Check for updates

AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2025WR042446

Assessing the Added Value of Earth Observations for Multi-Model Kilometer-Scale Hydrological Modeling Across Contrasting Basins

Ehsan Modiri[1], Oldrich Rakovec[1,2], Pallav Kumar Shrestha[1], Almudena García-García[3,4], Leandro Avila[5], Katie Blackford[6], Elizabeth Cooper[6], Nicolas Cortes-Torres[7], Federico Di Paolo[8], Bram Droppers[9], Paolo Filippucci[10], Milan Fischer[11], Nathaly Güiza-Villa[7], Matěj Orság[11], Pietro Stradiotti[12], Matthew Wiggins[6], Luca Brocca[10], Douglas Clark[6], Matteo Dall'Amico[8], Wouter Dorigo[12], Felix Frances[7], Stefan Kollet[5], Jian Peng[3,4], Niko Wanders[9], and Luis Samaniego[1,13]

[1]Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany, [2]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Prague, Czech Republic, [3]Department of Remote Sensing, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany, [4]Institute for Earth System Science and Remote Sensing, Leipzig University, Leipzig, Germany, [5]Institute of Bio- and Geosciences Agrosphere (IBG-3), Forschungszentrum Jülich GmbH, Jülich, Germany, [6]UK Centre for Ecology and Hydrology, Wallingford, UK, [7]Research Group of Hydrological and Environmental Modelling (GIMHA), Research Institute of Water and Environmental Engineering (IIAMA), Universitat Politècnica de València, Valencia, Spain, [8]Waterjade Srl, Pergine Valsugana, Italy, [9]Department of Physical Geography, Utrecht University, Utrecht, The Netherlands, [10]National Research Council of Italy, Research Institute for Geo-Hydrological Protection, Perugia, Italy, [11]Global Change Research Institute CAS, Brno, Czech Republic, [12]Department of Geodesy and Geoinformation, Vienna, Austria, [13]Institute of Environmental Science and Geography, University of Potsdam, Potsdam, Germany

Key Points:

- EO forcing improves simulations in the Po and Tugela, but adds little in the Rhine without discharge calibration
- For the three models tested with both, joint calibration with discharge and EO data is more reliable than EO-only calibration
- Best model setups are basin-specific, with EO snow helping in the Po and EO evapotranspiration helping in the Tugela

Correspondence to:

E. Modiri, O. Rakovec and L. Samaniego,
ehsan.modiri@ufz.de;
rakovec@fzp.czu.cz;
luis.samaniego@ufz.de

Citation:

Modiri, E., Rakovec, O., Shrestha, P. K., García-García, A., Avila, L., Blackford, K., et al. (2026). Assessing the added value of Earth observations for multi-model kilometer-scale hydrological modeling across contrasting basins. *Water Resources Research*, *62*, e2025WR042446. https://doi.org/10.1029/2025WR042446

Received 29 SEP 2025
Accepted 4 NOV 2026

Author Contributions:

Conceptualization: Ehsan Modiri, Oldrich Rakovec, Almudena García-García, Federico Di Paolo, Bram Droppers, Milan Fischer, Luca Brocca, Douglas Clark, Matteo Dall'Amico, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego

Abstract Earth observation (EO) products provide spatially and temporally detailed information on key water-cycle processes, yet their value for improving hydrological and land-surface model simulations remains uncertain. Here, we evaluate high-resolution EO data for daily streamflow simulations at ≈1 km resolution across contrasting hydroclimatic and data-density regimes in Europe and Africa, using up to five model structures (five in the Po, four in the Rhine and three in the Tugela). We explore two complementary pathways: (a) replacing conventional meteorological forcing with EO-based precipitation, and (b) estimating model parameters with EO constraints on soil moisture, evapotranspiration, and snow cover, either alone or jointly with streamflow. Model skill is quantified using the Kling–Gupta Efficiency (KGE) and its components, based on paired gauge comparisons. Results show that EO precipitation forcing improves streamflow simulations in the Po and Tugela river basins, but reduces skill in the Rhine. Calibration against EO variables alone yields mixed outcomes, whereas joint calibration with EO variables and streamflow, available for three of the models, gives more consistent improvements, particularly for TETIS in the Tugela and mHM in the Po. KGE decomposition reveals that improved variability alignment is the dominant contributor to these gains, with changes in bias and timing playing secondary roles. Despite differences across models and basins, our findings show that EO integration can enhance kilometer-scale hydrological simulations in specific model–basin configurations, particularly when combined with discharge information. This highlights both the potential and the current limits of available EO products as complementary constraints for operational hydrology, flood and drought early warning, and climate-resilient water management.

1. Introduction

Terrestrial Essential Climate Variables (ECVs), such as streamflow (Q), evapotranspiration (ET), and soil moisture (SM), are important for monitoring Earth's hydroclimate and guiding decisions on climate change adaptation and water management (Balsamo et al., 2015; Bojinski et al., 2014; Wood et al., 2011). Hydrological and land surface models are central tools for simulating and predicting these variables, providing decision-relevant information for applications ranging from flood forecasting to climate impact assessments (Clark et al., 2015). These models are advancing to high-resolution, kilometer-scale simulations to support such applications (Bierkens et al., 2015; Fatichi et al., 2016). The integration of high-resolution Earth Observation

© 2026 The Author(s). *Water Resources Research* published by Wiley Periodicals LLC on behalf of American Geophysical Union.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Data curation: Ehsan Modiri, Oldrich Rakovec, Pallav Kumar Shrestha, Almudena García-García, Leandro Avila, Katie Blackford, Elizabeth Cooper, Nicolas Cortes-Torres, Federico Di Paolo, Bram Droppers, Paolo Filippucci, Milan Fischer, Nathaly Güiza-Villa, Matěj Orság, Pietro Stradiotti, Matthew Wiggins, Luca Brocca, Douglas Clark, Matteo Dall'Amico, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego
Formal analysis: Ehsan Modiri, Oldrich Rakovec, Pallav Kumar Shrestha, Almudena García-García, Leandro Avila, Katie Blackford, Elizabeth Cooper, Nicolas Cortes-Torres, Federico Di Paolo, Bram Droppers, Paolo Filippucci, Milan Fischer, Nathaly Güiza-Villa, Matěj Orság, Pietro Stradiotti, Luca Brocca, Douglas Clark, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego
Funding acquisition: Luis Samaniego
Investigation: Ehsan Modiri, Oldrich Rakovec, Pallav Kumar Shrestha, Almudena García-García, Leandro Avila, Katie Blackford, Elizabeth Cooper, Nicolas Cortes-Torres, Federico Di Paolo, Bram Droppers, Paolo Filippucci, Milan Fischer, Nathaly Güiza-Villa, Matěj Orság, Pietro Stradiotti, Matthew Wiggins, Luca Brocca, Douglas Clark, Matteo Dall'Amico, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego
Methodology: Ehsan Modiri, Oldrich Rakovec, Pallav Kumar Shrestha, Almudena García-García, Leandro Avila, Nicolas Cortes-Torres, Federico Di Paolo, Bram Droppers, Paolo Filippucci, Milan Fischer, Nathaly Güiza-Villa, Matěj Orság, Pietro Stradiotti, Luca Brocca, Douglas Clark, Matteo Dall'Amico, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego
Project administration: Almudena García-García, Luis Samaniego
Resources: Ehsan Modiri, Oldrich Rakovec, Pallav Kumar Shrestha, Almudena García-García, Leandro Avila, Katie Blackford, Elizabeth Cooper, Nicolas Cortes-Torres, Federico Di Paolo, Bram Droppers, Paolo Filippucci, Milan Fischer, Nathaly Güiza-Villa, Matěj Orság, Pietro Stradiotti, Matthew Wiggins, Luca Brocca, Douglas Clark, Matteo Dall'Amico, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego
Software: Ehsan Modiri, Oldrich Rakovec, Pallav Kumar Shrestha, Leandro Avila, Katie Blackford, Elizabeth Cooper, Nicolas Cortes-Torres, Federico Di Paolo, Bram Droppers, Paolo Filippucci, Milan Fischer, Nathaly Güiza-Villa, Matěj Orság,

(EO) products has been proposed as a key strategy to enhance the performance of these models; however, their specific added value and optimal integration methods remain underexplored (Sheffield et al., 2018). A central unresolved question is whether EO products provide robust improvements across models and basins, or whether their benefits depend on the dominant hydrological processes, model structure, and the availability of discharge information.

Current practices often test single products in isolation, lack a consistent experimental design, and rarely benchmark multiple models across diverse hydroclimatic conditions. Moreover, most EO applications in hydrology rely on sequential assimilation methods (Liu et al., 2012; Reichle, 2008), while less attention has been paid to the value of EO data when used as calibration targets to constrain model parameters. This gap limits our understanding of how different EO streams contribute to improved representation of hydrological processes and restricts the practical benefits of high-resolution modelling (Miralles et al., 2019; Sheffield et al., 2018).

Significant challenges persist in the implementation and evaluation of high-resolution hydrological modelling. The hydrological community still requires consistent benchmarking protocols for comparing multiple models at the increasingly used 1 km scale (Best et al., 2015; Collier et al., 2018; Rakovec et al., 2016). Benchmarking is critical for quantifying the added value of increased spatial detail and ensuring that the advantages of finer resolution justify the computational costs (Wood et al., 2011). High-resolution EO products provide spatially distributed constraints on key processes, including soil moisture, evapotranspiration, snow cover area (SCA), and precipitation (Dietz et al., 2012; Funk et al., 2015; Sun et al., 2018; Kumar et al., 2014; Miralles et al., 2019). Parameter estimation through calibration against these observed variables can enhance model accuracy by accounting for uncertainties in both the structural model and the input data (Refsgaard, 1997). Integrating these variables into hydrological and land surface models offers the potential to reduce structural and input uncertainties, improve the cross-basin transferability of parameters, and enhance the robustness of hydrological simulations. In this study, we treat EO integration strictly as a parameter estimation strategy with EO constraints, where EO products are explicitly used as calibration targets to inform parameter optimisation rather than relying on sequential data assimilation updates.

To address these challenges, we conduct a multi-model, multi-basin evaluation of EO integration strategies across contrasting hydroclimatic conditions (Duan et al., 2006; Merz et al., 2009). The EO products were selected following the recommendations of García-García et al. (2026). We focus on three elements: benchmarking model skill at 1 km resolution, integrating EO products as calibration targets, and assessing consistency across diverse model structures. Building on prior multi-model intercomparisons (Droppers et al., 2024; Modiri et al., 2025; Smith et al., 2012), we test how different EO constraints affect model performance. We hypothesise that EO products used as calibration targets can improve kilometre-scale hydrological simulations by constraining processes observable from space and relevant to the dominant runoff-generation mechanisms of a basin, particularly when combined with discharge information. Rather than seeking a universal calibration strategy, this study investigates how the effectiveness of EO-constrained calibration depends on the interaction between model structure, calibration target, and basin characteristics. We therefore evaluate EO integration within a coordinated, multi-model, multi-basin benchmarking framework designed to identify context-dependent responses and process sensitivities.

Three key questions guide our analysis. First, we ask what predictive contribution EO products provide when introduced through EO-based precipitation forcing or EO-based calibration. Second, we examine how different EO calibration configurations perform across models, identify the best-performing combinations of EO constraints, and analyse which processes they most strongly affect. Third, we assess whether these findings are consistent across basins, model structures, hydroclimatic regimes, and observation-network contexts.

Together, these analyses evaluate when EO-based calibration enhances model performance. We also introduce a benchmarking protocol for incorporating EO data into hydrological modelling. The results provide guidance for using EO products as complementary constraints in hydrological simulation, forecasting, flood and drought early warning, and water resources assessment. More broadly, this study provides a coordinated multi-model evaluation of EO-based calibration across three contrasting hydroclimatic regions and identifies where available EO products add value to kilometre-scale hydrological prediction.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Pietro Stradiotti, Matthew Wiggins, Luca Brocca, Douglas Clark, Matteo Dall'Amico, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego
Supervision: Oldrich Rakovec, Luca Brocca, Douglas Clark, Matteo Dall'Amico, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego
Validation: Ehsan Modiri, Oldrich Rakovec, Pallav Kumar Shrestha, Almudena García-García, Leandro Avila, Katie Blackford, Elizabeth Cooper, Nicolas Cortes-Torres, Federico Di Paolo, Bram Droppers, Paolo Filippucci, Milan Fischer, Nathaly Güiza-Villa, Matěj Orság, Pietro Stradiotti, Matthew Wiggins, Luca Brocca, Douglas Clark, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego
Visualization: Ehsan Modiri, Oldrich Rakovec
Writing – original draft: Ehsan Modiri, Oldrich Rakovec, Almudena García-García, Luis Samaniego
Writing – review & editing: Ehsan Modiri, Oldrich Rakovec, Pallav Kumar Shrestha, Almudena García-García, Leandro Avila, Katie Blackford, Elizabeth Cooper, Nicolas Cortes-Torres, Federico Di Paolo, Bram Droppers, Paolo Filippucci, Milan Fischer, Nathaly Güiza-Villa, Matěj Orság, Luca Brocca, Douglas Clark, Wouter Dorigo, Felix Frances, Stefan Kollet, Jian Peng, Niko Wanders, Luis Samaniego

2. Study Design and Data

2.1. Study Regions

Three study basins were selected to benchmark modeled terrestrial ECVs: the Po and Rhine river basins in Europe and the Tugela basin in South Africa. The selection was guided by the need to cover a wide range of hydroclimatic regimes, physiographic characteristics, and data availability. The Rhine basin is characterized by a mixed nival–pluvial regime and dense observational coverage (Stahl et al., 2016), while the Po basin represents a Mediterranean-influenced system with regulated lakes and substantial human modifications (Montanari, 2012). In contrast, the Tugela basin exemplifies a subtropical, data-scarce environment with steep gradients and inter-basin water transfers (Wade et al., 2024). This combination of basins enables evaluation of EO integration under both well-monitored and challenging conditions, thereby strengthening the comparative scope and diagnostic utility of the benchmarking framework. The spatial extent of these regions is shown in Figure 1.

2.2. Data Sources and Earth Observation Products

Table 1 summarizes the EO and atmospheric reanalysis data sets used for meteorological forcing, calibration, and evaluation. Our analysis utilizes a multi-source, hybrid precipitation product (referred to as HYPER-P) (Filippucci et al., 2025) created by downscaling and merging ERA5 reanalysis precipitation estimates (Muñoz Sabater et al., 2021), IMERG satellite precipitation observations (Huffman et al., 2015), and soil moisture-derived precipitation from the SM2RAIN algorithm applied to satellite ASCAT observations (Brocca et al., 2019). This composite approach leverages the complementary characteristics of the three data sets. ERA5 provides physically consistent and spatially complete precipitation fields, but precipitation is a model-derived forecast variable and is not directly constrained by assimilated rain-gauge observations. Its estimates can therefore deviate from actual precipitation, particularly in regions where the observing system weakly constrains precipitation processes. To reduce this limitation, observational data from IMERG and SM2RAIN are incorporated. IMERG provides precipitation estimates derived from satellite-observed cloud and rainfall signals, contributing high-temporal-resolution information on near-instantaneous precipitation dynamics. In contrast, SM2RAIN infers rainfall indirectly from temporal variations in satellite-observed soil moisture, offering an accumulated precipitation perspective linked to land-surface response processes. The integration of these complementary sources improves the realism and reliability of precipitation estimates, particularly in data-sparse regions. This combined product is publicly hosted and reproducible. Crucially, to clarify the configuration of our precipitation forcing, this engine must be distinguished from the NSIDC SMAP soil moisture product listed subsequently; the latter was kept completely separate from the HYPER-P precipitation generation and was utilized strictly as an independent spatial validation and calibration target within our experimental optimisation framework. Because the SM2RAIN component within HYPER-P relies exclusively on satellite ASCAT observations, utilizing the independent SMAP radiometer product as a calibration target avoids structural circularity or artificial autocorrelation in our model optimisation.

Additional EO data sets used for model calibration include high-resolution soil moisture retrievals from NSIDC SMAP (1 km resolution) (Lakshmi & Fang, 2023), and evapotranspiration estimates from MODIS Terra products (Running et al., 2021). We also incorporated MODIS satellite-derived SCA products from 2002 to 2022 (Notarnicola et al., 2013), which have already been used in the literature to calibrate and to update hydrological models (Andreadis & Lettenmaier, 2006; Parajka & Blöschl, 2008). This medium-resolution product, available through Eurac Research, has been extensively evaluated in the literature. While other high-resolution SCA products, such as those from Sentinel-2, Landsat-8, and Sentinel-1, are available, we chose the MODIS product for its long-term consistency and established performance in our target regions. The EO products were selected following the intercomparison of García-García et al. (2026). That study evaluated 7 precipitation, 5 snow cover, 6 surface soil moisture and 6 evapotranspiration products over Europe against observational references. It recommended two merged 1 km precipitation products, one of which is the ERA5–IMERG–SM2RAIN merge used here as HYPER-P, the MODIS and Sentinel-2/Landsat-8 snow cover products, which reached the highest classification accuracy, and the NSIDC SMAP 1 km soil moisture product. The evapotranspiration products performed similarly at the flux sites; among them, MODIS-Terra has the finest spatial resolution (500 m), at an 8-day time step. The forcing and calibration targets of this study are taken from these recommendations. The evaluation covered Europe only, so for the Tugela basin the same products are used without an equivalent regional evaluation.

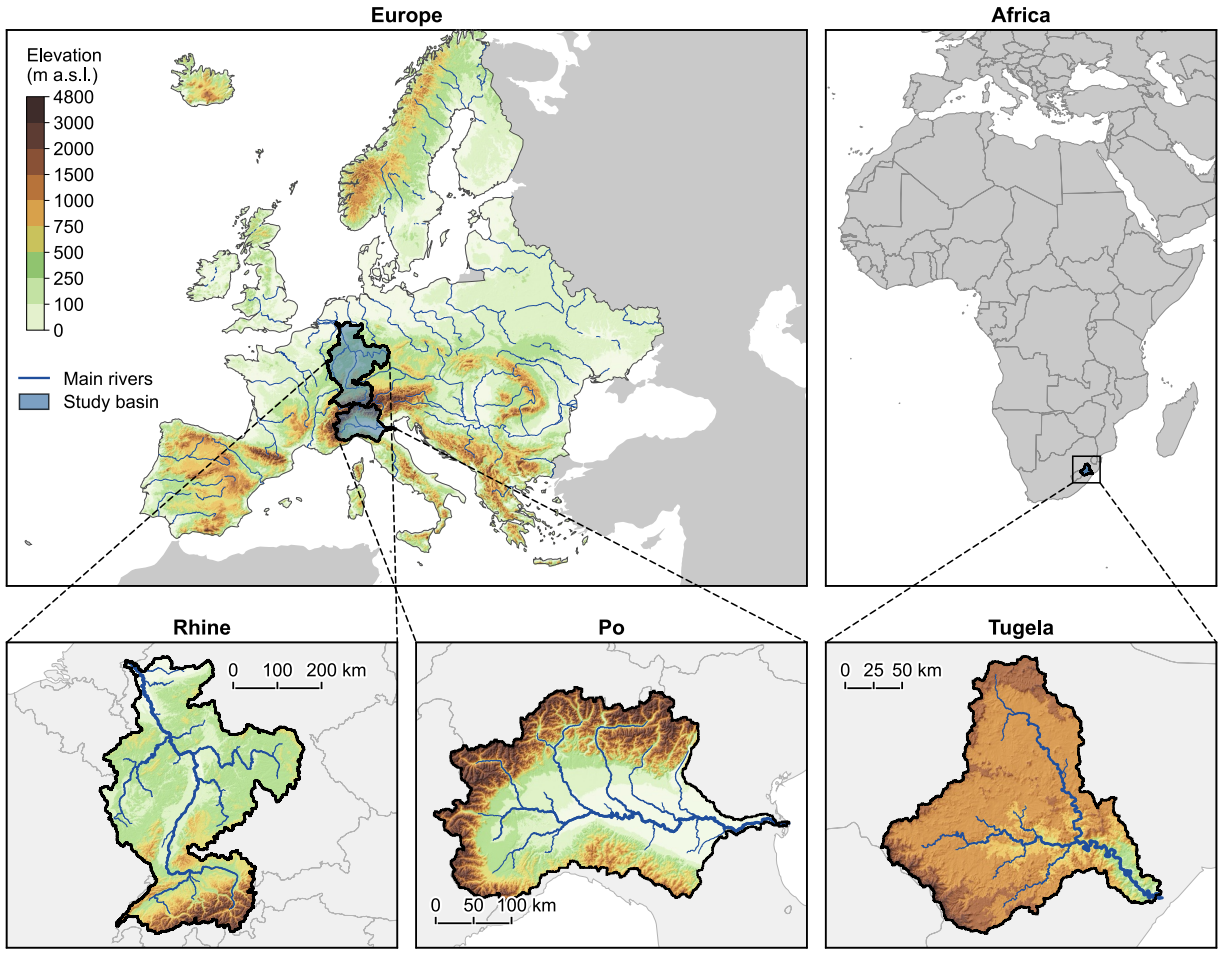


Figure 1. Three study regions used for model benchmarking: The Po River basin (Italy), the Rhine River basin (central Europe, upstream of the Netherlands), and the Tugela River basin (South Africa). The map shows the spatial extent of each region.

