This manuscript presents a coupled vegetation–hydrology modelling framework that integrates the individual-based forest model FORMIND with the distributed hydrological model mHM through the FINAM coupling framework. The topic is relevant to Geoscientific Model Development, and the motivation for coupling stand-scale vegetation dynamics with catchment-scale hydrology is generally clear. However, in its current form, I have several concerns regarding the methodological justification, reproducibility, and robustness of the model evaluation.

In particular, the large spatial-scale mismatch between FORMIND and mHM, the representativeness of the 9 ha forest samples within 4 km × 4 km hydrological grid cells, and the parameter optimization/evaluation strategy require further clarification and, where feasible, additional sensitivity analysis. Overall, I recommend **major revision**. The Introduction would benefit from substantial reorganization and clearer positioning, while the Methods require additional details and stronger justification of several key modelling choices. Detailed comments are provided below.

> Thank you very much for your thorough review and the helpful comments and suggestions. Below we respond to all your comments in detail and explain how we would like to address these in a revised manuscript.

Major comments:

Lines 40–50. The discussion of the limitations of conventional hydrological models is useful, but the current wording appears somewhat too general. Some relatively simple hydrological models may indeed rely on prescribed or static LAI and empirical PET formulations, and may not explicitly distinguish different components of evapotranspiration. However, these limitations do not apply to all hydrological models. Many more process-based or integrated hydrological models are able to represent dynamic vegetation properties, and some explicitly simulate evaporation and transpiration as separate processes. I therefore suggest avoiding an overly broad characterization of “conventional hydrological models.” Instead, the authors could distinguish between simpler conceptual or empirical hydrological models and more process-based or integrated models, and provide examples showing which models use static LAI or empirical PET formulations and which models explicitly represent dynamic vegetation and separate evaporation and transpiration processes. This additional review would help clarify the specific model limitations that motivate the proposed coupling framework and better position the methodological contribution of this study.

> Thank you for this comment. We agree that the characterization of “conventional hydrological models” was too broad. We will adjust the passage in the introduction to distinguish three groups: 1) conceptual and empirical models, which represent vegetation only implicitly; 2) distributed process-based models such as mHM and VIC, which commonly prescribe static or climatological LAI; and 3) ecohydrological models such as LPJ-Guess or RHESSys, which simulate vegetation dynamics and, in part, separate evaporation and transpiration.

Lines 64–77. These two paragraphs contain a considerable amount of model-specific and setup-specific information that may be more appropriate for the Methods section. For example, details such as the 20 m × 20 m patch size of FORMIND, its previous applications, the typical grid resolution of mHM, and particularly the choice of a 4 km × 4 km resolution in this study are methodological details rather than part of the general motivation. In the Introduction, I suggest retaining only the information needed to explain why FORMIND and mHM are complementary and why coupling these two models is scientifically or methodologically useful. The detailed descriptions of the individual models and the specific model setup could then be moved to the Methods section. This would also provide more

space in the Introduction for positioning the proposed framework relative to existing coupled vegetation–hydrology modelling approaches.

> Thank you for this assessment. We will reduce the model and setup-specific information in the introduction.

Lines 78–88. The listed questions/objectives are not fully parallel in either content or sentence structure. Question (1) concerns model coupling, while Questions (2)–(3) focus on model behavior and evaluation. Question (4) is then introduced separately as a further application of the coupled framework. The grammar is also inconsistent: “We therefore ask” is followed by “(1) how...”, but then switches to “(2) examine...” and “(3) analyse...”. I suggest revising this paragraph to distinguish the main research questions from the subsequent application and to use a more consistent parallel structure.

> Thank you. We will revise the structure and content of the questions.

Section 2.2/2.3. The use of a 4 km × 4 km mHM grid appears relatively coarse compared with the much finer FORMIND representation and may smooth important spatial heterogeneity in topography, soils, and land cover. The authors should further justify this resolution choice and discuss the limitations associated with using such a coarse hydrological grid. If computationally feasible, a sensitivity test using one or more finer mHM resolutions would be valuable to assess whether the main conclusions are robust to spatial resolution.

