Response

We thank the Editor and Dr. Anneli Guthke for their positive assessment of our work and for the constructive comments, which will help us improve the manuscript. Below we provide a point-by-point response to the reviewer comments. Reviewer comments are reported in black, while our responses are shown in blue.

We hope that the revised manuscript and responses satisfactorily address all comments and that the manuscript can now be considered for publication in GMD.

Yours sincerely,

Francesca Moschini, on behalf of all coauthors

Summary:

This study assesses a ubiquitous but often overlooked problem: hydrological models are built for a specific purpose (e.g., flood forecasting), and at some point "misused" for other tasks (e.g., drought prediction, water resources management), often without specific re-training and re-evaluation. Since any model is just a coarse abstraction of reality and suffers from model structural errors, compensation for model error happens within the allowed parameter ranges, and unphysical behavior can emerge across compartments, processes and variables. So if trained for streamflow only, a hydrological model might perform poorly on other components of the water balance, and this is the target of the presented analysis in this manuscript. The authors investigate different model setups of a specific distributed model, LISFLOOD, on the Po River Basin, with respect to streamflow prediction performance, but also through diagnostic evaluation of other fluxes.

Overall evaluation:

The authors reveal interesting contradictions between performance, parameter estimation and water balance closure when training different versions of LISFLOOD. The manuscript is very well structured and a pleasure to read. While the conclusions of the study are supported by the findings, unfortunately, the manuscript left me somewhat "uninspired" – I had hoped for more insights. Yet, the findings are worthwhile reporting and the analysis itself is nicely done, so my recommendation is to still consider this manuscript for publication, albeit not the most forward-directed paper.

We thank the reviewer for the positive assessment of our manuscript and for the constructive review comments, which will help us strengthen the analysis.

We appreciate the reviewer's broader point regarding the level of insight and forward-looking implications of the study. In the revised manuscript, we will therefore strengthen some parts of the discussion, particularly concerning the implications for model calibration, parameter

constraints, and the use of diagnostic frameworks such as the Budyko analysis for evaluating hydrological model realism.

Specific comments:

Abstract: “Their realism in simulating water balance components is crucial for building trust across different use cases.” Thank you for this statement, this is so true but often overlooked. Glad to see this issue addressed explicitly in this study.

We thank the reviewer for this supportive remark. It encourages us that the motivation of the study resonates with the community, and we will preserve this framing in the revised version.

l. 4/5: “Therefore, alternative setups can benefit specific applications by improving the representation of relevant water balance components.” I know what you mean, but I feel this sentence is too dense. Please invest one or two sentences more to explain alternative setups and how they could eventually lead to improved representation of water balance.

Thanks for this comment, we have rephrased the sentence to give it more clarity. We will add the following considerations from line 4 in the revised manuscript.

“Alternative setups, such as the choice of input data, model structure, and calibration methods, can influence how water is partitioned across model states and fluxes. Modellers can adjust these choices to improve the representation of the water-balance components most relevant for a given application, even when this comes at some cost to overall streamflow performance.”

l. 17: “… in a flexible and target-driven way,…” This might almost be a philosophical question, but doesn’t that approach contradict your motivation to “build trust across different use cases”? I’ve long been debating with myself and others whether models should be built and tested goal-oriented (this was the word I used in the past, see Guthke (2017)) or open-purpose. Intuitively, a model should do the right things for the right reasons and combine internal realism with best performance, but empirically, we observe that this is not the case, and hence the (also justified) advice to optimize and evaluate the model on those aspects it will be applied for. In that sense, is there any problem with a flood forecasting system that predicts stream discharge really well, but struggles with evapotranspiration? Who cares (to play devil’s advocate here)? I’m curious to read more about the authors’ perspective on this dilemma in the manuscript. – Ok, from reading the introduction I understand that LISFLOOD is a specific case where a model is used beyond its intended purpose; this piece of information would be helpful to integrate into the abstract. To tell the story that (at latest) when a model is asked to perform other tasks (especially: water resources management) than it was trained on, it needs to be reevaluated and maybe retrained with a focus on a realistic water balance.

We thank the reviewer for this thoughtful comment, which engages directly with one of the core conceptual questions underlying the motivation of this study. We have addressed the specific

suggestions regarding the abstract first, and then we offer our perspective on the broader question raised (following which we will extend the discussion in the revised manuscript).

