Water Resources Research

RESEARCH ARTICLE
10.1029/2025WR042375

Assessing the Impact of Geological Map Detail on Process-Based and Data-Driven Hydrological Models

Thiago V. M. do Nascimento[1,2], Julia Rudlang[3], Sebastian Gnann[4], Jan Seibert[2], Markus Hrachowitz[3], and Fabrizio Fenicia[1]

[1]Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland, [2]Department of Geography, University of Zurich, Zurich, Switzerland, [3]Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands, [4]Chair of Hydrology, Faculty of Environment and Natural Resources, University of Freiburg, Freiburg, Germany

Key Points:

- We quantify how the use of geological detail influences streamflow predictions in ungauged basins using process-based and data-driven models
- Increasing geological detail leads to consistent improvements in model performance under space–time evaluation
- Detailed geology improves streamflow signature representation where subsurface processes matter, even when NSE gains are small

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Thiago V. M. do Nascimento,
thiago.nascimento@eawag.ch

Citation:

do Nascimento, T. V. M., Rudlang, J., Gnann, S., Seibert, J., Hrachowitz, M., & Fenicia, F. (2026). Assessing the impact of geological map detail on process-based and data-driven hydrological models. *Water Resources Research*, *62*, e2025WR042375. https://doi.org/10.1029/2025WR042375

Received 29 SEP 2025
Accepted 5 MAY 2026

Author Contributions:

Conceptualization: Thiago V. M. do Nascimento, Fabrizio Fenicia
Data curation: Thiago V. M. do Nascimento, Julia Rudlang, Markus Hrachowitz, Fabrizio Fenicia
Formal analysis: Thiago V. M. do Nascimento
Funding acquisition: Markus Hrachowitz, Fabrizio Fenicia
Investigation: Thiago V. M. do Nascimento
Methodology: Thiago V. M. do Nascimento, Fabrizio Fenicia
Project administration: Markus Hrachowitz, Fabrizio Fenicia

Abstract Although large-sample hydrology data sets are increasingly used to advance predictions in ungauged basins, the influence of landscape data quality on model regionalization remains insufficiently explored. This study investigates whether geological catchment attributes derived from maps of increasing detail—global, continental, and regional—improve parameter transfer and model regionalization. To ensure robustness across model approaches, we applied both a semi-distributed process-based hydrological model using hydrological response units (HRUs) and a data-driven Long Short-Term Memory (LSTM) model. The analysis covered a total of 130 catchments in the Moselle (Central Europe) and Garonne (southwestern France) basins. We conducted five model experiments differing only in the representation of geological information: a benchmark without geology, a benchmark with random geology classes, and configurations based on the global-, continental-, and regional-scale geological maps. Model performance was evaluated using a modified Nash-Sutcliffe (NSE) metric for daily streamflow, as well as Pearson correlation and relative bias for three streamflow signatures: baseflow index, slope of the flow duration curve, and half-flow date. Across both basins and modeling frameworks, increasing geological detail consistently improved predictive performance under space–time evaluation. While differences in NSE were modest, improvements were pronounced for streamflow signatures: only models using the more detailed geological information, especially the regional map, consistently reproduced spatial variability in baseflow and flow regime characteristics. These findings highlight the importance of integrating high-quality geological data into hydrological modeling, particularly for improving predictions in ungauged basins through more reliable parameter transfer and regionalization.

1. Introduction

In recent years, the availability of large-sample hydrology (LSH) data sets has considerably increased (e.g., Addor et al., 2017; Do et al., 2018; Kratzert et al., 2023; do Nascimento et al., 2024). These data sets usually provide long time series of streamflow and meteorological forcing data, together with a standardized set of catchment attributes describing the landscape, such as topography, soils, vegetation, and geology. Their emergence has enabled systematic evaluation of hydrological models across large numbers of catchments and plays a central role in advancing prediction in ungauged basins (PUB) (Hrachowitz et al., 2013; Kratzert et al., 2019; Nearing et al., 2024; Shen et al., 2023).

Catchment attributes available in these LSH data sets are typically derived from globally or continentally available geospatial data sets and are summarized into basin-averaged indices or class fractions (Addor et al., 2017; Do et al., 2018; Kratzert et al., 2023). Their inclusion is motivated by the understanding that landscape characteristics influence catchment response (Winter 2001), and by the assumption that basin-averaged attributes provide a sufficiently informative representation of spatially heterogeneous landscapes to allow meaningful relationships with hydrological behavior to be identified. This assumption has been shown to hold to varying degrees across a wide range of large-sample studies, depending on the processes considered and the spatial hererogeinity of the used data (Addor et al., 2018; Bassi et al., 2024; Fenicia & McDonnell, 2022; Klotz et al., 2025; Kratzert et al., 2019; Kuentz et al., 2017; Rudlang et al., 2025; Wagener et al., 2007; Wu et al., 2021).

Particularly in PUB applications, catchment attributes play a central role by enabling extrapolation beyond gauged basins, either by informing parameter regionalization in conceptual models (Fenicia et al., 2022; Hrachowitz et al., 2013; Pool et al., 2021) or by providing static input information in data-driven models such as Long

Supervision: Jan Seibert, Markus Hrachowitz, Fabrizio Fenicia
Validation: Thiago V. M. do Nascimento
Writing – original draft: Thiago V. M. do Nascimento
Writing – review & editing: Thiago V. M. do Nascimento, Julia Rudlang, Sebastian Gnann, Jan Seibert, Markus Hrachowitz, Fabrizio Fenicia

Short-Term Memory (LSTM) networks (Kratzert et al., 2019) or hybrid approaches (Shen et al., 2023). While these implementations differ fundamentally in structure, they all rely on the assumption that catchment attributes encode spatially transferable information that is relevant to hydrological processes controlling streamflow generation.

Despite this widespread use, it remains largely unclear which catchment attributes contribute to predictive skill in such models, which components of the hydrograph they influence, and whether improvements in predictive performance attributed to the inclusion of specific landscape attributes reflect physically meaningful information rather than incidental correlations (Heudorfer et al., 2025; do Nascimento et al., 2025). This knowledge gap is particularly relevant when models are evaluated under extrapolation conditions, where observations are not available.

The predictive value of catchment attributes in hydrological models depends on two distinct factors. First, many attributes are derived as highly abstracted statistics summarizing complex and heterogeneous landscapes, which may limit their ability to represent key hydrological controls (Floriancic et al., 2022; Holt & McMillan, 2025; do Nascimento et al., 2025; Tarasova et al., 2024). Second, the informativeness of attributes also depends on the quality of the underlying geospatial data sets. This is particularly critical for subsurface attributes, where uncertainty, coarse spatial resolution, and conceptual simplifications can propagate into model predictions (Gnann et al., 2021; Holt & McMillan, 2025). Despite the central role of subsurface processes in streamflow, the role of geological data uncertainty and simplification for hydrological model predictions, although widely acknowledged (Addor et al., 2018; Kratzert et al., 2019; Kuentz et al., 2017; Rudlang et al., 2025), remains poorly explored.

A recent study by do Nascimento et al. (2025) investigated catchment attributes in Europe derived from geological maps with different levels of detail—global, continental, and regional—and showed that their correlations with streamflow signatures vary markedly with map detail. In the Moselle basin in particular, depending on the map detail used, geology shifted from being among the least correlated attribute groups to the most strongly correlated as map detail increased, surpassing other climate, soil and topography attributes. While such correlation-based analyses provide valuable insights into the potential information content of landscape data, they do not reveal how this information propagates through hydrological models, nor whether its apparent value persists under extrapolation conditions relevant for PUB.

In this study, we systematically investigate how geological input data of varying informational content influence streamflow predictions in ungauged basins. Here, the term geological detail does not exclusively refer to spatial resolution, but also accounts for the number of lithological classes these maps include and how finely they delineate geological boundaries. Specifically, we assess whether potential performance improvements associated with more detailed geological information are robust across different modeling paradigms and hydro-climatic contexts.

Here, we compare two contrasting hydrological modeling approaches: (a) a semi-distributed, bucket-type process-based hydrological model following Fenicia et al. (2022), and (b) a data-driven LSTM model implemented following the setup of Kratzert et al. (2019). This comparison is informative because the two model classes are fundamentally different modeling approaches. Process-based models reflect hypotheses about dominant processes and how landscape variability influences them. In our implementation, landscape variability is represented solely through geological variability, allowing us to isolate the effect of different geological discretizations. In contrast, LSTM models learn relationships directly from data and are less constrained by a priori assumptions. By training LSTM models on all available landscape attributes, we assess whether they can compensate for limited or absent geological information by exploiting correlations with other attributes.

The specific aims of this study are to:

1. Quantify the impact of geological data detail on overall streamflow predictive performance under ungauged conditions.
2. Identify which aspects of the simulated hydrograph are most affected by geological data detail, using a set of hydrological signatures.
3. Assess whether the effects of geological data detail are consistent across model classes, comparing process-based and machine-learning approaches.
4. Evaluate the regional robustness and transferability of these effects across contrasting catchments

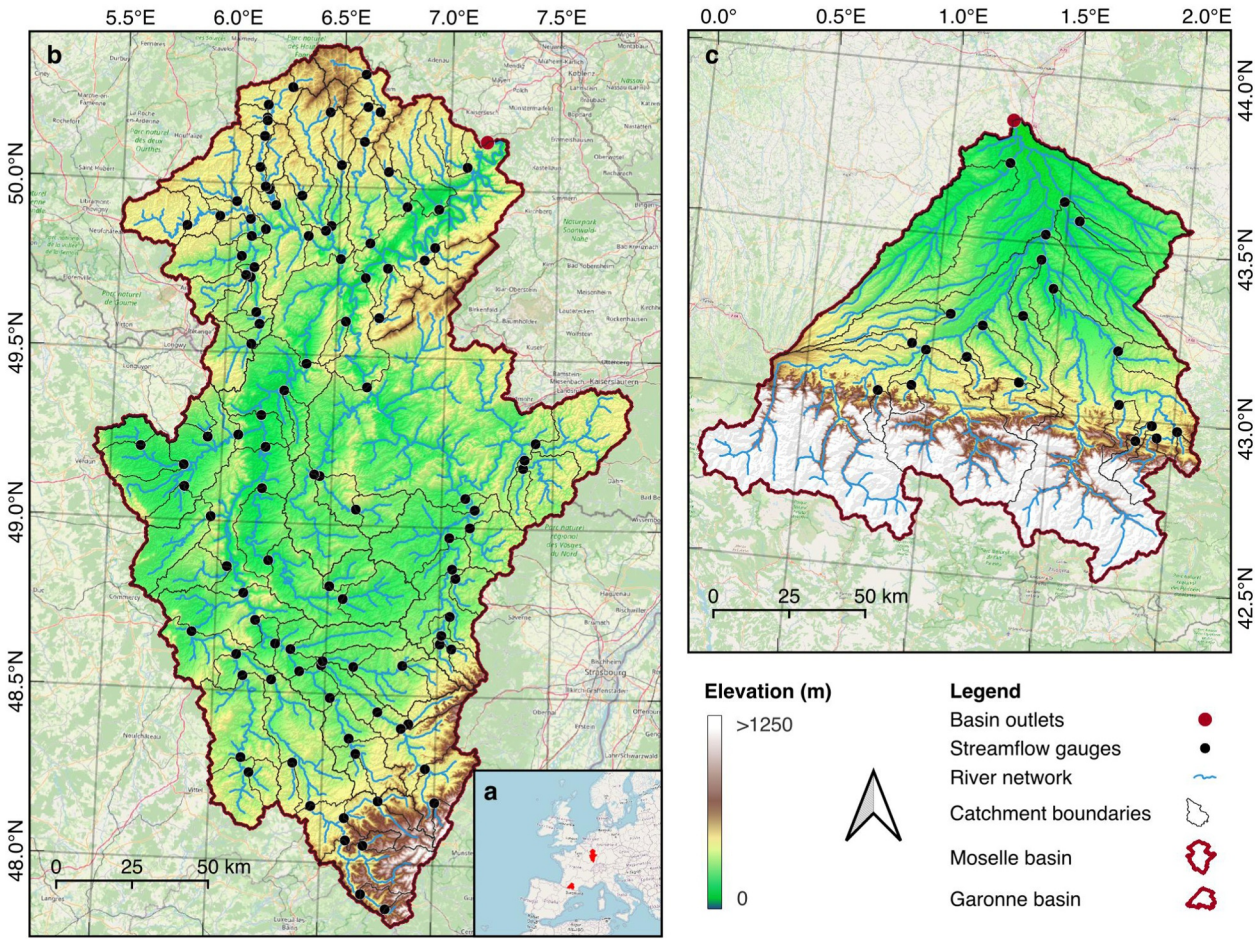

Figure 1. Location of the Moselle (b) and Garonne (c) river basins within Europe (a). Catchment boundaries, river networks, and topography are shown, together with streamflow gauge locations (black). Basin outlets are indicated by red circles. Base map data © OpenStreetMap contributors, obtained via QGIS and available at https://www.openstreetmap.org (last access: 25 January 2026).