> Thank you for this comment. We chose a 4 km × 4 km resolution following the setup of the German Drought Monitor, for which mHM has been extensively evaluated (Boeing et al., 2022). Several aspects limit the effect of this choice on our results. First, mHM derives its effective parameters from input data at 100 m × 100 m resolution via multiscale parameter regionalisation (MPR; Samaniego et al., 2010), which accounts for subgrid heterogeneity in soil properties and land cover and yields parameters that are transferable across scales (Kumar et al., 2013). Second, in the coupled framework the forested areas are not simulated at the 4 km scale: the snow, upper-soil and vegetation processes within forests are calculated on the 20 m × 20 m FORMIND patches, so that fine-scale heterogeneity in soil moisture and forest structure is preserved. The coarse grid mainly affects the non-forested fraction and the routing, which are identical in both configurations. Third, the computational cost of the coupled framework scales with the number of grid cells, and a 1 km grid would require about 16 times more FORMIND instances.
>
> In the revised manuscript, we will add a sentence in Sect. 2.3 explaining the choice of grid resolution.

Sections 2.3–2.4.1. The spatial upscaling of the FORMIND simulations would benefit from further justification. For each 4 km × 4 km mHM grid cell, the authors simulate a representative 300 m × 300 m forest area, and the initialization is taken from the nearby inventory with the highest forest fraction; within that inventory, the 300 m × 300 m tile with the highest tree density is selected. Please clarify why this sampling strategy is considered representative of the forested fraction of the full mHM grid cell, and whether selecting the highest-density tile could systematically influence simulated forest structure or transpiration.

> Thank you, this is an excellent point. We have revised the sampling strategy. The initialization is now based on the German National Forest Inventory (BWI), whose sampling points are arranged on a 4 km × 4 km grid and are designed to provide an unbiased, representative sample of forest structure at the landscape scale. The mHM grid is aligned with the BWI grid, so that a BWI sample point is located at each corner

of every grid cell. Each sample plot was extrapolated to 1 ha, and this hectare is divided among the four adjacent grid cells. Each cell is thus initialized with one quarter hectare from each of its four corner plots, resulting in a 1 ha initial inventory based on four independent sample points. This inventory is used to initialize a 9 ha forest area per cell, which is sufficient to capture the stochasticity of FORMIND.

All results in the revised manuscript will be based on the new sampling strategy.

Section 2.4.1. The 5% forest-fraction threshold used to determine whether a 4 km × 4 km cell is treated as a forest cell should be justified. Since this threshold controls where the FORMIND coupling is activated, it may influence the spatial extent of the coupled simulation and the resulting catchment-scale fluxes. The authors should explain the basis for selecting 5%, and, if feasible, provide a sensitivity analysis using alternative threshold values to assess the robustness of the results.

Thank you for pointing that out. The 5 % threshold was chosen to avoid running FORMIND instances for 4 km x 4 km cells with negligible forest cover. Forest in the Selke catchment is spatially aggregated. Therefore, varying the forest threshold has little influence on the coupled domain.

But in the process of implementing a new sampling strategy, we will avoid the forest threshold.

Section 2.4.4. The parameter optimization procedure is not described in sufficient detail for reproducibility. Since the optimized configurations are central to the comparisons in Section 3.3, please specify which mHM parameters were optimized, their parameter ranges, the optimization algorithm, and the objective function used. The calibration period is reported as 2005–2015, but the corresponding evaluation/validation setup should also be clarified.

Thank you. This is in line with comments from the other reviewers and will be addressed in our revision. We will add a table including the optimized parameters, their parameter ranges as well as information on the algorithm and functions.

We will add this table to the Supplementary Information.

Sections 3.1–3.2. The use of eddy covariance observations to evaluate ET and GPP at the stand scale is a strength of the study. However, at the catchment scale, AET and soil saturation are mainly compared between FORMIND-mHM and stand-alone mHM, without an independent observational benchmark. The authors may consider including remotely sensed ET and, if available, soil moisture/saturation products to provide an additional catchment-scale evaluation of these variables, rather than relying only on differences between the two model configurations.

Thank you for this suggestion, which was also raised by the other reviewers. We will add a catchment-scale evaluation of ET and GPP against remotely sensed MODIS products for the Selke catchment. The products will be aggregated to the 4 km model grid and compared with FORMIND-mHM (and, for ET, with stand-alone mHM), at monthly resolution, for the mean seasonal cycle and annual sums. As these products are themselves model-based, we will treat them as an independent but not authoritative reference.