2.3. Experimental Design and EO Integration Strategy

We designed a suite of experiments to quantify the added value of different EO constraints in kilometer-scale hydrological simulations (Table 2). The framework progresses stepwise from baseline simulations using conventional forcing data to EO-forced and EO-constrained calibration experiments, allowing the contribution of individual EO data streams and their combinations to be evaluated separately.

Table 1
Summary of the EO and Atmospheric Reanalysis Data Sets Used for Precipitation Forcing, Calibration and Evaluation

Data set	Variable type	Resolution	Primary source
CHIRPS	Precipitation Forcing	≈5 km/Daily	Funk et al. (2015)
EMO-1[a]	Precipitation Forcing	≈1 km/Daily	Salamon et al. (2026)
ERA5[b]	Temperature Forcing	≈30 km/Daily	Hersbach et al. (2020)
HYPER-P	Precipitation Forcing	≈1 km/Daily	Filippucci et al. (2025)
MODIS Eurac	Snow Cover Area Target	≈250 m/Daily	Notarnicola et al. (2013)
MODIS Terra	Evapotranspiration Target	≈1 km/8-Day	Running et al. (2021)
NSIDC SMAP	Soil Moisture Target	≈1 km/Daily	Lakshmi and Fang (2023)

[a]For JULES, EMO 6-hr sub-daily precipitation fields were utilized instead of daily aggregates to preserve sub-daily process features. [b]For JULES, additional sub-daily fields (temperature, humidity, radiation fields, wind speed, and pressure) were extracted from ERA5-Land. CLM also integrated specialized ancillary data sets.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2
Experimental Design Definition, Including the Category, Meteorological Data Set, Variables, and Objective Functions

Category	Experiment	Precipitation forcing	Calibration strategy & targets
Baseline	exp10	EMO-1/CHIRPS[a]	Forward run; no calibration.
	exp20/21	EMO-1/CHIRPS[a]	Benchmarked Q calibration[c] (≈5 km for exp20; ≈1 km parameter transfer for exp21).
	exp30	HYPER-P	Forward run; no calibration.
EO-only[b]	exp31	HYPER-P	SM
	exp32	HYPER-P	ET
	exp33	HYPER-P	SCA
	exp34	HYPER-P	SM + ET
EO + Q[b]	exp41	HYPER-P	SM + Q
	exp42	HYPER-P	ET + Q
	exp44	HYPER-P	SM + ET + Q

[a]Meteorological forcing: EMO-1 for European basins (Rhine, Po), CHIRPS for Tugela. [b]EO targets scored with SPAEF. [c]Calibration resolution, parameter transfer and transfer mechanism are given for each model in Table C1 (Appendix C).

The baseline configuration (`exp10`) uses default parameterizations, while discharge-calibrated experiments (`exp20/21`) retain the same meteorological forcing but optimize parameters against observed streamflow. The EO-forcing experiment (`exp30`) isolates the role of precipitation by replacing conventional precipitation forcing with EO-based precipitation, evaluated in forward, uncalibrated mode.

Subsequent EO-constrained calibration experiments (`exp31–exp34`) use EO-based precipitation forcing and integrate soil moisture, evapotranspiration, and snow cover area, alone or in combination, as calibration targets. In these cases, model parameters are adjusted to minimize mismatches between simulated and observed EO variables, while using the same EO precipitation forcing as `exp30`. Finally, the joint-calibration experiments (`exp41/42/44`) combine EO constraints with observed discharge to evaluate the additional benefit of discharge information in EO-constrained calibration.

To guide the reader through the experiments, we define a hierarchy of calibration configurations. At the highest level, the experiments are grouped into three categories: baseline and non-EO-calibrated simulations (`exp10–exp30`), EO-only calibration (`exp31–exp34`), and joint EO + Q calibration (`exp41–exp44`). Within these categories, configurations optimizing one variable (`exp31`, `exp32`, and `exp33`) are classified as univariate calibrations, configurations optimizing two variables (`exp34`, `exp41`, and `exp42`) as bivariate calibrations, and the simultaneous optimisation of soil moisture, evapotranspiration, and discharge in `exp44` as multivariate joint calibration.

The objective is not to derive a universally optimal EO calibration strategy, but to evaluate how different EO constraints interact with model structure and basin characteristics within a consistent benchmarking framework.

2.4. Hydrological Model Ensemble and Experimental Availability

This study employs a diverse ensemble of five hydrological and land-surface models to assess the value of EO data across different model structures and parameterisation schemes (Droppers et al., 2024; Modiri et al., 2025). The multi-model approach allows us to consider structural uncertainties inherent in hydrological modelling and helps identify model-dependent and basin-dependent responses. Rigorous data quality control was implemented, requiring complete daily simulation coverage for the period 2007–2020, with no missing dates. The models used are listed below, along with a brief description. A comprehensive technical description of model implementations, parameterisations, and benchmarking procedures is provided in Modiri et al. (2025). That companion study is a multi-scale intercomparison of these and further land surface and hydrological models with harmonised forcing in the same three basins; it benchmarks their discharge performance and the consistency of their fluxes across spatial resolutions from about 10 to 1 km, and so describes the state of the models before EO data are introduced. Its discharge-calibrated configurations (`exp20/21`) serve as a baseline here. Each model was configured and

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
Model-Experiment Availability Matrix

Model	exp10	exp20/21	exp30	exp31	exp32	exp33	exp34	exp41	exp42	exp44
CLM[a]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
JULES[a]	✓	✓	✓	✓	✓	–	✓	–	–	–
mHM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
PCR-GLOBWB	✓	✓	✓	✓	✓	–	✓	–	–	–
TETIS[a]	✓	✓	✓	✓	✓	–	✓	✓	✓	✓
Total per exp	**5**	**5**	**5**	**5**	**5**	**2**	**5**	**3**	**3**	**3**

Note. Check marks indicate the 41 available combinations across five hydrological models and 10 experimental configurations. [a]indicates basin-specific availability. CLM and JULES were not applied in the Tugela basin; TETIS was not applied in the Rhine basin.

calibrated by its own development team. Differences in performance between models therefore reflect both model structure and team-specific choices in configuration and calibration (Melsen et al., 2019), and should be read with this in mind.

Community Land Model (CLM) integrated within the Community Climate System Model (CCSM3), simulates comprehensive land surface processes including biophysics, hydrology, and vegetation dynamics (Dickinson et al., 2006; Niu et al., 2005). Version 3.5 incorporates critical hydrological enhancements, including a TOPMODEL-based runoff parameterisation and explicit frozen soil processes.

Joint UK Land Environment Simulator (JULES) is a process-based land surface model which is designed to simulate coupled water, energy and carbon cycle interactions across multiple spatial scales (Best et al., 2011; Clark et al., 2011). Its extensive application in large-scale water cycle studies, including the FP6-WATCH and eartH2Observe projects, and regional assessments (Martínez-De La Torre et al., 2019), demonstrates its capability for continental-scale hydrological applications.

The mesoscale Hydrologic Model (mHM) represents a grid-based distributed hydrological model that simulates the complete terrestrial water cycle from rainfall interception to flood routing (Kumar et al., 2013; Samaniego et al., 2010). Its Multiscale Parameter Regionalisation (MPR) scheme is designed to transfer parameters across spatial scales, which makes it suitable for multi-resolution applications.

PCR-GLOBWB is a global-scale grid-based hydrological model that simulates daily water fluxes while explicitly incorporating anthropogenic water use, including irrigation and domestic demands (van Beek & Bierkens, 2009; Wada et al., 2014). Its extensive application in global water resource assessments using default parameterisations provides a contrasting large-scale model structure for evaluating EO-constrained calibration.

TETIS is a spatially distributed ecohydrological model developed by the Instituto de Ingeniería del Agua y Medio Ambiente (IIAMA) of the Universitat Politècnica de València (UPV) since 1995, which accounts for spatial heterogeneity in its inputs, parameters, and state variables. The model simulates the vertical and horizontal movement of water and has been tested under various spatial scales and climatic conditions worldwide, making it suitable for a range of hydrological applications (Francés et al., 2007; Vélez et al., 2009).

The experimental design encompasses 41 distinct model-experiment combinations across the five-model ensemble, as depicted in Table 3. All simulation outputs are publicly accessible via the open science catalogue (https://4dhydro.eu/catalog/). This matrix reflects the varying capabilities and computational constraints of different models, with some supporting all experimental configurations while others are limited to specific calibration approaches. This uneven availability is therefore accounted for when interpreting inter-experiment and inter-model comparisons (Modiri et al., 2025).

3. Methods

3.1. Benchmarking Protocol and Performance Metrics

Modeled discharge is benchmarked against observed streamflow at a daily time step to quantify predictive skill. We require a minimum overlap of 2 years between simulated and observed records to ensure robustness:

$$T_o \geq 730 \text{ days}, \quad \text{where } T_o \text{ is the overlap period.} \tag{1}$$

Daily performance is primarily evaluated with the Kling–Gupta Efficiency (KGE; Gupta et al., 2009) and its three components: correlation, variability ratio, and bias ratio.

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \tag{2}$$

$$r = \text{corr}(Q_s, Q_o), \quad \alpha = \frac{\sigma_s}{\sigma_o}, \quad \beta = \frac{\mu_s}{\mu_o}, \tag{3}$$

where Q_s and Q_o are simulated and observed discharge, r is the Pearson correlation coefficient, α is the ratio of standard deviations, β is the ratio of means, μ is the mean and σ the standard deviation.

3.2. Performance Contrasts and Classification

We analyze performance contrasts (ΔKGE) alongside absolute benchmark skill and component-wise attribution metrics to compare experiments at paired gauges. Contrasts are defined as paired differences in KGE at the gauge level, with a target experiment expX compared against a reference experiment ref, in order to isolate forcing or calibration effects. These contrasts are interpreted strictly alongside absolute KGE baselines to avoid overstating relative improvements in poorly performing basins. Because ΔKGE alone is difficult to interpret physically, decomposed KGE components (r, α^*, β^*) are subsequently used to attribute the dominant hydrological drivers of performance change.

$$\Delta\text{KGE} = \text{KGE}_{\text{expX}} - \text{KGE}_{\text{ref}}. \tag{4}$$

Two reference frameworks are considered: ref = `exp10` to quantify the overall added value relative to baseline simulations, and ref = `exp30` to isolate the effect of calibration while holding the EO precipitation forcing constant.

For each gauge i, the paired contrast ΔKGE_i is computed. Following previous benchmarking studies (Knoben et al., 2019), we set a tolerance $t = 0.10$ to represent a practically meaningful change in performance. Gauges are classified as

$$C(\Delta\text{KGE}) = \begin{cases} \text{better}, & \Delta\text{KGE} > t, \\ \text{same}, & |\Delta\text{KGE}| \leq t, \\ \text{worse}, & \Delta\text{KGE} < -t. \end{cases} \tag{5}$$

This *tolerance band* ($\pm t$) is used to assign qualitative labels in the 1 : 1 scatter summaries. For the complementary win-rate analysis, we apply the strict sign of the contrasts without a tolerance band:

$$\mathcal{B} = \{i : \Delta\text{KGE} > 0\}, \quad \mathcal{S} = \{i : \Delta\text{KGE} = 0\}, \quad \mathcal{W} = \{i : \Delta\text{KGE} < 0\}, \tag{6}$$

with percentages

$$p_{\text{better}} = \frac{100\,|\mathcal{B}|}{N}, \qquad p_{\text{same}} = \frac{100\,|\mathcal{S}|}{N}, \qquad p_{\text{worse}} = \frac{100\,|\mathcal{W}|}{N}, \tag{7}$$

where N is the number of paired gauges. This provides a complementary assessment of gains and losses independent of tolerance thresholds.

3.3. Earth Observation Calibration Objective

The EO targets used for calibration include SM, ET and SCA. Their evaluation relies on complementary performance metrics. SM and ET are assessed with the SPAtial EFficiency (SPAEF) metric (Koch et al., 2018),

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4
Weights Used in the Unified Objective J: Larger EO Weights Give Greater Emphasis to EO Calibration Targets

Experiment	w_{SM}	w_{ET}	w_{SPF}	w_Q
exp31 (SM)	1	0	0	0
exp32 (ET)	0	1	0	0
exp33 (SPF)	0	0	1	0
exp34 (SM + ET)	1	1	0	0
exp41 (SM + Q)	5	0	0	1
exp42 (ET + Q)	0	5	0	1
exp44 (SM + ET + Q)	2.5	2.5	0	1

which balances spatial correlation, variability, and distributional similarity between model fields X and EO fields Y:

$$\mathrm{SPAEF}(X,Y) = 1 - \sqrt{(\alpha-1)^2 + (\beta-1)^2 + (\gamma-1)^2}, \tag{8}$$

$$\alpha = \mathrm{corr}(X,Y), \qquad \beta = \frac{CV_X}{CV_Y} = \frac{\sigma_X/\mu_X}{\sigma_Y/\mu_Y}, \tag{9}$$

$$\gamma = \sum_b \min\big(p_X(b), p_Y(b)\big). \tag{10}$$

where α is the spatial correlation, β the ratio of coefficients of variation, and γ the histogram intersection between distributions. SPAEF is computed per time slice and averaged over time to yield a single score.

Snow requires a harmonization step because EO products provide SCA while models simulate snow *depth* (SD). Both are converted to a binary snow-presence flag (SPF), where 1 = snow and 0 = no snow:

$$\mathrm{SPF}_X(\mathbf{x},t) = \mathbf{1}[\,\mathrm{SD}_X(\mathbf{x},t) > 0\,], \qquad \mathrm{SPF}_Y(\mathbf{x},t) = \mathbf{1}[\,\mathrm{SCA}_Y(\mathbf{x},t) > 0\,]. \tag{11}$$

The agreement between modeled and observed SPF is then quantified with classification accuracy (CA):

$$\mathrm{CA} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\big(\mathrm{SPF}_X^i = \mathrm{SPF}_Y^i\big), \tag{12}$$

where n is the number of space–time pixels.

Streamflow is evaluated with $\mathrm{KGE}(Q_s, Q_o)$. To enable consistent calibration across all targets, these metrics are combined into a single *unified objective* function:

$$\begin{aligned} J &= 1 - \frac{1}{W}\big(w_{SM}\,\mathrm{SPAEF(SM)} + w_{ET}\,\mathrm{SPAEF(ET)} + w_{SPF}\,\mathrm{CA(SPF)} + w_Q\,\mathrm{KGE}(Q)\big), \\ W &= \sum w_k. \end{aligned} \tag{13}$$

The calibration algorithm minimizes J; therefore, higher component scores correspond to lower objective-function values. Experiment-specific calibration setups are realised by varying the weights $w_k \geq 0$, as detailed in Table 4. This unified objective is used for CLM, mHM and TETIS. PCR-GLOBWB and JULES are calibrated within their own frameworks, which do not optimize Equation 13. PCR-GLOBWB uses differentiable parameter learning (Tsai et al., 2021): a neural network maps static grid-cell attributes to four soil parameters and one groundwater parameter and is trained by minimizing the mean squared error (MSE) between simulated and observed values. A deep-learning surrogate of PCR-GLOBWB provides the gradients needed to train the network (Droppers et al., 2025). For `exp20` the network was trained against discharge over 2006–2008, with a 1-year warm-up, at 3′ (≈5 km), and the learned parameters were then applied at 30″ (≈1 km). JULES uses a data-

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

assimilation framework that minimizes a cost function with two terms, one that penalizes departures of the parameters from their default values and one that measures the misfit to the observations. For soil moisture, this framework adjusts the parameters of the soil pedotransfer functions; for evapotranspiration, which is relatively insensitive to the soil parameters in JULES, it adjusts the ratio of the roughness length for heat to that for momentum.

For example, SM and ET are equally weighted in `exp34`, while EO targets are emphasized relative to discharge in mixed EO + Q experiments, with ratios of 5 : 1 (`exp41/42`) or 2.5 : 2.5 : 1 (`exp44`).

The objective function weights for the joint calibration experiments were chosen to impose a clear hierarchy of optimisation priorities, following established multi-objective configurations in satellite-streamflow integration (Koch et al., 2018). Higher weights assigned to EO targets emphasize high-resolution spatial and temporal patterns. (Q) is retained as a discharge constraint to preserve hydrological plausibility and water-balance consistency without dominating the EO constraints. This asymmetric weighting is a deliberate design choice aligned with the aim of testing EO data utility. Alternative weighting schemes would explore different trade-offs and may lead to different parameter sets; therefore, the interpretation of joint-calibration performance is conditional on the weighting structure used here.

3.4. Component Analysis and Normalization

To attribute performance differences to individual KGE components, we analyze both raw and normalized values. Normalization ensures comparability across components by mapping ratio terms to a bounded, symmetric *closeness-to-one* scale (0, 1]:

$$\alpha^* = C(\alpha) = \exp(-|\ln \alpha|), \tag{14}$$

$$\beta^* = C(\beta) = \exp(-|\ln \beta|). \tag{15}$$

This transformation achieves its optimum of 1 when the ratio equals unity, decays smoothly and symmetrically for both overestimation and underestimation, and remains bounded away from zero. Because the correlation component is already bounded, we retain it in its original form.

The component metrics for attribution analysis are then computed as:

$$\Delta r = r_{\text{expX}} - r_{\text{ref}}, \quad \Delta\alpha^* = \alpha^*_{\text{expX}} - \alpha^*_{\text{ref}}, \quad \Delta\beta^* = \beta^*_{\text{expX}} - \beta^*_{\text{ref}}. \tag{16}$$

3.5. Statistical Analysis and Effect Sizes

All statistical inference is performed on paired gauge-level contrasts. For each region, model, and experiment combination, we report the median ΔKGE and its interquartile range (IQR) across gauges. The Wilcoxon signed-rank test (Wilcoxon, 1945) is applied to assess whether ΔKGE differs significantly from zero (two-sided), and corresponding confidence intervals are calculated. The test treats all gauges as independent, although nested gauges share upstream area, and the p-values are not corrected for multiple comparisons; they are therefore used as descriptive indicators rather than confirmatory tests (Wilks, 2016). To attribute changes to individual KGE components, we further employ ordinary least squares (OLS) regression of the form $\Delta\text{KGE} = a + b\Delta c + \epsilon, \quad c \in \{r, \alpha^*, \beta^*\}$.

Statistical significance is indicated using conventional notation:

$$S(p;N) = \begin{cases} \text{(not shown)}, & N < 10, \\ ***, & p < 10^{-3}, \\ **, & 10^{-3} \le p < 10^{-2}, \\ *, & 10^{-2} \le p < 0.05, \\ \text{ns (non-significant)}, & p \ge 0.05. \end{cases} \tag{17}$$

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.6. Data Processing and Experiment Design

All contrasts are computed for paired gauges, requiring that the same station be present in both the reference and target experiments; unpaired records are excluded.

For analysis, `exp20` (Q calibration at ≈5 km) and `exp21` (parameter transfer to ≈1 km) are combined and reported jointly as `exp2`. How each model reaches the ≈1 km evaluation scale in `exp2` is summarized in Table C1 (Appendix C). mHM, TETIS, PCR-GLOBWB and JULES are calibrated at ≈5 km and their parameters are then used at ≈1 km. mHM transfers them with its MPR framework, which supports scale-consistent parameter transfer across resolutions, and TETIS with a conceptual scale-effective parameter approach. PCR-GLOBWB and JULES apply the calibrated parameters at the finer grid without a dedicated scaling scheme. CLM is calibrated directly at its ≈1 km grid, so no parameters are transferred. Direct transfer of parameter sets from coarser to finer resolutions can affect routing, infiltration, and other process representations in conceptual or semi-distributed models. Consequently, models without explicit scale-consistent formulations may show structural scale effects when parameters are transferred across grid scales; this distinction is considered when interpreting multi-model performance contrasts.

4. Results and Discussion

4.1. Overall Performance

The evaluation across the three study regions reveals distinct patterns in how EO data integration affects hydrological model skill, with strong basin dependence and clear model-specific responses (Figure 2).