To address these aims, we analyze two river basins, the Moselle and the Garonne, comprising a total of 130 internal catchments. Previous studies have shown that geology seems to exert a strong control on the spatial variability of streamflow in both basins (Fenicia et al., 2022; Fenicia & McDonnell, 2022; Hellebrand et al., 2007; do Nascimento et al., 2025; Pfister et al., 2017). Despite this common control, the two basins exhibit contrasting hydrological behavior. The Garonne is characterized by stronger topographic gradients and a greater influence of snow, leading to distinct hydrograph signatures, such as differences in flow duration curve (FDC) slope and half-flow date, as discussed later in the manuscript.

The paper is structured as follows: Section 2 introduces the study area and data used. Section 3 describes the methods applied. Section 4 presents the main results. Section 5 discusses the results, and Section 6 summarizes the main conclusions.

2. Data

2.1. Study Area

This study focuses on two river basins in western and central Europe: sub-regions of the Moselle and of the Garonne basins. Figure 1 shows the locations of both basins within the European context, together with the 130 gauged catchments used in the analysis and their underlying catchment boundaries and river network.

The Moselle basin is shown in Figure 1b. We consider the portion upstream of Cochem, approximately 50 km upstream from the confluence with the Rhine River in Germany. The basin spans parts of northeastern France,

western Germany, Belgium and Luxembourg and covers an area of 27,100 km^2, with elevation ranging from 60 to 1,424 m. Its land use is mainly forests, followed by agriculture, and pastures (Fenicia & McDonnell, 2022; do Nascimento et al., 2025). Annual precipitation ranges between 800 and 1,500 mm yr^{-1}, while potential evapotranspiration (PET) is more uniform across the basin, ranging from 700 to 850 mm yr^{-1} (do Nascimento et al., 2025). Soils are heterogeneous, with coarse-textured substrates in the south, medium-textured soils in the north, and finer materials concentrated in the central area. Geologically, the area is dominated by sedimentary and metamorphic formations (Fenicia & McDonnell, 2022).

The Garonne basin is shown in Figure 1c. We consider the portion upstream of Verdun-sur-Garonne, corresponding to a contributing area of approximately 13,730 km^2. This choice reflects data availability and ensures consistency with the catchment-scale modeling framework adopted in this work. The Garonne catchment spans elevations ranging from 90 to 3,191 m and integrates contrasting physiographic units draining the northern slopes of the Pyrenees Mountains and the southern slopes of the Massif Central (Caballero et al., 2007; Martin et al., 2016).

The climate of the Garonne basin is predominantly influenced by Atlantic conditions, characterized by relatively high winter precipitation and warm, humid summers, while Mediterranean influences become more pronounced toward the southeastern parts of the wider basin. Mean annual precipitation ranges between 600 and 1,200 mm yr^{-1} (Martin et al., 2016), while PET ranges from 800 to 960 mm yr^{-1}. About 50% of the basin is covered by agriculture (irrigated or not) and 28% is covered by forests, while soils and geological formations reflect the transition between the Pyrenean domain, the Massif Central, and the intervening lowland plain.

Compared to the Moselle, the Garonne basin is subject to a higher influence of snow and human influence, including flow regulation and agricultural and industrial water use (Caballero et al., 2007; Fenicia & McDonnell, 2022). Human influences are not represented explicitly in the models used in this study. However, since all geological experiments are conducted under identical modeling assumptions and forcing data, the relative differences in performance between experiments remain informative.

The catchment delineations and attributes as well as the meteorological forcing time series data used in this study were obtained from the EStreams data set (do Nascimento et al., 2024) version 1.4. In this version, all forcing variables were obtained from the E-OBS ensemble mean product with a spatial resolution of 0.1° in both latitude and longitude (Cornes et al., 2018). For the 22 catchments located in the Garonne basin, previous analyses have shown that E-OBS precipitation estimates can differ considerably in this region (Clerc-Schwarzenbach & do Nascimento, 2026) when compared to the Catchment Attributes and Meteorology for Large-sample Studies for France data set (CAMELS-FR) (Delaigue et al., 2024). We therefore applied a simplified bias correction to the E-OBS precipitation time series for these 22 catchments, using CAMELS-FR as a reference data set. Details of the bias-correction procedure are provided in Text S2 in Supporting Information.

Daily streamflow data were obtained using the streamflow catalog for Europe provided within the framework of the EStreams project (www.estreams.eawag.ch, last access: 22 December 2025). Although the daily streamflow data cannot be redistributed, the catalog gives the necessary information to access daily streamflow data from their respective data providers. Because data quality varies across the catchments included in EStreams, we applied the following selection criteria to identify catchments suitable for this study:

- Located within the Moselle or Garonne basins.
- High-quality catchment delineations, as defined by do Nascimento et al. (2024).
- Catchment area larger than 50 km^2.
- At least 14 years of daily streamflow data (not necessarily consecutive) recorded between 1988 and 2015.
- Average streamflow not exceeding 10 mm day^{-1} or runoff ratio between 0.1 and 1.1, as such values may indicate water balance inconsistencies or major anthropogenic impacts.
- Visual inspection of the observed streamflow to exclude series with anomalous behavior.

Following these criteria, a total of 108 catchments in the Moselle basin and 22 catchments in the Garonne basin were selected for subsequent analyses. The complete list of selected catchments is provided in Tables S1 and S2 in Supporting Information S1.

2.2. Geological Maps

In this study we used the same geological maps used by do Nascimento et al. (2025): a global, a continental, and a regional map. These maps differ in spatial resolution and geological information content, reflecting differences in data sources, classification schemes, and intended scale of application. Rather than emphasizing individual lithological units, we focus on how differences in geological detail translate into contrasting representations of subsurface permeability at the basin scale. Table S3 in Supporting Information S1 presents an overview of major lithological classes represented in each map, their sources, the number of original distinct lithological classes, and their subsequent reclassification into relative permeability classes (see Section 2.3). The original geological maps were obtained from the following sources:

- The global scale geological map used was derived from the Global Lithological Map (GLiM) (Hartmann et al., 2012). Importantly, GLiM is based on a compilation of national and continental geological maps that were harmonized into a unified lithological classification. The underlying sources include the Bureau de Recherches Géologiques et Minières (BRGM, 2003) for France; Trurnit et al. (2003) for Germany; the Instituto Geológico y Minero de España (IGME, 1994) for Spain; and data sets from the One Geology Europe Consortium (Surface geological maps of Europe, 2010, available at http://www.onegeology-europe.org/, last accessed 17 January 2011) for Belgium and Luxembourg (Hartmann et al., 2012).
- The continental scale geological map was obtained from the International Hydrogeological Map of Europe (IHME), version 11, available at www.bgr.bund.de, last access: 01 January 2026, with a scale of 1:1,500,000 (Duscher et al., 2019; Günther & Duscher, 2019).
- The regional scale geological maps for the Moselle are an aggregation of four different geological maps sourced from different providers: France: BD LISA database (version 1, niveau 2, ordre 1, scale: 1:250,000, downloaded at https://bdlisa.eaufrance.fr, last access: 11 December 2021); Germany: Geologische Übersichtskarte der Bundesrepublik Deutschland (GÜK200) (scale: 1:200,000, downloaded at www.bgr.bund.de, last access: 11 December 2021); Luxembourg: The map was obtained from the "Administration de la gestion de l'eau" (at a scale of 1:250,000, and available at https://eau.gouvernement.lu/fr.html, last access: 11 December 2021); and Belgium: Information from the continental-scale IHME database (Duscher et al., 2019; Günther & Duscher, 2019). This means that when these country-based maps were concatenated, their lithological classes were translated into a harmonized product. See (Fenicia & McDonnell, 2022) for further details.
- The regional scale geological map for the Garonne was obtained from two different providers: France: BD LISA database (version 1, niveau 2, ordre 1, scale: 1:250,000, downloaded at https://bdlisa.eaufrance.fr, last access: 01 January 2026); and Spain: the IHME database (Duscher et al., 2019; Günther & Duscher, 2019). The concatenation and harmonization followed the same procedure as for the Moselle.

2.3. Reclassification of the Geological Maps

We reclassified the geological information into three relative permeability classes (i.e., low, medium, and high) following the approach introduced by Fenicia et al. (2022) and do Nascimento et al. (2025) for both the Moselle and the Garonne. The details of this reclassification, including the association of each geological class with its corresponding relative permeability class, are shown in Table S3 of Supporting Information S1. While the same three permeability classes are used in all experiments, their areal coverage varies across geological data sets and between basins.Table 1 summarizes the percentage of basin area assigned to each permeability class for the global, continental, and regional geological maps in both basins. The spatial distribution of these three permeability classes is shown in Figure 2 across the Moselle and the Garonne basins.

Together, Figure 2 and Table 1 illustrate that increasing geological detail does not merely refine spatial resolution, but can lead to markedly different representations of basin-scale relative permeability composition. For example, while the distributions of permeability are very similar for the continental and regional map for both the Moselle and Garonne, particularly the areal fraction of the low permeability class estimated from the global map is with <4% considerably lower in both basins, compared to the estimates obtained from the continental and regional maps that each exceed 19%.

Table 1
Areal Coverage (%) of Relative Permeability Classes for the Moselle and Garonne Basins Derived From the Global, Continental, and Regional Geological Maps

Basin	Geological map (%)	Low permeability (%)	Medium permeability (%)	High permeability (%)
Moselle	Global	4.0	65.0	31.0
	Continental	22.7	54.8	22.5
	Regional	19.0	56.0	25.0
Garonne	Global	3.0	68.0	29.0
	Continental	26.0	57.0	17.0
	Regional	39.0	47.0	14.0

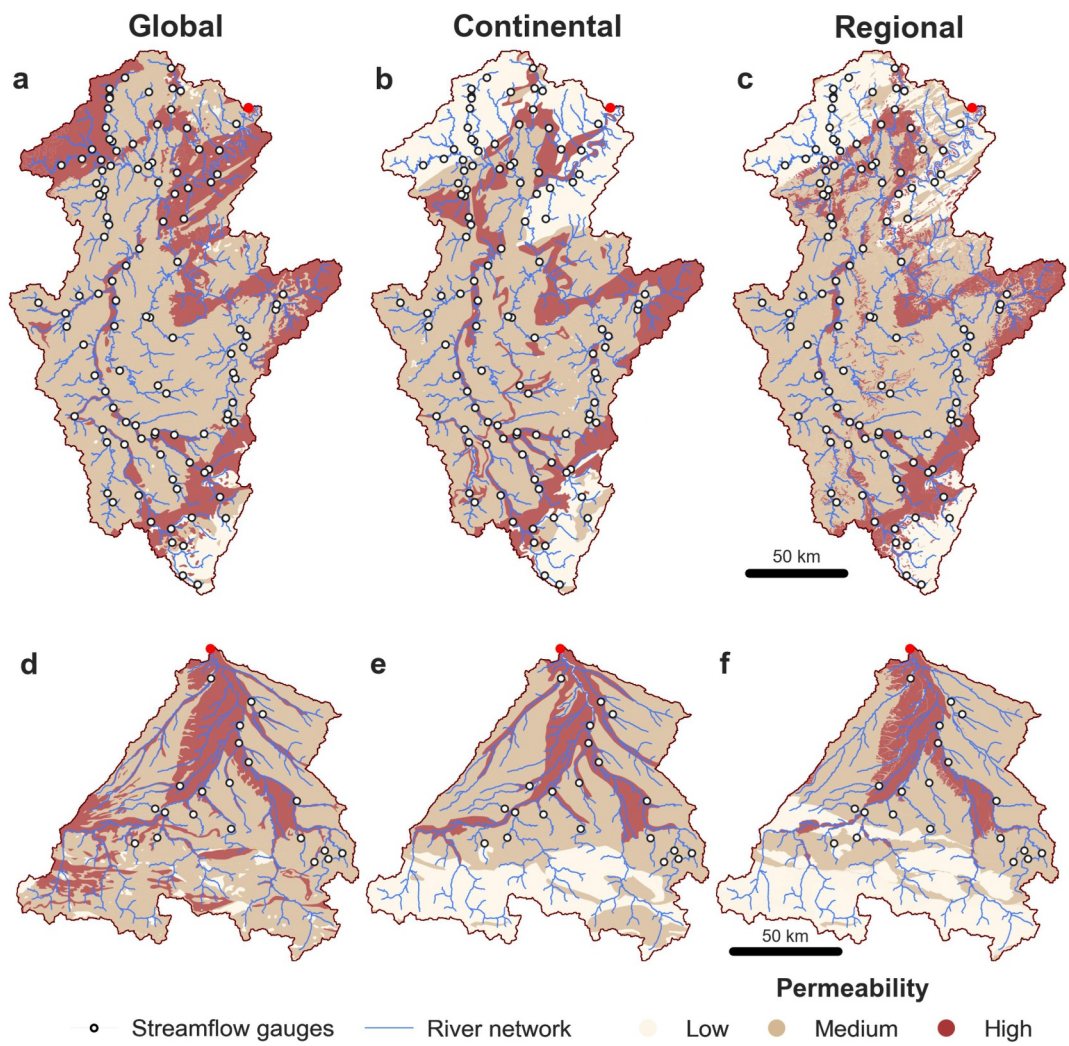

Figure 2. Maps of relative permeability classes used in this work for the Moselle (a–c) and the Garonne (d–f) basins, derived from global-, continental- and regional-scale geological maps. The river network is shown in blue, the streamflow gauges of the catchments are represented by white dots, and the basin outlet by larger red circles.