Regarding the abstract, we will integrate the reviewer's suggestion (from line 9), making explicit that the LISFLOOD model is used beyond its original intended purpose of flood forecasting:

"We present one such exercise in a representative European case study using a physically-based hydrological model (LISFLOOD), as calibrated and set up for the European Flood Awareness System (EFAS). Originally developed for flood forecasting, LISFLOOD is increasingly also employed for drought monitoring and water resources management."

Moreover, we revised the sentence mentioned by the reviewer (L. 17, "... in a flexible and target-driven way...") and extended the end of the abstract to highlight the broader implications of our work, as follows:

"We propose diagnostic criteria to support the evaluation of physically-based distributed models, across different applications, while preserving consistency in the representation of long-term water-balance components. Finally, we argue that such hydrological auditing becomes increasingly relevant as the added value of physically-based models lie in their ability to provide diagnostically useful and physically-consistent representations of internal water-balance processes."

Regarding the general perspective and discussion, we agree with the reviewer that when a model originally calibrated for one purpose (e.g., flood forecasting) is repurposed for others (e.g., water resources management), its realism in representing the full water balance must be re-evaluated, and the model may need to be re-calibrated with new targets beyond streamflow.

This connects directly to the philosophical question the reviewer rightfully raises: "is there any problem with a flood forecasting system that predicts discharge well but struggles with evapotranspiration?". In this work we argue that, in a physically-based model, certain internal fluxes and states (e.g., soil moisture, evapotranspiration, long-term storage) should remain within physically plausible bounds and not deviate too far from theoretical expectations, particularly in the long-term water balance, regardless of the calibration target. If they do not, the model is producing the right streamflow for the wrong reasons, and the modeller has no principled way of knowing whether the model will still be reliable under non-stationary conditions, in ungauged basins, or for any application outside the narrow calibration target. In the revised manuscript, we will extend the discussion on these points and on the broad implications of our work.

This study shows that the operational LISFLOOD setup tends to favour streamflow at the expense of other water-balance components. That finding raises a more provocative question, which we think the community should urgently engage with: why use a physically-based model at all if its internal fluxes and non-calibrated states are unreliable? With the rapid advances in deep-learning approaches in hydrology, this question becomes increasingly pressing; if a black-

box model can predict the target variable as well as or better than a physically-based one, the value of the physical model rests precisely on its internal realism. Audits like the one we present are, in our view, the way to defend (or challenge) that value.

Conversely, when a physically-based model is set up correctly, internal-realism diagnostics can be informative in their own right. A physically-based model run on a leaky catchment with lack of knowledge on its groundwater processes, or with poor-quality forcing, or with a missing process (e.g., unmodelled irrigation withdrawals), should fail the water-balance check, and that failure is itself useful information, pointing the modeller toward data quality issues, incorrect parametrisation, or structural gaps in the model rather than masking them through compensatory calibration. In such a case, we argue that deep learning models trained to reproduce streamflow accurately can perform their intended task (reproducing the target variable) well, but could not provide the same useful information on data quality or processes that a well-constrained physically-based model could provide. We will add a few sentences in the discussion of the revised manuscript to provide more general insights on these broad implications of such a hydrological auditing for physically-based models.

l. 74: It would be great to round off the introduction (and the manuscript) with a claim to develop a diagnostic methodology that applies to distributed models in general, if possible; in the current form, the analysis relies heavily on the architecture of LISFLOOD. While such a demo case is certainly interesting, the impact of the study could be increased if the authors invested some effort into more general considerations.

We thank the reviewer for this relevant suggestion, which could make our work more impactful. While our parameter-interplay analysis is necessarily tied to LISFLOOD's specific structure, much of the diagnostic workflow we propose (multi-scale streamflow metrics, FDC signatures, and the cell-based Budyko-distance diagnostic) is in fact model-agnostic and can be applied to any (semi-)distributed hydrological model.

We strongly agree that a more general framing of how this diagnostic can be applied to distributed hydrological models is needed. We will revise the manuscript to make the generalisable elements of our diagnostic framework more explicit and to indicate how they could be transferred to other distributed models.