In the Mediterranean-influenced Po basin, mHM consistently achieves the highest skill, with median KGE values between 0.4 and 0.6 across most configurations. Its performance peaks remarkably in the snow-related experiment (`exp33`: SCA) and remains strong under joint EO + Q calibration configurations (`exp41/42/44`), indicating that process-relevant EO constraints can improve model performance. In this case, snow dynamics in Alpine-fed headwaters appear to provide an informative constraint on runoff generation. TETIS shows more variable behavior but performs well under bivariate joint calibration targets (`exp41/42`), while CLM demonstrates modest but steady improvements. The transition to EO precipitation forcing (`exp30`) produces contrasting model responses in the Po basin. JULES shows the clearest positive response to EO precipitation, with nearly 60% of gauges improving and stronger variability representation under Mediterranean conditions. The paired median contrast reaches $\Delta\mathrm{KGE} = +0.14$. CLM exhibits more moderate basin-wide gains (+0.04), while TETIS remains close to neutral (−0.01), indicating limited sensitivity to forcing replacement alone. mHM maintains comparatively stable behavior, with more than half of gauges remaining within the no-material-change band and a modest median contrast of −0.04, suggesting that forcing-only replacement neither strongly improves nor strongly degrades performance without recalibration. In contrast, PCR-GLOBWB shows systematic performance degradation under EO forcing (−0.17, 63% worse). This degradation is likely the result of the model's relatively simple evapotranspiration scheme and related precipitation sensitivity. More broadly, across the entire ensemble, the direct transfer of rigid parameters from coarser to finer resolutions (5 km to 1 km) may introduce inherent challenges. For any model structure that operates without explicit scale-invariant parameter estimation or effective regionalization routines, such cross-scale transfer creates structural-scale artifacts. The abrupt resolution shift alters the spatial topology of routing pathways and the effective representation of sub-grid soil-column processes; consequently, relying on unscaled parameter sets across mismatched grids introduces systemic performance penalties that reflect resolution incompatibility rather than the model's true process-representation capabilities.

The Rhine presents a more challenging case for EO forcing integration. Here, mHM again delivers the most consistent overall skill, with median KGE values generally exceeding 0.5 across configurations. However, replacing EMO-1 with EO precipitation consistently degrades performance across all models. Rather than a simple failure of satellite data quality, this response exposes a classic manifestation of compensating-error dynamics. Because the dense, gauge-based EMO-1 network already provides highly refined precipitation inputs in this data-rich basin, forcing the models to calibrate using comparatively less accurate EO precipitation (`exp31–34`) may therefore introduce forcing errors that the models partly compensate for through changes in soil and routing parameters. CLM shows the strongest performance drop (median $\Delta\mathrm{KGE} = -0.25$, 62% of gauges worsening), while JULES and PCR-GLOBWB also decline substantially. mHM remains comparatively stable, with nearly

10.1029/2025WR042446

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

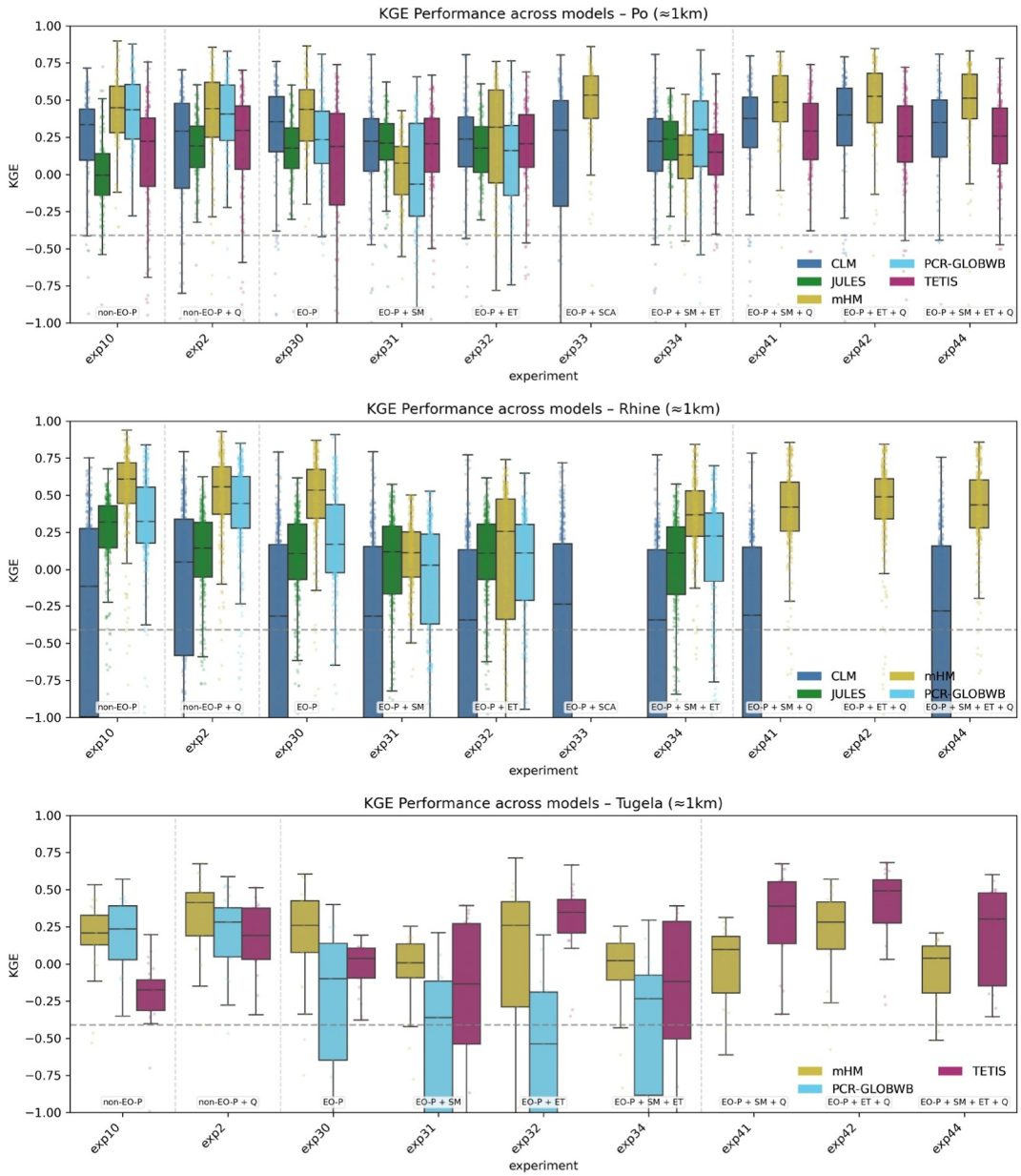


Figure 2. KGE performance across hydrological models and EO integration strategies at ≈1 km resolution for the Po, Rhine, and Tugela basins. Box plots show KGE distribution across gauges for each model and experiment. Colors indicate models; availability varies by basin and experiment. Experiments progress from the non-EO baseline (`exp10`), through discharge calibration and parameter transfer (`exp2`), EO precipitation forcing (`exp30`), EO-only calibration (`exp31`–`exp34`), and joint EO + Q calibration (`exp41`, `exp42`, and `exp44`).

half of the gauges remaining within the no-material-change band and a more moderate median contrast (−0.06). Overall, the limited improvement under EO-only calibration suggests that, in well-instrumented basins such as the Rhine, replacing high-quality gauge-based forcing with EO precipitation can introduce errors that calibration alone cannot fully compensate for.

In the data-scarce Tugela basin, EO integration produces the clearest relative improvements, although the absolute predictive skill remains lower than in the European basins (baseline performance is generally weak (median KGE < 0.5)). TETIS emerges as the strongest performer, particularly in evapotranspiration-focused experiments (`exp32`, `exp42`), with median KGE values approaching 0.5 and a clear dominance of better-performing gauges under EO-constrained calibration. The shift to EO precipitation forcing alone already

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

substantially benefits TETIS, with 71% of gauges improving and a paired median contrast of $\Delta KGE = +0.19$. However, this marked performance increase highlights a critical distinction between EO data acting as a fine-scale refinement and a macro-scale volumetric rescue action to restore a broken baseline. In the convective, semi-arid regime of the Tugela, actual evapotranspiration heavily dominates the annual water balance. Because the baseline KGE for this basin sits notably low—often dipping below zero—the initial a priori parameter sets were fundamentally detached from the catchment's physical reality. Rather than indicating highly precise fine-scale spatio-temporal optimisation, this large relative improvement demonstrates the capacity of satellite-derived data to provide a coarse corrective lens, rescuing an otherwise unconstrained baseline in data-scarce regions. Conversely, mHM remains comparatively neutral, whereas PCR-GLOBWB responds negatively to nearly all EO-constrained configurations, again showing the impact of its relatively simple evapotranspiration scheme in this highly arid region. Joint EO + discharge calibration (`exp41/42/44`) yields the largest overall gains, where combining ET constraints with discharge provides vital structural corrections to unconstrained parameters rather than the incremental fine-tuning of already stable model configurations.

Across regions, the results highlight distinct model capabilities. Across the experiments tested here, mHM shows the most stable performance in the three basins, which is consistent with the scale-aware parameter regionalisation of its MPR scheme. TETIS demonstrates strong adaptability, serving as an effective corrective tool in subtropical and Mediterranean settings by rescuing poorly constrained baseline parameters during calibration with EO targets. CLM and JULES show moderate gains, primarily in the Po basin, whereas PCR-GLOBWB results show large forcing sensitivities. For PCR-GLOBWB, the forcing response (`exp30`) is obtained with its default parameters, while its calibrated results come from a different calibration framework (Section 3.3); its ranking relative to the other models should therefore be read with care.

Taken together, the results demonstrate that the effectiveness of EO-constrained calibration is strongly conditioned by both hydroclimatic setting and model structure. Rather than identifying a single universally superior configuration, the framework reveals how different EO targets influence distinct hydrological processes across contrasting environments. In the Po basin, snow-related constraints and joint calibration strategies were most beneficial. In contrast, in the Rhine, replacing high-quality gauge-based forcing proved difficult without additional compensatory calibration. In the semi-arid Tugela basin, evapotranspiration constraints combined with discharge produced the clearest improvements. These contrasting responses highlight the importance of process relevance and regional data characteristics when designing EO-constrained calibration strategies. Component attribution confirms that variability alignment drives most improvements, with bias correction contributing to TETIS gains, while correlation plays a secondary role. This diagnostic decomposition suggests that while hyper-resolution remote sensing constraints can assist in regularising volumetric water storage variations and bounding vertical flux envelopes, such as evapotranspiration and upper-layer soil moisture, their influence may be constrained by underlying routing configurations. Incorporating EO data tends to adjust regional mass volumes and spatial variability scales; however, it appears less effective at mitigating systematic timing discrepancies that may stem from simplified hydraulic routing formulations or uncalibrated river network propagation velocities. Consequently, these remote sensing constraints seem primarily suited for adjusting hydrological mass distributions rather than resolving downstream routing and timing characteristics.

Overall, EO data integration can enhance kilometre-scale hydrological modelling when aligned with basin hydroclimatic characteristics and model structure. This conditional picture is in line with earlier studies that calibrated single models against remotely sensed snow cover (Parajka & Blöschl, 2008), evapotranspiration and soil moisture (Kunnath-Poovakka et al., 2016; Rajib et al., 2016) or spatial patterns of several satellite products (Dembélé et al., 2020): the constrained variables were generally better represented, whereas the gains in streamflow were smaller and depended on the catchment and on whether discharge was kept in the objective (Finger et al., 2015; Mei et al., 2023; Nijzink et al., 2018). Within this experimental design, joint calibration strategies provide the most reliable pathway for improvement in the data-scarce Tugela, while snow-affected basins such as the Rhine require more robust compensatory approaches before EO forcing can be safely adopted. A detailed summary across all configurations is provided in Appendix Figure A1.

Given that aggregated ΔKGE spatial maps can mask compensating errors among physical sub-components, our metric narrative follows a strict multi-component hierarchy. After presenting the aggregate overview, Section 4.5 and Appendix B explicitly deconstruct these findings. We map the independent movements of timing, variability,

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 5
Gauge-Level EO-P Forcing Responses Across the Study Basins (exp30 vs. exp10), Summarized Using Paired Median Changes and Better/Same/Worse Classification Statistics

Basin	Model	Median ΔKGE	% Better	% Same/% Worse
Po	CLM	+0.04	33	46/21
	JULES	+0.14	59	29/12
	mHM	−0.04	19	53/28
	PCR-GLOBWB	−0.17	6	31/63
	TETIS	−0.01	31	30/39
Rhine	CLM	−0.25	16	22/62
	JULES	−0.15	11	32/58
	mHM	−0.06	14	48/38
	PCR-GLOBWB	−0.11	5	41/54
Tugela	mHM	−0.02	43	19/38
	PCR-GLOBWB	−0.36	14	10/76
	TETIS	+0.19	71	19/10

Note. Threshold for *same*: $|\Delta KGE| \leq 0.10$.

and bias via regressions (Figure 7) and geometric radar profiles (Figure 8) to pinpoint exactly which physical processes drive the aggregated performance shifts.

4.2. Forcing Data Impact

The following analysis combines paired performance contrasts with gauge-level classification statistics and component attribution to distinguish operational performance variations from their underlying hydrological drivers. Figure 3 summarizes how precipitation forcing influences model skill and points to the variable utility of EO precipitation. The dashed separators indicate the progression from baseline to EO forcing (`exp30`), EO-only calibration (`exp31`–`exp34`), and joint EO + Q calibration (`exp41/42/44`). The `exp30` nodes isolate the forcing-only effect, while subsequent segments show whether calibration with EO targets tends to offset or reinforce the initial response.

In the Po basin, the curves are separated by model at `exp30`. JULES shows notable gains (median $\Delta KGE = +0.14$), CLM and TETIS remain near neutral to slightly positive, mHM dips modestly (−0.04), and PCR-GLOBWB declines. Within the EO-only calibration family, JULES continues to show incremental increases, CLM oscillates around zero, mHM declines most when constrained by soil moisture (`exp31/34`), and PCR-GLOBWB remains negative. Once discharge is added (EO + Q), TETIS shifts to a positive value, mHM returns toward neutral, CLM stays close to zero, and PCR-GLOBWB shows minor adjustments but remains lower than the other configurations. The Rhine basin shows a different pattern. At `exp30`, all models shift below zero, with PCR-GLOBWB experiencing the largest drop (−0.11), CLM and JULES showing moderate declines, and mHM the smallest (−0.06). EO-only calibration (`exp31-34`) does not resolve the deficit; for mHM, the decline deepens before easing slightly. Joint EO + Q calibration produces only modest recovery, most notably for mHM, but the curves remain negative overall. This suggests that forcing replacement can be detrimental in the Rhine's nival–pluvial regime unless alternative compensatory strategies are applied.

In the Tugela, forcing replacement produces more favorable outcomes for specific structures. At `exp30`, TETIS improves (median $\Delta KGE = +0.19$, with 71% of gauges improving), mHM lies slightly below zero, and PCR-GLOBWB is negative (−0.36). EO-only calibration further supports TETIS, while mHM drifts more negatively before stabilizing. Under EO + Q calibration, TETIS retains its positive median values across the basin, whereas mHM returns to neutral, and PCR-GLOBWB remains low. This trajectory indicates that the utility of EO precipitation in semi-arid, data-scarce basins is dependent on the model configuration.

The configuration of the baseline precipitation forcing helps explain these contrasts. In the Rhine, EMO-1 achieves relatively high skill due to dense gauge coverage, limiting the opportunity for satellite-derived products to add structural value. In the Po, EO precipitation may provide localized adjustments, particularly near the

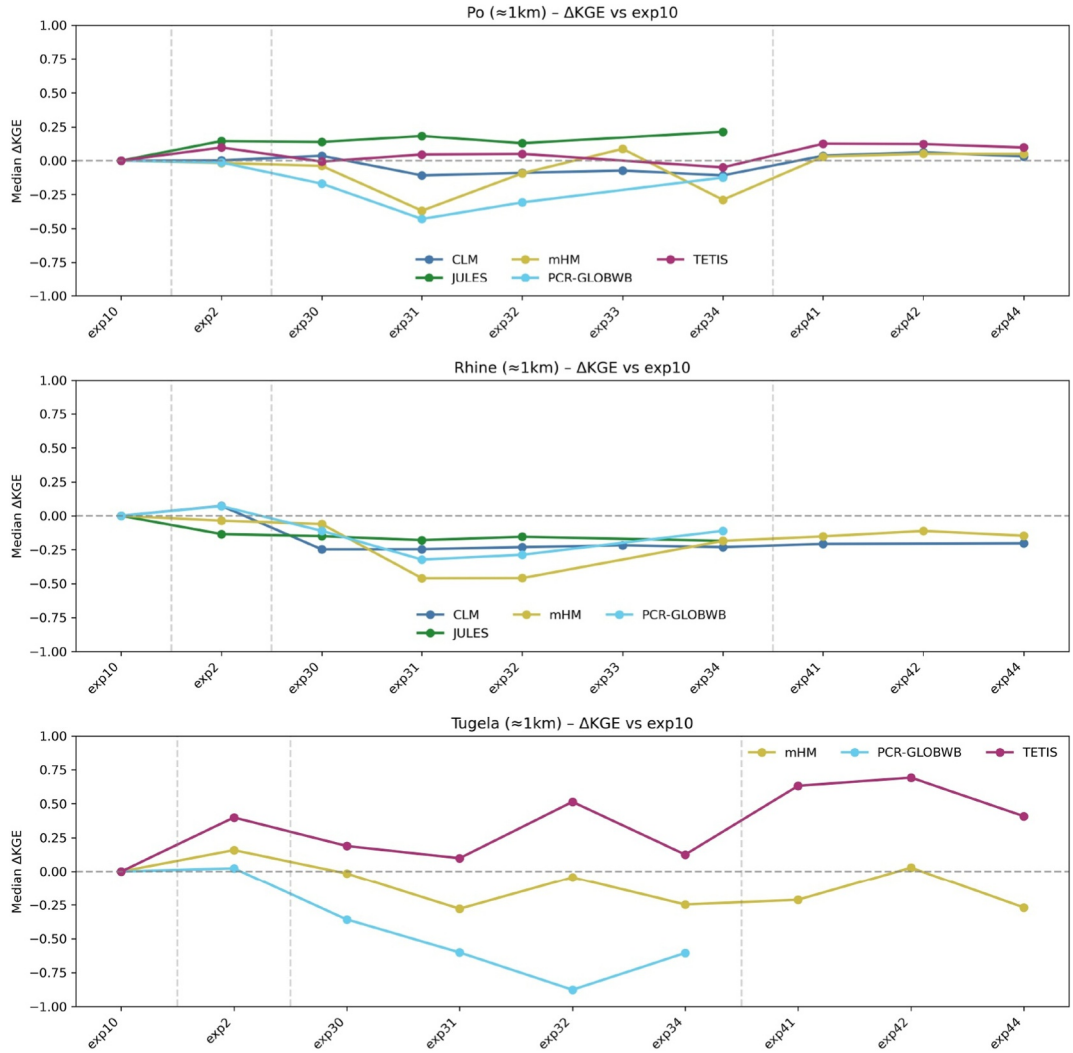


Figure 3. Median change in performance (ΔKGE) relative to the `exp10` baseline at ≈1 km resolution for the Po, Rhine, and Tugela basins. Solid lines show basin-wide median ΔKGE for each model across experiments, and markers indicate individual configurations. The horizontal dashed line marks no change relative to `exp10`; positive and negative values indicate improvement and deterioration, respectively. Vertical dashed lines separate experiment families: discharge calibration and parameter transfer (`exp2`), EO precipitation forcing (`exp30`), EO-only calibration (`exp31`–`exp34`), and joint EO + Q calibration (`exp41`, `exp42`, and `exp44`). The `exp30` configuration isolates the forcing-only effect of replacing the conventional precipitation forcing with EO-based precipitation.

river mouth where gauge density is lower. In the Tugela, the CHIRPS-based baseline forcing is more strongly affected by sparse in situ observations, which limit their utility for baseline calibration and reduce baseline skill. Detailed median values and gauge-level classifications for `exp30` are provided in Table 5.

