameters obtained from soil profile data via PTFs) for all 22 profiles. This would help to evaluate the added value of calibration for both SM and GWRCH.

A: We agree that a comparison between calibrated and uncalibrated models provides valuable insights into the added value of calibration and assess how well the model performs with default or literature-based inputs, identifys which profiles or parameters benefit most from calibration.

In the revised manuscript, we run a reference simulation with measured and standard literature values before we start the calibration. So we could evaluate the added value of calibration.

Transferability and scientific significance

R: The most critical point of the study is that the experimental setup appears highly site-specific and dependent on data availability and chosen parameters. As currently presented, the results have very limited transferability and limited practical/scientific added value. This should be discussed more clearly with respect to the intended audience (modellers, foresters, soil scientists, hydrologists). Key results should either be presented with appropriate caution or supported more robustly, especially in terms of why such a complex and expensive setup is necessary for plot-scale GWRCH estimation.

A: We acknowledge that the experimental setup is highly site-specific, relying on detailed soil profile characterization, vegetation data and continuous soil moisture monitoring what limits the direct transferability of the results on other sites. However, the study was designed to explore the feasibility and sensitivity of plot-scale Groundwater recharge estimation under realistic field conditions. The complexity of the setup reflects the challenges of capturing recharge dynamics in structured, heterogeneous forest soils, where simplified approaches may fail to represent key processes.

In the revised version of the manuscript, we added a chapter to the transferability of the results, where we discussed limitations of generalizability, audience relevance and justification of your experimental design and modeling approach.

Figures

R: The quality (resolution) of all figures is currently insufficient, making them difficult to interpret.

A: This could be due to the upload process; we checked that and improved the quality where required.

Scope and title

R: A substantial portion of the manuscript deals with soil moisture simulation and calibration, whereas groundwater recharge is adressed only briefly. You may wish to reconsider the title or rebalance the manuscript content.

A: The aim was to highlight the differences in groundwater recharge, based on calibration. But you are right that substantial portion of the manuscript deals with soil moisture simulation and calibration, so we added some words to the title as follows:

“Effects of spatial soil moisture variability in forest plots on model parameterization and simulated groundwater recharge estimates”.

Specific comments

R: L52: The motivation for the importance of GWRCH needs to be further elaborated. As presented in the first paragraph, it remains vague.

A: We revised that part and add some facts to the motivation

R: L76–77: This represents only one of many possible approaches.

A: We revised that part and add some facts to other approaches for reducing parameter uncertainty such as sensitivity analyses, ensemble simulations, Bayesian calibration frameworks

R: L79–81: This is quite a bold statement and is not universally true (e.g., cases of over-calibration without validation).

A: We revised that part and add some facts to the risks of no independent validation or if the calibration is too heavily tailored to a single data set

R: L101: Why are plot-scale GWRCH estimations needed rather than multi-site or gridded estimations, which are typically required for practical applications by stakeholders?

A: We added a discussion in the revised manuscript (Chapter: 4.5 Transferability and Generalizability of the Findings) emphasizing that plot-scale measurements serve both to quantify first-order recharge variability and to support the development and validation of multi-site or gridded modeling approaches, bridging the gap between detailed process studies and practical large-scale applications.

R: L115: Please add the Latin names of species, full soil texture classifications, and the averaging period for the meteorological data. Additionally, include soil profile data in the Appendix to assess heterogeneity.

A: We added this information to the revised manuscript

R: L135–136: If stagnant conditions at this site are well known due to shallow bedrock, what GWRCH can realistically be expected, and how is this addressed in the model setup?

A: Indeed, the presence of redoximorphic features and perching horizons on the plot area indicates locally restricted vertical percolation due to low subsoil permeability which reduces percolation to deeper groundwater. But the absence of a permanent water table also suggests that this is for limited times.

To address this limitation more explicitly, we revised the model setup. In particular, we improved the representation of near-surface storage dynamics and vertical percolation by adjusting soil hydraulic parameters and refining the layering scheme. Additionally, we implemented bypas flow to better represent flow conditions.

R: L144: Do you mean ±3%?

A: Yes, we changed it.

R: L149–150: Please elaborate on the calibration process in more detail.

A: The factory calibration is based on laboratory measurements in reference media (typically quartz sand and deionized water) across a range of volumetric water contents and temperatures. The resulting calibration curves are stored internally in the sensor electronics and automatically applied to the raw TDT signal, converting it into volumetric water content and temperature values. In this study, we relied on this manufacturer calibration rather than performing site-specific calibration, because our focus was on relative soil moisture dynamics and temporal patterns rather than absolute water content values. Previous studies have shown that the factory calibration of SMT100 sensors provides sufficiently robust results for monitoring soil moisture dynamics in a variety of soils (Sprenger et al., 2015; Demand et al., 2019).

We acknowledge that site-specific calibration (e.g., by gravimetric sampling or soil-specific calibration functions) could reduce systematic bias in absolute values, especially in soils with atypical texture or bulk density. However, given the study objectives, the factory calibration was considered adequate.