For soil saturation, no suitable independent product is available at the scale of our catchment (458 km^2): satellite products have a coarse resolution and represent only the uppermost soil layer, whereas the simulated saturation refers to the full soil column. The German Drought Monitor is itself based on mHM. We therefore use stand-alone mHM, whose soil moisture was evaluated against in situ observations in

Germany (Boeing et al., 2022), as a reference for the dynamics that the coupling is not expected to change (seasonal cycle, timing).

Section 3.3. Since the manuscript explicitly includes a calibration step for the optimized configurations, I suggest also reporting model performance over an independent evaluation period. This would help assess the out-of-sample performance and transferability of both mHM and FORMIND-mHM, rather than relying mainly on performance over the calibration or full simulation period.

Thank you for this suggestion. We calibrated mHM over 2005–2015 and will present the water balance and performance metrics for the independent evaluation period, 2016–2020. We will report KGE with its components r, α and β at gauge Meisdorf separately for the calibration period and the evaluation period in an extended version of Table 1. The configurations with default parameters contain no fitting to observations and serve as a calibration-free baseline.

We will add the extended table to the Supplement.

Section 4.1. The discussion of “cross-domain parameter inference” may be somewhat stronger than what is directly demonstrated in the present analysis. The results show sensitivity of discharge to WUE and a relationship between GPP and discharge, which clearly indicates potential for cross-domain parameter constraint. However, a formal parameter-inference or identifiability analysis is not performed. I suggest slightly moderating the terminology or clarifying this distinction.

Thank you. We will replace the term ‘parameter inference’ with ‘parameter restriction’ to avoid any misunderstanding.

Minor comments:

Lines 90–97. This paragraph mainly describes the study area, observational datasets, and model evaluation strategy, and may be more appropriate for the Methods section.

We will move or cut this paragraph.

Section 2.4.2. The statement that soil parameters were “adjusted according to the local soil type” is too vague for reproducibility. I would suggest specifying which soil parameters were adjusted and how the corresponding values were determined in Supplementary Information.

Thank you. The soil texture was determined from soil samples taken at the site Hohes Holz, and the maximum saturation, field capacity and permanent wilting point were derived from it using the texture-class look-up table of Maidment (1995). As the soil moisture exponent is functionally similar to the Brooks-Corey λ (pore size distribution), the parameter was estimated according to the pore size distribution estimates by Rawls et al (1982). Root distribution was estimated from literature data (Leuschner et al., 2004).

The specifics will be added to the Supplementary Information.

Figure 3. The day-of-year color coding is somewhat difficult to interpret. The colorbars are narrow and the seasonal gradient is not very distinct in the scatter points. Please consider using a clearer seasonal colormap and/or a more prominent colorbar with more intuitive month labels.

We will improve the readability of Figure 3.

Section 3.4. It would be helpful to briefly remind readers what the two tested parameters represent physically and how they enter the respective models. In particular, please clarify which processes and model responses are directly controlled by WUE in FORMIND and by the canopy interception constant in mHM. This would make the sensitivity results easier to interpret.

The WUE in FORMIND determines the plants' water demand based on the maximum possible Gross Primary Production under current climate conditions. Therefore, it influences transpiration and GPP. The canopyInterceptionFactor is estimated effective maximum interception capacity and determines the size of the canopy storage and therefore, affects evaporation from the canopy.

We will clarify this in the text in Sect 3.4.

Section 4.2. Since computational cost is highlighted as an important limitation of the coupled framework, please consider providing quantitative runtime information for the main model configurations, together with the computing platform and resources used (e.g., HPC system and number of cores), preferably in the Supplementary Information.

Will be added.

References

The references can be found in the reviewed main text. The only new reference used in our response is the following:

Leuschner, C., Hertel, D., Schmid, I., Koch, O., Muhs, A. & Hölscher, D. (2004). Stand fine root biomass and fine root morphology in old-growth beech forests as a function of precipitation and soil fertility. Plant and Soil, 258. https://doi.org/10.1023/B:PLSO.0000016508.20173.80

Rawls, W. J., Brakensiek, D. L., & Saxton, K. E. (1982). Estimation of Soil Water Properties. Transactions of the ASAE, 25(5), 1316–1320.