In the Po basin (Figure 4), models show distinct scatter patterns in their response to EO precipitation. CLM and mHM align closely with the 1 : 1 line (Pearson correlations of 0.93 and 0.85, respectively), with the densest hex-bins concentrated between 0.3 and 0.6 on both axes, indicating that the forcing change largely preserves gauge ranking. JULES displays a broader cloud ($r = 0.42$) that is slightly shifted above the 1 : 1 line, consistent with a general uplift in timing and volume metrics. This behavior is linked to JULES's internal representation of canopy interception and surface resistance schemes, which appear capable of effectively utilizing the temporal distribution of the satellite precipitation product. TETIS scatters more widely ($r = 0.59$), reflecting spatially variable forcing effects across varying topographic zones. By contrast, PCR-GLOBWB shows a diffuse pattern ($r = 0.88$) with density concentrated below the 1 : 1 line, pointing to a performance decline under the uncalibrated satellite forcing.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

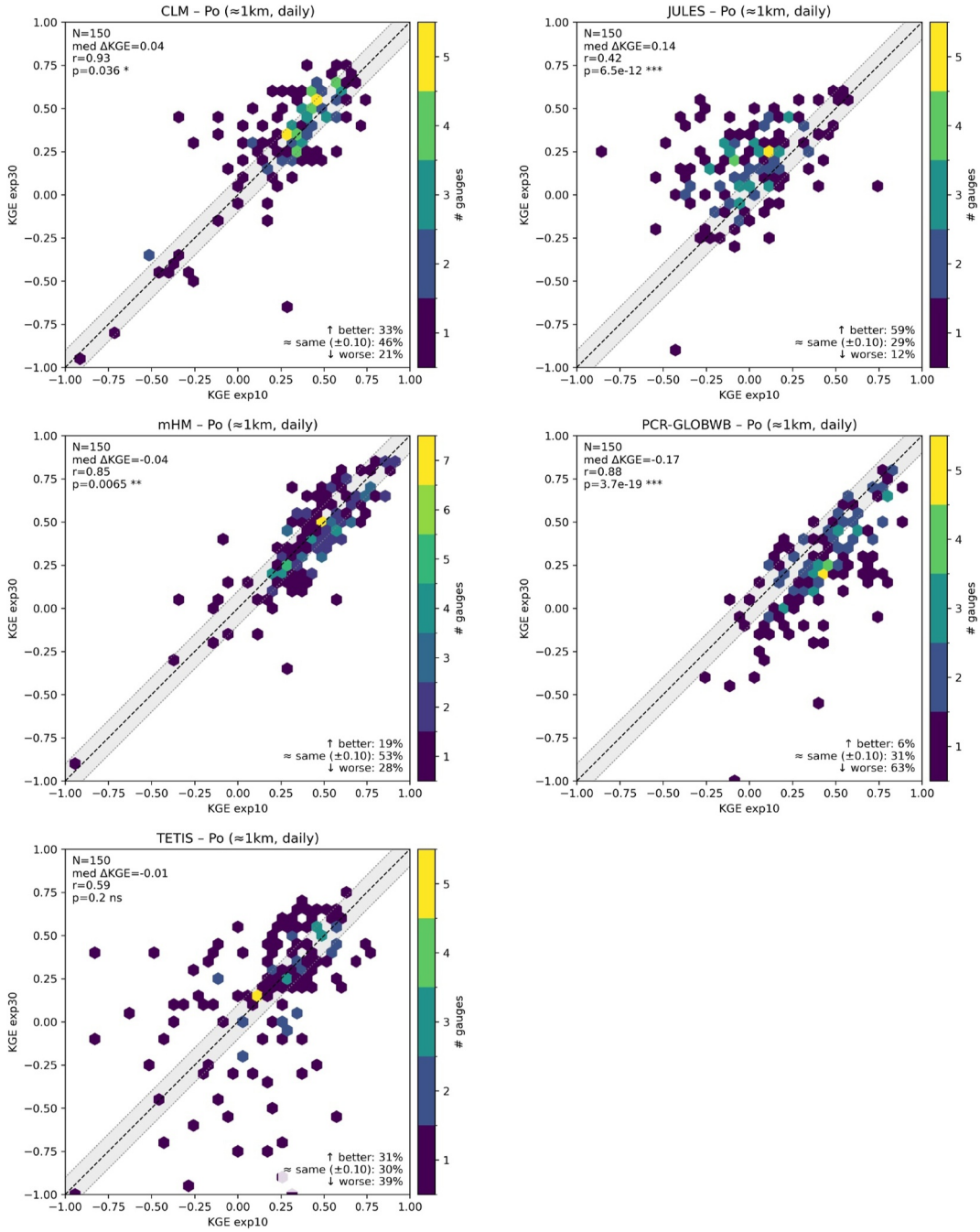


Figure 4. Gauge-wise skill for EO precipitation forcing without recalibration (`exp30`) relative to the baseline configuration (`exp10`) at ≈1 km resolution in the Po basin. Each hexagon aggregates co-located gauges, with color indicating the number of gauges. The black 1 : 1 line indicates no change, and gray dashed lines show the ±0.10 *no-material-change* band. In-panel text reports the number of gauges (N), median change, Pearson correlation, and Wilcoxon p-value with significance codes (***, **, *, ns). Percentages indicate the shares of gauges classified as better, unchanged, or worse using the ±0.10 tolerance.

A basin-wide comparison of these effects is summarized in Table 5, which reports median performance differences and the proportions of gauges classified as *better*, *same*, or *worse* for each model–basin combination.

The cross-basin comparison highlights distinct regional responses to EO precipitation forcings (Table 5).

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

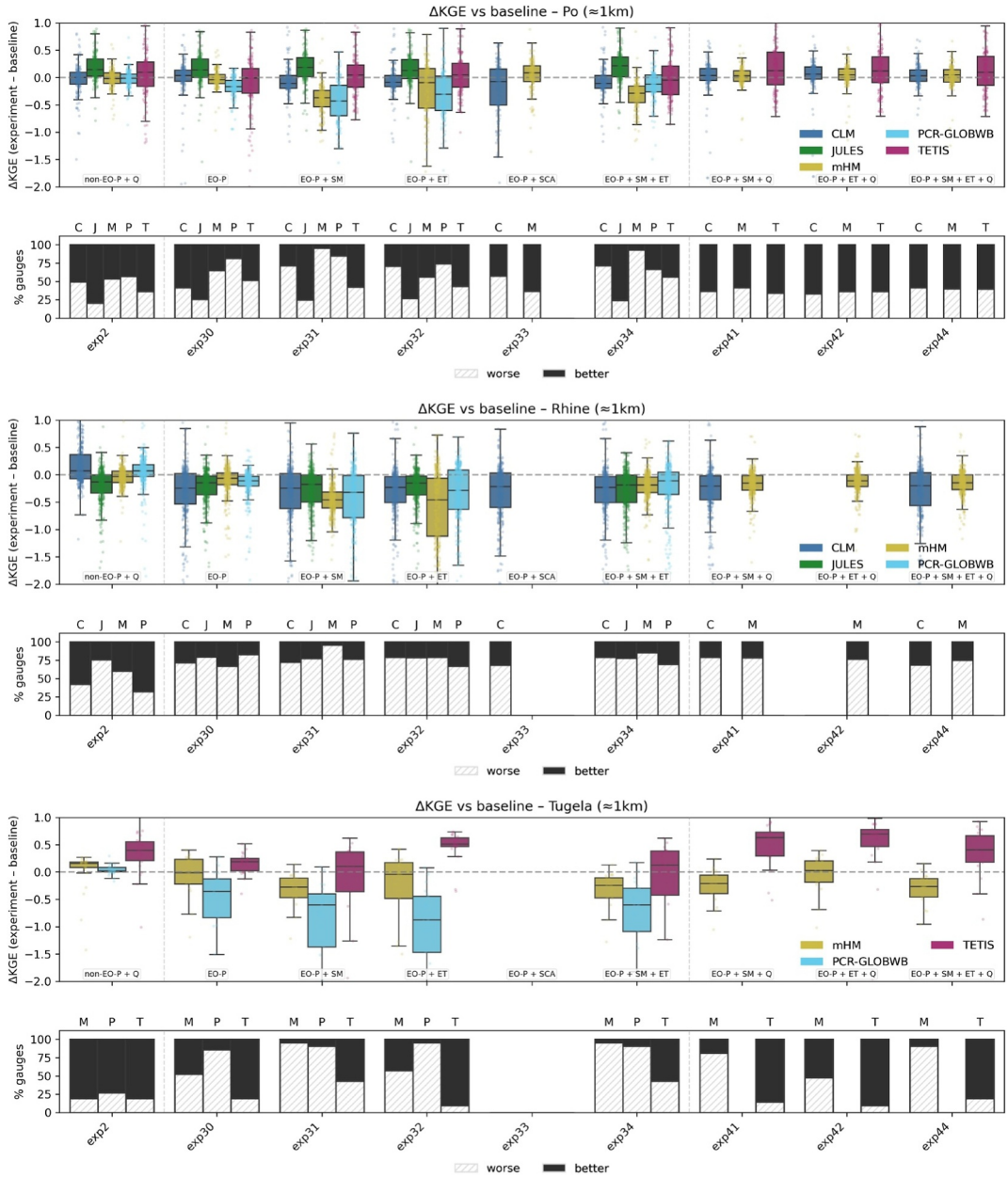


Figure 5. Added value of EO-based calibration relative to the baseline configuration: gauge-wise ΔKGE (`experiment–exp10`) across models at ≈1 km resolution for the Po, Rhine, and Tugela basins. For each experiment, the upper panel shows the distribution of ΔKGE for each model (box: interquartile range; whiskers: central range; dots: gauges), and the lower stacked bars show the percentage of gauges classified as better, unchanged, or worse. Vertical dashed lines separate experiment families: discharge calibration and parameter transfer (`exp20/21`), EO precipitation forcing (`exp30`), EO-only calibration (`exp31–exp34`), and joint EO + Q calibration (`exp41`, `exp42`, and `exp44`). The contrast between `exp31–exp34` and `exp41/exp42/exp44` shows the added value of combining EO constraints with discharge information.

In the Po basin, substituting EMO-1 with merged EO precipitation improves JULES and yields modest gains for CLM. mHM remains near-neutral, indicating that forcing-only replacement does not systematically modify its performance without recalibration. This resilience is consistent with the parameter regionalization of mHM, although it was only tested for the forcing change considered here. TETIS is largely neutral, while PCR-GLOBWB displays a notable downward shift (−0.17). These outcomes demonstrate that in Mediterranean-influenced settings, the underlying model structure and prior parameterization configurations condition the response to EO forcing.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2025WR042446

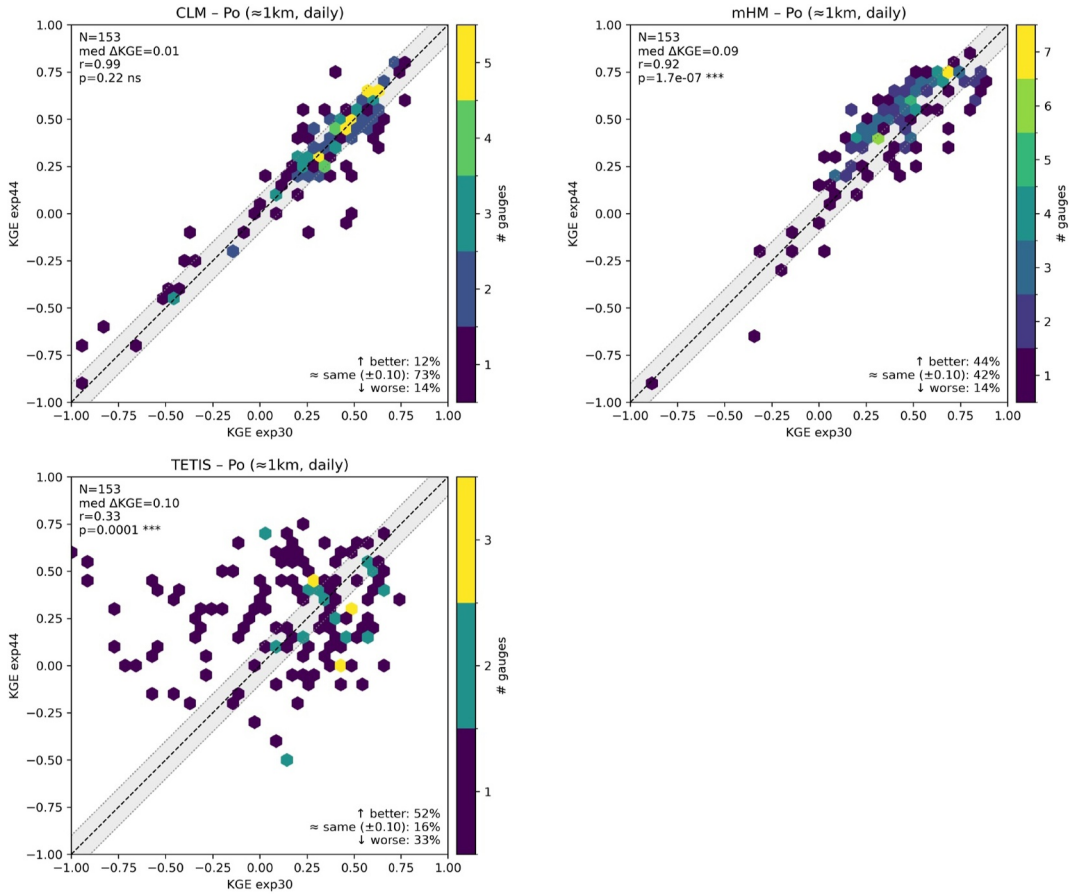


Figure 6. Gauge-wise skill for joint EO + Q calibration (`exp44`: SM + ET + Q) relative to the EO precipitation forcing without recalibration (`exp30`: EO-P) at ≈1 km resolution in the Po basin. Each hexagon aggregates co-located gauges, with color indicating the number of gauges. The black 1 : 1 line indicates no change, and gray dashed lines show the ±0.10 *no-material-change* band. In-panel text reports the number of gauges (*N*), median change, Pearson correlation, and Wilcoxon *p*-value with significance codes (***, **, *, ns). Percentages indicate the shares of gauges classified as better, unchanged, or worse using the ±0.10 tolerance.

In the Rhine basin, EO precipitation forcing leads to a reduction in performance across all models, indicating that forcing replacement alone introduces errors in this nival–pluvial system. CLM shows the largest drop (−0.25), while JULES and PCR-GLOBWB also decline. mHM remains more stable, with nearly half of the gauges classified as unchanged. This generalized decline implies that forcing the models to ingest satellite precipitation fields in a data-rich basin alters internal parameter dynamics. When forced with these inputs, the model algorithms may adjust soil moisture states and routing mechanisms to configurations that are less representative of actual field behavior, simply to match observed outlet hydrographs. This suggests that the dense, gauge-based EMO-1 forcing already characterizes the precipitation patterns in the Rhine basin well, minimizing the opportunity for alternative satellite forcing inputs to improve spatial variability features without introducing parameter distortion.

In the semi-arid Tugela basin, responses to EO precipitation vary by model structure. TETIS shows a positive response to the EO forcing, with the majority of gauges improving, suggesting that satellite products capture convective spatial precipitation features that are under-represented in sparse in situ station networks. Conversely, mHM remains neutral, with a balanced distribution of gains and losses across gauges, reflecting a steady response to changes in meteorologic forcing. PCR-GLOBWB exhibits a downward shift, indicating lower flexibility when utilizing uncalibrated EO precipitation fields under high-intensity semi-arid regimes. These results indicate that EO precipitation can improve performance metrics in data-scarce subtropical systems, provided the model formulation can capture spatial information without introducing internal flux mismatches.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

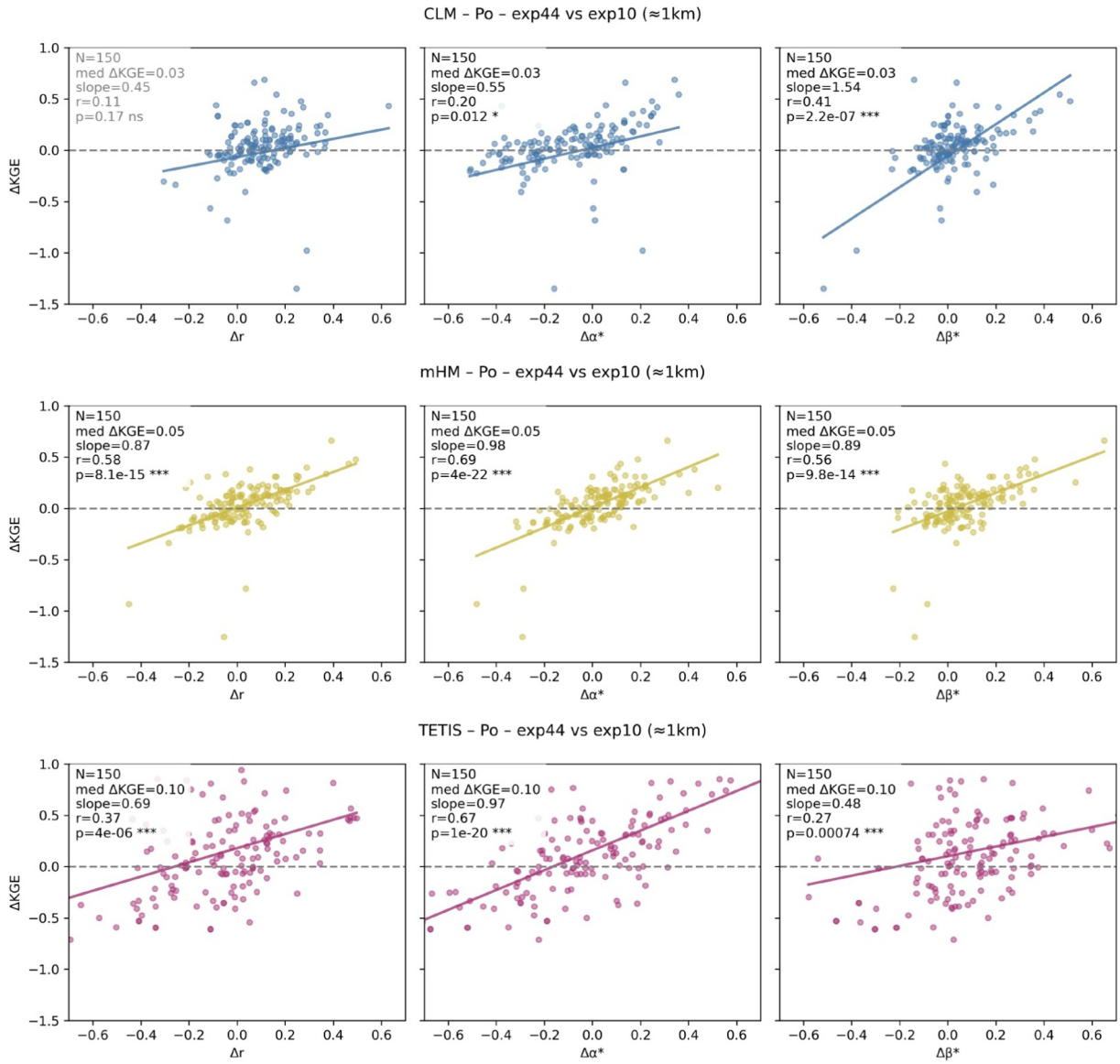


Figure 7. Attribution of performance changes in the Po basin: ΔKGE for joint EO + Q calibration (`exp44`: SM + ET + Q) relative to the baseline (`exp10`) at ≈1 km resolution. Columns show Δr (timing), $\Delta\alpha^*$ (variability), and $\Delta\beta^*$ (bias). Each dot represents one gauge; solid lines show OLS fits. Annotations report the number of gauges (N), median ΔKGE, fit slope, Pearson correlation, and p value with significance codes (***, **, *, ns).

Across regions, PCR-GLOBWB displays a consistent negative shift with uncalibrated inputs, suggesting difficulties when digesting the spatial patterns of the EO precipitation product. In contrast, JULES and TETIS exhibit basin-specific variations, while mHM preserves stable results across the studied hydroclimates. These findings support the interpretation that EO forcing effects are dependent on both basin and model structure, showing positive effects in some hydroclimatic contexts and adjustments in others. Joint calibration approaches (EO + Q) appear useful for counteracting initial performance drops and managing parameter interactions across scales. Detailed gauge-wise scatter plots for the Rhine and Tugela are provided in Appendix A (Figures A2 and A3).

4.3. EO Data Integration Benefit

In this workflow, EO integration refers specifically to calibration against satellite-derived products. Although such frameworks are occasionally described under the umbrella of data assimilation, the methodology applied here does not involve sequential state updating or filtering; rather, the spatial EO products serve as explicit calibration targets for parameter estimation with EO constraints. Figure 5 illustrates how this parameter

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

estimation strategy affects model skill across basins, tracking adjustments from the initial forcing-only configuration (`exp30`) through EO-only calibration targets (`exp31–exp34`) and joint multi-variable configurations (`exp41/42/44`).

In the Po basin, parameter estimation against spatial targets provides moderate adjustments beyond forcing, particularly when coupled with discharge observations. Relative to the uncalibrated configuration (`exp30`), JULES maintains its positive performance shifts, with its distribution moving slightly higher above zero and an increased share of gauges classified as improved. TETIS also responds favorably, showing narrower performance spreads and higher median values. Conversely, mHM experiences minor downward shifts under configurations constrained primarily by soil moisture, but returns toward neutral under joint EO + Q calibration. This pattern suggests that incorporating point-based streamflow observations helps stabilize parameter structures by anchoring vertical flux volumes to downstream hydrograph conditions. CLM exhibits minimal variation throughout, while PCR-GLOBWB displays consistently lower values in the univariate calibration experiments but higher values in the bivariate calibration experiments.

