

Response to Reviewer #2 (RC2)

In this paper, the authors examine to which Earth Observation (EO) data can constrain hydrological model calibration and influence the representation of soil moisture drought across different model structures. Four hydrological and land-surface models (CLM, JULES, mHM, PCR-GLOBWB) are forced with harmonised meteorological inputs over the Rhine River basin and calibrated using three strategies: discharge-only (baseline), EO-only (satellite-derived soil moisture and/or evapotranspiration), and hybrid (EO combined with discharge). Simulations are evaluated against the ESA CCI Soil Moisture product and against in-situ observations from a single ISMN station (Niederwerth), across three major drought events (2015, 2018, 2019), using an Inter-Model Agreement Index (IMAI) together with several temporal and spatial skill metrics. The authors conclude that EO-constrained calibration improves several dimensions of drought representation, exposing structural differences among models. Overall, I found this study timely and relevant, as it addresses whether EO data add value to multi-model calibration for drought monitoring purposes, and how this affects inter-model consistency. The experimental design is ambitious, combining four models, three calibration strategies and nine experimental configurations, multiple evaluation metrics, and three drought events. Nevertheless, I have substantial concerns that need to be addressed before this manuscript can be considered for publication. In particular: (1) the manuscript never establishes why inter-model coherence should be expected or is a desirable property to begin with, given that the four models tested differ fundamentally in structure and process representation; (2) differences attributed to “model structure” are not disentangled from parameter equifinality, since – from my understanding of the manuscript – a single calibrated parameter set per model and objective function is used; (3) several conclusions drawn from the point-scale evaluation – limited to a single station (Niederwerth) – are generalised well beyond what the data supports; and (4) many spatial-pattern comparisons rely on qualitative visual inspection of maps (with claimed differences that are, in several instances, difficult to actually see) rather than on quantitative metrics. For these reasons, my overall recommendation is major revisions before this manuscript can be considered for publication in HESS.

We thank the reviewer for the thorough and constructive assessment of our manuscript. We addressed the major comments, the itemised methodological gaps, and every minor comment and suggested edit below. Checking specific statements against the reported values in our tables identified two genuine inconsistencies, which we corrected (L340–341, L365).

Major comments

M1. *The manuscript repeatedly treats inter-model coherence (or “agreement”) as a desirable outcome without ever justifying why this should be expected or sought after, given that CLM, JULES, mHM and PCR-GLOBWB differ fundamentally in structure, process representation and numerical implementation (L64–65, L261, L483). Unless a modular modelling system is used (e.g., Noah-MP, SUMMA), interpreting and achieving inter-model consistency seems an extremely difficult task. This point resurfaces in the Conclusions without a mechanistic explanation of why the model structures actually differ (L490: how different are the ET parameterisations across models? How is water stored and released from the soil column in each model?). Relatedly, the terminology used for this concept is inconsistent throughout the manuscript – e.g., “(IMAI)” (L13), “ensemble agreement” (L20), “ensemble coherence” (L14). I recommend that the authors explicitly motivate, early in the Introduction, why multi-model coherence (as opposed to individual-model realism) is a meaningful target for this analysis, and that they adopt one consistent terminology throughout the manuscript.*

On motivation: operational drought-monitoring products (e.g., the German Drought Monitor,

cited in Sect. 1) increasingly rely on ensembles of models or indicators, and consensus among structurally independent models is often used, implicitly or explicitly, as a proxy for confidence in a drought classification – an intuition with a long history in the multi-model climate-ensemble literature (Knutti et al., 2010; Tebaldi and Knutti, 2011), who establish both why agreement is treated as informative and why it does not by itself guarantee accuracy. IMAI is not treated as a target to be maximised for its own sake in this study; it is used as a diagnostic to test whether EO constraints move the ensemble toward or away from this operationally relevant consensus behaviour, and whether resulting shifts in agreement reflect genuine structural information or an artefact of discharge-driven homogenisation being relaxed. A paragraph motivating IMAI explicitly on these grounds, with this literature and an explicit caveat – consistent with our response to M2 below – that agreement is diagnostic and not by itself evidence of accuracy, is added to the Introduction.

