Response

We thank the Editor and the reviewer for their 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

The authors evaluated LISFLOOD's performance in simulating streamflow, evapotranspiration, and overall water balance in the Po River Basin using six model setups. With the use of the Budyko framework, the authors demonstrated that the current model setup in EFAS performs best in streamflow simulation, but tends to underestimate ET and have a relatively poor water balance representation compared to other setups. The findings in this paper are crucial for future LISFLOOD configuration for different purposes and introduce an interesting and effective Budyko-based diagnostic framework. I recommend a minor revision with the following comments.

We thank the reviewer for the constructive and detailed review, and for highlighting the importance of our analysis as fundamental for future LISFLOOD setup and developments.

Major comments:

1 In the authors' different model setups, could you clarify why the maximum soil depth is set to be 3m? Can any data support this?

We thank the reviewer for this important question which helps us clarify this choice in the manuscript. The choice of a uniform maximum total soil depth of 3 m was motivated by both consistency with data used for the LISFLOOD setup and by recent findings on the role of deeper soil storage in large-scale hydrological modelling.

First, the original 5 km resolution LISFLOOD setup derived soil depth from the European Soil Database, with values between 140 mm and 3200 mm across the European domain (Laguardia and Niemeyer, 2008). The selected 3 m upper bound is therefore consistent with soil thicknesses the LISFLOOD soil-moisture scheme has historically been parameterized to handle.

Second, recent work on the distributed hydrological model wflow_sbm (Mendoza et al., 2025) has noted that the conventional 2 m soil depth, guided by global soil datasets maximum soil thickness, may be too shallow to capture deeper groundwater dynamics and their influence on catchment processes. Mendoza et al. (2025) reported improved discharge performance when the soil column is extended beyond 2 m. In our study, the 3 m value represents a moderate extension of this 2 m baseline, allowing to test the hypothesis that additional unsaturated storage allows more evapotranspiration without departing radically from the conventional setup. The 3 m setup is applied as a uniform constant for the total soil depth across the three layers (sd1 + sd2 + sd3) over the whole basin, by design. The spatial variability visible in Figure 2 for sd2 and sd3 individually reflects only the partitioning of this fixed 3 m total between the two deeper layers as a function of vegetation root depth, following Burek et al. (2013), while the top layer (sd1) is fixed at 50 mm. Holding the total depth uniform is a deliberate experimental choice that isolates the effect of soil column thickness from the spatial heterogeneity in water-table depth driving the BP-WT and NOBP-WT setups, allowing us to disentangle the two controls on the modelled water balance.

We will clarify this point in the revised manuscript, by adding the following text after line 204:

“The third soil depth configuration is estimated as for the second, except in that it limits the maximum soil depth to 3 m, consistently with the upper range of soil depths historically used in operational LISFLOOD setups, based on the European Soil Database (Laguardia and Niemeyer, 2008). This design choice represents a moderate extension of the 2 m baseline commonly used in global soil datasets, which has been shown to be too shallow to capture deeper groundwater dynamics (Mendoza et al., 2025).“

2 The authors only used the KGE of streamflow as the objective function. I would suggest the authors use ET or Budyko as an additional constraint to see whether it can help make the prediction accurate on both streamflow, ET, and water closure.

We thank the reviewer for this valuable suggestion, which is well aligned with one of the perspectives we propose in the discussion (Section 4, lines 574–582). We agree that a multi-objective calibration using ET or Budyko-based constraints is a promising path to improve the consistency of simulated ET and water-balance closure, potentially at some cost to streamflow-based performance (KGE). However, the scope of this study is diagnostic: we aim to demonstrate that calibration produces based only on streamflow can produce internally inconsistent water balances in LISFLOOD, understand to what extent this may occur, and identify which model components are driving the effect (preferential flow, soil depth, infiltration and saturation-excess runoff). This diagnostic step is, in our view, a prerequisite for understanding how the model deviates from expected physical behaviour, how its structure or parameterization might be improved, and finally for designing an effective calibration strategy.

Moreover, our experiments suggest that improving the realism of internal fluxes in a physics-based model does not necessarily imply a substantial deterioration in streamflow performance. For example, the NOBP-3M setup shows improved consistency with the Budyko framework while maintaining streamflow performance comparable to the benchmark at monthly and annual scales. This indicates that Budyko-based diagnostics can help guide the selection or adjustment of model configurations depending on the intended application (e.g., flood forecasting versus water-balance studies), even before implementing multi-objective calibration.