Specifically, the cell-based Budyko-distance analysis can be applied to any model output, provided that the assumptions underlying the Budyko framework are respected (see line 283). It serves as a simple diagnostic to check whether a model is Budyko-compliant or deviating from the curve for identifiable physical reasons (e.g., snow accumulation, irrigation, leaky catchments). In the Discussion we have also added a reference to Wang et al. (2026), who recently proposed an event-type-based diagnostic framework that complements our long-term, flux-based perspective by focusing on which processes drive errors at the event scale. We believe that their study, focused on streamflow events, could be complementary to the long-term fluxes diagnostic we proposed.

We will add a sentence to the introduction from line 69:

“While we apply the approach to LISFLOOD, the diagnostic workflow we propose, based on multi-scale streamflow metrics, FDC signatures, and a cell-based Budyko-distance diagnostic, is intended as a first step towards a methodology applicable to distributed hydrological models more broadly”

And we will revise parts of the discussion from line 587 as reported below:

“In particular, the cell-based evaluation can help the modeller to spot deviations from the expected Budyko value. A natural next step in developing this diagnostic tool would be to classify areas a priori according to whether higher, lower, or comparable AET is expected with respect to the Budyko empirical AET. For example, in mountainous regions, karstic, or impervious areas we expect a lower AET than the Budyko value, whereas a higher AET is expected when significant water sources exist beyond precipitation (e.g., irrigation) or where the landscape features high water retention or strong groundwater influences. Combined with the four FDC-based signature metrics on streamflow, this classification can shed light on how the model partitions water among storage, AET, and runoff, and on which part of the flow-duration curve discrepancies concentrate.

The evaluation of modelled fluxes and states against such functional relationships is model-agnostic and can be performed on any hydrological model, whereas the link between observed discrepancies and the underlying model structure or parameters is necessarily model-specific. This separation between general diagnostics and model-specific interpretation aligns with the broader perspective advocated by Gleeson et al. (2021) and Gnann et al. (2023), and is, in our view, the natural direction in which our work can be extended.

Recent studies have moved in this direction. For example, (Wang et al., 2026) introduce an event-type-based multi-dimensional diagnostic framework that evaluates timing and magnitude errors for streamflow events of different types (e.g., snow-related events, rainfall on dry or wet soils) and uses explainable machine learning to identify the relative importance of different sources of error, providing a route to diagnose which processes drive event-type-specific model failures. While Wang et al. (2026) focus on streamflow events, our diagnostic considers both streamflow and AET against the Budyko expectation; combining the two perspectives, event-scale streamflow diagnostics and long-term flux-based diagnostics, could provide a more complete picture of where and why a model fails.

Beyond diagnostics, another way to include the Budyko framework, and the BD, could be employed in the constraining of the parameter space prior calibration, ensuring that only physically meaningful values are considered. Alternatively, it can be integrated as a metric during calibration or used in post-calibration evaluation to guide parameter selection/tuning, reducing equifinality, and identify combinations that underestimate AET.”

References:

Wang, Z., Tarasova, L., and Merz, R.: Event-type-based multi-dimensional diagnostics of process limitations in hydrological models, Water Resour. Res., 62, 2026.

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., Taylor, R., Scanlon, B., Rosolem, R., Rahman, S., Oshinlaja, N.,770 Maxwell, R., Lo, M.-H., Kim, H., Hill, M., Hartmann, A., Fogg, G., Famiglietti, J. S., Ducharne, A., de Graaf, I., Cuthbert, M., Condon, L., Bresciani, E., and Bierkens, M. F. P.: GMD perspective: The quest to improve the evaluation of groundwater representation in continental-to global-scale models, Geosci. Model Dev., 14, 7545–7571, 2021.

Gnann, S., Reinecke, R., Stein, L., Wada, Y., Thiery, W., Müller Schmied, H., Satoh, Y., et al.: Functional Relationships Reveal Differences in the Water Cycle Representation of Global Water Models, Nature Water, 1, 1079–90, 2023.

l. 379: “All the experiments including the ByPass do not show any significant difference in the frequency distribution of the calibrated parameters as identified by the Kolmogorov-Smirnov test…” I find this somewhat surprising, because, yes, the distribution seems more stable than without ByPass, but even with, I (visually) identify distinct pattern shifts. For example, the benchmark model seems to favor other values than the alternative models with ByPass concerning GWPercValue, LZThreshold, bInfilt, SnowMeltCoef.