On terminology: the manuscript inconsistently used “IMAI”, “ensemble agreement”, and “ensemble coherence” for the same concept (e.g., in the Abstract and Conclusions). This is standardised on “IMAI” (or “inter-model agreement”, when referring to the concept rather than the specific metric) throughout.

M2. *A central issue is that differences in model responses to EO calibration are consistently attributed to structural differences, yet a single meta-objective function is used to calibrate each model, without any exploration of parameter equifinality. This means the analysis cannot actually disentangle whether the reported differences arise from genuine structural (i.e., process representation) differences, from the specific parameter set selected during calibration, or from a combination of both. This concern directly undermines several strong statements in the manuscript – e.g., L443 (“the progression Baseline $\rightarrow$ EO $\rightarrow$ EO+Q is consistent across all drought events, demonstrating that EO calibration exposes pre-existing structural uncertainty rather than creating new uncertainty”) and L492 (“rather than calibration artefacts”) – which are not supportable without an equifinality analysis (e.g., using multiple behavioural parameter sets per model). I also disagree with the framing in L489, which equates high inter-model agreement with forecast quality/sharpness; ensemble spread (statistical consistency) is a distinct and equally important attribute that is not addressed here. I recommend that the authors either conduct a limited equifinality analysis (even for a subset of models and events) or substantially soften/remove the structural-uncertainty claims made throughout the manuscript.*

A single, deterministically optimised parameter set is used per model and experiment (Sect. 3.4), with no behavioural parameter ensemble or Monte Carlo exploration; genuine structural (process-representation) differences therefore cannot be statistically separated from parameter-set-specific effects within the present design.

One nuance worth noting, without disputing the core point: the four models compared here are independently developed codes with fundamentally different governing equations and process representations, not different parameter sets within a single model. The classical equifinality problem (Beven, 2006) concerns multiple parameter sets of the *same* model structure achieving similar fit – a narrower problem than cross-model structural diversity. This distinction does not remove the reviewer’s concern: a different, equally plausible parameter set for any one of our models could, in principle, behave more like another model, so we retain the reviewer’s recommended qualification throughout, as follows.

Every claim that attributed an outcome to “structural” differences or “structural uncertainty” without qualification is softened, including the two instances flagged directly:

- The statement that “the progression Baseline $\rightarrow$ EO $\rightarrow$ EO+Q is consistent across all drought events, demonstrating that EO calibration exposes pre-existing structural uncertainty rather than creating new uncertainty” (L443–445) is reworded to state that this progression is *con-*

sistent with EO calibration exposing pre-existing structural differences, with an explicit note that, absent a parameter-equifinality analysis, part of the observed divergence cannot be ruled out as reflecting the specific optimum selected for each model rather than its structure alone.

- The statement that observed divergence reflects genuine differences “rather than calibration artefacts” (L491–492) is qualified in the same way.

The framing at L489 that equates high agreement with forecast quality is also revised: the new Discussion section (Implications for Operational Drought Monitoring, Sect. 5.2) now states that inter-model agreement (ensemble consistency) and ensemble spread/sharpness (a distinct statistical property) are not interchangeable, and that IMAI is used only as a consistency diagnostic (see also M1).

The calibration design used in this study partially mitigates this concern, without requiring a new dedicated analysis. CLM’s calibration uses a GLUE-based sampling procedure (Beven and Binley, 1992), which by construction retains a population of behavioural parameter sets rather than a single optimum, offering a route to a behavioural-ensemble diagnostic that is not currently exploited in this manuscript; we note this explicitly as a direction for a future equifinality analysis rather than a check already performed. mHM’s multiscale parameter regionalisation (MPR) framework (Samaniego et al., 2010; Kumar et al., 2013) constrains the effective parameter space through physically based transfer functions rather than independent grid-cell calibration, which the regionalisation literature understands to reduce equifinality relative to unconstrained calibration; this argument, with the corresponding citations, is added explicitly to a new “Limitations and Scope” subsection of the Discussion (Sect. 5.3), together with the other scope limitations of the study. Every structural-uncertainty claim not supported by this partial evidence is qualified as above, following the reviewer’s recommendation.

