https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

What do machine learning models learn about floods?

Emma Ford[1,2,3], Manuela I. Brunner[4,5,6], Hannah M. Christensen[1], Elizabeth Lewis[7], Simon Moulds[8], Lorna Shires[7], and Louise Slater[2]

[1]Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK
[2]School of Geography and the Environment, University of Oxford, Oxford, UK
[3]UK Centre for Ecology & Hydrology (UKCEH), Wallingford, UK
[4]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[5]WSL Institute for Snow and Avalanche Research SLF, Davos Dorf, Switzerland
[6]Climate Change, Extremes and Natural Hazards in Alpine Regions Research Center CERC, Davos Dorf, Switzerland
[7]School of Engineering, University of Manchester, Manchester, UK
[8]Department of Civil and Environmental Engineering, Imperial College London, London, UK

Correspondence: Emma Ford (emma.ford@hertford.ox.ac.uk)

Abstract. Machine-learning (ML) models are increasingly used to estimate flood magnitudes and infer their controls through post-hoc feature importance analysis. However, feature importance does not directly measure hydrological process importance nor infer physical causality. The extent to which feature importance is robust to changes in ML model type (e.g., decision-tree versus neural networks), evaluation setting (e.g., held-out catchments versus held-out years), and predictor information (e.g., static versus dynamic attributes) is an important area of research if ML-based inference is to support flood-generation understanding. Therefore, we systematically evaluate predictive skill and feature importance across 518 catchments in Great Britain, using two ML model types: (i) an event-based Random Forest (RF) trained directly on flood magnitude events; and (ii) a sequence-based Long Short-Term Memory (LSTM) network trained on continuous rainfall-runoff observations, which is subsequently evaluated on the same held-out events as the RF. We test performance and feature importance for the flood magnitude events across a spectrum of evaluation settings, predictor changes, training seeds, and sequence lengths. We assess performance at both the pooled and the within-catchment level, in order to distinguish between inter- and intra-catchment skill. The RF achieved higher predictive skill in every directly comparable model fit (e.g., pooled R^2 = 0.86 versus 0.68 for the LSTM, under temporal evaluation), reflecting that its event-focused structure is suitable for addressing the task at hand. Notably, for both models pooled skill substantially exceeded within-catchment skill, indicating that much of the apparent skill reflects differences between catchments. The two models failed for different reasons in predicting the largest flood magnitudes, as the RF displayed a magnitude-dependent bias, while the LSTM showed a consistent underestimation. Both models relied predominantly on short-timescale precipitation, with temperature and potential evapotranspiration being utilized over longer timescales. Static attributes performed two separable functions, namely suppressing a baseflow-index structured bias in predicted flood magnitude and shifting antecedent precipitation usage toward shorter timescales. Our findings demonstrate that similar predictive skill can be supported by different feature importance structures, and emphasize the need to test feature importance robustness before inferring process conclusions in future work.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

1 Introduction

Flood magnitudes are generated by complex interactions between multiple environmental controls, including local meteorological forcing, antecedent catchment wetness, and spatially heterogeneous catchment properties (Berghuijs et al., 2019, 2016; Shen and Chui, 2023). Quantifying the relative contributions of these controls is required for both improving the understanding of flood-generating mechanisms and advancing flood estimation. Flood-generating mechanisms may show nonlinear behavior, interact with each other, vary between catchments, and change over time (Tramblay et al., 2025; Blöschl et al., 2019; Tarasova et al., 2023). Thus, accurately estimating flood magnitude and determining which hydroclimatic processes support these estimates are intertwined tasks of prediction and inference.

Flood estimation has traditionally relied on a range of modeling approaches, including process-based, data-driven, and, more recently, hybrid modeling (Slater et al., 2023). Within process-based approaches, rainfall-runoff models explicitly represent hydrological process interactions, such as storages and fluxes (Clark et al., 2015; Fatichi et al., 2016). However, in the case of floods of high magnitude, performance can degrade when important processes are poorly constrained by observations or simplified in the model structure. This may include processes such as snow dynamics, groundwater contributions, and subcatchment rainfall heterogeneity (Brunner et al., 2021; Wang et al., 2026). Furthermore, explicit representation of physical mechanisms in process-based models does not necessarily ensure accurate flood prediction or that all processes responsible for flood generation are adequately represented (Beven and Freer, 2001; Kirchner, 2006).