3. Methods

3.1. Models

To enable a structured comparison between modeling approaches, we employed a process-based model and a data-driven LSTM model. Below, we summarize their key commonalities and differences.

Common feature:

- Both models are trained using multi-catchment, shared-parameter calibration strategies, commonly referred to as "regional models" (Kratzert et al., 2019). To avoid confusion with the spatial scale of geological data sets (global, continental, regional), we do not use this term further in this study.

Key differences:

- Representation of spatial heterogeneity: In the process-based model, parameters are HRU-specific, and spatial heterogeneity is represented through the varying proportions of HRUs across catchments (Fenicia et al., 2022). In our implementation, HRUs are defined solely based on geology to isolate the effect of geological classification on model results.
- Use of geological information: In the LSTM, geological information is provided implicitly alongside all other static catchment attributes, rather than explicitly structuring the model.
- Training strategy: The process-based model is calibrated separately for the Moselle and Garonne basins, whereas the LSTM uses a single shared-weight network trained jointly across all calibration catchments in both basins (Kratzert et al., 2019).

Further details about both models are provided in the following sections.

3.2. Process-Based Hydrological Model

3.2.1. Model Setup

The process-based model is based on a semi-distributed HRU-based bucket-type hydrological model previously implemented in the Moselle (Fenicia et al., 2022) and other regions in Central Europe (Dal Molin et al., 2020). Time series of precipitation (mm d^{-1}), mean temperature (°C) and PET (mm d^{-1}) were used as input data, discretized based on the catchments delineation. Each catchment was further discretized into HRUs to represent landscape heterogeneity. This means that each of the 108 catchments in the Moselle and 22 catchments in the Garonne was divided into up to three HRUs reflecting the respective relative geological permeability classes "Low", "Medium" and "High" present in each individual catchment. Importantly, the same catchment-scale forcing time series were applied to all HRUs and aggregated according to HRU area fractions, so that only the representation of geological heterogeneity differed between experiments.

Unlike previous applications (i.e., Dal Molin et al., 2020; Fenicia et al., 2022), river routing was omitted to simplify the model and focus on the comparative influence of geological input data. While routing could improve absolute performance, it is not expected to affect the relative ranking of geological maps. Streamflow at each outlet was computed by averaging HRU outputs weighted by their relative area.

Each HRU was represented by the same conceptual structure, composed of four reservoirs:

- Snow reservoir (WR): accounts for snow accumulation and melting, using a degree day method. Moreover, following previous implementations, the snow reservoir was not stratified into elevation bands (Fenicia & McDonnell, 2022).
- Unsaturated zone reservoir (UR): receives snowmelt and rainfall from the WR reservoir as input, and controls both the evaporation, and the partitioning of precipitation into infiltration and runoff processes.
- Fast reservoir (FR): receives a fraction of the UR outflow and controls the generation of fast runoff components and hydrograph peaks. A lag function was applied to offset peak timing, following the approach by Dal Molin et al. (2020).
- Slow reservoir (SR): receives the remaining fraction of the UR outflow and represents the groundwater storage and baseflow generation.

The spatial organization of the model structure is illustrated in Figure 3, while a detailed description of model parameters, water balance equations and constitutive relationships are described in Tables S4 to S6 in Supporting

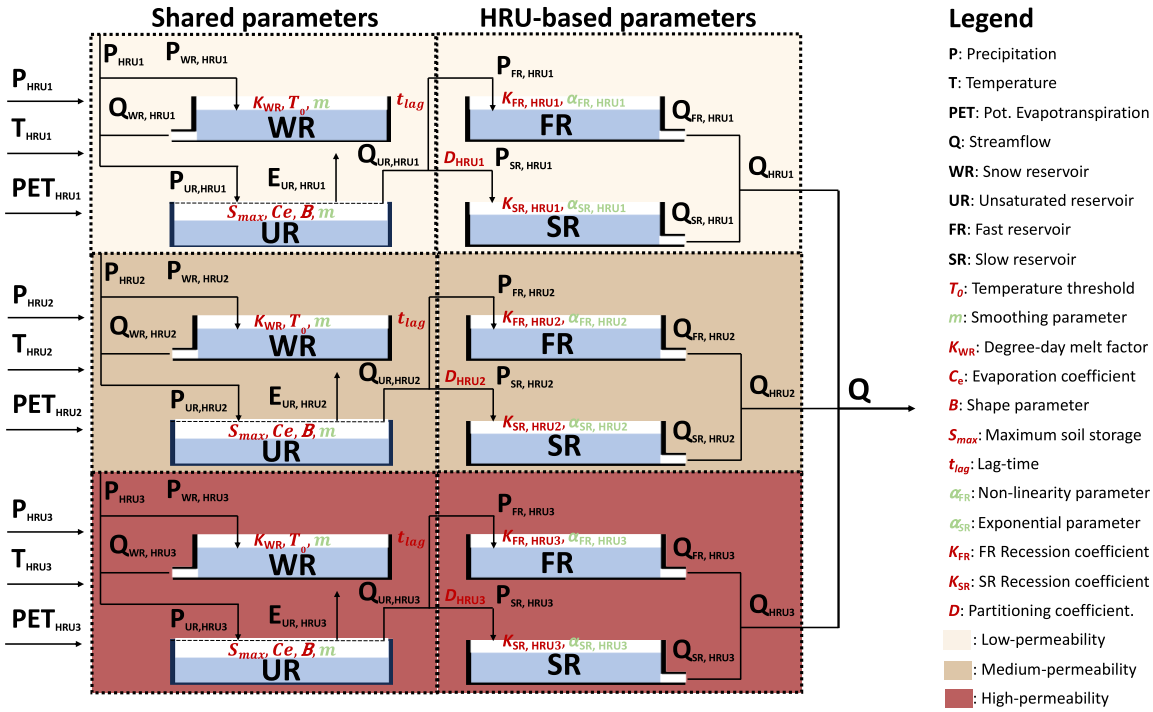

Figure 3. Organization of the model spatial structure. The three HRUs are explicitly shown. Calibrated parameters are shown in red, and fixed parameters in green. Reservoirs on the left share parameters, whereas those on the right have HRU-specific parameters, labeled with their index (e.g., HRU1).

Information S1. As shown in Figure 3, six parameters are shared across all HRUs and catchments, reflecting process components that are not expected to vary systematically with subsurface permeability (e.g., snow and evapotranspiration processes) (Fenicia & McDonnell, 2022; Fenicia et al., 2022). In contrast, parameters controlling subsurface storage, partitioning, and drainage processes were defined separately for each HRU, allowing process dynamics to differ between geological permeability classes (i.e., D, K_{FR} and K_{SR}). This means that we used a total of up to 15 calibrated parameters in our study (Figure 3).

HRU-specific parameters were calibrated using identical prior ranges across all permeability classes, with no explicit constraints imposed on their relative magnitude (Table S4 in Supporting Information S1). This choice was intentional. Rather than imposing distinct HRU behaviors through parameter ranges, we treated the link between geological classes and hydrological function as a hypothesis to be tested (see Fenicia et al., 2022). While such a setup would be underconstrained in a single-catchment context, the multi-catchment calibration introduces strong constraints by requiring consistent performance across catchments with varying HRU proportions. If geological classes exert a systematic control on streamflow dynamics (e.g., baseflow), the calibration will naturally differentiate parameter values among HRUs to reproduce these patterns. In this sense, the approach provides a stringent test: if HRUs representing low, medium, and high permeability consistently assume distinct functional roles, this pattern emerges from the data and supports our underlying hypothesis.

3.2.2. Calibration and Evaluation

The process-based model was implemented and calibrated using SuperflexPy, an open-source, Python-based hydrological modeling framework that enables the flexible construction of conceptual rainfall–runoff models using modular reservoir components (Dal Molin et al., 2021; Fenicia et al., 2011).

To evaluate model performance under conditions relevant to PUB, we applied a cross-evaluation strategy across space and time (Dal Molin et al., 2020; Fenicia et al., 2016, 2022; Klemeš, 1986). For each basin, catchments were divided into folds, and in each run one fold was used for model calibration, producing one shared parameter set per fold. The calibration fold was rotated across runs so that parameter sets were always evaluated on catchments not used for calibration. In this setup, all catchments share the same parameters for low-, medium-, and high-permeability HRUs, while differences in landscape characteristics are captured solely through the varying

proportions of HRUs across catchments. These parameters values were then transferred without modification to the remaining catchments not used for calibration. This multi-site parameter transfer strategy has been applied in previous studies (Fenicia et al., 2016, 2022; Gao et al., 2014), although we acknowledge that other strategies for model evaluation in ungauged catchments are also possible (Pool et al., 2021).

In the Garonne basin, 22 catchments were divided into two folds of 11 catchments each. Calibration was performed alternately on one fold, with the other used for evaluation. In the Moselle basin, containing 108 catchments, a more stringent evaluation was possible. Catchments were divided into seven folds of about 16 catchments each, with calibration performed on one fold and evaluation on the remaining six folds. The calibration fold was rotated so that each catchment was associated with six independent evaluation simulations, and the ensemble mean was used for performance assessment. Accordingly, the reported evaluation results summarize the out-of-sample simulations across all folds, so that all catchments appear in the final assessment despite being excluded from calibration in each individual evaluation run. To ensure that each fold was representative of hydro-climatic diversity in the study area, catchments were ranked by drainage area, randomly assigned to folds, and visually inspected for spatial balance (see Figures S2 and S3 in Supporting Information S1). Text S1 in Supporting Information S1 provides the definition of each fold and further details on their composition. For the Moselle basin, this "unbalanced" cross-evaluation design, in which calibration is performed on a limited subset of catchments and evaluated on the majority, reflects challenging PUB conditions, where streamflow information is limited.

Model performance was evaluated under three modes: (a) calibration, (b) space evaluation, defined as testing on catchments not used for calibration within the calibration period, and (c) space–time evaluation, defined as testing on different catchments and period. Calibration and space evaluation were performed for the period 01.10.1991 to 30.09.2001, while space-time evaluation from 01.10.2001 to 30.09.2015. Importantly, in the Moselle basin, this time split choice led to 16 catchments with streamflow data available only from 2002 onwards to be excluded from calibration and used exclusively for evaluation (Text S1 in Supporting Information S1).

The model operated with a daily time step and used an additional two-years warm-up period. Parameter optimization sought to maximize the basin-averaged Nash–Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970) of the square-root transformed daily streamflow ($\sqrt{Q}$) time series as the objective function, thereby reducing the influence of high flows (see Fenicia et al., 2022).

3.3. LSTM Model

3.3.1. Model Setup

We used a single layer LSTM, consistent with previous implementations in several regions (Acuña Espinoza et al., 2025; Gauch et al., 2021; Klotz et al., 2025; Kratzert et al., 2019). The network consisted of one hidden layer with 128 hidden states, a sequence length of 365 days and a batch size of 256. Optimization was performed using the Adam algorithm (Kingma & Ba, 2017), with an initial learning rate of 10^{-3}, reduced to 5×10^{-4} and 10^{-4} after 10 and 20 epochs, respectively. A dropout rate of 40% was applied to the output layer to mitigate overfitting.

The LSTM was trained using catchment-average daily meteorological forcings as dynamic inputs, including mean precipitation (mm d^{-1}), average shortwave radiation (W m^{-2}), maximum, mean, and minimum air temperature (°C), and vapor pressure (Pa) (Table S8 in Supporting Information S1). In addition, we used static catchment attributes characterizing long-term average landscape and climate properties (Table S9 in Supporting Information S1). Specifically, 16 attributes derived from the EStreams data set (do Nascimento et al., 2024) were used: basin area, mean, minimum, and maximum elevation, mean slope, depth to bedrock, aridity index, precipitation seasonality, high and low precipitation duration, mean normalized difference vegetation index, fraction of precipitation falling as snow, and mean soil fractions of sand, silt, clay, and soil organic carbon. The normalized upstream capacity attribute was derived using Equation 9 in Salwey et al. (2024). It measures how the storage capacity of reservoirs relates to the average yearly rainfall over their underlying catchment. A value of 1 means the catchment reservoirs can hold the equivalent of 1 year's accumulated rainfall. Geological information was represented by the three permeability-related attributes derived from the different geological data sets, depending on the experiment (regional, continental, or global), as defined above and described in do Nascimento et al. (2025). As a result, the "low", "medium", and "high" permeability fractions differ across experiments while

the remaining 17 static attributes remain unchanged. Details on the experimental design and attribute selection are provided in Section 3.4.

3.3.2. Calibration and Evaluation

LSTM calibration, validation, and evaluation were implemented using the NeuralHydrology python package (Kratzert et al., 2022). To evaluate LSTM performance under PUB conditions, we adopted the same standard k-fold space-time cross-evaluation strategy used by Kratzert et al. (2019). This strategy differs from the inverted k-fold design used for the process-based model (Section 3.2.2), reflecting differences in training requirements and computational constraints of deep learning models (Kratzert et al., 2024). Model calibration was conducted for the period from 01.10.1991 to 30.09.1998, validation from 01.10.1998 to 30.09.2001, and space-time evaluation for an independent period from 01.10.2001 to 30.09.2015.

For clarity, the space–time evaluation used here corresponds to what is commonly termed as out-of-sample evaluation in the machine-learning literature (e.g., Heudorfer et al., 2025; Kratzert et al., 2019). We retain the term space–time evaluation to remain consistent with established hydrological testing frameworks for process-based models.