The Rhine basin exhibits limited variations under calibration adjustments. Given the performance drops observed under uncalibrated satellite forcing, calibration against EO targets alone does not reverse the overall decline. CLM and JULES remain negative relative to the baseline, while mHM shows its lowest values under soil moisture constraints, then experiences minor adjustments when evapotranspiration data are added. PCR-GLOBWB continues to display lower metrics across all univariate calibrations, with an increase under the bivariate calibration. The introduction of discharge constraints yields only minor improvements over baseline performance, with mHM shifting toward neutral, while the remaining structures show little change. These responses suggest that in the Rhine's nival–pluvial regime, spatial parameter regularization alone is insufficient to compensate for the systematic errors introduced by the underlying input forcing data.

In the Tugela basin, the selection of calibration targets significantly impacts model skill. Incorporating EO targets alone supports the positive shifts observed in TETIS, resulting in higher median performance, tighter gauge distributions, and a higher proportion of improved gauge classifications than the forcing-only experiment. These variations are linked to changes in the representation of spatial flux variability under semi-arid conditions. By contrast, mHM drifts downward under EO-only targets before experiencing a partial recovery under joint EO + Q configurations, suggesting that discharge constraints help counteract parameter drift induced by satellite-derived targets. PCR-GLOBWB again shows low metrics for the univariate calibrations and higher metrics for the bivariate calibration, suggesting parameter overfitting in calibrations with only a single target variable.

To isolate the effects of parameter adjustments, the multi-variable joint experiment (`exp44`: SM + ET + Q) is compared against the uncalibrated forcing configuration (`exp30`: EO-P). These contrasts seek to quantify the operational value of multi-variable calibration relative to forcing changes alone. Parallel comparisons for the Rhine and Tugela basins are detailed in Appendix A (Figures A4 and A5).

Figure 6 contrasts the isolated effect of satellite forcing (`exp30`) with the multi-variable joint calibration framework (`exp44`) within the Po basin. CLM remains essentially unmodified by the transition from satellite forcing to multi-variable joint calibration. The gauge distributions cluster closely along the 1 : 1 line (with a Pearson correlation coefficient of 0.99), showing a minimal median change ($\Delta\mathrm{KGE} = +0.01$) that is not statistically significant ($p = 0.22$). Approximately 12% of the gauges improve, while nearly three quarters remain within the ± 0.10 tolerance band, indicating that joint calibration yields little additional shift for this model structure beyond the forcing change.

Conversely, mHM shows a statistically significant performance increase under the multi-variable joint calibration framework, with 44% of gauges improving and 14% worsening. The paired median contrast shows a positive shift of $+0.09$ $\left(p = 1.7 \times 10^{-7}\right)$, while a high correlation coefficient (0.92) indicates that the parameter adjustments preserve the overall gauge ranking established under the satellite forcing while shifting the total distribution upward. This result indicates that mHM's multiscale regionalization workflows successfully internalize the complementary information provided by combining vertical spatial targets (soil moisture and evapotranspiration) with localized downstream hydrographs. TETIS exhibits a median positive change of $+0.10$ $\left(p = 1 \times 10^{-4}\right)$, with more than half of the gauges showing improvements relative to the forcing-only experiment. However, a lower correlation coefficient (0.33) indicates substantial reordering of gauge behavior, reflecting a spatially heterogeneous calibration response across the basin.

10.1029/2025WR042446

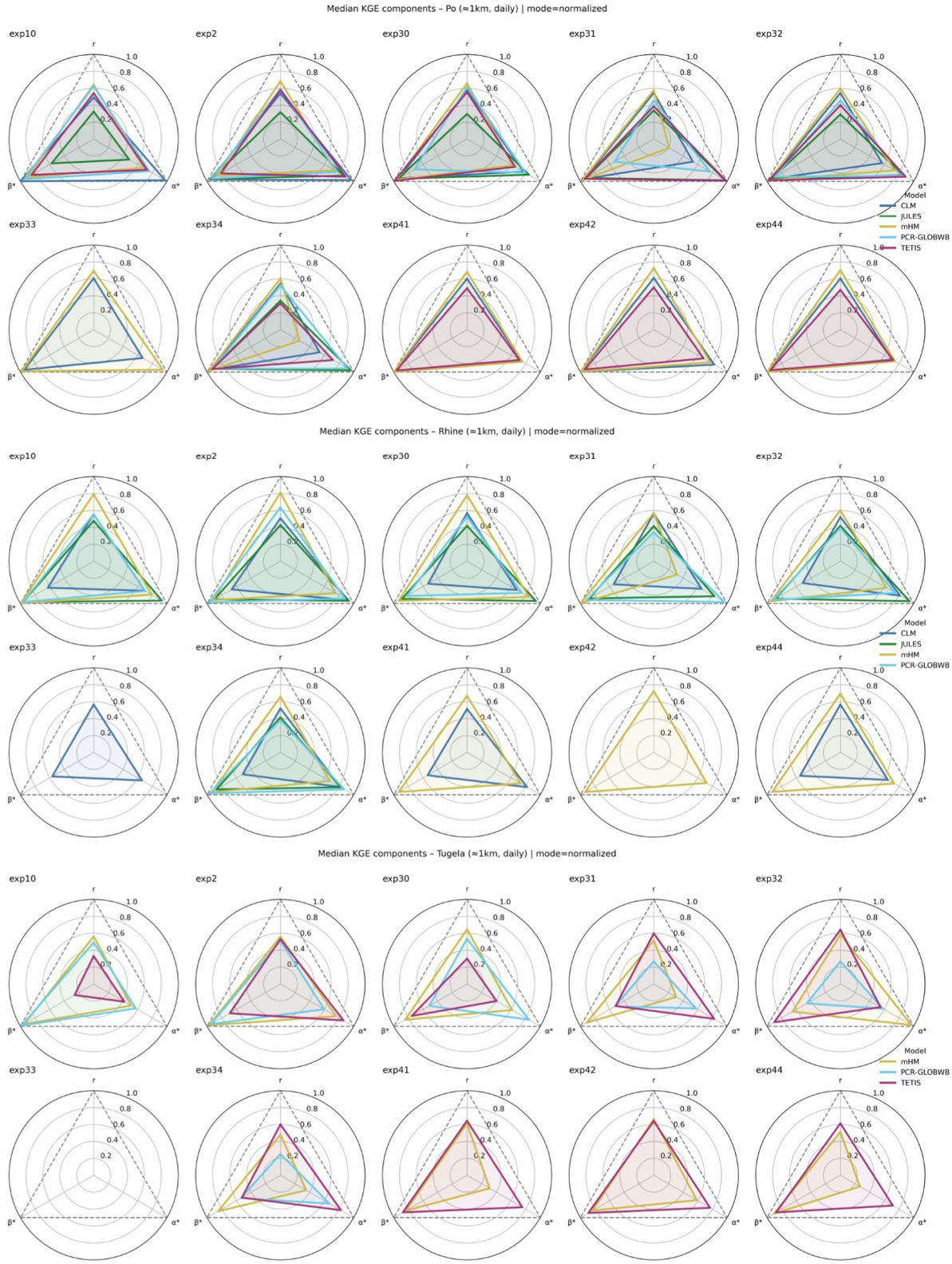


Figure 8. Component-wise KGE evaluation using radar charts at ≈1 km resolution. Axes show correlation (r), normalized variability (α^*), and normalized bias (β^*). Polygons show basin-median component scores across gauges for each model and experiment. The dashed triangle indicates the ideal performance profile (1, 1, 1); departures from this profile identify whether performance is limited mainly by correlation, variability, or bias.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 6
Top-Performing Configurations by Basin, Ranked by Median KGE at ≈1 km Resolution, and Their Improvement Relative to the Best Baseline configuration, Defined as the Maximum Median KGE Among `exp10` and `exp2` Across Models

Basin	Best model	Experiment	Median KGE	Best baseline (model/exp.; KGE)	$\Delta_{vs\ baseline}$
Po	mHM	`exp33`: SCA	0.53	mHM/`exp10` (0.46)	+0.07
Rhine	mHM	`exp10`: baseline	0.61	mHM/`exp10` (0.61)	0.00
Tugela	TETIS	`exp42`: ET + Q	0.49	mHM/`exp2` (0.39)	+0.10

Taken together, the transition from uncalibrated satellite forcing (`exp30`) to joint multivariate calibration (`exp44`) in the Po basin results in minimal operational modifications to CLM, a stable, rank-preserving performance increase for mHM, and a structurally heterogeneous yet generally positive shift for TETIS.

4.4. Best Performing Configurations

The best-performing configurations are shown in Table 6.

The highest-performing configurations vary across basins, suggesting an interaction between regional hydroclimatic conditions and the internal model formulation. The top setup in the Po is mHM–`exp33`, which shifts above the best baseline by +0.07 KGE. This outcome suggests that the utility of EO targets is conditioned by regional process relevance. Targeting spatial snow information (via SCA) in a nival-influenced, Alpine-fed headwater catchment appears to introduce structural value that extends beyond the initial forcing replacement, modifying seasonal persistence metrics and spatial accumulation patterns. This alignment indicates that constraining a model component explicitly linked to the primary driver of regional runoff variability can be more effective than implementing broad calibration strategies that omit basin-specific weightings.

In the Rhine basin, by contrast, the baseline configuration (mHM–`exp10`) retains the highest median KGE value (0.61), and no alternative utilizing EO targets exceeds this baseline. This behavior is consistent with the interpretation that substituting the forcing data with satellite precipitation fields introduces errors in nival–pluvial systems that are difficult to mitigate via single- or multi-variable parameter estimation workflows.

In the Tugela, the leading configuration is TETIS–`exp42` (ET + Q), which exceeds the best baseline by approximately +0.10 KGE. In this context, joint calibration using both spatial evapotranspiration and discharge observations provides a noticeable performance adjustment, likely because spatial ET targets help regularize vertical flux variability while discharge constraints anchor temporal routing and volume metrics under data-scarce, convective conditions.

Taken together, these variations imply that a singular configuration may not be suitable across all domains. The Po basin shows adjustments under an EO-only framework (`exp33`: SCA), the Rhine basin retains its baseline metrics, and the Tugela basin shows improvements under a joint framework (`exp42`: ET + Q). The optimal configuration appears dependent on both basin hydroclimate and model selection, as summarized in Table 7.

In the Po basin, performance leadership varies across experiment configurations. mHM achieves the highest median value under baseline and forcing-only setups (`exp10`, `exp2`, and `exp30`), with a baseline peak of 0.46. When looking at EO-only calibration, CLM yields the highest relative metrics when soil moisture targets are active (`exp31`: 0.23 and `exp34`: 0.21), whereas mHM leads under the evapotranspiration target configuration (`exp32`: 0.30). Within the joint EO + Q calibration family, mHM achieves the highest absolute skill under ET + Q (`exp42`: 0.52) and SM + ET + Q (`exp44`: 0.50), shifting above its baseline performance. Although CLM leads the model subset under SM + Q (`exp41`: 0.33), all configurations constrained primarily by satellite-derived soil moisture (`exp31`, `exp34`, `exp41`) exhibit lower absolute performance metrics than corresponding experiments, suggesting that these relative advantages do not translate to substantive performance increases.

In the Rhine basin, mHM retains the highest median KGE across nearly all evaluated configurations. It reaches its maximum value at the baseline configuration (`exp10`: 0.61) and remains the leading model configuration under uncalibrated forcing changes (`exp30`) and all joint EO + Q implementations (`exp41`, `exp42`, and `exp44`). Among the EO-only calibration targets, mHM yields the highest scores under evapotranspiration (`exp32`) and combined soil moisture and evapotranspiration (`exp34`). However, it shows a lower median value than JULES

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 7
Best Model per Experiment and Basin (Highest Median KGE at ≈1 km)

Experiment	Basin	Best model	Median KGE
exp10	Po	mHM	0.46
	Rhine	mHM	0.61
	Tugela	PCR-GLOBWB	0.25
exp2	Po	mHM	0.46
	Rhine	mHM	0.54
	Tugela	mHM	0.39
exp30	Po	mHM	0.44
	Rhine	mHM	0.53
	Tugela	mHM	0.26
exp31	Po	CLM	0.23
	Rhine	JULES	0.12
	Tugela	mHM	0.00
exp32	Po	mHM	0.30
	Rhine	mHM	0.25
	Tugela	TETIS	0.33
exp34	Po	CLM	0.21
	Rhine	mHM	0.37
	Tugela	mHM	0.02
exp41	Po	CLM	0.33
	Rhine	mHM	0.38
	Tugela	TETIS	0.36
exp42	Po	mHM	0.52
	Rhine	mHM	0.49
	Tugela	TETIS	0.49
exp44	Po	mHM	0.50
	Rhine	mHM	0.43
	Tugela	TETIS	0.31

(0.12) under soil-moisture-only targets (exp31). This specific configuration response occurs within a range of low performance metrics, indicating that the relative shift is of limited practical significance. Overall, the Rhine basin results show a consistent trend for mHM: integrating alternative EO targets does not lead to increases above the baseline.

In the Tugela basin, performance rankings change following the introduction of spatial calibration targets. mHM calibrated against discharge represents the highest baseline configuration (exp2: 0.39), and the uncalibrated forcing swap (exp30) does not change this relative order. Within the EO-only family, TETIS displays the highest metrics under evapotranspiration targets (exp32: 0.33) and maintains relatively stable values under soil moisture (exp31) and combined soil moisture and evapotranspiration (exp34) constraints, a trend during which mHM and PCR-GLOBWB values shift downward. Under joint EO + Q frameworks, TETIS yields higher relative metrics across all configurations (exp41/42/44), with the ET + Q combination (exp42: 0.49) reaching the highest overall value for the basin. This shift suggests that, under semi-arid, convective regimes, incorporating satellite-derived evapotranspiration constraints alongside gauge discharge data is an effective configuration, while the relative performance of mHM appears centered on the baseline setups.

4.5. Attribution of Performance Drivers

In Figure 7, the three models exhibit distinct attribution patterns when transitioning from the baseline configuration (exp10) to the joint multi-variable framework (exp44). To avoid the limitations of evaluating model

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

performance solely through an aggregated index, we explicitly dissect the total KGE metric into its three independent physical sub-components: timing (Δr), variability ($\Delta\alpha^*$), and volumetric bias ($\Delta\beta^*$). For CLM, the median change in performance is minor ($\Delta\text{KGE} = +0.03$). This adjustment is primarily associated with the bias component ($\Delta\beta^*$), which shows a statistically significant relationship ($r = 0.41$, slope = 1.54, $p = 2.2 \times 10^{-7}$). The variability component ($\Delta\alpha^*$) plays a weaker secondary role ($r = 0.20$, slope = 0.55), whereas the timing component (Δr) has a negligible link to the total performance change ($r = 0.11$, slope = 0.45, not significant). For mHM, the change in KGE is modest but evenly distributed across components (median $\Delta\text{KGE} = +0.05$). The variability component is the strongest contributor ($r = 0.69$, slope = 0.98), followed by timing ($r = 0.58$, slope = 0.87) and bias ($r = 0.56$, slope = 0.89). OLS slopes approaching unity suggest that adjustments across all three components translate directly into overall KGE changes, resulting in a distributed performance increase across the basin while largely maintaining the original gauge rankings. For TETIS, the median increase in KGE is the largest among the evaluated models (+0.10). The response is dominated by the variability component ($r = 0.67$, slope = 0.97), with lower contributions coming from timing ($r = 0.37$, slope = 0.69) and bias ($r = 0.27$, slope = 0.48). A wider scatter across gauge coordinates indicates a higher degree of reordering in gauge performance than observed in mHM, though the net basin shift remains positive.

In general, transitioning from the baseline configuration (`exp10`) to the joint multi-variable calibration (`exp44`) in the Po basin modifies KGE via different mechanisms depending on the model structure. For CLM, the change is mainly associated with bias reduction, whereas mHM shows a more distributed response across timing, variability and bias. For TETIS, the largest improvement is due to variability representation. From a process-based hydrological perspective, this component-wise attribution provides a critical insight into what satellite observations actually offer to hyper-resolution models. Because the systematic performance gains across the ensemble are heavily front-loaded into variability alignment and volumetric bias adjustments, while temporal correlation remains relatively static or highly resistant to change, the data delivers a clear diagnostic message: EO data helps us close the volumetric water balance (ensuring the correct mass and variance within the storage "bucket"), but has limited impact on fixing landscape routing or wave propagation dynamics. In short, satellites fix volumes, not flow paths. When we constrain a model with remote sensing streams such as soil moisture or evapotranspiration, we optimize the vertical partitioning of fluxes, yet the underlying speed at which water moves through the hillslope and channel networks remains rigidly dictated by the model's structural routing limitations. Recognizing that satellite data acts as a volumetric corrective lens rather than a temporal dynamic synchronizer represents a core boundary condition that must guide future hyper-resolution calibration and data integration frameworks. Component attribution results for the Rhine and Tugela basins are provided in Appendix B (Figures B1 and B2).

4.6. Component-Wise Evaluation

The radar charts provide a multivariate summary of absolute normalized performance by plotting the KGE components—timing (r), variability (α^*), and bias (β^*)—as shown in Figure 8. The dashed triangle marks the optimal performance coordinate system at (1, 1, 1), where specific directional changes and axis extensions reflect modifications in simulation characteristics. This visualization highlights how individual components adjust and balance across configurations, complementing the attribution analyses that report net metric changes (ΔKGE) relative to a baseline reference. The evaluated configurations span the baseline (`exp10`), streamflow-only calibration (`exp2`), uncalibrated satellite forcing (`exp30`), EO-only calibration targets (`exp31–34`: SM, ET, SCA, SM + ET), and joint multi-variable configurations (`exp41/42/44`).

In the Po basin, initial baseline performance profiles are heterogeneous across the ensemble; TETIS exhibits balanced axes, whereas JULES appears constrained along specific dimensions. Experiment `exp2` results in minor, broad-based adjustments across the axes. Calibration against EO targets modifies specific component responses through asymmetrical shape variations: soil moisture targets stretch the variability (α^*) axis, evapotranspiration data expands the bias (β^*) coordinates (particularly for mHM), and snow cover data introduces an incremental outward pull consistent with localized alpine snow processes. The joint EO + Q configurations yield the most balanced profile geometries across the majority of models, suggesting that the discharge constraint helps anchor the timing axis while consolidating the water-balance modifications provided by the spatial EO data. Among these configurations, `exp42` (ET + Q) shifts the overall geometry closest to the optimal boundary.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In the Rhine basin, initial performance profiles are positioned relatively close to the optimal boundary under the baseline (`exp10`). Introducing the uncalibrated satellite forcing (`exp30`) or calibrating against EO targets alone (`exp3x`) typically induces an inward collapse of specific axes or a general deformation of the shape. This behavior suggests that altering forcing data or optimizing parameters without streamflow constraints can destabilize the internal parameter balance of pre-calibrated model structures within data-rich basins. The joint EO + Q calibration configurations (`exp41/42/44`) show the most noticeable shape recovery, with the multivariate joint calibration (`exp44`) yielding the most balanced radial profiles across the constituent parameters. The mHM expands across all three components, though the final geometries remain comparable to the initial baseline profiles.

In the Tugela basin, which poses significant modeling challenges, the baseline configurations exhibit compressed geometries with timing coordinates restricted to 0.4 – 0.6. Calibrating against EO targets alone leads to pronounced, model-specific axis extensions. For instance, TETIS achieves an expanded and balanced profile under `exp32` (ET), indicating that a targeted remote sensing constraint can shift the bias-variability axes even in data-scarce regions. The joint EO + Q strategies build upon these variations, extending the profiles across multiple dimensions. The ET + Q combination (`exp42`) is generally effective, supporting the utility of a joint-constraint strategy for introducing multi-component modifications in semi-arid hydroclimatic regimes.

Collectively, the radar charts illustrate that the effects of EO data integration are revealed through targeted, directional changes in shape rather than in total polygon area, which is heavily conditioned by the initial baseline performance. Evaluating the relative shifts along individual axes suggests that EO-only targets adjust isolated error sources by selectively extending specific components, such as soil moisture, modifying the variability representation or evapotranspiration, and extending the bias axis. In contrast, joint EO + Q configurations simultaneously expand polygons across multiple dimensions, showing concurrent adjustments in timing, bias, and variability. Among these configurations, `exp42` (ET + Q) frequently provides the most balanced axis extensions across regions and models. Model-specific differences persist; mHM and TETIS achieve cohesive expansions under joint constraints. Conversely, where performance drops occur, most notably for PCR-GLOBWB across components or under uncalibrated satellite forcing in the Rhine basin, the radar profiles reveal an asymmetric collapsing of the axes. These geometric adjustments demonstrate that remote sensing data does not uniformly scale model capabilities across all dimensions, but rather acts as an axis-specific corrective lens that modifies the balance of individual performance components depending on the physical process being constrained.