We thank the reviewer for this careful observation, which prompted us to revisit, extend and clarify the statistical analysis to add more insights. In doing so, we noticed that the significance threshold for the original Kolmogorov–Smirnov (KS) test had been mistakenly set at $\alpha = 0.04$ instead of the intended $\alpha = 0.05$. We have re-run the test at the correct significance level. In addition to this small correction, as further explained below, we included an adjustment of the KS test for multiple comparisons (based on the Benjamini–Hochberg correction) and added the results of a complementary statistical test.

Before describing the updated results, it is worth clarifying one point about Figure 5 itself. The figure represents calibrated parameter values, which are continuous, through a discrete heat map, i.e., showing frequency values per bin. This discretisation is useful for overall readability, but it can be misleading when interpreting visual shifts in central tendency, because close values that fall on either side of a threshold (bin separator) are placed in different bins even though the underlying difference between them may be small. Conversely, values that cross the same threshold appear visually identical even when the underlying distributions differ. Figure 5 is therefore informative about the overall behaviour of the calibrated parameters, but the apparent magnitude of any individual shift should be interpreted cautiously and cross-checked against a quantitative statistical test. This is the motivation for the additional statistical analysis that we will add in the revised manuscript and we report below.

In the revised analysis we kept the KS test and added a second statistical test based on the 2-Wasserstein distance (WD) proposed by Schefzik et al. (2021), already applied in a hydrological

context by Ficchì et al. (2026). The two tests are complementary: KS is sensitive to localised differences between empirical CDFs, while the Wasserstein test captures bulk distributional shifts and differences in location, size, and shape. Including both gives a more complete picture, since agreement between the two strengthens the evidence of a real shift, while disagreement is itself informative. We ran the tests across thirteen parameters tested against each of five alternative setups (PowerPrefFlow is treated separately, since it is not active in the no-ByPass configurations) and applied the Benjamini–Hochberg correction to control the false discovery rate at the 0.05 level on both sets of p-values (Benjamini and Hochberg, 1995).

The Benjamini–Hochberg correction is appropriate here because our analysis is exploratory: we aim to characterise which parameters shift across setups, so an overly conservative correction is not required. BH offers a good balance between false-discovery control and statistical power, less restrictive than family-wise approaches such as Bonferroni or Holm (Benjamini and Hochberg, 1995). The same correction was applied to both tests on the same 65 comparisons.

We report here below two new figures supporting this analysis that will be added in the revised manuscript in the Appendix section. Moreover, considering the new results, we have updated Section 3.3 “Calibrated parameters”

Figure A shows the distribution of paired shifts (alternative setup − benchmark) across the sub-catchments for the six parameters where at least one alternative differs significantly from the benchmark under either test. Each blue point is one catchment; the red bar is the median shift, the green dot the mean, and the dashed line the no-shift reference. Bold setup labels indicate significance under the Kolmogorov–Smirnov test; an asterisk indicates significance under the 2-Wasserstein test, both Benjamini–Hochberg corrected at $\alpha = 0.05$ over the 65 benchmark-versus-alternative comparisons.

Figure B uses the same format to show the seven parameters with no significant shift under either test (SnowMeltCoef, GwPercValue, LZThreshold, GwLoss, LakeMultiplier, adjust_Normal_Flood, ReservoirRnormqMult).

The results show some significant differences detected by both statistical tests which involve mainly bInfilt, CalChanMan1, CalChanMan2, UpperZoneTimeConstant, LowerZoneTimeConstant, and QSplitMult, whereas most other parameters show no robust differences. Where the two tests agree (bInfilt, CalChanMan1, CalChanMan2, and QSplitMult in the no-ByPass setups) we have robust evidence of systematic shifts. Where they disagree, the disagreement reflects the complementary sensitivity of the two tests: the Wasserstein test flags small bulk shifts in bInfilt for the ByPass-enabled setups and in CalChanMan1 for NOBP, while the KS test alone detects tail-localised differences in UpperZoneTimeConstant and LowerZoneTimeConstant in the no-ByPass setups.