Differently to the process-based model, the LSTM model was trained simultaneously on the Garonne and Moselle basins. This approach was adopted because the Garonne basin alone contains only 22 catchments, which may affect the performance of a regional LSTM model (Kratzert et al., 2024). The 130 available catchments were partitioned into five spatial folds. For each fold, 60% of catchments were used for calibration, 20% for validation, and 20% withheld for evaluation, with random assignment constrained such that each catchment appeared in the validation and evaluation sets only once across all folds.

For each fold, one LSTM model with shared weights was trained using the calibration catchments. Validation catchments were used exclusively for model selection, while evaluation catchments were not used in any stage of model training or tuning. As a result, each catchment produced exactly one independent evaluation simulation. The reported evaluation results therefore correspond to the combination of these out-of-sample evaluation simulations across all folds, which is why the final assessment includes the full set of catchments. To account for stochasticity in weight initialization, each LSTM configuration was trained using five different random seeds. The resulting simulations were combined into an ensemble by averaging simulated discharge across seeds at each time step. This ensemble-mean approach follows previous studies showing improved robustness and reduced sensitivity to random initialization (Acuña Espinoza et al., 2025; Gauch et al., 2021; Kratzert et al., 2019).

Finally, model parameters were optimized by maximizing the basin-averaged NSE loss computed on daily streamflow (Q). Note that differently to the process-based model implementation, here we used a non-transformed Q to be coherent with the standardized implementation of LSTM models (Acuña Espinoza et al., 2025; Gauch et al., 2021; Kratzert et al., 2019). Calibration was performed for 30 epochs, and the epoch yielding the highest validation performance was withheld.

3.4. Model Experiments

We designed five model experiments to evaluate how subsurface heterogeneity influences the catchment hydrological response. The experiments differ exclusively in the representation of geological information, hence in the proportion of HRUs, while all other aspects of model structure, forcing data, calibration, and evaluation are kept identical. This controlled experimental design is applied consistently to both the process-based model and the LSTM model. The geological categories used here are defined by the relative permeability classes based on the global, continental, and regional geological maps (Figure 2).

3.4.1. Geological Configurations

Here we defined five model experiments as follows:

- No geology experiment (ben): a benchmark experiment in which subsurface properties are represented uniformly within each catchment, without any spatial differentiation of geological classes.
- Random geology experiment (ran): a benchmark experiment in which the permeability classes fractions were assigned randomly to each catchment. The total number of permeability classes are identical to those in the

Table 2
Overview of Model Experiments and Representation of Subsurface Heterogeneity

Model	Experiment	Geological fractions	Description
Process-based (Moselle)	PM_{ben}	None	Single HRU
	PM_{ran}	Random	HRUs weighted by geological fractions
	PM_{glo}	Global map	
	PM_{con}	Continental map	
	PM_{reg}	Regional map	
Process-based (Garonne)	PG_{ben}	None	Single HRU
	PG_{ran}	Random	HRUs weighted by geological fractions
	PG_{glo}	Global map	
	PG_{con}	Continental map	
	PG_{reg}	Regional map	
LSTM (Moselle and Garonne)	L_{ben}	None	No geological static attributes
	L_{ran}	Random	Static attributes (fractions)
	L_{glo}	Global map	
	L_{con}	Continental map	
	L_{reg}	Regional map	

geology-based experiments, but the spatial information is not physically meaningful. This experiment serves to assess whether performance changes arise from informative geological structure rather than from the mere introduction of heterogeneity.

- Global geology experiment (glo): geological heterogeneity is represented using permeability classes derived from the global geological map.
- Continental geology experiment (con): as above with permeability classes derived from the continental geological map.
- Regional geology experiment (reg): as above with permeability classes derived from the regional geological map.

Table 2 summarizes the experiments and illustrates how the respective geological fractions are implemented in each model class.

3.4.2. Implementation in the Process-Based Model

In the process-based model, geological heterogeneity is represented explicitly through HRUs. In the uniform experiment (PM_{ben} or PG_{ben}), each catchment is represented by a single HRU. In the remaining experiments, three HRUs are defined per catchment, corresponding to their respective permeability classes. The single-HRU model included nine calibration parameters, which were shared among all catchments in the respective basin (K_{WR}, T_0, S_{max}, C_e, B, $t_{lag,}$ D, K_{FR} and K_{SR}) (Figure 3). The three-HRU included 15 calibration parameters (Figure 3), from where six were shared among all catchments (and HRUs) (K_{WR}, T_0, S_{max}, C_e, B and t_{lag}), while the other nine were shared only among catchments according to their respective HRU (D, K_{FR} and K_{SR}) (Figure 3). All geology-based experiments use an identical model structure and parameterization and differ only in the fractional representation of permeability classes; therefore, there is no additional computational cost among the random, global, continental, and regional configurations. Relative to the benchmark without geological differentiation, calibration time increases due to the larger parameter set, but this increase is identical across all geology-based experiments.

3.4.3. Implementation in the LSTM Model

In the LSTM model, geological information is incorporated implicitly through static catchment attributes provided as inputs to the network. In the benchmark experiment (L_{ben}), the LSTM is provided with the same set of static catchment attributes described in Section 3.3.1, excluding the three geological permeability fractions (17

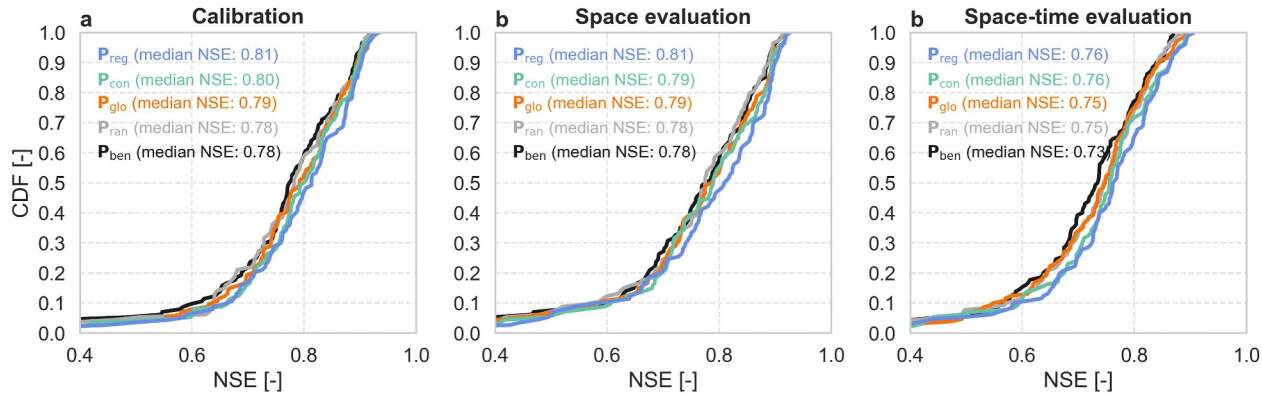

Figure 4. Cumulative distribution functions of NSE values for the five process-based model experiments evaluated under (a) calibration, (b) space evaluation, and (c) space–time evaluation. Note that the limits of the *x*-axis were clipped to the range 0.4–1.0 to facilitate visualization of the results.

catchment attributes in total). In the remaining experiments geological heterogeneity is represented by the three static attributes describing the fractional coverage of the permeability classes within each catchment, totaling 20 catchment attributes. All other static and dynamic inputs, network architecture, and training strategy remain unchanged across experiments.

3.5. Performance Assessment

The assessment was carried out both in the time and in the signature domain. In the time domain, the NSE of the ensemble-mean $\sqrt{Q}$ was used as evaluation performance metric. In the signature domain, model assessment was based on the Pearson's correlation coefficient (r) and in the relative bias of the observed and simulated values of three streamflow signatures: the baseflow index (BFI) (Ladson et al., 2013), which was computed using three filter passes and alpha set at 0.925; the slope of the FDC, which was computed using Equation 3 from Sawicz et al. (2011); and the mean half-flow date (HFD), which represents the date on which the cumulative streamflow reaches half of the annual discharge. These signatures were chosen for their relevance to groundwater contributions and their sensitivity to geological characteristics (do Nascimento et al., 2025). All hydrological signatures were computed using the NeuralHydrology Python package (Kratzert et al., 2022), ensuring consistent and reproducible signature estimation across all catchments and model experiments.

To formally assess whether increasing geological detail led to significant improved model performance across catchments, we applied paired Wilcoxon signed-rank tests (Wilcoxon, 1945) to compare experiments. For NSE, tests were performed directly on paired catchment-wise NSE values. For hydrological signatures, tests were performed on paired absolute errors relative to the observed signature values, such that lower values indicate better reproduction. We focused on comparisons between the regional experiment and the other experiments. To account for multiple testing, p-values were adjusted using the Holm correction, and significance was evaluated at the 5% level based on the adjusted p-values.

4. Results

4.1. Model Calibration and Evaluation Overall Performance

Figures 4 and 5 present the cumulative distributions functions (CDFs) of NSE values for the process-based and the LSTM model respectively. For both model classes, results are presented considering the two basins jointly. For example, P_{reg} refers to the concatenated results from PM_{reg} and PG_{reg}. Basin-specific results are provided in Figures S8 and S9 in Supporting Information S1. The results for each individual model class are presented in detail below.

4.1.1. Process-Based Model

In calibration (Figure 4a), model performance of the process-based model increased systematically with the higher level of geological detail for both basins. Median NSE values were 0.78 for P_{ben} and P_{ran}, followed by 0.79

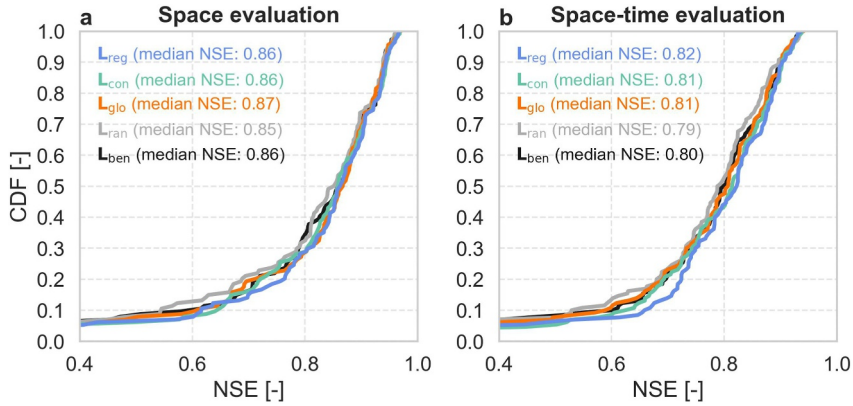

Figure 5. Cumulative distribution of the NSE for the five geological experiments evaluated under (a) space evaluation, and (b) space–time evaluation for the long short-term memory (LSTM) model. Note that Calibration results are omitted, as the LSTM is expected to achieve high in-sample performance irrespective of experiment, making calibration metrics largely uninformative. Note that the limits of the x-axis were clipped to the range 0.4 to 1.0 to facilitate visualization of the results.

for P_{glo}, 0.80 for P_{con}, and 0.81 for P_{reg}. This ordering is consistent across most of the distribution, with P_{reg} achieving the highest NSE values for the majority of catchments also when basin were evaluated individually (Tables S11 and S12 in Supporting Information S1). Intersections among the CDF curves arise because NSE values at a given quantile do not correspond to the same catchments across model configurations, and because calibration is performed jointly for subsets of catchments rather than independently for each catchment.

In space evaluation (Figure 4b), NSE values decreased slightly compared to calibration, but the relative ranking remained consistent. The experiments P_{con} (0.79) and P_{reg} (0.81) showed consistently the highest performance across all catchments.

In space-time evaluation (Figure 4c), as expected, performance further decreased for all experiments in comparison to calibration and space evaluation. Median NSE values preserved a similar ranking as in calibration, with now P_{ran} (0.75) performing slightly better than the benchmark experiments (0.73). Performance improved progressively for P_{glo} (0.75), P_{con} (0.76) and P_{reg} (0.76). Consistently, P_{con} and P_{reg} achieved performance gains across all the distribution, including in the lower end of the CDF curve, compared with the other experiments. These differences were consistent with the formal significance tests, which confirmed significantly higher NSE for P_{reg} relative to all other experiments (Table S13 in Supporting Information S1).

4.1.2. LSTM Model

The CDFs of NSE values for the LSTM experiments indicate more limited sensitivity to the different geological configurations compared to the process-based models. Particularly during the space evaluation phase (Figure 5a), LSTM performance remained higher than the process-based model across all experiments, with median NSE values varying from 0.85 for L_{ran} to 0.87 in L_{glo}. Although medians were arguably close, differences between experiments were more pronounced in the lower end of the CDF, with the L_{con} and L_{reg} performing better than the other experiments.

In space-time evaluation (Figure 5b), performance decreased and variability increased across all experiments compared to space evaluation. Median NSE values remained relatively high, ranging from 0.79 for L_{ran} to 0.82 in L_{reg}. Again, although differences between experiments were small in terms of median and upper-quartile performance, pronounced differences emerge in the lower tails of the distributions. In particular, L_{glo} shows wider negative spread, while L_{reg} improves the lower tail in comparison to all other experiments (Figure 5b). Such results are consistent with the formal significance tests, which confirmed that L_{reg} significantly outperformed all other experiments despite the small absolute gains (Table S13 in Supporting Information S1).