4.7. Robustness and Limitations

Our results are internally consistent across multiple diagnostics (medians, gauge-wise contrasts, win–rates, and component attribution), but several factors limit their overall robustness and external validity. (a) *Incomplete coverage*: Only three models completed the joint multi-variable family (`exp41/42/44`), and not all EO-only targets (`exp31–34`) are available for every model–basin pair, which constrains cross-model generalisations across the ensemble. (b) *Uneven sampling*: Basin-wise gauge sample sizes differ; for instance, the Tugela basin contains $N = 21$ paired gauges, reducing statistical power and yielding borderline p-values for some localized effects (e.g., `exp44` vs. `exp30` for TETIS). (c) *EO record length and support*: EO time series are shorter than the historical streamflow records, featuring partially overlapping periods and spatial supports that differ from the grid configurations. This mismatch introduces potential representativeness errors and seasonally biased learning signals. (d) *Forcing sensitivity and parametric compensation:* The merged satellite precipitation data modifies performance selectively, showing improvements in some basin–model combinations (e.g., Po for TETIS and JULES) but performance decreases in others (e.g., the Rhine basin). This dependency highlights how sub-optimal forcing error structures trigger compensating errors; parameter estimation algorithms may constrain parameter structures into unphysical spaces to match outlet discharge, leading to model simulations that preserve systematic biases when driven by uncalibrated inputs. (e) *Objective function weighting*: The unified objective function uses fixed relative weights (specifically 5 : 1 in `exp41/42`) to prioritize remote sensing constraints while using point-based streamflow as a bounding physical anchor. However, this asymmetric configuration represents a single operational point within a broader multi-objective trade-off space, meaning that the observed benefits of joint calibration remain conditional upon this specific priority hierarchy. (f) *Metric scope*: Inference is based primarily on daily KGE; although decomposed into timing, bias, and variability components, it compresses continuous hydrograph behavior and may under-represent high-flow extremes or human-induced river regulation. (g) *Study scope and resources*: Not all model structures participated in the full joint calibration family, and the analysis was

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich Gmbh, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

centered on three specific regions of interest. (h) *Structural scale penalty*: The direct transfer of parameter structures across spatial resolutions (from 5 km to 1 km) introduces an artificial performance penalty for models that lack explicit scale-invariant or scale-effective regionalization mechanisms. This spatial topology mismatch induces a structural-scale artifact that can distort local routing and infiltration dynamics, independent of the model's true process-representation capabilities. (i) *Calibration frameworks*: PCR-GLOBWB (MSE-based parameter learning) and JULES (data-assimilation cost function) do not optimize the unified objective of Equation 13, so their calibrated results are not strictly comparable with those of CLM, mHM and TETIS. Because the main results are paired contrasts within each model, differences between the calibration frameworks affect the comparison between models, but not the added value of EO data estimated for each model.

Overall, the conclusions are more robust when multiple models are consistent, and sample sizes are large (e.g., the Rhine and Po basins for mHM). Caution is warranted in data-scarce contexts and in model–experiment slots with limited sample sizes or missing asset combinations. Robustness could be enhanced by broadening model participation across all joint configurations, evaluating additional domains with distinct hydroclimates, extending EO records via gap-filled alternatives, and testing sensitivity to objective-function weights and alternative forcings. Furthermore, the localized intersection between model structure and regional data quality heavily influences performance. Consequently, global transferability remains conditioned by the regional accuracy of the chosen EO products or by specific land surface configurations across different hydroclimates.

5. Conclusions

This study assessed how EO data affect daily river discharge simulations at kilometer-scale resolution across three contrasting hydroclimatic domains: the Po, Rhine, and Tugela basins, using an ensemble of up to five model structures (five in the Po, four in the Rhine and three in the Tugela). By systematically comparing baseline simulations (`exp10` and `exp2`), uncalibrated EO precipitation forcing (`exp30`), EO-only calibration targets (`exp31–34`), and joint calibration frameworks incorporating both discharge and EO products (`exp41/42/44`), we quantified the predictive contribution of EO integration using gauge-wise performance metrics and component attribution.

Uncalibrated EO precipitation forcing (`exp30`) improves performance selectively, with gains for JULES and CLM in the Po basin and TETIS in the Tugela basin, while consistently reducing skill in the Rhine basin. This contrast shows that a replacement of conventional precipitation forcing with EO-based precipitation is not universally beneficial and may require parameter recalibration. Within this experimental design, the most stable improvements arise from joint calibration with discharge and EO products (`exp41/42/44`, available for three models), which help reduce forcing-related performance losses and stabilize simulations, especially in data-scarce basins such as the Tugela.

EO-only calibration (`exp31–34`) produces more variable outcomes that depend strongly on model structure and basin characteristics. Configurations targeting a regionally dominant process, such as snow cover constraints (`exp33`) in the Alpine-influenced Po basin, can outperform broader EO-only calibration strategies. This outcome suggests that EO constraints should be selected according to process relevance rather than applied as an unweighted set of available data streams.

Component analysis indicates that performance changes are primarily driven by variability alignment (α^*) and bias changes (β^*), while the correlation component (r) plays a secondary role. This diagnostic result suggests that the EO constraints used here primarily affect water-balance-related quantities and flow variability, while their influence on timing remains limited. This limitation is consistent with the fact that routing dynamics, catchment lag, and wave propagation are governed mainly by model structure and are not directly constrained by the EO targets used in this study. Consequently, EO integration appears better suited to improving water balance and variability-related aspects of streamflow simulations than to resolving routing and timing errors. Where performance deteriorates, most notably in the Rhine basin, losses are likewise associated mainly with variability and bias errors rather than with changes in correlation.

These findings have practical implications for kilometer-scale modeling workflows. In data-scarce regions, combining spatial EO data with point discharge records (`exp41/42/44`) offers a promising pathway to improve model performance, with evapotranspiration constraints (`exp42`) and snow information being particularly valuable when physically relevant. For operational implementation, our results indicate that models should be

recalibrated after changes in precipitation forcing, that KGE components should be analyzed alongside aggregate skill metrics, and that multi-model ensembles remain important for diagnosing structural uncertainty.

Future work should test joint EO–discharge frameworks across a wider range of models and basins, and further progress will depend on improving the length, consistency, and gap-filling soil moisture, evapotranspiration, and snow records, and on evaluating alternative weighting schemes and adaptive calibration strategies. Additionally, explicitly accounting for human influences, such as abstraction and reservoir regulation, and inter-basin transfers, will be important in strongly managed basins. Finally, extending the evaluation to extreme events and sub-daily dynamics would help determine whether EO-constrained calibration can improve not only water-balance and variability metrics, but also the timing and routing behavior required for operational forecasting.

Appendix A: Regional Performance Comparisons

This appendix provides detailed regional performance comparisons that complement the primary manuscript analysis. Figure A1 presents a performance overview across all three study regions at ≈1 km resolution, while Figures A2 and A3 show gauge-wise forcing effects for the Rhine and Tugela basins, respectively. Figures A4 and

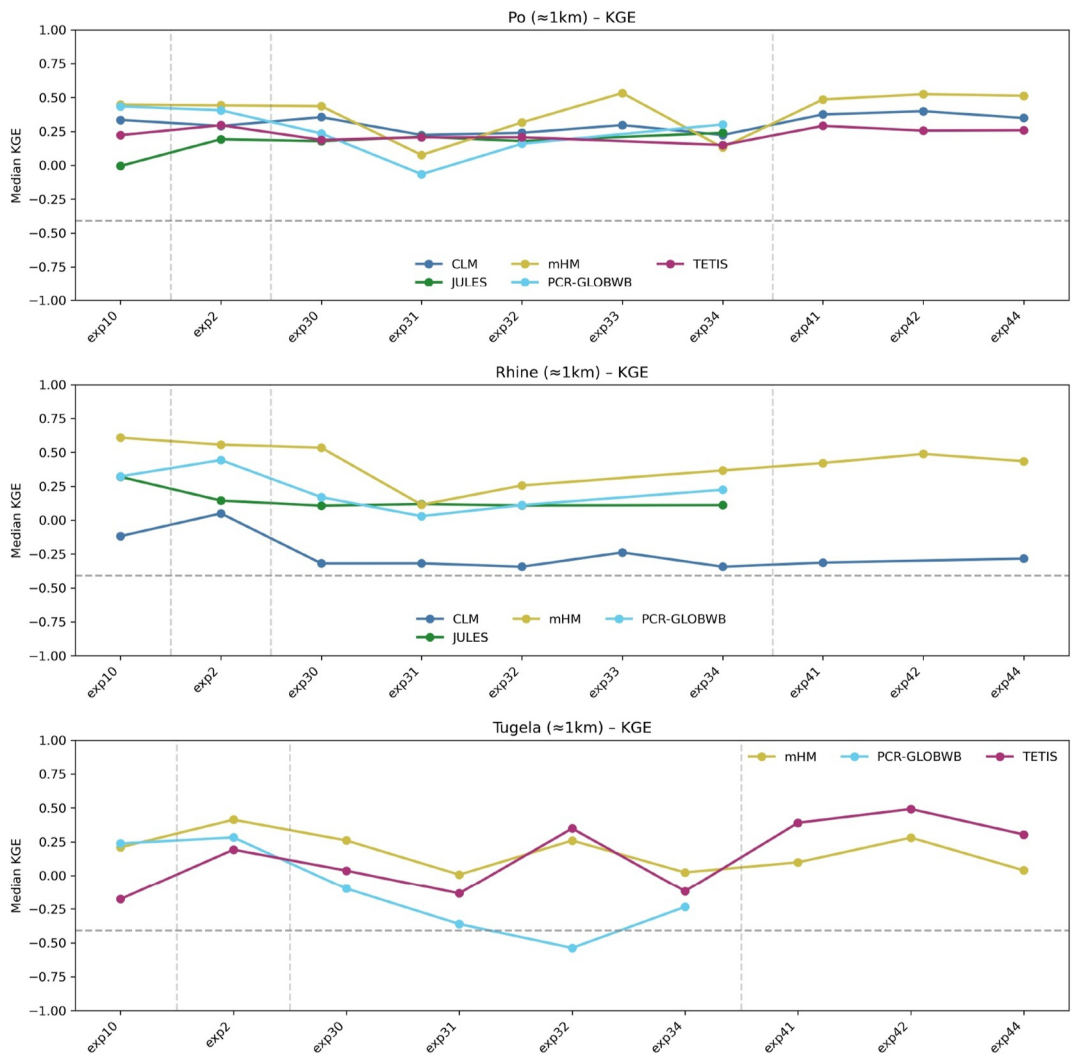


Figure A1. Overall model performance across the three study regions showing median KGE values at ≈1 km resolution. Top: Po River basin with five models across multiple experimental configurations. Middle: Rhine River basin displaying results for four models. Bottom: Tugela River basin showing performance for the three available models. The horizontal dashed line indicates the KGE threshold of −0.41 for acceptable performance.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

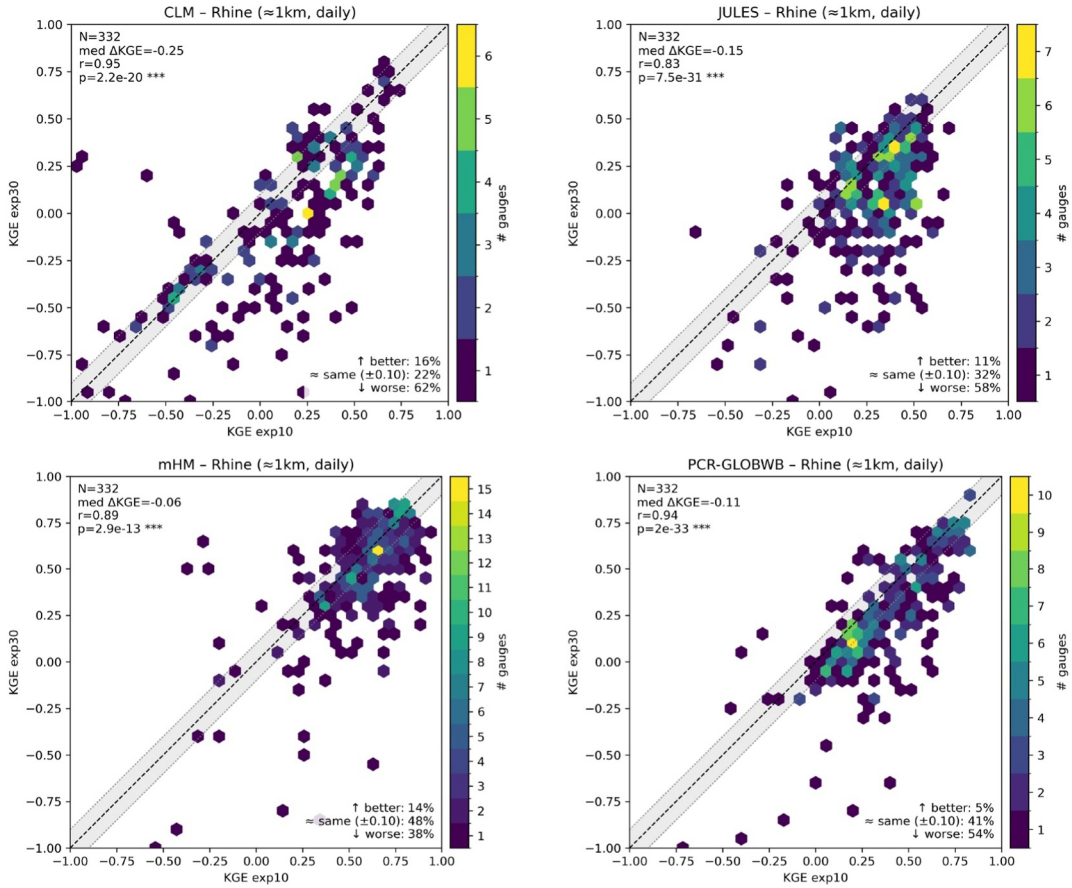


Figure A2. Gauge-wise skill under EO-forcing-only (exp30) versus baseline (exp10) at ≈1 km, in Rhine. Each hexagon aggregates co-located gauges (color = count). The black 1 : 1 line indicates no change; gray dashed lines show the ±0.10 *no-material-change* band. In-panel text lists N, median, ΔKGE, Pearson r, and Wilcoxon *p* with significance codes (***, **, *, ns). Percentages denote *better/same/worse* shares using the ±0.10 tolerance.

A5 demonstrate the relative shifts between joint EO + Q calibration and forcing-only configurations in these regions.

The Rhine basin results (Figure A2) reveal systematic performance reductions under uncalibrated satellite precipitation forcing across the evaluated models, with CLM showing the largest negative variations. PCR-GLOBWB, JULES, and mHM exhibit linear alignments with the 1 : 1 reference line, but their densest hex-bins are positioned below it, indicating a coherent basin-wide skill reduction despite preserved gauge order. In contrast, the Tugela basin results (Figure A3) exhibit model-dependent responses; TETIS produces a compact cluster positioned above the 1 : 1 line, pointing to localized skill increases, whereas other model configurations reflect performance decreases.

The joint multi-variable comparison indicates that joint calibration frameworks do not fully compensate for the initial forcing-related modifications in the Rhine basin. In the Tugela basin, TETIS shows notable adjustments, although statistical inference remains constrained by the gauge network's limited sample size.

The consolidated performance assessment demonstrates significant regional variations in model effectiveness and experimental sensitivity. In the Po River basin, mHM consistently achieves the highest median KGE values across most experiments, with performance ranging from approximately 0.4 to 0.6 depending on the experimental configuration. The model shows stable metrics in `exp33` and maintains comparable results through the EO-enhanced experiments (`exp41, exp42, exp44`). TETIS exhibits variable performance, with higher results in `exp41` and `exp42`, while CLM shows moderate adjustments across the experiments. PCR-GLOBWB maintains stable performance in EO-constrained experiments but at lower absolute values.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

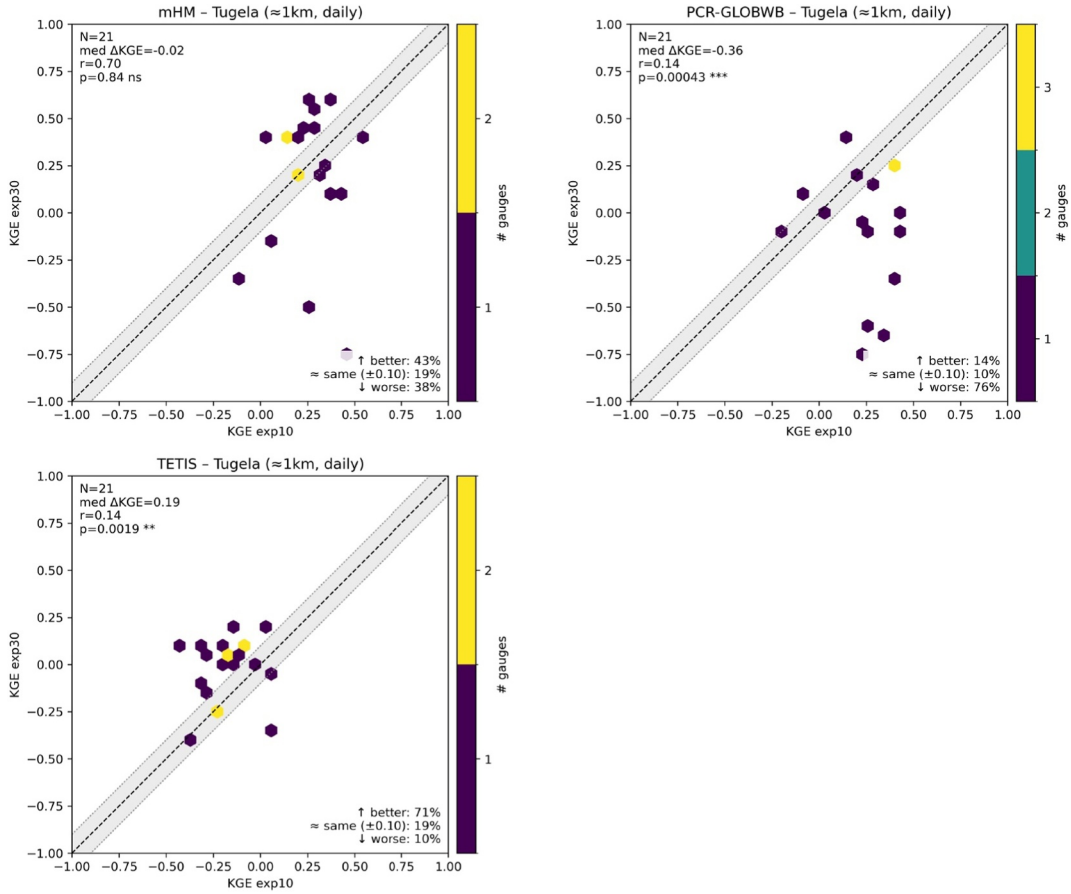


Figure A3. Gauge-wise skill under EO-forcing-only (exp30) versus baseline (exp10) at ≈1 km, in Tugela. Each hexagon aggregates co-located gauges (color = count). The black 1 : 1 line indicates no change; gray dashed lines show the ±0.10 *no-material-change* band. In-panel text lists N, median ΔKGE, Pearson r, and Wilcoxon *p* with significance codes (***, **, *, ns). Percentages denote *better/same/worse* shares using the ±0.10 tolerance.