Multi-objective calibration could be a logical next step building upon the structural diagnostics presented here. Future work could investigate whether jointly optimizing the accuracy on streamflow and Budyko/ET metrics can preserve strong streamflow performance while improving the physical consistency of evapotranspiration and other water balance components.

We will extend the discussion by adding the following sentences in the revised manuscript (after line 534 of the original manuscript, i.e., after the sentence: “Careful consideration is therefore needed when expanding and/or constraining the parameter set and/or modifying the preferential flow process.”):

“The aforementioned considerations highlight the need for a deeper investigation of how the LISFLOOD model simulates the vadose zone, identifying which processes should be constrained or structurally modified to ensure that the model achieves the right response (streamflow) for the right reasons, that is, with internally consistent states and fluxes (Beven and Cloke, 2012; Wagener and Pianosi, 2019). We argue that such hydrological auditing or model structural diagnosis should precede, rather than be replaced by, the introduction of additional calibration targets, whether in the form of extra parameters to be calibrated or of external constraints such as ET or Budyko-based metrics added to the objective function.”

3 The authors evaluated the Budyko distance. The deviation from the Budyko equation is attributed to the model configuration. Is there any missed representation of the model in terms of anthropogenic activities, urbanization, or snow processes that can cause this deviation?

We thank the reviewer for this comment, which allows us to clarify better this point.

Anthropogenic and urbanization effects: grid cells with substantial human influence were excluded from the Budyko analysis, as described in Section 2.5.2, from line 280, specifically, cells with more than 80% irrigated crops, more than 80% sealed surface, or 100% open water or rice cultivation. The retained domain is shown in grey in Figure A2 (Appendix A). The Budyko deviations discussed in the paper therefore cannot be attributed to these factors.

Snow processes: snow-affected cells were not masked, and the influence of snow is visible in Figure B11, where Alpine grid cells show lower AET than cells at lower elevations. This is consistent with the well-documented effect of snow fraction on Budyko-based partitioning, where higher snow fractions are associated with greater annual runoff and reduced AET (Liu et al., 2015; Hwang et al., 2022; Zaerpour et al., 2024) However, Figure B12, the corresponding map for runoff deviations from Budyko, does not show a coherent elevation-driven pattern. If snow processes were the dominant driver of Budyko deviations, we would expect Alpine cells to show systematically higher Q/Pr than the Budyko prediction; instead, the spatial patterns of deviation in Figure B12 align primarily with the boundaries of the calibrated sub-catchments and with soil-depth distributions, not with elevation or snow fraction.

This reinforces our central argument: even in regions where snow processes should exert a strong influence on the AET/runoff partitioning, the dominant signal in the Budyko deviations comes from the model configuration and calibrated parameters rather than from physiographic controls.

We will add this observation in the Discussion, after ln 490:

"The prominent influence of the calibrated parameters over the physical conditions is apparent from Figure B12, where the pattern of the Budyko distance outlines the boundaries of the calibrated sub-catchments rather than the natural characteristics that would be expected to drive deviations from the Budyko theoretical relationship. Snow, for example, is known to influence the runoff/evaporation ratio, favouring runoff and reducing evaporation losses (Zaerpour et al., 2024). This relationship is visible in Figure B11, where Alpine grid cells show lower AET than cells at lower elevations. However, Figure B12, the corresponding map for runoff deviations from Budyko, does not show a coherent elevation-driven pattern. If snow processes were the dominant driver of Budyko deviations, we would expect Alpine cells to show systematically higher Q/Pr than the Budyko prediction; instead, the spatial patterns of deviation in Figure B12 align primarily with the boundaries of the calibrated sub-catchments and with soil-depth distributions, rather than with elevation or snow fraction."

4 The authors compared the model-simulated Budyko relationship with the theoretical Budyko curve. I was wondering if, since there are PET and AET datasets available, the author could compare against the observed Budyko relationship?

We thank the reviewer for this suggestion. The use of EO-based AET and water cycle products is in fact already identified in the discussion (Section 4, lines 547–552) as a promising direction, and we cite Brocca et al. (2024) as an example of an EO-based reconstructed water cycle specifically for the Po basin. We did not extend the analysis to include such a comparison in this paper for two reasons.