4.2. Models' Ability to Reproduce Streamflow Signatures

Sections 4.2.1 and 4.2.2 below present the signatures analyses for the process-based and LSTM models respectively. The signature analyses combine results from the Moselle and Garonne basins, as they exhibit the

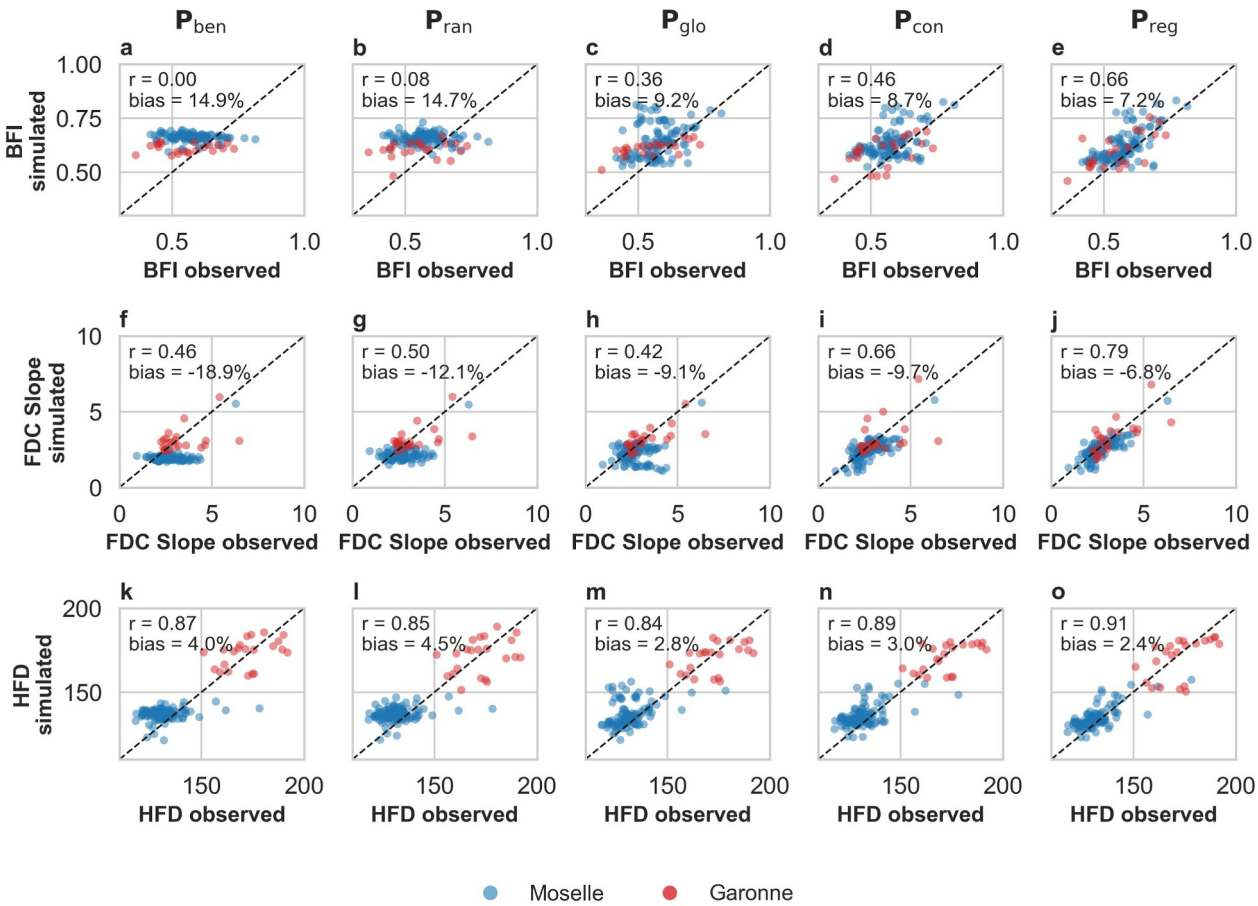

Figure 6. Scatter plots of simulated (*y*-axis) versus the observed (*x*-axis) streamflow signatures (BFI, flow duration curve slope and HFD) in space-time evaluation for the process-based model. The first column represents the P_{ben} model experiment, followed by P_{ran}, P_{glo}, P_{con} and P_{reg}. Circle colors indicate catchments: Moselle in blue and Garonne in red. Each subplot also shows the r and bias metrics calculated jointly for both basins.

same overall pattern. For completeness, basin-specific results are provided in Tables S14 and S15 in Supporting Information S1.

4.2.1. Process-Based Model

The benchmark experiment P_{ben} performed poorly in reproducing observed streamflow signatures. Except for mean HFD, which showed relatively high correlations ($r = 0.87$), BFI and FDC were poorly represented, with correlations of $r = 0.00$ and $r = 0.46$ respectively, and systematic biases reaching up to −18.9% (Figures 6a, 6f and 6k). The high HFD correlation likely results from evaluating both basins jointly, as Table S14 in Supporting Information S1 shows that the process-based model has limited skill in reproducing HFD when the Moselle basin is evaluated individually, although performance remains higher than for the other signatures. Overall, simulated signatures exhibited limited spatial variability across catchments. Considering that the model uses spatially uniform parameter values, this suggests that the observed spatial variability in signatures is driven more by landscape characteristics than by climate.

The random geology experiment (P_{ran}) performed qualitatively similarly to P_{ben} (Figures 6b, 6g and 6l). Correlations remained low for BFI ($r = 0.08$) and moderate for FDC slope ($r = 0.50$), while biases reached respectively 14.7 and −12.1%. This indicates that introducing heterogeneity without physically meaningful spatial organization did not consistently improve signature reproduction in PUB settings. Consequently, any improvements observed in subsequent experiments cannot be attributed to increased model complexity alone.

Introducing geological information from the global map (P_{glo}) led to mixed results compared to P_{ben}. The systematic bias is reduced across all three signatures: 9.2% (BFI), −9.1% (FDC), and 2.8% (HFD) relative to the

benchmark and random experiments. Correlations increased for BFI (r = 0.36) and remained high for HFD (r = 0.84). However, correlation for the FDC slope reduced to 0.42. Indeed, visual inspection of Figure 6h indicates that a subset of catchments in the Moselle exhibits nearly constant simulated FDC slopes, which weakens the linear relationship with observations. Overall, P_{glo} provides some limited evidence that the inclusion of geological information improves regionalization.

The continental geology experiment (P_{con}) produced a more pronounced improvement across all three signatures, particularly for the FDC slope (Figure 6i), where correlation increased to r = 0.66 alongside a continued reduction in bias (−9.7%). Improvements were also observed for BFI (r = 0.46 and bias = 8.7% in Figure 6d) and HFD (r = 0.89 and bias = 3.0% in Figure 6n). This indicates that increased geological detail begins to translate into improved spatial differentiation of subsurface-driven signatures, in addition to reduced bias.

The regional geology experiment (P_{reg}) achieved the highest performance across all experiments. Correlations were highest for BFI (r = 0.66), HFD (r = 0.91), and FDC slope (r = 0.79), while systematic biases were simultaneously minimized (BFI: 7.2%, HFD: 2.4%, FDC slope: −6.8%). Simulated signatures clustered tightly around the 1:1 line (Figures 6e, 6j and 6o), indicating that increased geological detail leads to markedly improved model's ability to capture both the magnitude and spatial variability of subsurface-controlled streamflow signatures across catchments. P_{reg} is generally able to capture the differing hydrological behavior of the Moselle and Garonne basins. This interpretation was also supported by the formal significance tests: P_{reg} significantly reduced errors for FDC slope and HFD relative to all other experiments, while for BFI the contrast with P_{glo} narrowly failed to reach the 5% significance threshold after correction (Table S13 in Supporting Information S1). Finally, in terms of BFI, the two basins span a similar range; for FDC slope, the Garonne generally exhibits higher values. The HFD shows the greatest contrast, with the two catchments forming distinct clusters.

4.2.2. LSTM Model

Compared to P_{ben}, the benchmark LSTM experiment (L_{ben}) was better able to reproduce streamflow signatures variability under space-time evaluation. Correlations ranged from r = 0.33 (bias = −17.0%) for the slope of the FDC (Figure 7f), to r = 0.55 (9.6%) for BFI (Figure 7a), to r = 0.92 (2.0%) for HFD (Figure 7k). This improvement arises because, while P_{ben} received no information about landscape spatial variability, L_{ben} was provided with static landscape attributes (Section 3.3.1), except for geological ones. The ability of L_{ben} to capture spatial variability in signatures indicates that the model can leverage these static attributes and potentially compensate for missing information when predictors are correlated.

The random geology experiment (L_{ran}) resulted in a degradation of performance relative to L_{ben} for BFI, which decreased to r = 0.47 (bias = 11.1%) (Figure 7b). For the slope of the FDC (r = 0.44 and bias = −16.4%) (Figure 7g) and for HFD mean (r = 0.92 and bias = 1.9%) (Figure 7l), their values remained very similar to L_{ben}. This is somewhat consistent with the findings of Figure 6, and reinforces that adding randomly structured geological attributes does not improve (and may even hinder) signature reproduction.

The use of static attributes derived from the global map (L_{glo}) led to improvements in all signatures related to the previous two experiments. Correlation for BFI increased from r = 0.55 (L_{ben}) to r = 0.61 accompanied by a more consistent bias = 9.5% (Figure 7c). Similarly, for the slope of the FDC the metrics increased to r = 0.60 and bias = −18.8% (Figure 7h). For the HFD mean (with r = 0.95 and bias = 1.9%) (Figure 7m) the values remained more comparable to those obtained for L_{ben}.

The continental geology experiment (L_{con}) produced more consistent improvements, particularly for the FDC slope (Figure 7i), where correlation increased to r = 0.71, accompanied by a reduction in bias to −16.7%. Performance for BFI improved modestly (r = 0.63 and bias = 8.8%) (Figure 7d), while HFD remained largely similar (r = 0.93 and bias = 1.6%) (Figure 7n).

The regional geology experiment (L_{reg}) achieved the best overall performance among the LSTM experiments (Figures 7e, 7j and 7o). Correlations were highest for all three signatures (BFI: r = 0.67; HFD: r = 0.95; FDC slope: r = 0.75), while relative biases were simultaneously reduced for FDC slope (bias = −17.0%) and HFD mean (1.6%). Although improvements relative to the benchmark are more modest than for the process-based model, these results indicate that detailed geological information provides complementary value to the LSTM by refining both the magnitude and spatial structure of simulated streamflow signatures. This pattern was also supported by the formal significance tests: L_{reg} significantly reduced errors for HFD relative to all other

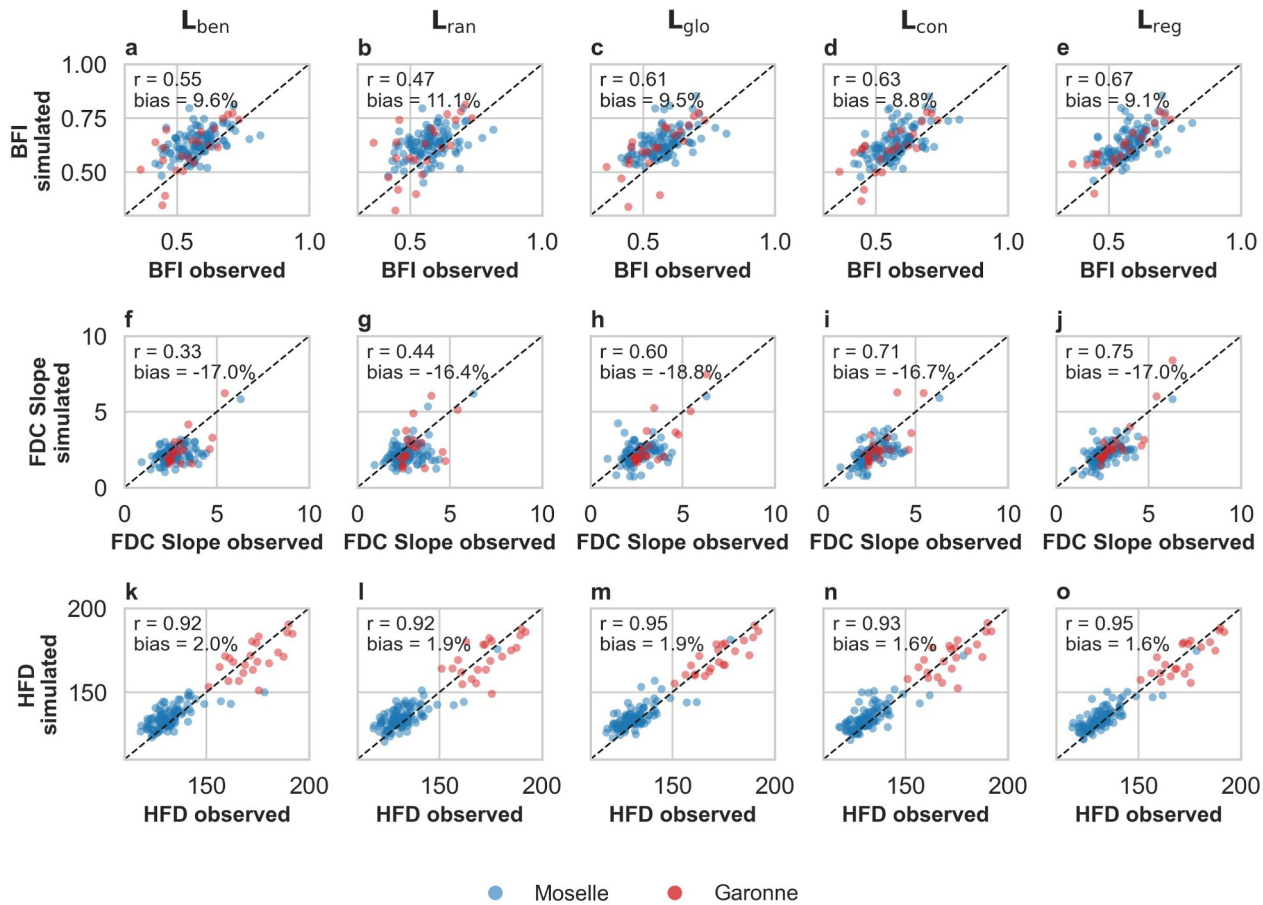

Figure 7. Scatter plots of simulated (*y*-axis) versus the observed (*x*-axis) streamflow signatures (BFI, flow duration curve slope and HFD) in space-time evaluation for the long short-term memory model. The first column represents the L_{ben} model experiment, followed by L_{ran}, L_{glo}, L_{con} and L_{reg}. Circle colors indicate catchments: Moselle in blue and Garonne in red. Each subplot also shows the r and bias metrics calculated jointly for both basins.

experiments, whereas significant improvements for FDC slope were observed only relative to L_{ran} and L_{glo}, and for BFI only relative to L_{ran} (Table S13 in Supporting Information S1).