We added an short paragraph to this topic.

R: L165: Add the north direction and satellite/aerial background imagery to better assess spatial heterogeneity.

A: We revised the figures, however without including satellite/aerial background imagery due to the poor visibility of details.

R: L193: It is unclear where the parameter calibration ranges come from. For example, cvpd values below 1–1.5 result in unreliable transpiration and should be justified. The same applies to other sensitive parameters as well.

A: Thank you for pointing out the recommended range for the parameter cvpd. We acknowledge that values between 1.5 and 2.5 kPa are generally considered physiologically plausible thresholds, as transpiration becomes unreliable below this range due to stomatal regulation. In our initial calibration we tested a slightly broader interval to explore model sensitivity and potential site-specific adaptations.

Following your suggestion, we restricted the calibration range to 1.5–2.5 kPa in the revised version of the manuscript.

We also checked the value ranges of the other parameters during the revision of the manuscript and adjusted them where necessary.

R: L195: Please include a sensitivity analysis in the Appendix and specify its boundary conditions and variables of interest.

A: A preliminary sensitivity assessment was performed earlier in the context of a Master's thesis, which provided useful guidance for designing the current model setup. However, the underlying assumptions, implementation details and documentation of that work were not developed to a standard suitable for publication and several methodological aspects would require substantial revision before they could be transparently presented in the Appendix. For this reason, we decided not to include the analysis in the manuscript.

Importantly, the sensitivity assessment was only one of several sources informing the selection of parameters and calibration ranges. Parameter choice was additionally guided by published BROOK90 documentation, site-specific measurements and expert judgement based on the physical characteristics of the study area. These complementary lines of evidence ensured that key processes were adequately represented, even without presenting the preliminary sensitivity analysis. We therefore believe that the current parameterisation strategy is well justified, while a fully documented and reproducible sensitivity analysis remains an important task for future work.

R: L203: A warm-up period of only 3 months is, in my experience, insufficient for this model setup. A minimum of 6 months (preferably 1 year) is recommended. Please either extend the warm-up period or demonstrate that a longer warm-up does not significantly affect soil moisture at the start of spring 2023.

A: We agree that the 3 months warm up period is a bit too short. As mentioned before, we extended the warm-up period to 15 months during the revision of the paper.

R: L220: This setup violates one of the core assumptions of the KS test—independence of samples—since the bootstrapped subset contains data from the full set. You should choose a different metric or modify the setup.

A: We agree that the classical two-sample KS test assumes independent samples, which is violated when comparing a bootstrap subset to the full data set containing that subset. In the revised manuscript, we therefore no longer use the KS statistic, instead we performed a Monte Carlo subsampling to identify the number of soil profiles required for a representative estimation of groundwater recharge.

R: L230: How were raw observations post-processed? Were sensor issues, spikes, or errors encountered? (e.g., see Kienhorst L1–10 cm, red straight line after 23 Sept.)

A: We did only very little post-processing of the raw data; the graphs show more or less the original measurements. We only performed a visual inspection and physical plausibility checks of the soil moisture time series, so we could identify sudden changes that cannot be explained by known rainfall or limited percolation or values that are not within physically meaningful limits.

We added a short comment to that in the figure caption

R: L245: The spikes observed in Tharandt—do they indicate sensor errors or preferential/bypass flow in some plots?

A: These spikes or very high soil moisture readings are due to waterlogging following heavy rainfall events caused by poorly permeable soil layers. These conditions are consistent with field observations, where we were also able to observe this phenomenon of stagnant water in some soil profiles. Therefore, we believe these are not erroneous measurements.

R: L260 (270): What exactly is meant by "simulations" here—an average of the 30 'best' simulations and their uncertainty bandwidth?

A: Yes, it is an average of the 30 'best' simulations and their uncertainty bandwidth.

We added explanation to that in the figure caption.

R: L263: There appears to be a substantial underestimation of variance. Please provide the full KGE decomposition to better assess this and discuss it, as it is crucial for estimating downward water flows and, consequently, GWRCH.

A: We agree that variance underestimation is a relevant aspect when evaluating downward water fluxes and groundwater recharge. So we added a discussion to the single components of KGE in terms of the interpretation of model performance and its impact on groundwater recharge, the values can be found in the supplement Table S6 to S8 and S10 to S12.

R: L289: To my knowledge, ksnvp is one of the least sensitive parameters in the BROOK90 model, as snow evaporation rarely exceeds 1% of annual precipitation. It is therefore unclear why it is included in the calibration list.

A: It is correct that ksnvp is generally considered as a low-sensitivity parameter, particularly in temperate regions where snow evaporation contributes minimally to the overall water balance.

In our initial calibration list, ksnvp was included for completeness, as part of a broader vegetation parameter set. However, we acknowledge that its inclusion may not be justified given the limited influence on simulated outputs under the climatic conditions of our study sites.

Therefore, we excludeksnvp in the new calibration during the revision oft he manuscript and focus on parameters with greater hydrological relevance.