M3. *Several statements in the Results generalise findings from the point-scale evaluation well beyond what is supported by the data, since this evaluation is limited to a single ISMN station (Niederwerth) within the Rhine basin (L391, L394). For instance, L391 states that results are “consistent with [the fact] that streamflow integration effectively improves point-scale dynamics”, when this was only demonstrated for soil moisture at Niederwerth. I suggest rewording these statements to explicitly reflect the limited spatial representativeness of the point-scale analysis (see my proposed rewording for L391 under “Suggested edits”).*

The sentence at L391 is changed to: “...consistent with [the fact] that streamflow integration effectively improves soil moisture simulations at the Niederwerth site” (rather than the current general statement about “point-scale dynamics”). At L394–398, the statement that ensemble convergence under EO+Q calibration “indicates this hybrid approach produces more similar model performance, potentially improving ensemble robustness where model structure permits” is likewise revised to make clear that this convergence is observed only at Niederwerth and should not be read as a basin-representative statement. A sentence stating explicitly that all point-scale interpretations in Sect. 4.4 and 4.5 (Taylor diagram, ISMN panel) are local to Niederwerth and not extrapolated to the basin scale is added at the start of Sect. 4.4, echoing the caveat already present in Sect. 2.3 for the dataset description itself.

M4. *The description of the experimental and evaluation framework (Section 3) omits several details needed to assess and reproduce the results. Specifically: What optimisation algorithm was used, and what is the maximum number of iterations allowed per calibration experiment? (L191) What time step is associated with the model evaluations? (L195) Which KGE formulation (e.g., Gupta et al., 2009; Kling et al., 2012; Pool et al., 2018; Tang et al., 2021) was used during calibration? (L198) At what time scale (I assume monthly, given the use of SMI) are the Taylor-diagram metrics in Figure 6 computed? This description currently appears in the Results and*

should be moved to the Methods section (L379), together with the material presented around L400. I recommend including a table with the mathematical definitions of all evaluation metrics used (PCC, KGE, SPAEF, SSIM, SACC, sCoV, IMAI) and their associated time scales, placed immediately after Equation 1 (L213, L221). How are the nine experimental configurations aggregated into the three calibration categories analysed in the Results? Showing only the ensemble mean of each category risks losing important information (L147; L284, L362). Please clarify this procedure and consider showing the individual-experiment results behind each category-level summary, at least as supplementary material.

Optimisation algorithm and iteration budget (L191). Calibration in this study follows the model-specific algorithms already established and documented for this experimental framework in our companion benchmarking paper (Modiri et al., 2025): CLM uses a GLUE-based sampling procedure (Beven and Binley, 1992), mHM uses the Dynamically Dimensioned Search algorithm (DDS; Tolson and Shoemaker, 2007), and PCR-GLOBWB uses a differentiable, deep-learning-based parameter-learning approach minimising MSE (Tsai et al., 2021; Droppers et al., 2025). JULES instead followed a distinct, model-specific data-assimilation approach – a cost-function-based optimisation of the model’s soil parameters – rather than the GLUE/DDS/deep-learning-based approaches used by the other three models, consistent with the broader multi-model calibration protocol described in the same companion paper. A table listing the calibration algorithm used for each model in the EO-constrained experiments (exp31–exp44) reported here is added to Sect. 3.4; iteration budgets follow the companion papers and are not repeated here. mHM’s calibration periods and warm-up lengths (verified directly against the production namelists) are added as a new Appendix.

Time step of model evaluation (L195). The underlying model simulations and the SPAEF calculation used during calibration (Sect. 3.4, “SPAEF is computed at each time step”) run at a daily time step, consistent with the daily HYPER-P/EMO-1 forcing and daily discharge data. All evaluation metrics reported in the Results (SACC, SSIM, IMAI, sCoV, the ISMN comparison in Table 8, and the Taylor diagrams in Figure 6) are instead computed on monthly-aggregated SMI or anomaly fields, as already stated individually in the relevant subsections (e.g., the Table 8 caption, and the SACC definition in Sect. 3.5). Both time steps are now stated once, explicitly and together, in Sect. 3.1.