4.3. Spatial Patterns of Model Performance During Space-Time Evaluation

Spatial patterns of NSE model performance during space–time evaluation reveal distinct model behavior across the Moselle and Garonne basins using the process-based experiments. Figure 8 shows the spatial distribution of the NSE for the P_{ben}, P_{glo}, P_{con} and P_{reg}. Equivalent maps for the LSTM model are provided in Figure S11 in Supporting Information S1.

For the Moselle basin, notable spatial contrasts emerge in the northwestern region, where P_{glo} (Figure 8b) consistently performed worse than all other experiments, including P_{ben} (Figure 8a). This area is classified as predominantly high permeability in the global geological map, in contrast to the continental and regional maps, which both classify it as low permeability. The LSTM model also exhibited slightly reduced performance in this region, although the decline is markedly less pronounced than for the process-based experiments (Figure S11 in Supporting Information S1). In addition, a small group of headwater catchments in the Moselle consistently showed low NSE values across experiments (Figure 8a). These catchments also exhibited poor performance in the LSTM experiments. Further diagnostics indicate mismatches in simulated flow variability and baseflow contribution, suggesting that these deficiencies are likely driven by local data limitations or unresolved process representations rather than geological information alone (see Text S4 in Supporting Information).

In contrast, no comparable spatially coherent region of systematically degraded performance is observed in the Garonne basin with the global map (Figure 8f). Despite differences in permeability classification among

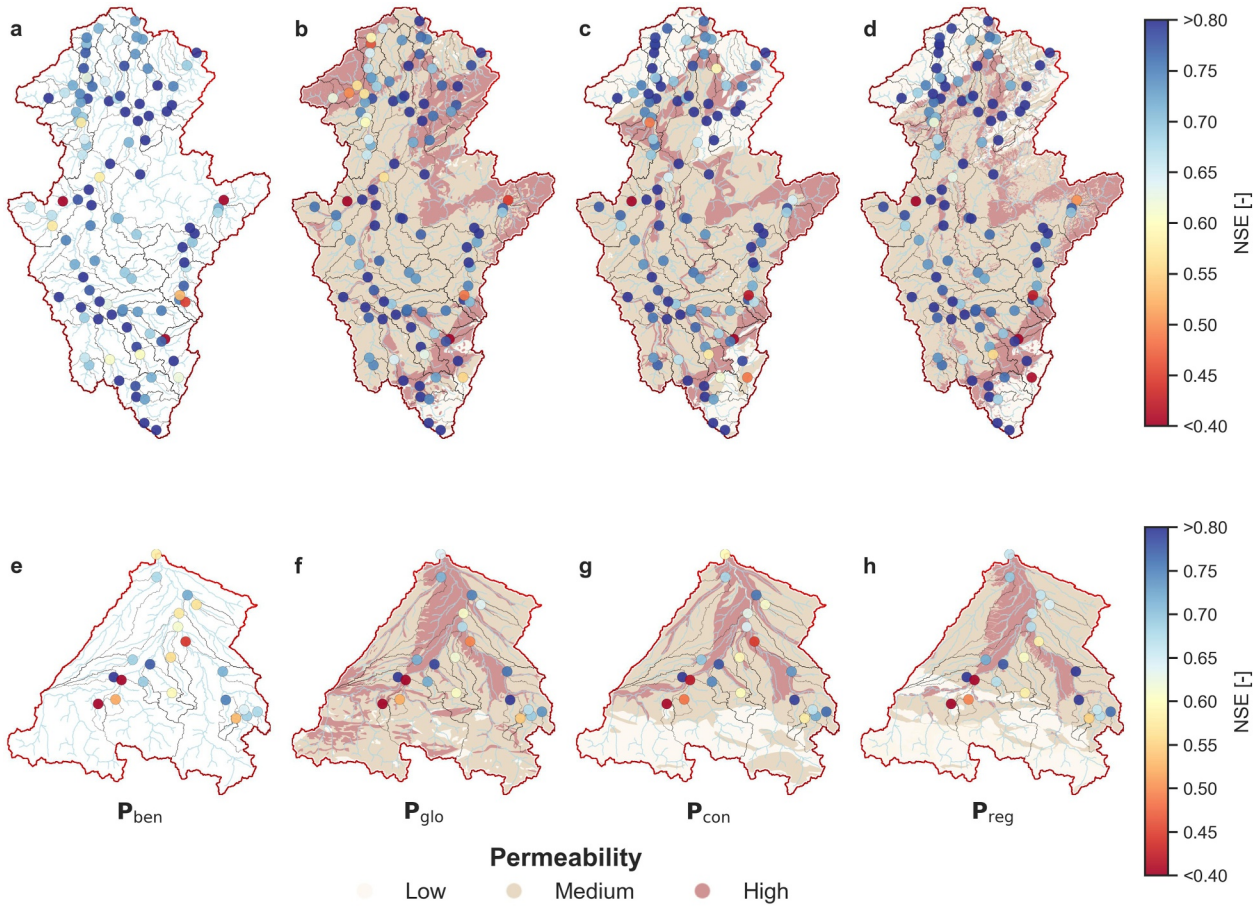

Figure 8. Model performance (NSE) in space-time evaluation for four different model experiments: P_{ben} (a, e), P_{glo} (b, f), P_{con} (c, g) and P_{reg} (d, h). Colors indicate NSE values, and the background shading represents the permeability classes.

geological data sets (Figures 8f–8h), P_{glo} exhibited a consistent improvement relative to P_{ben} across most of the basin, with P_{reg} achieving the highest NSE values overall. However, three neighboring catchments in the central–western part of the basin display persistently low NSE values across all experiments. These catchments are characterized by shallow rooting depths and substantial discrepancies between observed and simulated mean streamflow (Figure S12 in Supporting Information S1). Notably, reduced performance is also observed for the LSTM model in this area, indicating that these differences are most likely related to forcing data limitations or local anthropogenic influences rather than model structure or geological representation (see Text S4 in Supporting Information S1).

5. Discussion

5.1. When and How Does Geological Detail Improve Hydrological Predictions?

The aim of this study was to assess how increasing geological detail affects streamflow predictions in ungauged basins. We focused on two basins, the Moselle and the Garonne, where previous work has shown that geology strongly influences streamflow. Our results demonstrate that geological information adds value to both process-based and data-driven models, with the magnitude of this effect depending on the level of geological detail, the model type, and the dominant hydrological processes being evaluated, as reflected by the streamflow signatures.

Differences in hydrograph reproduction measured by NSE were modest but consistent (Figures 4 and 5). However, these modest NSE differences coincided with substantial disparities in the reproduction of specific streamflow signatures, particularly under space–time evaluation (Figures 6 and 7). The baseflow-related signatures here considered were best reproduced using the most detailed geological information, highlighting that the

information content of geological attributes governs how baseflow response can be transferred robustly in space and time. This interpretation was further supported by the formal significance tests, which confirmed clear improvements for FDC slope and HFD for both model classes, while BFI improvements were strongest and most systematic for the process-based model.

The apparent mismatch between small NSE improvements and large signature differences can be explained by the dominant role of climate in controlling most spatio-temporal variability (Figures S12 and S13 in Supporting Information S1). Accounting for additional landscape-driven variability produces only limited gains in overall hydrograph fit, which are muted in aggregated metrics like NSE. Signatures that isolate specific hydrograph aspects—particularly those linked to subsurface storage and flow partitioning—are more sensitive to landscape properties, revealing systematic benefits of increased geological detail.

These results reinforce evidence that aggregated metrics can mask compensating errors and provide limited diagnostic insight (Bouaziz et al., 2021; Euser et al., 2013; Gupta et al., 2024; Hrachowitz et al., 2014; Hulsman, Hrachowitz, & Savenije, 2021; Hulsman, Savenije, & Hrachowitz, 2021; Loritz et al., 2017; Pool et al., 2021; Ruzzante et al., 2026; Shafii & Tolson, 2015; Vis et al., 2015; Williams, 2025). More broadly, they highlight that geological information is most valuable when evaluation metrics are aligned with the processes it controls.

Streamflow signatures exhibited a clear hierarchy of sensitivity to geological detail. BFI and FDC slope were generally the most responsive, whereas HFD differences were smaller in magnitude, despite still showing significant improvements in several comparisons. This reflects a gradual shift from stronger landscape control to stronger climatic control across the signatures: BFI is primarily governed by subsurface storage and transmissivity, FDC integrates both landscape and climate influences, and HFD reflects climatic forcing timing. The consistently high correlations for HFD across all experiments, similar to results in previous studies (e.g., Addor et al., 2018; Beck et al., 2017; Gauch et al., 2021), further support this interpretation. This hierarchy indicates that geological detail enhances transferability mainly for processes strongly controlled by subsurface storage and baseflow, while signatures dominated by climate show limited sensitivity.

5.2. Differences Between Process-Based and Data-Driven Models and Transferability to Other Models

Using both process-based and data-driven models allows testing the generality of these findings across distinct modeling paradigms. The process-based model relies exclusively on geological information to represent spatial variability, whereas the LSTM incorporates a broader set of static landscape attributes together with dynamic forcing inputs.

In the process-based model, increasing geological detail from benchmark and random experiments to continental and regional representations led to systematic increases in median NSE, reduced fractions of low-performing catchments, and substantial improvements in storage-related (BFI) and regime-related (FDC slope) signatures. Geological information directly structures subsurface representation through HRU organization and is the sole mechanism for representing landscape variability, so increases in detail translate directly into improved process representation.

In contrast, geological information exerted a weaker and more selective influence on LSTM behavior and signature reproduction. Although the regional geology configuration still significantly improved NSE and some streamflow signatures, the effects were less systematic than in the process-based model. The already high predictive skill achieved without explicit geological input indicates that other static attributes (e.g., soils and topography), together with forcing variables, already encode substantial spatial differentiation, consistent with previous LSTM studies (Heudorfer et al., 2025; Kratzert et al., 2019). Consequently, we acknowledge that the marginal benefit of geological detail depends on its complementarity to the information already provided by other attributes. This also helps explain why the random-geology experiment had a slightly more noticeable effect in the process-based model than in the LSTM: in the former, random HRU fractions still introduce structural heterogeneity, whereas in the latter non-informative geological inputs can be partly compensated for by other predictors. At the same time, it also remains to be tested whether the sensitivity to geological detail persists in larger-domain LSTM applications (Klotz et al., 2025; Nearing et al., 2024), where training across many catchments (Kratzert et al., 2024) and using broad attribute sets, along different forcing inputs (Kratzert et al., 2021), could change how strongly individual static attributes influence learned behavior.

Previous work has also investigated the ability of LSTM models to make use of catchment attributes in their predictions (Bassi et al., 2024; Heudorfer et al., 2025). In a previous study, Bassi et al. (2024) showed that about five landscape features would be sufficient to reconstruct streamflow variability in most catchments. However, their findings also suggest that the static attributes commonly used in LSH data sets capture only limited landscape information. Moreover, Heudorfer et al. (2025) found a limited ability of LSTM models to generalize from certain time-variable physiographic attributes: scenario-based experiments with a synthetic 50% reduction in forest cover showed only limited and inconsistent responses in predicted streamflow. This suggests that landscape information might not always be used in a meaningful way by LSTM models.

In our case, the Moselle and Garonne basins are regions where geological controls on streamflow generation are well established, particularly for storage- and baseflow-related processes (Fenicia & McDonnell, 2022; Hellebrand et al., 2007; do Nascimento et al., 2025; Pfister et al., 2017). In this context, geological attributes seem to represent physically relevant static descriptors of the catchment structure that are consequently used by the data-driven model.