The corrected analysis confirms most of the original claims but refines a few. GwPerc and GwLoss, originally listed in the manuscript as shifting parameters, do not pass the corrected KSs

and Wasserstein test for different reasons: GwPercValue calibrations in the no-ByPass setups appear to be constrained by the upper bound of the calibration range (2.0), which censors the magnitude of detectable shifts. GwLoss shows small but visually consistent downward shifts, but the within-catchment variability is large compared to the shift magnitude, and the signals do not remain significant after correction. The findings are still visible in the figure (the red median bars sit consistently on one side of zero).

In contrast, CalChanMan1, LowerZoneTime costant and UpperZoneTimecostant are added to the list of parameters with systematic shifts in the no-ByPass setups, which was not noted in the original manuscript.

The adjusted statistical analysis partially agrees with the parameter changes identified by the reviewer. Significant changes in bInfilt are confirmed: both BP-WT and BP-3M show statistically significant shifts under the Wasserstein test, in the direction the reviewer described, i.e., the calibrated parameter values change when soil depth is modified with the ByPass active. This is true especially in terms of central tendency, as confirmed by exploiting the decomposition of WD into location, scale, and shape components of Schefzik et al. (2021), which shows a prominent change in location (> 58%). GwPercValue, LZThreshold, and SnowMeltCoef do not reach significance (based on either the WD or KS test) after correction. The visual patterns the reviewer noted for these 3 parameters are present in the figures (small shifts in median, asymmetric spread) but are within the noise that the tests can resolve at our sample size.

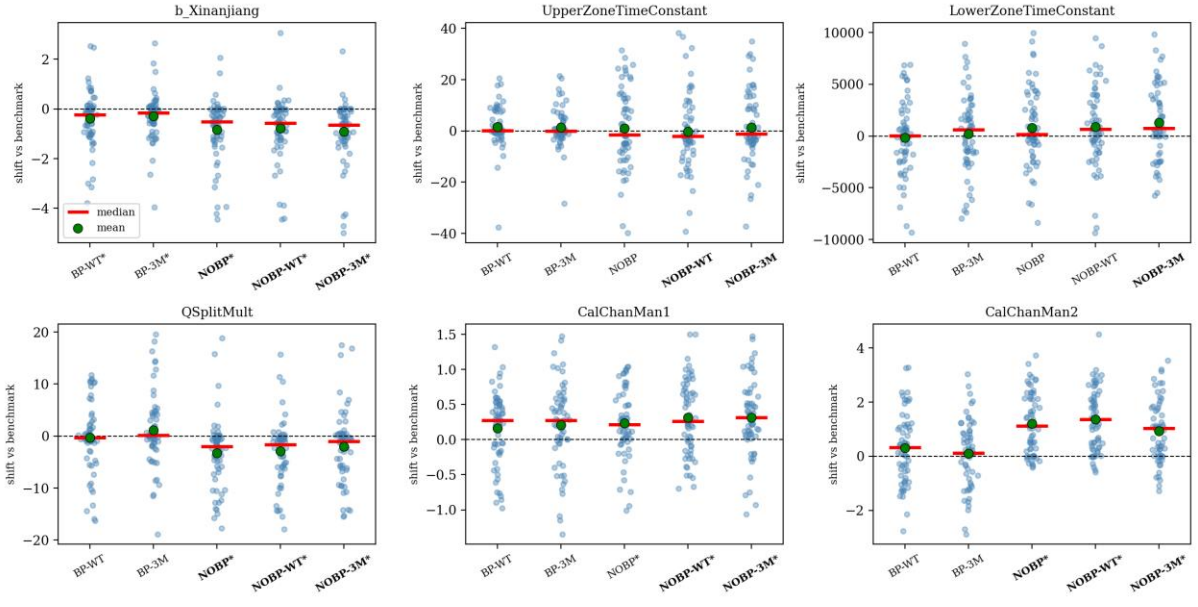

Figure A. Distribution of paired shifts in calibrated parameters (alternative setup − benchmark configuration) across the sub-catchments for the six parameters where at least one alternative differs significantly from the benchmark under at least one of the two statistical tests (KS and WD-based test with BH correction).

