

Response to Reviewer #1 (RC1)

The authors implemented Earth Observation data to constrain a set of different hydrological models and used three different strategies to evaluate the improvements in drought representation at both spatial and temporal scales. The topic is interesting and addresses current important gaps and research questions, which fit for the journal and are relevant for the community. The paper is well structured and clearly written. I have one major comment in its current format and recommend some minor revisions to improve clarity and readability.

We thank the reviewer for the constructive assessment of our manuscript. We addressed each comment below.

Major comment

RC1.1 *Methodology is mostly clear; however, while the experiments mention calibration, the validation strategy is not entirely clear. In particular, if the models (experiments) are evaluated over a different period after calibration, and in general, how the models are independently validated. Later in the manuscript, the model’s performance is evaluated at a point scale but over the same period used for calibration, which makes it difficult to fully assess the model’s predictive capability. It would be helpful if the authors could clarify the validation strategy.*

Two of the nine experiments are not calibrated at all: exp10 and exp30 (part of the Baseline category, Table 1) are forward simulations with default, non-optimised parameters run over the full native record, so no calibration/validation split applies to them by construction. For the calibrated experiments (exp20, and the EO-only/hybrid experiments exp31/32/34/41/42/44), parameters were calibrated over variable-specific observation-availability periods, and simulations using those parameter sets were subsequently evaluated over the common 2007–2020 analysis period. The calibration periods and warm-up lengths for the EO-constrained experiments are listed below.

Experiment	Calibration period	Warm-up
exp31 (EO-SM)	2016-01-15 – 2022-09-15	730 days
exp32 (EO-ET)	2009-01-15 – 2021-12-15	730 days
exp34 (EO-SM+ET)	2016-01-15 – 2021-12-15	730 days
exp41 (EO+Q+SM)	2016-01-15 – 2022-09-15	730 days
exp42 (EO+Q+ET)	2009-01-15 – 2021-12-15	730 days
exp44 (EO+Q+SM+ET)	2016-01-15 – 2021-12-15	730 days

This choice is consistent with the evaluation philosophy adopted throughout this experimental framework (Modiri et al., 2025) and with the findings of Shen et al. (2022), who showed that calibration over the full available record can provide robust parameter estimates compared with the conventional split-sample approach of Klemes (1986). Nevertheless, we now explicitly clarify that the evaluation presented here should not be interpreted as a formal, independent validation.

The mHM calibration periods and warm-up lengths above are now also added to the manuscript as a new Appendix C. The complete per-experiment calibration configuration for all four models, including the corresponding CLM, JULES, and PCR-GLOBWB periods, is documented in Modiri et al. (2025) (the companion quantifying-added-value study). For the ISMN point-scale comparison at Niederwerth specifically, all categories (including the default-parameter and discharge-calibrated Baseline configurations) are further restricted to their common 2010–2018 overlap window, constrained by station data availability; the multi-model spatial drought analysis presented in this manuscript instead harmonises all model outputs to their common 2007–2020 overlap period. Reported performance therefore reflects skill over each model’s full available (or calibrated) record

rather than a strictly independent validation period. This explanation, with the corresponding citations and the model-specific periods, is now added to Section 3.4 (Calibration Formulation and Optimisation Scheme). We revised the point-scale evaluation text (Sect. 4.4) with the same clarification for the ISMN comparison at Niederwerth.

Minor comments

RC1.2 *Line 119-120, if the data source does not provide usable information, why is it mentioned?*

TERENO and FLUXNET, together with the corresponding map panel in Figure 1, are mentioned to document network coverage relative to the Rhine basin and to explain why the point-scale evaluation is restricted to a single ISMN station. The sentence is revised to state this explicitly: these sources are mentioned to document the in-situ network coverage considered and why they were screened out (no station with sufficient temporal/spatial coverage within the basin and study period), not because they contributed data to the analysis.