Geological attributes are already included in several LSH modeling studies for regionalization and prediction (Dal Molin et al., 2020; Fenicia et al., 2022; Kratzert et al., 2019; Loritz et al., 2024; Nijzink et al., 2025). However, our results suggest that the benefits of including geological information (and their level of detail) extend beyond a single modeling framework and may plausibly apply to other regionalization approaches (and model configurations) for both process-based models (Pool et al., 2021), as well as to hybrid and physics-guided learning frameworks (Shen et al., 2023).

Beyond these conceptual considerations, some methodological limitations remain. In the process-based model, uniform HRU parameterization and simplified routing may constrain performance in shallow-soil catchments. Future work could adopt methods to regionalize S_{max} spatially (e.g., Oorschot et al., 2024; Tempel et al., 2024; Wang et al., 2024), improving representation of catchment-specific storage capacities. Forcing uncertainties, particularly in parts of the Garonne basin, may also have affected model skill (Clerc-Schwarzenbach & do Nascimento, 2026).

A further point concerns the process-based calibration setup, which used identical prior parameter ranges across low-, medium-, and high-permeability HRUs in combination with multi-catchment calibration. This design was intentional, as it allowed functional differentiation among HRUs to emerge naturally from the calibration process, constrained by the need to reproduce diverse hydrographs across catchments (Text S3 in Supporting Information S1). However, in a more traditional PUB setting, such as calibrating individual catchments and transferring parameters to nearby basins (Pool et al., 2021), identical parameter ranges across HRU types may prove less effective.

While these considerations may influence absolute performance, they are unlikely to alter the main conclusion that geological detail provides conditional, process-relevant value in both process-based and data-driven models.

5.3. Conditions Under Which Geological Information Is Informative and Transferability to Other Regions

Random geology experiments provide an important control. Across both model classes, randomly structured geological information did not consistently improve the overall performance outside the calibration domain and in some cases significantly degraded signature reproduction. This demonstrates that improvements arise from geological information that meaningfully represents spatial variability relevant to dominant hydrological processes (do Nascimento et al., 2025), not from increased model complexity alone.

Spatial patterns of performance further illustrate this point. In the Moselle basin, the northwest region showed systematic underperformance under less detailed global geology compared to continental and regional representations, and occasionally even compared to the benchmark. This spatially coherent degradation is plausibly linked to differences in permeability classification. Such results emphasize that the value of geological information is conditional on both data quality and alignment with dominant processes (Araki et al., 2025; Fenicia & McDonnell, 2022; Gnann et al., 2021; Holt & McMillan, 2025; do Nascimento et al., 2025).

Previous studies have shown that dominant controls on hydrograph variability vary considerably regionally, reflecting differences in climate, topography, soils, vegetation, and subsurface properties (Beck et al., 2015; Carlier et al., 2018; do Nascimento et al., 2025; Pfister et al., 2017; Rudlang et al., 2025; Schneider et al., 2007;

Zomlot et al., 2015). In regions where climate or surface processes dominate, increasing geological detail may provide limited additional benefit, particularly for metrics that integrate overall hydrograph behavior. It is also noteworthy that the required level of geological detail depends on catchment size and geological heterogeneity. For geologically uniform or small catchments, whose spatial extent is below the resolution of available maps, it is also expected that differences in performance would be diminished (see do Nascimento et al., 2025).

These considerations suggest that the value of geological detail is conditional rather than universal. Improvements are substantial when subsurface processes are dominant and geological attributes meaningfully encode spatial variability, but may be smaller or negligible elsewhere. This reinforces the broader principle in LSH that landscape attributes are most informative when aligned with dominant hydrological processes at relevant scales (Carlier et al., 2018; Fenicia & McDonnell, 2022; Gnann et al., 2021; Holt & McMillan, 2025; do Nascimento et al., 2025).

Future studies should assess whether the observed sensitivity to geological detail persists across larger regions, different climates, and for other hydrological metrics, particularly where the dominant controls on flow generation differ from the Moselle and Garonne basins.

6. Summary and Conclusions

LSH data sets have been widely used in hydrological modeling and PUB, yet the extent to which the quality and level of detail of catchment attributes included in such data sets influence model regionalization remains insufficiently understood. This study assessed how the level of detail in geological catchment attributes affects hydrological model regionalization and prediction in ungauged catchments. Using a controlled set of process-based and data-driven model experiments applied to 130 catchments across the Moselle and Garonne basins, we draw the following conclusions:

1. Geological data detail has a consistent effect on overall streamflow prediction performance: Across both basins and model classes, increasing geological detail consistently improved performance under space–time evaluation. Differences in aggregated metrics such as NSE were modest, whereas hydrological signatures revealed clearer and more systematic improvements.
2. Performance differences are most pronounced in storage- and regime-related components of the hydrograph: Signatures reflecting subsurface processes, particularly BFI and FDC slope, were most sensitive to geological detail, whereas timing-related signatures, such as mean HFD, showed weaker, but still detectable improvements. This hierarchy highlights that geological information primarily enhances the transferability of processes strongly controlled by subsurface storage.
3. The effects of geological detail were consistent across two independent model types, but were stronger and more systematic in the process-based model: Both the process-based and the data-driven models exhibited similar sensitivities to geological input detail, which indicates that the influence of geological information on the regionalization employed in this study is not specific to a particular modeling paradigm. However, the effect was weaker in the data-driven model, suggesting partial compensation through other static attributes.
4. These effects were robust across two independent regions: In both the Moselle and Garonne basins, increased geological detail consistently improved model predictions, reinforcing the general relevance of process-aligned geological information. However, improvements depend on whether geological attributes meaningfully represent dominant hydrological processes. In regions where climate or surface processes dominate, the same level of geological detail may provide limited additional benefit.

Overall, this study demonstrates that geological attributes can improve hydrological predictions in ungauged basins when their level of detail aligns with dominant processes and modeling objectives. Future research should assess whether these benefits extend to other modeling approaches, including hybrid frameworks, and should also explore analogous experiments focusing on other components of the streamflow hydrograph and corresponding catchment attributes. Such studies could broaden the usefulness of catchment attributes for better predictions in ungauged catchments.

Conflict of Interest

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

Availability Statement

The used version of the EStreams data set and catalog (v1.4) is stored online at a Zenodo repository (https://doi.org/10.5281/zenodo.17598150) (do Nascimento et al., 2024). Regional geology catchment attributes for the Moselle and Garonne are available online (https://doi.org/10.5281/zenodo.18392387). All code used in this study to derive the figures and analysis provided in this work is available online at a GitHub repository (https://github.com/thiagovmdon/estreams_superflexpy) (do Nascimento, 2026b). Model parameters, configuration files, forcing meteorological inputs, calibration codes and all simulations for the process-based and the data-driven are available in a Zenodo repository (https://doi.org/10.5281/zenodo.18392387) (do Nascimento, 2026a). Due to redistribution restrictions, daily streamflow is not available, but can be obtained from national and regional agencies responsible for hydrometric monitoring in each country and using the EStreams catalog (www.estreams.eawag.ch, last access: 22 December 2025). For Belgium (Wallonie), data were sourced from the (SPW, 2023). For Germany, data sets were provided by both the (BFG, 2023) and the regional authority of Rhineland-Palatinate (MKUEM et al., 2023). French data were obtained from the national hydrometric portal (BanqueHydro, 2024). Finally, streamflow data for Luxembourg were obtained from the (NGGL, 2023).

Acknowledgments

This project was funded by a "Money Follows Cooperation" project (Project No. OCENW.M.21.230) between the Netherlands Organization for Scientific Research (NWO) and the Swiss National Science Foundation (SNSF). This work was further supported by the TU Delft Climate Action Research and Education seed funds. We thank the editor and the three reviewers for their constructive comments, which helped to clarify and improve the manuscript. Open access publishing facilitated by ETH-Bereich Forschungsanstalten, as part of the Wiley - ETH-Bereich Forschungsanstalten agreement via the Consortium Of Swiss Academic Libraries.

References

Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., & Ehret, U. (2025). Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. *Hydrology and Earth System Sciences*, *29*(5), 1277–1294. https://doi.org/10.5194/hess-29-1277-2025

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, *54*(11), 8792–8812. https://doi.org/10.1029/2018WR022606

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*.

Araki, R., Holt, A., Hammond, J. C., Husic, A., Coxon, G., & McMillan, H. K. (2025). Continental-scale prediction of hydrologic signatures and processes. *EGUsphere*, 1–42. https://doi.org/10.5194/egusphere-2025-6156

BanqueHydro. (2024). *Hydro portail*. France. Retrieved from https://www.hydro.eaufrance.fr/

Bassi, A., Höge, M., Mira, A., Fenicia, F., & Albert, C. (2024). Learning landscape features from streamflow with autoencoders. *Hydrology and Earth System Sciences*, *28*(22), 4971–4988. https://doi.org/10.5194/hess-28-4971-2024

Beck, H. E., de Roo, A., & van Dijk, A. I. J. M. (2015). Global maps of streamflow characteristics based on observations from several thousand catchments. *Journal of Hydrometeorology*, *16*(4), 1478–1501. https://doi.org/10.1175/JHM-D-14-0155.1

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., et al. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201–6217. https://doi.org/10.5194/hess-21-6201-2017

BFG. (2023). Bundesanstalt für Gewässerkunde, Germany. Retrieved from https://www.bafg.de/DE/Home/homepage_node.html

Bouaziz, L. J. E., Fenicia, F., Thirel, G., de Boer-Euser, T., Buitink, J., Brauer, C. C., et al. (2021). Behind the scenes of streamflow model performance. *Hydrology and Earth System Sciences*, *25*(2), 1069–1095. https://doi.org/10.5194/hess-25-1069-2021

BRGM. (2003). Bureau de Recherches Géologiques et Minières [Dataset]. *Carte Géologique de la France à 1/1 000 000, BRGM, Paris. (Version 6ème. édition révisée)*. https://infoterre.brgm.fr/rapports/84-AGI-273-LOR.pdf

Caballero, Y., Voirin-Morel, S., Habets, F., Noilhan, J., LeMoigne, P., Lehenaff, A., & Boone, A. (2007). Hydrological sensitivity of the Adour-Garonne river basin to climate change. *Water Resources Research*, *43*(7). https://doi.org/10.1029/2005WR004192

Carlier, C., Wirth, S. B., Cochand, F., Hunkeler, D., & Brunner, P. (2018). Geology controls streamflow dynamics. *Journal of Hydrology*, *566*, 756–769. https://doi.org/10.1016/j.jhydrol.2018.08.069

Clerc-Schwarzenbach, F., & do Nascimento, T. V. M. (2026). Evaluating E-OBS forcing data for large-sample hydrology using model performance diagnostics. *Hydrology and Earth System Sciences*, *30*(1), 119–140. https://doi.org/10.5194/hess-30-119-2026

Cornes, R., van der Schrier, G., van den Besselaar, E. J. M., & Jones, P. (2018). An ensemble version of the E-OBS temperature and precipitation datasets. *Journal of Geophysical Research: Atmospheres*, *123*(17), 9391–9409. https://doi.org/10.1029/2017jd028200

Dal Molin, M., Kavetski, D., & Fenicia, F. (2021). SuperflexPy 1.3.0: An open-source Python framework for building, testing, and improving conceptual hydrological models. *Geoscientific Model Development*, *14*(11), 7047–7072. https://doi.org/10.5194/gmd-14-7047-2021

Dal Molin, M., Schirmer, M., Zappa, M., & Fenicia, F. (2020). Understanding dominant controls on streamflow spatial variability to set up a semi-distributed hydrological model: The case study of the Thur catchment. *Hydrology and Earth System Sciences*, *24*(3), 1319–1345. https://doi.org/10.5194/hess-24-1319-2020

Delaigue, O., Guimarães, G. M., Brigode, P., Génot, B., Perrin, C., Soubeyroux, J.-M., et al. (2024). CAMELS-FR dataset: A large-sample hydroclimatic dataset for France to explore hydrological diversity and support model benchmarking. *Earth System Science Data Discussions*, 1–27. https://doi.org/10.5194/essd-2024-415

Do, H. X., Gudmundsson, L., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata. *Earth System Science Data*, *10*(2), 765–785. https://doi.org/10.5194/essd-10-765-2018

do Nascimento, T. V. M. (2026a). Data from "Assessing the Impact of Geological Map Detail on Process-Based and Data-Driven Hydrological Models" (Version 0.1) [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.18392387

do Nascimento, T. V. M. (2026b). thiagovmdon/estreams_superflexpy: Code from "Assessing the Impact of Geological Map Detail on Process-Based and Data-Driven Hydrological Models" (Version v1.0) [Computer Software]. *Zenodo*. https://doi.org/10.5281/ZENODO.20178686

do Nascimento, T. V. M., Rudlang, J., Gnann, S., Seibert, J., Hrachowitz, M., & Fenicia, F. (2025). *How do geological map details influence the identification of geology-streamflow relationships in large-sample hydrology studies? Hydrology and Earth System Sciences*, *29*(24), 7173–7200. https://doi.org/10.5194/hess-29-7173-2025

do Nascimento, T. V. M., Rudlang, J., Höge, M., van der Ent, R., Chappon, M., Seibert, J., et al. (2024). EStreams: An integrated dataset and catalogue of streamflow, hydro-climatic and landscape variables for Europe. *Scientific Data 2024*, *11*(1), 879. https://doi.org/10.1038/s41597-024-03706-1