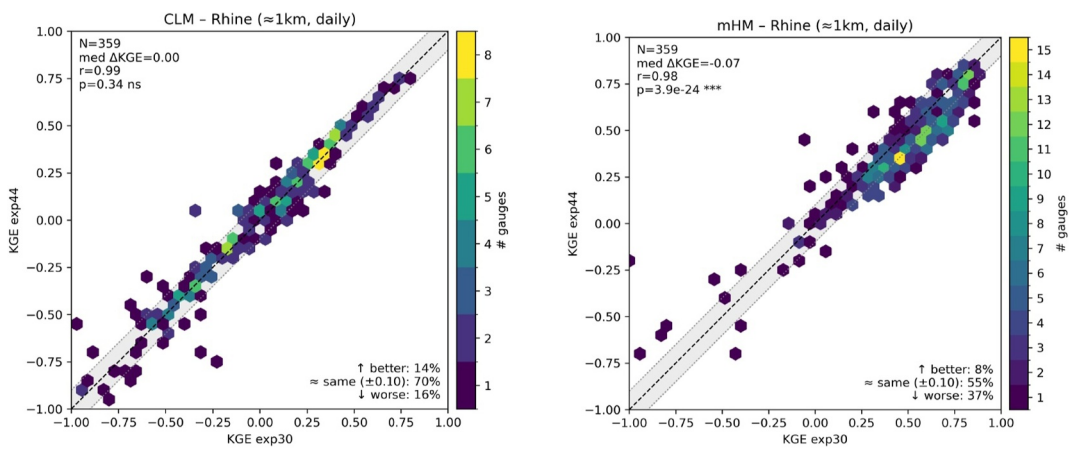


Figure A4. Gauge-wise skill under EO + Q (exp44) versus EO-forcing-only (exp30) at ≈1 km in Rhine. Each hexagon aggregates co-located gauges (color = count). The black 1 : 1 line indicates no change; gray dashed lines show the ±0.10 *no-material-change* band. In-panel text lists N, median ΔKGE, Pearson r, and Wilcoxon *p* with significance codes (***, **, *, ns). Percentages denote *better/same/worse* shares using the ±0.10 tolerance.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

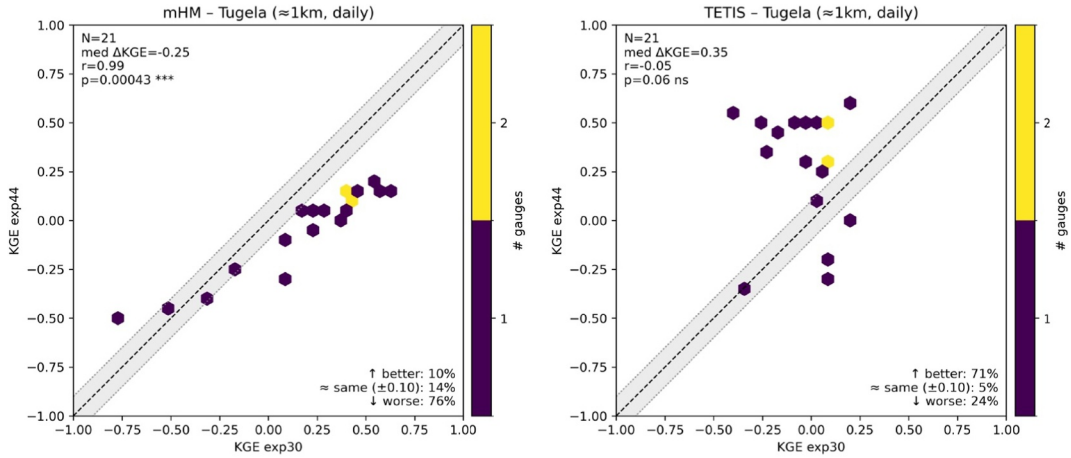


Figure A5. Gauge-wise skill under EO + Q (exp44) versus EO-forcing-only (exp30) at ≈1 km in Tugela. Each hexagon aggregates co-located gauges (color = count). The black 1 : 1 line indicates no change; gray dashed lines show the ±0.10 *no-material-change* band. In-panel text lists N, median ΔKGE, Pearson r, and Wilcoxon *p* with significance codes (***, **, *, ns). Percentages denote *better/same/worse* shares using the ±0.10 tolerance.

In the Rhine River basin, mHM demonstrates consistent performance, maintaining median KGE values above 0.5 across all experimental configurations. The model shows high stability throughout the experimental progression, reflecting its robust parameterization scheme and regionalization features. However, the performance metrics of EO-constrained experiments remain lower than those driven by the EMO-1 baseline. CLM exhibits substantial variability with generally lower performance, while JULES shows more stable performance patterns. PCR-GLOBWB performance metrics decrease when ingesting EO products.

The Tugela River basin presents a more challenging environment where most models perform below the 0.5 KGE threshold. TETIS emerges as the strongest performer in this region, particularly in `exp41` and `exp42`, achieving median KGE values approaching 0.5. mHM shows variable performance with improvements in later joint experiments, while PCR-GLOBWB maintains stable but lower performance levels throughout the experimental sequence. In this region, ET-calibrated configurations (`exp32` and `exp42`) yield the highest EO-constrained model performance.

These regional performance patterns highlight that model suitability is conditioned by basin characteristics and that experimental approaches vary in effectiveness across different hydroclimatic environments. Detailed model-specific performance analyses and experimental comparisons are presented in the subsequent figures.

For the majority of models, the gauge-wise comparisons in Figure A2 show a strong linear alignment to the 1 : 1 reference line, with CLM, JULES, and mHM achieving high correlation coefficients ($r = 0.95$, $r = 0.83$, and $r = 0.89$, respectively). However, the densest hex-bins lie clearly below the 1 : 1 line, indicating a coherent basin-wide reduction in KGE levels despite the preservation of original gauge ordering. PCR-GLOBWB also preserves the gauge ordering ($r = 0.94$), but more of its gauges fall below the tolerance band (54% worse; median $\Delta\text{KGE} = -0.11$), signaling a systematic performance degradation under the alternative satellite forcing. Table 5 provides the corresponding median values and class distributions for the Rhine basin across the defined performance categories.

In the Tugela basin (Figure A3), sample size is limited ($N = 21$ per model), meaning the hex-bin density highlights isolated clusters. TETIS shows a compact cloud positioned slightly above the 1 : 1 reference line, consistent with localized skill adjustments ($r = 0.14$, with a statistically significant *p*-value despite the modest effect size). mHM shows a moderate alignment ($r = 0.70$; *p*-value not significant) with gauge coordinates straddling the diagonal, indicating mixed local sensitivities. PCR-GLOBWB forms a diffuse cloud located primarily below the 1 : 1 line ($r = 0.14$, with a significant *p*-value), indicating a more consistent reduction in performance. Detailed median values and class percentages for these configurations are presented in Table 5.

Figure A4 and A5 demonstrate the relative shifts associated with joint EO + Q calibration compared to forcing-only configurations in these regions.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In Figure A4, the gauge cloud for CLM closely follows the 1 : 1 line ($r = 0.99$), yielding a median ΔKGE $= 0.00$ and a distributed class allocation (14% better, 70% same within the ± 0.10 threshold, and 16% worse). This change is not statistically significant ($p = 0.34$), suggesting that the joint framework leaves the model's performance under satellite forcing essentially unchanged. In contrast, mHM displays a consistent downward shift relative to the 1 : 1 line, showing a high degree of rank preservation ($r = 0.98$) but with a negative median shift (ΔKGE $= -0.07$, with 8% better, 55% same, and 37% worse). This negative variation is statistically significant ($p = 3.9 \times 10^{-24}$). Taken together, incorporating multivariate targets alongside alternative forcing in the Rhine basin yields no measurable change in CLM and a modest decrease in mHM performance, consistent with the basin's general resistance to EO-driven improvements.

In Figure A5, mHM exhibits a coherent, statistically significant performance decline moving from forcing-only to the joint framework; the cloud is positioned below the 1 : 1 reference line with a median ΔKGE $= -0.25$, strong rank preservation ($r = 0.99$), and 76% of gauges showing reduced metrics (10% better, 14% same). This variation is highly significant ($p = 4.3 \times 10^{-4}$), indicating that the joint configuration reduces mHM skill metrics relative to the forcing-only baseline in the Tugela basin.

In contrast, TETIS displays a large absolute positive effect size (median ΔKGE $= +0.35$), with 71% of gauges showing improvements and 24% showing declines. The coordinates show a weak, slightly negative correlation with the forcing-only run ($r = -0.05$), suggesting reordering across the gauge network. However, this adjustment is not statistically significant at the 5% level ($p = 0.06$); thus, while the absolute performance shift is substantial, the limited sample size restricts the statistical confidence required to assert a basin-wide, rank-consistent gain. Overall, transitioning from forcing-only setups to the joint multivariate calibration framework in the Tugela basin negatively affects mHM performance metrics, while TETIS shows a positive but statistically borderline trend.

Appendix B: Component Attribution Analysis

This appendix extends the component attribution analysis presented in the main body text to the Rhine and Tugela basins. The performance modifications observed in the Rhine basin (Figure B1) are primarily associated with adjustments in the variability and bias components rather than timing shifts. For CLM, both bias and variability show significant relationships with the overall performance change. For mHM, bias adjustments show the strongest link, followed by timing and variability.

For the Tugela basin (Figure B2), bias emerges as the main driver of performance modifications for both models: the gains of TETIS and the losses of mHM are both most strongly associated with changes in the bias component. Together with the Po results, these regional variations show that variability and bias, rather than timing, govern the outcomes of multi-variable data integration.

In Figure B1, CLM exhibits a median performance decrease (ΔKGE $= -0.20$; $N = 334$). Variations in the timing component do not show explanatory power (Δr: $r = 0.05$, slope $= 0.83$, $p = 0.34$, ns). Conversely, both variability and bias adjustments are significantly associated with ΔKGE: $\Delta\alpha^*$ ($r = 0.31$, slope $= 1.98$, $p = 6.6 \times 10^{-9}$) and $\Delta\beta^*$ ($r = 0.25$, slope $= 2.07$, $p = 2.7 \times 10^{-6}$). OLS slopes approaching a value of 2 suggest that modifications in mean or spread scale noticeably into composite KGE changes; the negative median shift indicates that variability and bias metrics generally reduced across the gauge network when transitioning to `exp44`, while temporal correlation remained relatively static.

For mHM, the net basin shift is similarly negative (median ΔKGE $= -0.15$). All three components are significantly associated with the overall performance change, with the strongest link observed for the bias component: $\Delta\beta^*$ ($r = 0.56$, slope $= 1.03$, $p = 1.4 \times 10^{-28}$), followed by timing Δr ($r = 0.40$, slope $= 1.22$, $p = 1.6 \times 10^{-14}$) and variability $\Delta\alpha^*$ ($r = 0.37$, slope $= 0.87$, $p = 3.4 \times 10^{-12}$). With slopes near unity, KGE changes for mHM are nearly proportional to individual component adjustments; the negative median implies that degradations in volumetric bias, accompanied by timing and variability degradations, govern the basin-wide trend.

For both models, the performance differences between `exp44` and `exp10` are predominantly driven by variability and bias adjustments rather than by correlation metrics alone. The statistically significant associations,

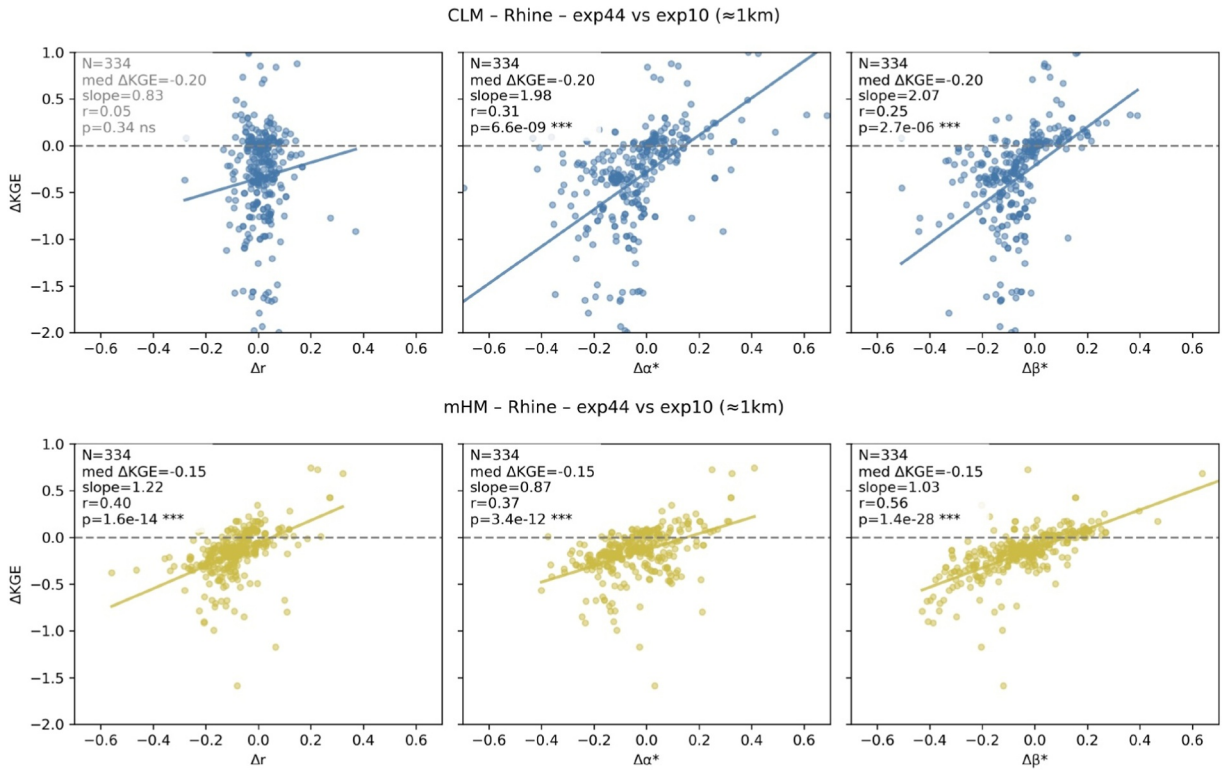


Figure B1. Attribution of performance drivers in the Rhine basin: ΔKGE (exp44 – exp10) versus Δr, $\Delta\alpha^*$, and $\Delta\beta^*$ at ≈1 km. Points denote gauges; lines are OLS fits; in-panel text gives N, median ΔKGE, slope, r, and p (***, **, *, ns). Both models show a negative median ΔKGE relative to the baseline. For CLM, degradation is significantly associated with changes in bias and variability (little role for timing). For mHM, bias is the primary driver (a strong, significant $\Delta\beta^*$ link), with additional contributions from timing and variability.

paired with negative median values, indicate consistent modifications of simulated mean and spread configurations under the full joint calibration framework relative to the baseline within this alpine-influenced basin.

In Figure B2, mHM exhibits a median performance reduction ($\Delta\text{KGE} = -0.27$) within a limited sample context ($N = 21$). Variations in timing and variability do not display explanatory power (Δr: $r = 0.25$, slope = 1.21, $p = 0.28$, ns; $\Delta\alpha^*$: $r = 0.26$, slope = 0.79, $p = 0.25$, ns). In contrast, the bias component ($\Delta\beta^*$) displays a statistically significant relationship with the KGE shift ($r = 0.49$, slope = 1.00, $p = 0.025$), indicating a near proportional mapping: gauges that alter their simulated flow volume tend to exhibit corresponding changes in overall KGE, with degraded bias accounting for a notable share of the median reduction.

For TETIS, the response reflects a substantial performance increase (median $\Delta\text{KGE} = +0.41$). The dominant contributor is similarly identified as the bias component: $\Delta\beta^*$ exhibits a strong statistical association with ΔKGE ($r = 0.85$, slope = 2.65, $p = 1.4 \times 10^{-6}$), indicating consistent gauge-to-gauge modifications and a sensitivity coefficient greater than unity, where adjustments in simulated volume correspond to larger shifts in composite KGE values. Neither timing ($r = 0.34$, $p = 0.13$, ns) nor variability ($r = 0.20$, $p = 0.38$, ns) accounts for additional variance.

When transitioning from the baseline (`exp10`) to the joint configuration (`exp44`) in the Tugela basin, volumetric bias ($\Delta\beta^*$) operates as a primary control on KGE modifications for both model structures; its degradation corresponds to mHM performance losses, whereas its improvement supports the positive shifts observed in TETIS. Timing and variability adjustments remain secondary and are not significant within this semi-arid hydroclimatic regime.

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

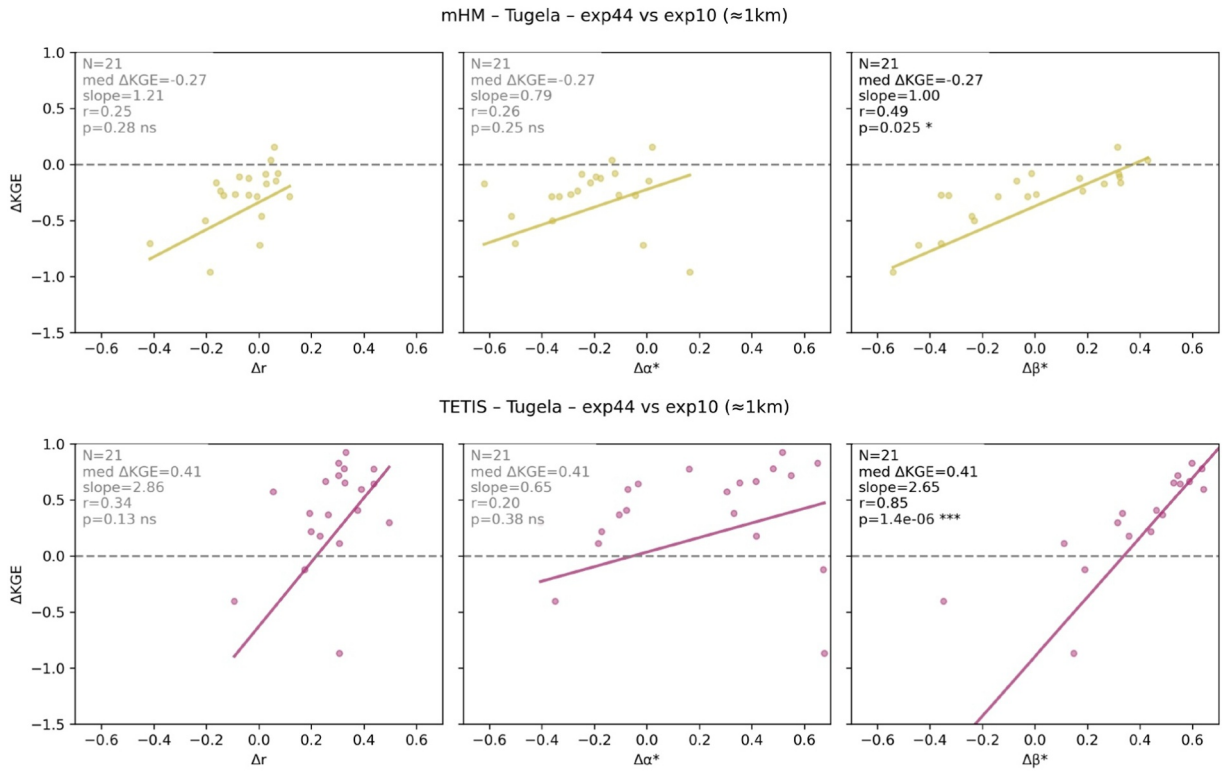


Figure B2. Attribution of performance drivers in the Tugela basin: ΔKGE (exp44 – exp10) versus Δr, $\Delta\alpha^*$, and $\Delta\beta^*$ at ≈1 km. Dots are gauges ($N = 21$ per model); lines are OLS fits; panels report N, median ΔKGE, slope, r, and p with significance codes. mHM shows a negative median ΔKGE, with a significant association to bias ($\Delta\beta^*$), indicating losses tied to degraded volume matching. TETIS displays a large positive median ΔKGE, driven predominantly by improvements in bias (a strong, significant $\Delta\beta^*$ link), while the relations with timing and variability are weaker and not statistically significant.

Appendix C: Calibration Resolution and Parameter Transfer

Table C1 states, model by model, the resolution at which the discharge calibration (`exp20`) was carried out, whether the calibrated parameters were transferred to the ≈1 km evaluation grid (`exp21`), and by which mechanism. The calibration algorithms and objective functions are described in Section 3.3 and, for `exp20/21`, in Modiri et al. (2025). In the EO-constrained experiments (`exp31–exp44`), mHM is calibrated directly at its ≈1 km grid (0.015625°).

Table C1
Calibration Resolution and Parameter Transfer in the Discharge-Calibrated Experiments (`exp20/21`)

Model	Calibration resolution (exp20)	Transferred to ≈1 km?	Transfer mechanism
CLM	≈1 km (0.0167°)	No	None; calibrated directly at the evaluation grid
JULES	≈5 km (0.0625°)	Yes	Calibrated parameter values applied at 0.015625°; no scaling scheme
mHM	≈5 km (0.0625°)	Yes	MPR: transfer-function parameters applied at 0.015625°
PCR-GLOBWB	≈5 km (3′)	Yes	Parameters from differentiable parameter learning applied at 30″; no scaling scheme
TETIS	≈5 km (0.0667°)	Yes	exp20-Calibrated corrector factors reused at 30″, with initial aquifer storage adjusted

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Acronyms

4DHydro	Hyper-resolution Earth Observations And Land-surface Modeling for A Better Understanding Of The Water Cycle
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station data
CLM	Community Land Model
ECMWF	European Center for Medium-Range Weather Forecasts
EMO	European Meteorological Observation
EO	Earth Observation
ERA-5	ECMWF ReAnalysis 5
ESA	European Space Agency
ET	Evapotranspiration
JULES	Joint UK Land Environment Simulator
KGE	Kling–Gupta Efficiency
mHM	mesoscale Hydrologic Model
MPR	Multiscale Parameter Regionalization
OLS	Ordinary least squares
PCR-GLOBWB	PCRaster Global Water Balance
SCA	Snow cover area
SM	Soil moisture
SPAEF	SPAtial EFficiency metric
SPF	Snow-presence flag
ECVs	Essential Climate Variables

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

All data sets generated and analysed in this study are archived in FAIR-compliant (Findable, Accessible, Interoperable, and Reusable) public repositories maintained by the individual modelling groups participating in the 4DHydro benchmarking initiative. These repositories provide persistent identifiers (DOIs), version control, and long-term access to the simulation outputs, flux data sets, and evaluation products used in this analysis. Given the multi-model, multi-experiment design of this study, the complete benchmarking archive comprises a large collection of individual data sets (including multiple spatial resolutions, calibration experiments, and basin configurations), each associated with its own repository record and DOI.