First, the diagnostic purpose of this contribution is to evaluate whether LISFLOOD reproduces the long-term climatological partitioning of precipitation between AET and runoff expected from the energy and water availability constraints encoded in the Budyko framework. The theoretical Budyko curve is the appropriate reference for this purpose, since it represents the climatological expectation under the framework's stated assumptions (closed water balance, negligible storage change, no anthropogenic disturbance). A comparison against observation-derived AET and PET products would shift the question from whether the model reproduces the expected long-term behaviour to whether the model agrees with a particular observational product. These are related but distinct questions.

Second, gridded AET and PET datasets such as ERA5-Land, GLEAM, MOD16, or FLUXCOM are themselves model (ERA5-Land) or algorithm-derived estimates (the latter three), each carrying its own assumptions, calibration choices, and biases. Using them as a reference would combine model-to-model differences with model-to-theory differences and require a parallel evaluation of the observational products to interpret the comparison meaningfully. This is a substantial undertaking, involving the choice of reference product, treatment of inter-product disagreement, overlap with the precipitation forcing, and propagation of observational uncertainty into the Budyko diagnostic. We consider this a study of comparable scope to the present one rather than an extension of it.

We have expanded the existing passage in Section 4 (from line 552) to make more explicit the link between the Budyko-based diagnostics presented here and a future EO-based evaluation.

“To address these limitations, combining GSA with empirical relationships, such as the Budyko framework, and incorporating spatially distributed Earth Observation (EO) products like soil moisture and evapotranspiration (ET) data offers a promising path forward, both for constraining AET in calibration and for complementing the theoretical Budyko comparison presented here with an observation-based one. The Po basin itself is an excellent example of EO-based reconstructed water cycle (Brocca et al., 2024). This approach can help better constrain and represent AET and other hydrological processes, even at the grid-cell level, hence improving realism while minimizing adverse impacts on streamflow simulations.”

Minor comments:

1 Does GLOFAS have the same model setup as EFAS? If not, which setup in the six evaluated models is the one used by GLOFAS?
GLOFAS uses the same LISFLOOD model setup as EFAS. In this study, we referred to the EFAS (v 5.0) and GloFAS (v4.0) setup as the “Benchmark” setup, i.e., the standard LISFLOOD soil-depth configuration with the ByPass mechanism enabled. Given the different spatial resolution of the two

models (1 arcmin of EFAS, 3 arcmin of GloFAS), some differences are expected, but the underlying high-resolution dataset used to create the input parameters are the same (Choulga et al. 2023). We will clarify this point in the revised manuscript.

2 Both left and right-hand sides of Eq. 2 have AWI; at least one of them is a typo.
Thanks for the comment. The notation, as noticed by the reviewer, is misleading, but not inheritably wrong. The AWI value is first calculated as water reaching the soil (precipitation and leaf drainage minus interception), we have called it AWI potential eq (1), AWI potential is then used to calculate the Preferential flow (eq 2), AWI is calculate as AWI potential minus the Preferential flow.
We have added the AWI potential equation and changed the naming to clarifying the different steps.
We thank the reviewer for pointing out this inconsistency and giving us the possibility to clarify the equations.

3 The authors have a couple of typos on "Xinanjiang" as "Xinjang" or other words. Please correct these.
We thank the reviewer for noticing the mispelling of Xinanjiang, we have corrected in the manuscript.

4 L420: "the energy limit (green line)": The energy limit is the black line. Please correct this.
We are thankful to the reviewer for noticing this oversight, we have changed “green line” to “black line”

5 313 has a typo:" does not simulate does not simulate."
We thank the reviewers for spotting the typo, we have corrected it.

Citation: https://doi.org/10.5194/egusphere-2026-423-RC2

References
Choulga, M., Moschini, F., Mazzetti, C., Grimaldi, S., Disperati, J., Beck, H., Salamon, P., and Prudhomme, C.: Technical Note: Surface680
Fields for Global Environmental Modelling, EGUsphere, https://doi.org/10.5194/egusphere-2023-1306, 2023.

Laguardia, G. and Niemeyer, S.: On the comparison between the LISFLOOD modelled and the ERS/SCAT derived soil moisture estimates,770
Hydrology and Earth System Sciences, 12, 1339–1351, https://doi.org/10.5194/hess-12-1339-2008, 2008.

Mendoza, R., van Verseveld, W., Seijger, C., and Weerts, A.: Data and model underlying the research of: Assessment of saturated hydraulic conductivity-depth relationships and extended soil column thickness in catchment hydrological modelling, 2025