KGE formulation (L198). The original Gupta et al. (2009) KGE formulation, already cited at the point of definition in Sect. 3.5, is used both for reporting streamflow skill and as the Q term in the calibration objective function (Eq. 1). This is now stated explicitly at both points of use (Sect. 3.4 and 3.5) rather than relying on the citation alone.

Time scale of Taylor-diagram metrics, and related material (L379, and material around L400). These methodological details were previously stated only in the Results (Figure 6’s caption). The description of the monthly, basin-aggregated computation underlying the Taylor diagrams is moved from the caption/Results text into Sect. 3.5, alongside the other metric definitions.

Table of metric definitions (L213, L221). A table consolidating the definitions of all evaluation metrics used (PCC, RMSE, KGE, SPAEF, SSIM, SACC, sCoV, IMAI), together with their associated time scales, is added at the end of Sect. 3.5, immediately before the Results.

Aggregation of nine experiments into three calibration categories (L147, L284, L362). The category-wise mean computation (Sect. 3.1: means computed from the three experiments listed per category in Table 1) is cross-referenced at these locations. Providing the individual-experiment results underlying each category-level summary (Figures 2, 4, 5 and Table 9) as supplementary material, beyond what is already provided at the experiment level in Appendix B for the IMAI diagnostic, requires access to the original gridded model output, which is not currently available (see our response to M5).

M5. *Throughout the Results, differences among models and calibration strategies in the spatial*

maps (Figures 2–5) are described qualitatively, and in several instances the described differences are difficult, or impossible, to actually see in the figures (L321–322, L341, L344). I recommend complementing these qualitative descriptions with quantitative summaries. For instance, you could report empirical probability distributions of the metric of interest computed across all grid cells, for each combination of model structure and calibration strategy, in place of (or in addition to) the current maps (see my comment on Figure 4). Similarly, statements linking model disagreement to topography (e.g., Alpine/glacierised terrain) would be considerably strengthened by explicitly correlating a model-disagreement index with topographic descriptors (elevation, slope) and glacier cover (L300), rather than relying on visual inspection of the maps.

The gridded model output underlying this analysis was produced on institutional HPC infrastructure that is no longer accessible following the end of the project, so full grid-cell distributions and a systematic topography correlation cannot be computed from the original data at this stage. As a partial substitute, the published sCoV maps for the Alpine-topography statement (Sect. 4.1, L300) were checked using an approximate colour-based pixel classification against the figures’ own discrete colour scale, which corroborates the qualitative pattern quantitatively (Alpine-headwater sCoV approximately 2–3× higher than lowland-reach sCoV across all three calibration categories); this is now reported alongside the qualitative description in Sect. 4.1 and 4.7, with the method disclosed. The remaining instances flagged as visually subtle (L321–322, L341, L344, Sect. 4.1–4.2), for which this check was not performed, are reworded to state explicitly that the described contrasts are qualitative, visual impressions rather than claims of clear or well-separated differences.

M6. *The manuscript would benefit from citing and building on previous studies that have systematically compared hydrological/land-surface model structures and their process representations when discussing inter-model consistency (Vano et al., 2012; Fenicia et al., 2014; Mendoza et al., 2015; Saavedra et al., 2022; Spieler & Schütze, 2024 and many others). I also strongly recommend including a summary table describing the main process representations (e.g., runoff generation, snow/glacier dynamics, evapotranspiration formulation, soil-moisture storage) of CLM, JULES, mHM and PCR-GLOBWB. This would substantially help readers interpret the structural differences reported later in the manuscript.*