R: L300: It is unclear whether the CV values represent summarized CV values for all parameters per site. Please explain this in more detail.

A: We computed the CV for each parameter (ksat, alpha, n, etc.) across the ensemble of accepted parameter sets and summarized these values as a site-level average. This provides an indication of parameter uncertainty and identifiability at each location.

We revised the methods section to clarify this calculation and added a note in the results to explain the interpretation of CV values in the context of model sensitivity and calibration robustness.

R: L301: The increased variation reflects heterogeneity in calibrated soil parameters within the model setup, not actual soil properties (unless laboratory verification was performed).

A: We agree that part of the observed variation may indeed originate from the calibration procedure and the parameter ranges applied, rather than representing fully verified soil properties. Our intention in presenting these CV values was to highlight the relative variability within the model setup, which reflects both the heterogeneity of the soils and the uncertainty inherent in parameter estimation.

To clarify this point, we revised the manuscript to explicitly state that the reported CVs represent variability in calibrated model parameters, which may differ from true soil heterogeneity.

R: L325–326: Please discuss these findings in the context of preferential flow (bypass flow) and how it is parameterized in the model.

A: As part of the revision, we adjusted the model setup to better represent the flows from precipitation to groundwater recharge. Among other things, this includes reviewing and modifying the previous assumptions regarding bypass flow. For the Tharandt site, we activated the bypass-flow routine, whereas it was deactivated for the Kienhorst site. This configuration ensures that water fluxes are represented in a manner consistent with the respective site conditions.

The influence of bypass flow on model performance and groundwater-recharge at the Tharandt plot is discussed in detail in the Results section."

R: L334–340: The cumulative GWRCH values seem much higher than expected for the studied sites. Please include a table or plot in the Appendix showing the monthly sums of all water balance components (as you are not covering full years). Also, relate GWRCH to seasonal/annual precipitation as a percentage.

A: We acknowledge that the reported groundwater recharge values may appear elevated without contextual information on the full water balance. To address this, we added a chapter and tables S15 and S16 in the supplement showing the annual sums of all major water balance components for each plot and year.

Moreover, we compared the calculated groundwater recharge rates with existing data for the sites to check their plausibility.

Furthermore, we also will express groundwater recharge as a percentage of seasonal/annual precipitation, which allows for a clearer assessment of recharge efficiency and hydrological plausibility.

R: L368–386: You do not demonstrate that the calibrated model reliably estimates all water balance components—only soil moisture at certain depths.

A: We agree that our calibration and validation approach was limited to soil moisture observations at specific depths, and therefore does not allow for a direct verification of all individual water balance components. Our intention was to assess the model's ability to reproduce soil moisture dynamics as a proxy for the underlying hydrological processes, rather than to provide a full closure of the water balance. We acknowledge this limitation and clarified it in the revised manuscript by changing the formulation.

R: L396–398: I disagree with this statement. It must be supported with model evidence. In many profiles, the model fails to reproduce high soil moisture peaks (likely from intense precipitation events). These peaks may represent preferential flow, which could be a major source of GWRCH (e.g. for Tharandt). Failure to simulate them may lead to significant underestimation of GWRCH during such events.

A: You are correct that high overall KGE scores do not necessarily imply that all aspects of soil moisture dynamics are represented equally well. In particular, we acknowledge that the model shows limitations in reproducing sharp soil moisture peaks following intense precipitation events. These peaks may indeed be linked to preferential flow processes, which are not explicitly represented in the current model structure and could contribute to groundwater recharge, especially at the Tharandt site.

We revised the manuscript to clarify this point and to explicitly acknowledge that the inability to reproduce preferential flow events may lead to an underestimation of recharge during such conditions.

R: L419–420: Given the model setup, emphasis on soil moisture, and limited vegetation parameters, the results are unsurprising.

A: ok

R: L427–429: This could become a valuable result if you analyze and discuss it in terms of the parameter list, parameter ranges, and degrees of freedom allowed during calibration. Would the conclusions change if different parameters or wider/narrower ranges were used?

A: The finding that model-based uncertainty exceeds plot-based variability could provide valuable insights into the sensitivity of the calibration setup. We acknowledge that different choices of parameters or wider/narrower ranges could potentially alter the relative contributions of model-based versus plot-based variability. However, a systematic exploration of alternative calibration setups was beyond the scope of this study.

We revised the manuscript to explicitly highlight this limitation and to discuss that the conclusions drawn here apply within the chosen calibration framework.

We agree that future work could benefit from a more detailed sensitivity analysis of parameter ranges and degrees of freedom, which would allow a deeper understanding of how calibration decisions influence uncertainty and variability in recharge estimates.

R: L464–466: The referenced GWRCH values are not reliable. I could not find any GWRCH data for Tharandt in Goldberg and Bernhofer (2007), and the Kienhorst values are taken from ArcEGMO simulations, which are often known to underestimate ET and therefore overestimate water available for GWRCH.

A: We checked that during the revision of the manuscript and added the right sources