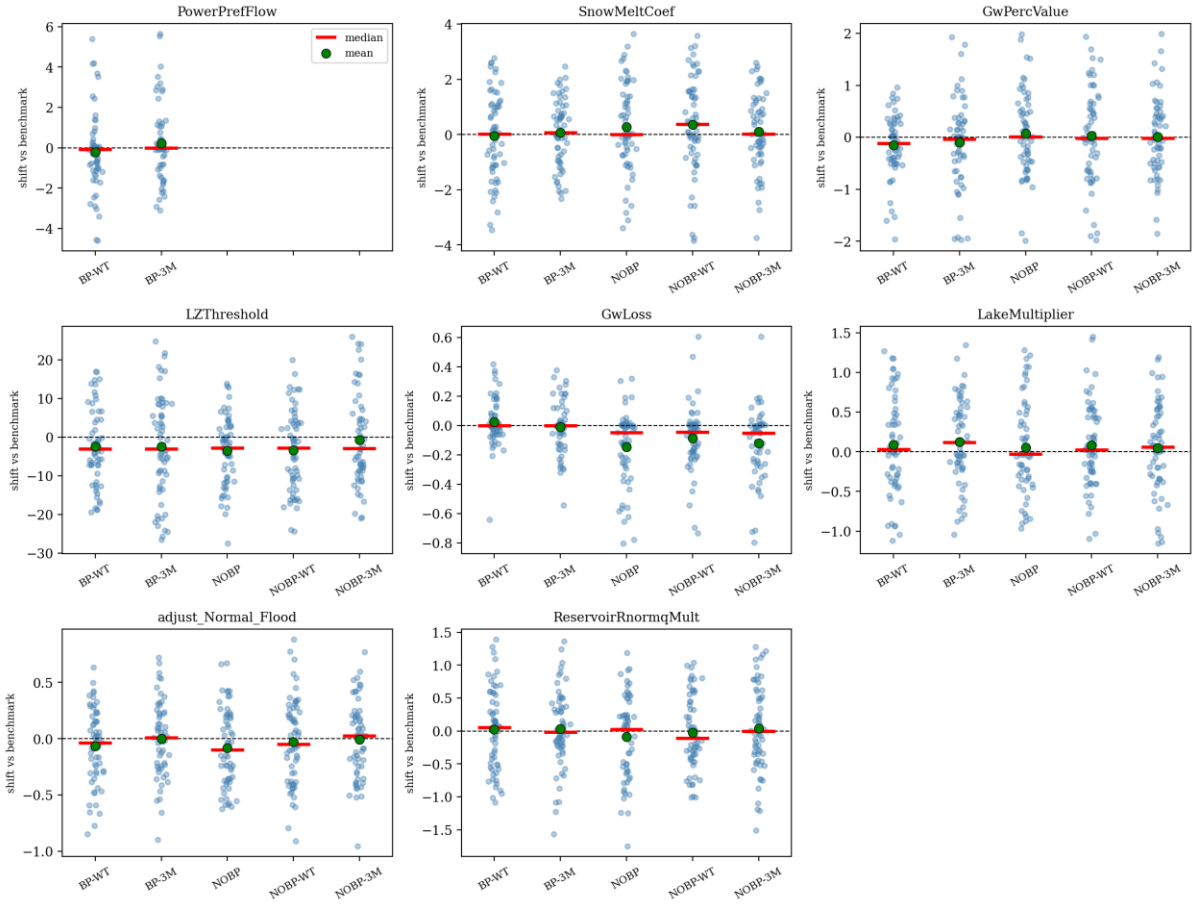

Figure B. Distribution of paired shifts in calibrated parameters (alternative setup − benchmark configuration) across the sub-catchments for the seven parameters where no significant differences are detected by either statistical test (KS and WD-based test with BH correction) between any setups with respect to the benchmark of the two statistical tests.

References:

Ficchì, A., Bavera, D., Grimaldi, S., Moschini, F., Pistocchi, A., Russo, C., Salamon, P., and Toreti, A.: Improving low and high flow simulations at once: An enhanced metric for hydrological model calibration, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2026-43, 2026.

Schefzik, R., Flesch, J., and Goncalves, A.: Fast identification of differential distributions in single-cell RNA-sequencing data with waddR,Bioinformatics, 37, 3204–3211, https://doi.org/10.1093/bioinformatics/btab226, _eprint: https://academic.oup.com/bioinformatics/articlepdf/37/19/3204/50338109/btab226.pdf, 2021.