Thank you for suggesting the extra literature. Following your recommendation, currently, Vano et al. (2012), Fenicia et al. (2014), Mendoza et al. (2015), Saavedra et al. (2022), and Spieler and Schütze (2024) are cited and discussed in the Introduction and/or Discussion of the revised manuscript, motivating why structural model differences are expected to matter for our results (linking directly to M1 and M2). Two further references are added alongside these: Prudhomme et al. (2014), who show structural spread across a multimodel ensemble specifically for hydrological drought hazard, directly supporting the drought-specific framing of our own inter-model comparison; and Bosshard et al. (2013), who quantify model structure as a substantial, non-negligible contributor to total uncertainty in ensemble hydrological-impact projections alongside forcing and parameter choices. Both are cited in the same Introduction paragraph as the structural-difference literature above. A summary table describing the main process representations of the four models (runoff generation, snow, and soil moisture/ET) is added to Sect. 3.2, compiled from the model reference papers already cited there for CLM (Dickinson et al., 2006; Niu et al., 2005), for JULES (Best et al., 2011; Clark et al., 2011), for mHM (Samaniego et al., 2010; Kumar et al., 2013), for PCR-GLOBWB (van Beek and Bierkens, 2009; Wada et al., 2014), and the more detailed technical descriptions in our companion studies (Modiri et al., 2025), with the two references cited directly in the table caption.

Minor comments

L46. *Why “in contrast”? Do the authors consider the uncertainties in precipitation datasets to be*

negligible? Please clarify or rephrase.

Precipitation uncertainty is not considered negligible; the sentence as written unintentionally implied this. It is rephrased to state that precipitation datasets are comparatively more mature but still carry non-negligible uncertainty, particularly for high-intensity events, whereas EO-based evaporation and soil-moisture products currently carry relatively larger uncertainties – a difference in degree, not in kind.

L64–65. *[Inter-model coherence desirability – see M1.]*

See our response to M1 above, which addresses this together with the related instances at L261 and L483.

L74. *Please clarify what “standard meteorological forcing” means – are the forcings identical across all model structures?*

Temperature, radiation, humidity, and wind speed are taken from the same ERA5-based fields for all four models (Sect. 2.2), and precipitation is identical within each baseline/EO-constrained group (EMO-1 for exp10/exp20; HYPER-P for exp30 and all EO-constrained experiments). “Standard meteorological forcing” is replaced with an explicit statement that all non-precipitation forcing variables are identical across models, cross-referencing Sect. 2.2.

L91. *Please specify which model structure(s) are being referred to here.*

This sentence (on structural differences in snow/glacier representation) was placed before the models are introduced in Sect. 3.2, which is why it could not name them. It is moved, with a forward reference, to Sect. 3.2, specifying that CLM, JULES, and PCR-GLOBWB each use distinct explicit snow/glacier schemes, while mHM represents sub-grid snow processes through its multiscale parameter regionalisation (MPR).

L97. *Please clarify whether three precipitation forcing datasets were used, given that two baseline configurations rely on EMO-1 while the main experiments use HYPER-P.*

Only two precipitation datasets are used in total: EMO-1 (exp10, exp20) and HYPER-P (exp30 and all EO-constrained experiments, exp31–exp44); HYPER-P itself is a composite of ERA5, IMERG, and SM2RAIN-derived precipitation (Sect. 2.2), which may be the source of the “three datasets” reading. An explicit sentence stating that exactly two precipitation products are used operationally in this study is added early in Sect. 2.2, with HYPER-P’s internal composition described immediately after.

L105. *Please state explicitly, at this point in the text, that the satellite benchmark is the ESA CCI Soil Moisture product (this is only specified a few sentences later).*

The identification of ESA CCI Soil Moisture as the satellite evaluation benchmark is moved earlier, to the first sentence introducing evaluation datasets in Sect. 2.3.

Figure 1. *Please add elevation to the study-area map, to help readers identify areas of complex topography.*

An elevation layer (hillshade) is added to Figure 1, so the Alpine relief in the southern part of the basin and the flat lowland reaches in the north are now visually apparent directly from the topography, alongside the country labels already shown (see also L299 below).

L132. *Please clarify what “conventional forcing” refers to.*

This refers to the EMO-1 precipitation forcing used in exp10/exp20, as opposed to the HYPER-P hybrid EO precipitation product. “Conventional forcing” is replaced with “EMO-1 precipitation forcing” at this location.