All data sets are federated and indexed through the 4DHydro Open Science Catalogue to facilitate transparent discovery and access, implemented following the SpatioTemporal Asset Catalog (STAC) standard: https://4dhydro.eu/catalog/

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Within the catalogue, data sets used in this study can be accessed under:

Products → WP5 Products → Hydrological Model Benchmarking.

The STAC framework provides structured metadata, direct repository links, and DOI references for each individual data set, enabling users to navigate from the catalogue landing page to the specific experiment, model configuration, and resolution required. Each catalogue record provides direct access to the underlying repository entry and associated DOI landing page for the corresponding data set. All data sets are publicly accessible via the repositories linked through the catalogue.

Examples of the repository records are given below; the complete list of records and DOIs, one per model, experiment, basin and resolution, is available through the STAC catalogue above.

- CLM simulations (Jülich DataPub): https://datapub.fz-juelich.de/slts/4DHydro/.
- JULES simulations (JASMIN): https://gws-access.jasmin.ac.uk/public/4dhydro/.
- mHM simulations (UFZ Data Management Portal): `exp10` https://doi.org/10.48758/ufz.15044 (Modiri et al., 2024a), `exp20` https://doi.org/10.48758/ufz.15172 (Modiri et al., 2024b) and `exp21` https://doi.org/10.48758/ufz.15173 (Modiri et al., 2024c).
- PCR-GLOBWB simulations (Utrecht University Yoda): https://doi.org/10.24416/UU01-KQQVAQ (Droppers et al., 2026).
- TETIS simulations (Zenodo, one record per experiment, basin and resolution), for example, https://doi.org/10.5281/zenodo.13253466 (`exp21`) (Cortés-Torres et al., 2024) and https://doi.org/10.5281/zenodo.15207599 (`exp44`) (Cortés-Torres et al., 2025).

The EO products used for calibration are available from their providers: NSIDC SMAP 1 km soil moisture https://doi.org/10.5067/U8QZ2AXE5V7B (Lakshmi & Fang, 2023) and MODIS-Terra evapotranspiration https://doi.org/10.5067/MODIS/MOD16A2GF.061 (Running et al., 2021). The HYPER-P precipitation data set is available at Zenodo via https://doi.org/10.5281/zenodo.15025514 (Filippucci et al., 2024), described in Filippucci et al. (2025). The gauge-level streamflow skill scores (KGE and its components) used for the figures and tables of this study are archived on Zenodo via https://doi.org/10.5281/zenodo.23472975 (Modiri, 2026b).

Model source codes used in this benchmarking study are publicly available through the following repositories:

1. CLM—https://github.com/HPSCTerrSys/CLM3.5 (Oleson et al., 2023)
2. JULES—https://code.metoffice.gov.uk/trac/jules/browser/main/branches/dev/douglasclark/vn7.7_4dhydro
3. mHM—https://github.com/mhm-ufz/mHM (Samaniego et al., 2023)
4. PCR-GLOBWB—https://github.com/UU-Hydro/PCR-GLOBWB_model (Sutanudjaja et al., 2023)
5. TETIS—https://gimha.upv.es/software/tetis (Francés et al., 2007; GIMHA, 2021)

The benchmarking and technical validation workflows developed for this study are available in the 4DHydro GitLab repository (Modiri, 2024), which includes post-processing scripts, evaluation metrics, and flux consistency diagnostics. The code that reproduces the figures and tables of this study, together with the benchmarking workflow, is archived on Zenodo via https://doi.org/10.5281/zenodo.23471091 (Modiri, 2026a).

Acknowledgments
The 4DHydro initiative is funded by the European Space Agency (ESA) under contract 4000141141/23/I-EF. The contribution of the TETIS team was partly supported by the Spanish Ministry of Science and Innovation through the research project TETISPREDICT (PID2022-141631OB-I00), and the JULES team was partly supported by the Natural Environment Research Council NE/X019063/1 (Hydro-JULES). This study utilises observational data from the Global Runoff Data Centre (GRDC). The research also benefited from access to European Meteorological Observations (EMO) data. Scientific results were computed at the EVE High-Performance Computing Cluster (HPC), a joint effort of Helmholtz Centre for Environmental Research – UFZ and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. We thank three anonymous reviewers and the Editor for constructive comments and suggestions. Open Access funding enabled and organized by Projekt DEAL.

References

Andreadis, K. M., & Lettenmaier, D. P. (2006). Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources*, *29*(6), 872–886. https://doi.org/10.1016/j.advwatres.2005.08.004

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., et al. (2015). ERA-Interim/Land: A global land surface reanalysis data set. *Hydrology and Earth System Sciences*, *19*(1), 389–407. https://doi.org/10.5194/hess-19-389-2015

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/jhm-d-14-0158.1

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., et al. (2011). The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes. *Geoscientific Model Development*, *4*(3), 677–699. https://doi.org/10.5194/gmd-4-677-2011

Bierkens, M. F., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., et al. (2015). Hyper-resolution global hydrological modelling: What is next? *Hydrological Processes*, *29*(2), 310–320. https://doi.org/10.1002/hyp.10391

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A., & Zemp, M. (2014). The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, *95*(9), 1431–1443. https://doi.org/10.1175/bams-d-13-00047.1

Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., et al. (2019). SM2RAIN–ASCAT (2007–2018): Global daily satellite rainfall data from ASCAT soil moisture observations. *Earth System Science Data*, *11*(4), 1583–1601. https://doi.org/10.5194/essd-11-1583-2019

Clark, D. B., Mercado, L. M., Sitch, S., Jones, C. D., Gedney, N., Best, M. J., et al. (2011). The joint UK land environment simulator (JULES), model description – Part 2: Carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, *4*(3), 701–722. https://doi.org/10.5194/gmd-4-701-2011

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., et al. (2015). Improving the representation of hydrologic processes in earth system models. *Water Resources Research*, *51*(8), 5929–5956. https://doi.org/10.1002/2015wr017096

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The international land model benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018ms001354

Cortés-Torres, N., Güiza-Villa, N., De León, D., Romero Hernandez, C. P., & Francés, F. (2024). 4DHydro results TETIS Po 7.5 arcsec daily exp21 1990–2022 [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.13253466

Cortés-Torres, N., Güiza-Villa, N., & Francés, F. (2025). 4DHydro results TETIS Tugela 7.5 arcsec daily exp44 2007–2022 [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.15207599

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, B. (2020). Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. *Water Resources Research*, *56*(1), e2019WR026085. https://doi.org/10.1029/2019wr026085

Dickinson, R. E., Oleson, K. W., Bonan, G., Hoffman, F., Thornton, P., Vertenstein, M., et al. (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. https://doi.org/10.1175/JCLI3742.1

Dietz, A. J., Kuenzer, C., Gessner, U., & Dech, S. (2012). Remote sensing of snow–a review of available methods. *International Journal of Remote Sensing*, *33*(13), 4094–4134. https://doi.org/10.1080/01431161.2011.640964

Droppers, B., Bierkens, M. F. P., & Wanders, N. (2025). A multi-resolution deep-learning surrogate framework for global hydrological models. *Water Resources Research*, *61*(4), e2024WR037736. https://doi.org/10.1029/2024WR037736

Droppers, B., Rakovec, O., Avila, L., Azimi, S., Cortés-Torres, N., De León Pérez, D., et al. (2024). Multi-model hydrological reference dataset over continental Europe and an African basin. *Scientific Data*, *11*(1), 1009. https://doi.org/10.1038/s41597-024-03825-9

Droppers, B., Wanders, N., & Bierkens, M. F. P. (2026). PCR-GLOBWB output – 4DHydro working package 5 (version 4.2) [Dataset]. *Utrecht University – Yoda*. https://doi.org/10.24416/UU01-KQQVAQ

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., et al. (2006). Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, *320*(1–2), 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., et al. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *Journal of Hydrology*, *537*, 45–60. https://doi.org/10.1016/j.jhydrol.2016.03.026

Filippucci, P., Brocca, L., Ciabatta, L., Mosaffa, H., Avanzi, F., & Massari, C. (2024). HYPER-P: HYdroclimatic PERformance-enhanced Precipitation at 1 km/daily over the Europe region from 2007 to 2022 (version 1.0) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.15025514

Filippucci, P., Brocca, L., Ciabatta, L., Mosaffa, H., Avanzi, F., & Massari, C. (2025). Development of HYPER-P: HYdroclimatic PERformance-enhanced precipitation at 1 km/daily over the europe-mediterranean region from 2007 to 2022. *Earth System Science Data*, *17*(10), 5221–5258. https://doi.org/10.5194/essd-17-5221-2025

Finger, D., Vis, M., Huss, M., & Seibert, J. (2015). The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments. *Water Resources Research*, *51*(4), 1939–1958. https://doi.org/10.1002/2014wr015712

Francés, F., Vélez, J. I., & Vélez, J. J. (2007). Split-parameter structure for the automatic calibration of distributed hydrological models. *Journal of Hydrology*, *332*(1), 226–240. https://doi.org/10.1016/j.jhydrol.2006.06.032

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, *2*(1), 1–21. https://doi.org/10.1038/sdata.2015.66

García-García, A., Stradiotti, P., Di Paolo, F., Filippucci, P., Fischer, M., Orság, M., et al. (2026). Intercomparison of Earth observation products for hyper-resolution hydrological modelling over Europe. *Remote Sensing of Environment*, *333*, 115131. https://doi.org/10.1016/j.rse.2025.115131

GIMHA. (2021). TETIS: V.9.1 [Software]. *(Grupo de investigación en modelación hidrólogica y ambiental distribuida (GIMHA), Instituto de Ingeniería del Agua y el Medio Ambiente (IIAMA), Universitat Politècnica de València)*. Retrieved from https://gimha.upv.es/software/tetis/

Gupta, H. V., Kling, H., Yilmaz, K. K., & 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. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Mu noz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., & Yoo, S.-H. (2015). NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). *Algorithm theoretical basis document (ATBD) version*, *4*(26), 30.

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Koch, J., Demirel, M. C., & Stisen, S. (2018). The SPAtial EFficiency metric (SPAEF): Multiple-component evaluation of spatial patterns for optimization of hydrological models. *Geoscientific Model Development*, *11*(5), 1873–1886. https://doi.org/10.5194/gmd-11-1873-2018

Kumar, R., Livneh, B., & Samaniego, L. (2013). Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme. *Water Resources Research*, *49*(9), 5700–5714. https://doi.org/10.1002/wrcr.20431

Kumar, S. V., Peters-Lidard, C. D., Mocko, D., Reichle, R., Liu, Y., Arsenault, K. R., et al. (2014). Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation. *Journal of Hydrometeorology*, *15*(6), 2446–2469. https://doi.org/10.1175/JHM-D-13-0132.1

Kunnath-Poovakka, A., Ryu, D., Renzullo, L., & George, B. (2016). The efficacy of calibrating hydrologic model using remotely sensed evapotranspiration and soil moisture for streamflow prediction. *Journal of Hydrology*, *535*, 509–524. https://doi.org/10.1016/j.jhydrol.2016.02.018

Lakshmi, V., & Fang, B. (2023). SMAP-derived 1-km downscaled surface soil moisture product, version 1 [Dataset]. *NASA National Snow and Ice Data Center Distributed Active Archive Center*. https://doi.org/10.5067/U8QZ2AXE5V7B

Liu, Y., Weerts, A., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., et al. (2012). Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences*, *16*(10), 3863–3887. https://doi.org/10.5194/hess-16-3863-2012

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Martínez-De La Torre, A., Blyth, E. M., & Weedon, G. P. (2019). Using observed river flow data to improve the hydrological functioning of the JULES land surface model (vn4.3) used for regional coupled modelling in Great Britain (UKC2). *Geoscientific Model Development*, *12*(2), 765–784. https://doi.org/10.5194/gmd-12-765-2019

Mei, Y., Mai, J., Do, H. X., Gronewold, A., Reeves, H., Eberts, S., et al. (2023). Can hydrological models benefit from using global soil moisture, evapotranspiration, and runoff products as calibration targets? Water. *Resource Research*, *59*(2), e2022WR032064. https://doi.org/10.1029/2022wr032064

Melsen, L. A., Teuling, A. J., Torfs, P. J., Zappa, M., Mizukami, N., Mendoza, P. A., et al. (2019). Subjective modeling decisions can significantly impact the simulation of flood and drought events. *Journal of Hydrology*, *568*, 1093–1104. https://doi.org/10.1016/j.jhydrol.2018.11.046

Merz, R., Parajka, J., & Blöschl, G. (2009). Scale effects in conceptual hydrological modeling. *Water Resources Research*, *45*(9). https://doi.org/10.1029/2009WR007872

Miralles, D. G., Gentine, P., Seneviratne, S. I., & Teuling, A. J. (2019). Land–atmospheric feedbacks during droughts and heatwaves: State of the science and current challenges. *Annals of the New York Academy of Sciences*, *1436*(1), 19–35. https://doi.org/10.1111/nyas.13912

Modiri, E. (2024). 4DHydro WP5 benchmarking code [Software]. https://codebase.helmholtz.cloud/4dhydro/wp5/benchmarking

Modiri, E. (2026a). Codebase for: Assessing the added value of Earth observations for multi-model kilometre-scale hydrological modelling across contrasting basins (version 1.1.0) [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.23471091

Modiri, E. (2026b). Derived datasets for: Assessing the added value of Earth observations for multi-model kilometre-scale hydrological modelling across contrasting basins (version v1.0) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.23472975

Modiri, E., Rakovec, O., & Samaniego, L. (2024a). 4DHydro mHM tier 2 simulations – exp10 [Dataset]. *Helmholtz Centre for Environmental Research – UFZ*. https://doi.org/10.48758/ufz.15044

Modiri, E., Rakovec, O., & Samaniego, L. (2024b). 4DHydro mHM tier 2 simulations – exp20 [Dataset]. *Helmholtz Centre for Environmental Research – UFZ*. https://doi.org/10.48758/ufz.15172

Modiri, E., Rakovec, O., & Samaniego, L. (2024c). 4DHydro mHM tier 2 simulations – exp21 [Dataset]. *Helmholtz Centre for Environmental Research – UFZ*. https://doi.org/10.48758/ufz.15173

Modiri, E., Rakovec, O., Shrestha, P. K., Droppers, B., Avila, L., Azimi, S., et al. (2025). A multi-scale intercomparison of land surface and hydrological models: Benchmarking performance and flux consistency across spatial resolutions [Preprint]. *ESS Open Archive*. https://doi.org/10.22541/essoar.174982768.80043676/v2

Montanari, A. (2012). Hydrology of the Po river: Looking for changing patterns in river discharge. *Hydrology and Earth System Sciences*, *16*(10), 3739–3747. https://doi.org/10.5194/hess-16-3739-2012

Muñoz Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., et al. (2018). Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, *54*(10), 8332–8362. https://doi.org/10.1029/2017wr021895

Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., & Gulden, L. E. (2005). A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *Journal of Geophysical Research*, *110*(D21). https://doi.org/10.1029/2005JD006111

Notarnicola, C., Duguay, M., Moelg, N., Schellenberger, T., Tetzlaff, A., Monsorno, R., et al. (2013). Snow cover maps from MODIS images at 250 m resolution, part 1: Algorithm description. *Remote Sensing*, *5*(1), 110–126. https://doi.org/10.3390/rs5010110

Oleson, K., Niu, G.-Y., Yang, Z.-L., Lawrence, D., Thornton, P., Lawrence, P., et al. (2023). CLM3.5 [Software]. https://github.com/HPSCTerrSys/CLM3.5

Parajka, J., & Blöschl, G. (2008). The value of MODIS snow cover data in validating and calibrating conceptual hydrologic models. *Journal of Hydrology*, *358*(3–4), 240–258. https://doi.org/10.1016/j.jhydrol.2008.06.006

Rajib, M. A., Merwade, V., & Yu, Z. (2016). Multi-objective calibration of a hydrologic model using spatially distributed remotely sensed/in-situ soil moisture. *Journal of Hydrology*, *536*, 192–207. https://doi.org/10.1016/j.jhydrol.2016.02.037

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., et al. (2016). Multiscale and multivariate evaluation of water fluxes and States over European River basins. *Journal of Hydrometeorology*, *17*(1), 287–307. https://doi.org/10.1175/JHM-D-15-0054.1

Refsgaard, J. C. (1997). Parameterisation, calibration and validation of distributed hydrological models. *Journal of Hydrology*, *198*(1–4), 69–97. https://doi.org/10.1016/s0022-1694(96)03329-x

Reichle, R. H. (2008). Data assimilation methods in the Earth sciences. *Advances in Water Resources*, *31*(11), 1411–1418. https://doi.org/10.1016/j.advwatres.2008.01.001

Running, S., Mu, Q., Zhao, M., & Moreno, A. (2021). MODIS/terra net evapotranspiration gap-filled 8-day L4 global 500 m SIN grid V061 [Dataset]. *NASA Land Processes Distributed Active Archive Center*. https://doi.org/10.5067/MODIS/MOD16A2GF.061

Salamon, P., Ramos Gomes, G. N., Sperzel, T., Radke-Fretz, M., Schweim, C., Ziese, M., et al. (2026). EMO: A high-resolution multi-variable gridded meteorological data set for Europe [Dataset]. *European Commission, Joint Research Centre*. https://doi.org/10.2905/0BD84BE4-CEC8-4180-97A6-8B3ADAAC4D26

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2008wr007327

Samaniego, L., Kumar, R., Zink, M., Cuntz, M., Mai, J., Thober, S., et al. (2023). mHM: V5.13.1 [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.8279545

Sheffield, J., Wood, E. F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A., & Verbist, K. (2018). Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions. *Water Resources Research*, *54*(12), 9724–9758. https://doi.org/10.1029/2017wr022437

Smith, M. B., Koren, V., Zhang, Z., Zhang, Y., Reed, S. M., Cui, Z., et al. (2012). Results of the DMIP 2 Oklahoma experiments. *Journal of Hydrology*, *418–419*, 17–48. https://doi.org/10.1016/j.jhydrol.2011.08.056

Stahl, K., Weiler, M., Freudiger, D., Kohn, I., Seibert, J., Vis, M., et al. (2016). *The snow and glacier melt components of the streamflow of the river Rhine and its tributaries considering the influence of climate change (CHR Report No. I-25)*. International Commission for the Hydrology of the Rhine Basin (CHR).

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K.-L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, *56*(1), 79–107. https://doi.org/10.1002/2017rg000574

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., et al. (2023). PCR-GLOBWB model [Software]. https://github.com/UU-Hydro/PCR-GLOBWB_model

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, *12*(1), 5988. https://doi.org/10.1038/s41467-021-26107-z

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

van Beek, R., & Bierkens, M. F. (2009). *The global hydrological model PCR-GLOBWB: Conceptualization, parameterization and verification* (Tech. Rep.). Department of Physical Geography, Utrecht University. Retrieved from https://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf

Vélez, J. J., Puricelli, M., López Unzu, F., & Francés, F. (2009). Parameter extrapolation to ungauged basins with a hydrological distributed model in a regional framework. *Hydrology and Earth System Sciences*, *13*(2), 229–246. https://doi.org/10.5194/hess-13-229-2009

Wada, Y., Wisser, D., & 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. https://doi.org/10.5194/esd-5-15-2014

Wade, M., O'Brien, G., & Jewitt, G. (2024). The use and protection of water resources in the uThukela River catchment, South Africa: A review of its past, present and future. *African Journal of Aquatic Science*, *49*(4), 283–297. https://doi.org/10.2989/16085914.2024.2404562

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometric Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Wilks, D. S. (2016). "the stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, *97*(12), 2263–2273. https://doi.org/10.1175/bams-d-15-00267.1

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L., Bierkens, M. F., Blyth, E., et al. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*(5). https://doi.org/10.1029/2010wr010090

19447973, 2026, 12, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR042446 by Forschungszentrum Juelich GmbH, Wiley Online Library on [26/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License