Discussion & conclusions: while the conclusions are supported by the findings of the study and the proposed ways forward (constraining certain parameters to shield against “abuse” during parameter calibration) are interesting, the study left me a bit “uninspired” – I would have hoped for a more comprehensive comparative analysis including more diverse model representations, and more insights into how to unify these apparently contradicting model uses.

(You might have noticed that usually I provide many more comments – this is a sign that the manuscript is very nicely written, structured in a clear manner and the analysis is well documented. Congrats!)

We thank the reviewer for this thoughtful comment. We agree that we have explored only a limited set of model configurations, both in terms of parameter variations and structural mechanisms (the ByPass activation/deactivation and different soil-depth choices). A broader range of hydrological process representations and parametrizations would certainly be worth investigating. However, the primary objective of this study is not to provide an exhaustive comparison of several alternative model structures and parametrizations, but rather to propose a diagnostic framework to audit hydrological modelling realism, and investigate the trade-offs between streamflow-oriented performance (the target of most existing hydrological models) and physically consistent water-balance partitioning, focusing on six exemplary alternative setups. We will clarify this framing more explicitly in the revised Abstract and Introduction, to avoid generating misleading expectations regarding the scope of the study. Moreover, as reported in our response to a previous comment (see point on L.74), we will revise the manuscript to make the generalisable elements of our diagnostic framework more explicit and to indicate how they could be transferred to other distributed models.

Another benefit of such a diagnostic framework, even when performed on a limited set of alternatives as ours, is to shed light on the most promising avenues for further research on model structural improvements. For example, one direction we did not explore, but that follows directly from our results, is the use of less complex but more robust representations of soil-water dynamics. The evapotranspiration equations in LISFLOOD are sensitive to the soil depth and to the amount of water allowed to infiltrate. The current allowed parameters range of soil depths and infiltration (binfilt) together with the current equations used by LISFLOOD to calculate AET can lead to unrealistic estimation of AET; hence, if we want to keep the current model structure while retaining physical realism, the infiltration and soil depths ranges need to be adjusted to work with the current AET equations to ensure reliable results . An alternative, in line with the arguments on defensible model complexity of Guthke (2017), would be to simplify the unsaturated-zone representation, with fewer interdependent parameters, for which we have limited information (e.g., good-quality and large-scale data) anyway. This would probably result in a more robust and defensible representation of the main fluxes and states governing the water balance, that could be validated on long-term Budyko AET. In the revised manuscript, we will include these arguments and include the reference on defensible model complexity (in the Discussion Section).

Among the configurations we did test, NOBP-3M emerges as the most balanced compromise between streamflow performance (in terms of KGE) and realism of the internal fluxes. However, our spatial analysis (Section 3) still reveals substantial inconsistencies that depend on the calibrated parameters, indicating that even the best of our six setups does not fully resolve the structural issues highlighted by the diagnostic analysis. This reinforces our point that further

structural variations, both within the current LISFLOOD architecture and toward simpler, more robust representations, are needed to move from a partially diagnosed model toward a fully consistent one. We will extend the discussion on this point in the revised manuscript.

Technical comments:

"Auditing" – this term is only used in the title and abstract. Make consistent use of it also in the body of the manuscript, or consider replacing these terms.

We thank the reviewer for noticing this. We agree that the term "hydrological auditing" is central to the study and should be used more consistently throughout the manuscript. In the revised version, we will introduce and refer to this concept more explicitly in the Introduction, Discussion, and Conclusions to better connect the broader motivation of the study with the presented analyses.

l. 116: Please define GPD.

We thank the reviewer for noticing the incorrect spelling of GDP (Gross Domestic Product), we have corrected it in the manuscript from GPD to GDP.

Eqs. 3 and 4: I guess the left-hand side of Eq. 3 should be named "Infiltr" to be consistent with Eq. 4.

Thank you for noticing this inconsistence, we corrected eq 4 from infiltr to Infiltration

l. 312: Remove one instance of "does not simulate"

We thank the reviewer for noticing the repetition. We removed the repetition of "does not simulate."

References:

Guthke, A. (2017). Defensible Model Complexity: A Call for Data-Based and Goal-Oriented Model Choice. Ground Water, 55(5).

Citation: https://doi.org/10.5194/egusphere-2026-423-RC1