RC1.3 *Lines 123 – 128, these lines, as written now, seem to fit better for the Method section.*

Agreed. These lines describe the depth mismatch between ESA CCI, ISMN, and the four models' uppermost soil-moisture layers, and how this mismatch is handled analytically (standardised SMI space for spatial comparison; illustrative point-scale diagnostic at Niederwerth) – methodological content rather than dataset description. The paragraph is moved from Section 2.3 (Reference Datasets) to Section 3.2, where the uppermost-layer depths per model are already listed.

RC1.4 *Although the current manuscript is well organized and structured, it does not contain a separate Discussion section; the interpretation of the results is generally clear. The authors might consider adding a brief discussion section of the limitations and broader implications of the findings directly into the main text, rather than including it in an appendix.*

A dedicated “Discussion” section (Sect. 5) is added between the Results and Conclusions, with three subsections: 5.1 synthesises the mechanistic interpretation previously distributed across the Results and Conclusions; 5.2 states the broader implications for operational drought monitoring, previously only in the Conclusions; and 5.3 (“Limitations and Scope”) consolidates the limitations already touched on in the text – quasi-independence of the ESA CCI benchmark, single-station point-scale evaluation, depth mismatch across models, validation strategy (RC1.1 above), the resource-driven unavailability of exp41/42/44 for JULES and PCR-GLOBWB, the sCoV pixel-classification method (RC2's M5), and equifinality (RC2's major comment 2) – into one place rather than leaving them scattered. The Conclusions are correspondingly reduced to three key messages (see RC1.7 below).

RC1.5 *The caption in Figure 3 explains the top panels, but the caption for the bottom panels is missing, or at least is not completely clear.*

The caption is extended with a sentence describing the bottom row: each panel shows the strict multi-event intersection of the drought coincidence agreement metric (a grid cell is flagged only where all three drought events, 2015/2018/2019, are jointly classified as drought), computed separately for the Baseline, EO-only, and EO+Q categories shown left to right, with the same colour scale and n_v masking convention as the top row.

RC1.6 *Figure 8 is not cited in the text.*

Figure 8 (spatial coefficient of variation, sCoV, across models for August 2015) is now cited explicitly in the first sentence referring to the sCoV maps in Section 4.7.

RC1.7 *The conclusions section is somewhat long; it could be improved by focusing on the key messages to provide greater clarity.*

The Conclusions are shortened, now that the new Discussion section (RC1.4) absorbs the mechanistic interpretation and the equifinality discussion previously repeated there. The revised Conclusions focus on three key messages: (1) EO-constrained calibration is associated with measurable,

though event- and model-dependent, gains in temporal and spatial drought representation; (2) this comes with a measurable reduction in multi-model spatial agreement, interpreted as consistent with exposing pre-existing structural differences rather than calibration failure, subject to the equifinality caveat detailed in the new Limitations and Scope subsection; and (3) soil moisture is the most consistent single EO constraint within this experimental framework, with the combined SM+ET and hybrid EO+Q configurations offering complementary but more variable behaviour.

References

- Klemes, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, 31(1):13–24.
- Modiri, E., Rakovec, O., Shrestha, P. K., García-García, A., Avila, L., Blackford, K., Cooper, E., Cortes-Torres, N., Di Paolo, F., Droppers, B., Filippucci, P., Fischer, M., Güiza-Villa, N., Orság, M., Stradiotti, P., Wiggins, M., Brocca, L., Clark, D., Dall’Amico, M., Dorigo, W. A., Francés, F., Kollet, S. J., Peng, J., Wanders, N., and Samaniego, L. E. (2025). Quantifying the added value of earth observation at the kilometre scale with a multi-model hydrologic and land surface ensemble. Authorea. Preprint, under review at *Water Resources Research*.
- Shen, H., Tolson, B. A., and Mai, J. (2022). Time to update the split-sample approach in hydrological model calibration. *Water Resources Research*, 58(3):e2021WR031523.