L137. *Do the authors mean “precipitation dataset choice” with “forcing choice” here? Please clarify.*

Yes – “forcing choice” refers specifically to the EMO-1 versus HYPER-P precipitation choice, not the full meteorological forcing stack (identical across models; see L74 above). It is replaced with “precipitation forcing choice” at this location.

L163. *How is generalisability across model conceptualisations and parameterisation philosophies actually assessed or verified in this study?*

This claim overstated what the present single-parameter-set, single-basin design (this manuscript covers only the Rhine basin) can verify (see also M2). The sentence is softened to state that the multi-model ensemble increases the robustness of comparative insight across structurally different models, without claiming that generalisability itself has been verified.

Table 2 (model–experiment availability). *Please clarify why JULES and PCR-GLOBWB are unavailable for certain experimental configurations.*

(This comment refers to Table 2, “Model-experiment availability”, which carries the bullet/dash notation; Table 1 lists the nine experimental configurations without an availability breakdown.) JULES and PCR-GLOBWB are unavailable for the joint EO+Q hybrid experiments (exp41, exp42, exp44) only; both are available for all baseline and EO-only configurations. The joint EO+Q configurations were optional extensions beyond the core baseline and EO-only experimental matrix defined for this study; completing all nine configurations for every model was contingent on the computational and time resources available within the project timeline, and this additional coverage was ultimately completed only for CLM and mHM. This is stated explicitly as a resource-driven limitation of experimental completeness, in Table 2’s footnote and in the new “Limitations and Scope” subsection of the Discussion (Sect. 5.3), rather than implying a structural incompatibility of the JULES or PCR-GLOBWB calibration frameworks.

L216. *“Best lag” between which variables? Please clarify.*

The best lag is the temporal shift (in months) that maximises the Pearson correlation between each model’s simulated soil-moisture (SMI) time series and the corresponding observed reference series (ESA CCI or ISMN) at a given location. This is stated explicitly at first use in Sect. 3.5.

L232 (SSIM equation). *This metric appears to depend on time; if so, please add “ t ” to the notation, for consistency with the other equations.*

SSIM is computed independently for each selected drought-event month. The time index t is added to the SSIM notation (Eq. 2) for consistency with the other time-dependent metric definitions.

L264. *Does the sCoV metric also depend on time? Please clarify, as per the previous comment.*

Yes, sCoV is likewise computed per time step (event month); the corresponding time index t is added to the sCoV notation (Eq. 6).

L272. *Is discharge (Q) also considered an “EO variable”? Please clarify the terminology.*

No – discharge is a conventional in-situ (gauge) observation, not an Earth Observation product; grouping it under “EO variable” in the “Critical Variable Identification” paragraph was inconsistent with the manuscript’s own terminology elsewhere. This paragraph is revised to refer to “SM, ET, and/or discharge (Q) constraints” rather than listing Q as an EO variable.

L297. *“Which differences” – please specify which model-structure differences are being referred to.*

This refers to the differences in topographic sensitivity, glacier-fed dynamics, and parameterisation choices listed in the same sentence, underlying the four model-specific response types identified immediately above it (intensification in CLM, persistence in JULES, localisation in mHM, dampening in PCR-GLOBWB). The sentence is restructured so the referent is unambiguous, and linked explicitly to the process-representation summary table introduced in M6.

L299. *Please indicate in Figure 1 the location of the Alpine and sub-Alpine regions mentioned in the text.*

See the Figure 1 comment above – the added hillshade makes the Alpine relief in the south and the flat lowland reaches in the north directly visible on the map, in combination with the existing country labels, without requiring additional text annotations.

L311 and L313. *This description is hard to follow for readers not familiar with European geography; please clarify which regions/countries are meant.*

The purely directional descriptions (“west of the basin”, “northern to southern and east of the basin”, “southern mountainous regions”) at these locations are replaced with explicit sub-basin/country references (e.g., the Swiss/southern-German Alpine headwaters, the Dutch/German lowland reaches), consistent with the labelling added to Figure 1.

L324–325. *Please clarify whether one month, or multiple months, per drought event were considered in this multi-event analysis.*

A single representative month is used per event (August 2015, August 2018, July 2019), as already stated for the spatial-pattern analysis in Sect. 4.1 (“August 2015 serves as a representative abnormal to moderate drought event”); the multi-event analysis in Sect. 4.2 applies the same single-month-per-event convention. This is now stated explicitly at the start of Sect. 4.2.