Duscher, K., Günther, A., Richts, A., Clos, P., Philipp, U., & Struckmeier, W. (2019). The GIS layers of the BInternational Hydrogeological Map of Europe 1:1,500,000^ in a vector format. https://doi.org/10.1007/s10040-015-1296-4

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, *17*(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development: Flexible FRAMEWORK FOR HYDROLOGICAL MODELING, 1. *Water Resources Research*, *47*(11). https://doi.org/10.1029/2010WR010174

Fenicia, F., Kavetski, D., Savenije, H. H. G., & Pfister, L. (2016). From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. *Water Resources Research*, *52*(2), 954–989. https://doi.org/10.1002/2015WR017398

Fenicia, F., & McDonnell, J. J. (2022). Modeling streamflow variability at the regional scale: (1) perceptual model development through signature analysis. *Journal of Hydrology*, *605*, 127287. https://doi.org/10.1016/j.jhydrol.2021.127287

Fenicia, F., Meißner, D., & McDonnell, J. J. (2022). Modeling streamflow variability at the regional scale: (2) Development of a bespoke distributed conceptual model. *Journal of Hydrology*, *605*, 127286. https://doi.org/10.1016/j.jhydrol.2021.127286

Floriancic, M. G., Spies, D., van Meerveld, I. H. J., & Molnar, P. (2022). A multi-scale study of the dominant catchment characteristics impacting low-flow metrics. *Hydrological Processes*, *36*(1), e14462. https://doi.org/10.1002/HYP.14462

Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S., & Savenije, H. H. G. (2014). Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China. *Hydrology and Earth System Sciences*, *18*(5), 1895–1915. https://doi.org/10.5194/hess-18-1895-2014

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

Gnann, S. J., McMillan, H. K., Woods, R. A., & Howden, N. J. K. (2021). Including regional knowledge improves baseflow signature predictions in large sample hydrology. *Water Resources Research*, *57*(2), e2020WR028354. https://doi.org/10.1029/2020WR028354

Günther, A., & Duscher, K. (2019). Extended vector data of the International Hydrogeological Map of Europe 1:1,500,000 (Version IHME1500 v1.2). Retrieved from https://www.bgr.bund.de/ihme1500

Gupta, A., Hantush, M. M., Govindaraju, R. S., & Beven, K. (2024). Evaluation of hydrological models at gauged and ungauged basins using machine learning-based limits-of-acceptability and hydrological signatures. *Journal of Hydrology*, *641*, 131774. https://doi.org/10.1016/j.jhydrol.2024.131774

Hartmann, J., Moosdorf, N., Hartmann, J., & Moosdorf, N. (2012). The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochemistry, Geophysics, Geosystems*, *13*(12), 12004. https://doi.org/10.1029/2012GC004370

Hellebrand, H., Hoffmann, L., Juilleret, J., & Pfister, L. (2007). Assessing winter storm flow generation by means of permeability of the lithology and dominating runoff production processes. *Hydrology and Earth System Sciences*, *11*(5), 1673–1682. https://doi.org/10.5194/HESS-11-1673-2007

Heudorfer, B., Gupta, H. V., & Loritz, R. (2025). Are deep learning models in hydrology entity aware? *Geophysical Research Letters*, *52*(6), e2024GL113036. https://doi.org/10.1029/2024GL113036

Holt, A., & McMillan, H. (2025). New predictors for hydrologic signatures: Wetlands and geologic Age across Continental scales. *Hydrological Processes*, *39*(2), e70080. https://doi.org/10.1002/hyp.70080

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., et al. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. *Water Resources Research*, *50*(9), 7445–7469. https://doi.org/10.1002/2014WR015484

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Hulsman, P., Hrachowitz, M., & Savenije, H. H. G. (2021). Improving the representation of long-term storage variations with conceptual hydrological models in data-scarce regions. *Water Resources Research*, *57*(4), e2020WR028837. https://doi.org/10.1029/2020WR028837

Hulsman, P., Savenije, H. H. G., & Hrachowitz, M. (2021). Learning from satellite observations: Increased understanding of catchment processes through stepwise model improvement. *Hydrology and Earth System Sciences*, *25*(2), 957–982. https://doi.org/10.5194/hess-25-957-2021

IGME. (1994). Instituto Geológico y Minero de Espana. Mapa geológico de España [Dataset]. scale 1:1,000,000, Madrid. https://datos.gob.es/es/catalogo/ea0010987-magna-3-0-mapa-geologico-de-espana-a-escala-1-50-000-3-serie1

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. *arXiv*. https://doi.org/10.48550/arXiv.1412.6980

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Klotz, D., Miersch, P., Do Nascimento, T. V. M., Fenicia, F., Gauch, M., & Zscheischler, J. (2025). EARLS: A runoff reconstruction dataset for Europe [preprint]. *Earth System Science Data Discussions*. https://doi.org/10.5194/essd-2024-450

Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. *Hydrology and Earth System Sciences*, *28*(17), 4187–4201. https://doi.org/10.5194/hess-28-4187-2024

Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022). NeuralHydrology — A Python library for Deep Learningresearch in hydrology. *Journal of Open Source Software*, *7*(71), 4050. https://doi.org/10.21105/joss.04050

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, *25*(5), 2685–2703. https://doi.org/10.5194/hess-25-2685-2021

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., et al. (2023). Caravan - A global community dataset for large-sample hydrology. *Scientific Data*, *10*(1), 61. https://doi.org/10.1038/s41597-023-01975-w

Kuentz, A., Arheimer, B., Hundecha, Y., & Wagener, T. (2017). Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences*, *21*(6), 2863–2879. https://doi.org/10.5194/hess-21-2863-2017

Ladson, A., Brown, R., Neal, B., & Nathan, R. (2013). A standard approach to baseflow separation using the Lyne and Hollick filter. *Australian Journal of Water Resources*, *17*(1). https://doi.org/10.7158/W12-028.2013.17.1

Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., et al. (2024). CAMELS-DE: Hydro-meteorological time series and attributes for 1582 catchments in Germany. *Earth System Science Data*, *16*(12), 5625–5642. https://doi.org/10.5194/essd-16-5625-2024

Loritz, R., Hassler, S. K., Jackisch, C., Allroggen, N., van Schaik, L., Wienhöfer, J., & Zehe, E. (2017). Picturing and modeling catchments by representative hillslopes. *Hydrology and Earth System Sciences*, *21*(2), 1225–1249. https://doi.org/10.5194/hess-21-1225-2017

Martin, E., Gascoin, S., Grusson, Y., Murgue, C., Bardeau, M., Anctil, F., et al. (2016). On the use of hydrological models and satellite data to Study the water budget of River basins affected by human activities: Examples from the Garonne Basin of France. *Surveys in Geophysics*, *37*(2), 223–247. https://doi.org/10.1007/s10712-016-9366-2

MKUEM. (2023). Ministerum für klimaschutz, umwelt, energie und mobilität: Rheinland-Pfalz, Germany. Retrieved from https://wasserportal.rlp-umwelt.de

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., et al. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, *627*(8004), 559–563. https://doi.org/10.1038/S41586-024-07145-1

NGGL. (2023). The national geoportal of the grand-dutchy of Luxembourg. Retrieved from https://map.geoportail.ludatareceived

Nijzink, J., Loritz, R., Gourdol, L., Zoccatelli, D., Iffly, J. F., & Pfister, L. (2025). CAMELS-LUX: Highly resolved hydro-meteorological and atmospheric data for physiographically characterized catchments around Luxembourg. *Earth System Science Data Discussions*, 1–34. https://doi.org/10.5194/essd-2024-482

van Oorschot, F., Hrachowitz, M., Viering, T., Alessandri, A., & van der Ent, R. J. (2024). Global patterns in vegetation accessible subsurface water storage emerge from spatially varying importance of individual drivers. *Environmental Research Letters*, *19*(12), 124018. https://doi.org/10.1088/1748-9326/AD8805

Pfister, L., Martínez-Carreras, N., Hissler, C., Klaus, J., Carrer, G. E., Stewart, M. K., & McDonnell, J. J. (2017). Bedrock geology controls on catchment storage, mixing, and release: A comparative analysis of 16 nested catchments. *Hydrological Processes*, *31*(10), 1828–1845. https://doi.org/10.1002/HYP.11134

Pool, S., Vis, M., & Seibert, J. (2021). Regionalization for ungauged catchments — Lessons learned from a comparative large-sample Study. *Water Resources Research*, *57*(10), e2021WR030437. https://doi.org/10.1029/2021WR030437

Rudlang, J. M., do Nascimento, T. V. M., van der Ent, R., Fenicia, F., & Hrachowitz, M. (2025). *Climate and landscape jointly control Europe's hydrology* (pp. 1–42). EGUsphere. https://doi.org/10.5194/egusphere-2025-6372

Ruzzante, S. W., Knoben, W. J. M., Wagener, T., Gleeson, T., & Schnorbus, M.2026). Technical note: High Nash–Sutcliffe Efficiencies conceal poor simulations of interannual variance in seasonal regimes. *Hydrology and Earth System Sciences*, *30*(8), 2337–2355. https://doi.org/10.5194/hess-30-2337-2026

Salwey, S., Coxon, G., Pianosi, F., Lane, R., Hutton, C., Singer, M. B., et al. (2024). Developing water supply reservoir operating rules for large-scale hydrological modelling. *Hydrology and Earth System Sciences*, *28*(17), 4203–4218. https://doi.org/10.5194/hess-28-4203-2024

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences*, *15*(9), 2895–2911. https://doi.org/10.5194/hess-15-2895-2011

Schneider, M. K., Brunner, F., Hollis, J. M., & Stamm, C. (2007). Towards a hydrological classification of European soils: Preliminary test of its predictive power for the base flow index using river discharge data. *Hydrology and Earth System Sciences*, *11*(4), 1501–1513. https://doi.org/10.5194/HESS-11-1501-2007

Shafii, M., & Tolson, B. A. (2015). Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research*, *51*(5), 3796–3814. https://doi.org/10.1002/2014WR016520

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth and Environment*, *4*(8), 552–567. https://doi.org/10.1038/s43017-023-00450-9

SPW. (2023). Service public de Wallonie: L'hydrométrie en Wallonie: Observations: Debit, Belgium. Retrieved from https://hydrometrie.wallonie.be/home/observations/debit.html?mode=announcement

Tarasova, L., Gnann, S., Yang, S., Hartmann, A., & Wagener, T. (2024). Catchment characterization: Current descriptors, knowledge gaps and future opportunities. *Earth-Science Reviews*, *252*, 104739. https://doi.org/10.1016/J.EARSCIREV.2024.104739

Tempel, N., Bouaziz, L., Taormina, R., van Noppen, E., Stam, J., Sprokkereef, E., & Hrachowitz, M. (2024). Catchment response to climatic variability: Implications for root zone storage and streamflow predictions. *Hydrology and Earth System Sciences*, *28*(20), 4577–4597. https://doi.org/10.5194/hess-28-4577-2024

Trurnit, P., Voges, A., & Wittekindt, H. (2003). Geologische Karte der Bundesrepublik Deutschland, scale 1:1,000,000, Bundesanst. für Geowiss. und [Dataset]. *Rohstoffe*. Hannover. https://numis.niedersachsen.de/trefferanzeige?docuuid=1C60DDA9-EF73-47B9-9ED7-FCD22B3226C1

Vis, M., Knight, R., Pool, S., Wolfe, W., & Seibert, J. (2015). Model calibration criteria for estimating ecological flow characteristics. *Water*, *7*(5), 2358–2381. https://doi.org/10.3390/w7052358

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography Compass*, *1*(4), 901–931. https://doi.org/10.1111/j.1749-8198.2007.00039.x

Wang, S., Hrachowitz, M., & Schoups, G. (2024). Multi-decadal fluctuations in root zone storage capacity through vegetation adaptation to hydro-climatic variability have minor effects on the hydrological response in the Neckar River basin, Germany. *Hydrology and Earth System Sciences*, *28*(17), 4011–4033. https://doi.org/10.5194/hess-28-4011-2024

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

Williams, G. P. (2025). Friends don't let friends use Nash-Sutcliffe Efficiency (NSE) or KGE for hydrologic model accuracy evaluation: A rant with data and suggestions for better practice. *Environmental Modelling and Software*, *194*, 106665. https://doi.org/10.1016/j.envsoft.2025.106665

Winter, T. C. (2001). The concept of hydrologic landscapes. *JAWRA Journal of the American Water Resources Association*, *37*(2), 335–349. https://doi.org/10.1111/j.1752-1688.2001.tb00973.x

Wu, S., Zhao, J., Wang, H., & Sivapalan, M. (2021). Regional patterns and physical controls of streamflow generation across the conterminous United States. *Water Resources Research*, *57*(6), e2020WR028086. https://doi.org/10.1029/2020WR028086

Zomlot, Z., Verbeiren, B., Huysmans, M., & Batelaan, O. (2015). Spatial distribution of groundwater recharge and base flow: Assessment of controlling factors. *Journal of Hydrology: Regional Studies*, *4*, 349–368. https://doi.org/10.1016/J.EJRH.2015.07.005