L340–341. *This sentence appears incomplete – please review.*

The sentence was indeed missing a verb: “mHM (0.52) and PCR-GLOBWB (0.47), with mHM demonstrating likely uniform spatial distribution and all models being weak in mountainous regions.” It is corrected to: “mHM (0.52) and PCR-GLOBWB (0.47) show comparable, moderate correlations, with mHM demonstrating a more uniform spatial distribution and all models remaining weak in mountainous regions.” (The separate, similarly incomplete sentence a few lines later, “EO-only calibration modest changes SACC across most models...”, is addressed under L349 in “Suggested edits” below.)

L348. *How do the authors know that the described spatial pattern is actually “coherent”? Please support this with a quantitative measure.*

“Coherent” was used qualitatively here without a supporting statistic. This is either supported with the corresponding IMAI/SSIM value for the EO+Q category at this event (available in Table 9/Appendix B) or the wording is softened where the quantitative support is marginal.

L365. *This statement does not appear to hold for the Baseline experiments, where PCR-GLOBWB provides better results – please check.*

Correct: in the Baseline category, PCR-GLOBWB (RMSE 0.184, $r = 0.722$) slightly outperforms CLM (RMSE 0.189, $r = 0.702$), so the statement that “CLM shows the closest correspondence with the ISMN series among the evaluated models” does not hold at Baseline. This sentence is corrected to state that CLM shows the most consistent *improvement* across the Baseline $\rightarrow$ EO-only $\rightarrow$ EO+Q progression (supported by Table 8) rather than claiming it has the closest correspondence overall, and notes explicitly that PCR-GLOBWB has the strongest Baseline performance among the models with hybrid-category data unavailable.

Figure 7 (L407). *The description of the quadrants in the ΔKGE – $\Delta SSIM$ plane should be moved to the figure caption rather than the main text.*

The quadrant description (upper-right/upper-left/lower-left/lower-right) is moved from the main text into Figure 7’s caption.

L467. *Please clarify what is meant by “structural transparency.”*

“Structural transparency” refers to EO constraints revealing pre-existing structural differences among models that discharge-only calibration otherwise masks through basin-integrated closure, rather than the EO constraints themselves introducing new uncertainty. This term is defined explicitly at first use, in the new Discussion section (Sect. 5.1), subject to the qualification described in M2, since this interpretation currently exceeds what can be fully supported without an equifinality analysis.

Suggested edits

Abstract (L6). “ $\sim 1, km$ resolution” $\rightarrow$ “ ~ 1 km resolution” (remove comma, add space).

- Done; checked the rest of the manuscript for the same issue – this was the only instance.
- L279.** “improvement” → “improvement/worsening”, since both temporal and spatial skill may be degraded.
- Done.
- L336.** Please confirm whether this should read “July/2019”.
- Confirmed and standardised to “July 2019” throughout.
- L349.** Consider rewording to “EO-only calibration yields modest...” for clarity.
- Revised to: “EO-only calibration yields modest changes in SACC across most models, while EO+Q calibration is particularly effective where main river channel dynamics dominate.”
- L391.** Reword to reflect the actual scope of the point-scale analysis, e.g.: “...consistent with [the fact] that streamflow integration effectively improves soil moisture simulations at the Niederwerth site” (rather than the current general statement about “point-scale dynamics”).
- Adopted as suggested (see also M3).
- L415.** Delete one of the two consecutive periods (“..”).
- Done.

References

- Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J. (2011). The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes. *Geoscientific Model Development*, 4(3):677–699.
- Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1-2):18–36.
- Beven, K. and Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, 6(3):279–298.
- Bosshard, T., Carambia, M., Goergen, K., Kotlarski, S., Krahe, P., Zappa, M., and Schär, C. (2013). Quantifying uncertainty sources in an ensemble of hydrological climate-impact projections. *Water Resources Research*, 49(3):1523–1536.
- Clark, D. B., Mercado, L. M., Sitch, S., Jones, C. D., Gedney, N., Best, M. J., Pryor, M., Rooney, G. G., Essery, R. L. H., Blyth, E., Boucher, O., Harding, R. J., Huntingford, C., and Cox, P. M. (2011). The joint uk land environment simulator (jules), model description – part 2: Carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, 4(3):701–722.
- Dickinson, R. E., Oleson, K. W., Bonan, G., Hoffman, F., Thornton, P., Vertenstein, M., Yang, Z.-L., and Zeng, X. (2006). The Community Land Model and its climate statistics as a component of the Community Climate System Model. *journal of Climate*, 19(11):2302 – 2324.
- Droppers, B., Bierkens, M. F. P., and Wanders, N. (2025). A multi-resolution deep-learning surrogate framework for global hydrological models. *Water Resources Research*, 61(4):e2024WR037736.
- Fenicia, F., Kavetski, D., Savenije, H. H., Clark, M. P., Schoups, G., Pfister, L., and Freer, J. (2014). Catchment properties, function, and conceptual model representation: is there a correspondence? *Hydrological Processes*, 28(4):2451–2467.

- Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, 377(1-2):80–91.
- Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, 23(10):2739–2758.
- Kumar, R., Livneh, B., and Samaniego, L. (2013). Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme. *Water Resources Research*, 49(9):5700–5714.
- Mendoza, P. A., Clark, M. P., Mizukami, N., Newman, A. J., Barlage, M., Gutmann, E. D., Rasmussen, R. M., Rajagopalan, B., Brekke, L. D., and Arnold, J. R. (2015). Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. *Journal of Hydrometeorology*, 16(2):762–780.
- Modiri, E., Rakovec, O., Shrestha, P. K., Droppers, B., Avila, L., Azimi, S., Salehi, H., Cortes-Torres, N., Güiza-Villa, N., Imhoff, R. O., Francés, F., Kollet, S. J., Rigon, R., Weerts, A. H., García-García, A., and Samaniego, L. E. (2025). A multi-scale intercomparison of land surface and hydrological models: Benchmarking performance and flux consistency across spatial resolutions. *ESS Open Archive*. Preprint, under review at *Water Resources Research*.
- Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., and Gulden, L. E. (2005). A simple TOPMODEL-based runoff parameterization (SIMPOT) for use in global climate models. *Journal of Geophysical Research: Atmospheres*, 110(D21).
- Prudhomme, C., Giuntoli, I., Robinson, E. L., Clark, D. B., Arnell, N. W., Döll, P., Fekete, B. M., Franssen, W., Gerten, D., Gosling, S. N., Hagemann, S., Hannah, D. M., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., and Wisser, D. (2014). Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment. *Proceedings of the National Academy of Sciences*, 111(9):3262–3267.
- Saavedra, D., Mendoza, P. A., Addor, N., Llauca, H., and Vargas, X. (2022). A multi-objective approach to select hydrological models and constrain structural uncertainties for climate impact assessments. *Hydrological Processes*, 36(1):e14446.
- Samaniego, L., Kumar, R., and Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46(5).
- Spieler, D. and Schütze, N. (2024). Investigating the model hypothesis space: Benchmarking automatic model structure identification with a large model ensemble. *Water Resources Research*, 60(7):e2023WR036199.
- Tebaldi, C. and Knutti, R. (2011). Mapping model agreement on future climate projections. *Geophysical Research Letters*, 38(23):L23701.
- Tolson, B. A. and Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43(1):W01413.
- Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, 12(1):5988.

- van Beek, R. and Bierkens, M. F. (2009). The global hydrological model PCR-GLOBWB: Conceptualization, parameterization and verification. Technical report, Department of Physical Geography, Utrecht University, Utrecht, the Netherlands. Accessed 2023-11-28.
- Vano, J. A., Das, T., and Lettenmaier, D. P. (2012). Hydrologic sensitivities of Colorado River runoff to changes in precipitation and temperature. *Journal of Hydrometeorology*, 13(3):932–949.
- Wada, Y., Wisser, D., and Bierkens, M. F. (2014). Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics*, 5(1):15–40.