Data-driven methods, such as machine learning (ML) models, provide an alternative approach to modeling by flexibly learning nonlinear relationships between meteorological forcing, catchment characteristics, and streamflow, directly from observations (Dawson and Wilby, 2001; Abrahart et al., 2012). Unlike process-based models, ML models learn predictive relationships without the need to provide equations, constraints, or parameterization based on physical theory. In recent years, large-sample deep learning (DL) models, trained on hundreds to thousands of catchments, have been demonstrated to exploit information shared across diverse locations (Kratzert et al., 2024), thus enabling improvement of streamflow prediction in ungauged settings (Kratzert et al., 2019b; Nearing et al., 2021). Critically, assuming that ML models can learn flexibly without physical constraints and successfully improve the prediction of streamflow, it is important to investigate and determine what exactly these models learn.

A prime example of this investigation is reflected in the literature on Long Short-Term Memory (LSTM) networks, a variant of the DL approach (Hochreiter and Schmidhuber, 1997). LSTMs have received attention for achieving high predictive skill for daily streamflow across large samples of catchments (Kratzert et al., 2018), including for Great Britain (GB) (Lees et al., 2021). Due to LSTMs' sequence-based learning architecture, temporal dependencies, and representation of hydrological memory, it is possible to gain implicit learning of previous hydrological states without requiring antecedent conditions to be feature-engineered by the modeler (Hochreiter and Schmidhuber, 1997; Kratzert et al., 2019b). Instead, the sequence length, also known as the *look-back window*, can be modified to test the model's dependency on different lengths of memory representation (Kratzert et al., 2018; Gauch et al., 2021a).

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

The predictive success of LSTMs for daily streamflow has motivated further research to scrutinize whether their learned representations reflect expected and physically meaningful hydrological relationships between the target variable and predictors (De la Fuente et al., 2024; Bayati et al., 2026; Yokoo et al., 2022). So far, there has been some evidence to indicate that LSTMs can learn hydrologically relevant structures, including storage-related memory and hydroclimate-dependent lags, and the use of static attributes for regional evaluation (Lees et al., 2022; De la Fuente et al., 2024; Kratzert et al., 2019b). Lees et al. (2022) revealed that soil moisture and snow states could be recovered from the internal states of regional LSTMs across GB, despite these quantities not being explicit training targets. Likewise, approaches designed to constrain or interrogate LSTM representations have identified interpretable temporal and regional differences in catchment response (De la Fuente et al., 2024).

However, recent large-sample evaluation has shown that highly skillful LSTMs can learn input–output relationships inconsistent with defensible hydrological behavior, including implausible short-lag responses to temperature and potential evapotranspiration (Bayati et al., 2026). Further, Li et al. (2022) found physiographic static catchment attributes could be replaced with random catchment-specific identifiers, retaining comparable performance for catchments seen in training. Consistent with this observation, Heudorfer et al. (2025) identified that conventional physiographic attributes contributed relatively little additional predictive information to unseen catchments (spatial out-of-sample prediction), with regional evaluation being supported to a greater extent by dynamic meteorological information. Nonetheless, this does not simply imply that catchment information is unimportant, nor that the role of static information remains the same under different evaluation settings or model configuration decisions. Heudorfer et al. (2026) further showed that improving the static information derived from measurement improved prediction in unseen basins, concluding that data quality was critical. Thus, predictive accuracy, hydrologically plausible internal representations, and process-relevant inference are related yet distinct properties of a fitted model that must be interpreted together carefully. Overall, the literature implies that the apparent contribution of an information source can depend on what information a predictor representation contains, what other information is available, and the evaluation setting under which the model is evaluated.

Decision-tree ensembles, such as Random Forest (RF) models, provide another ML model type commonly used in hydrology. Introduced by Breiman (2001a), RFs construct an ensemble of decision trees, each trained on a bootstrap sample of observations, with a randomly selected subset of predictors in each split. The predictions are then averaged across the ensemble of trees to produce a predicted value for the target variable (Scornet and Hooker, 2026). This structure means that RFs are able to map complex, nonlinear relationships, as well as the interactions between the target variable and predictors (Wu and Li, 2023), while reducing the likelihood of overfitting (Breiman, 2001a). In addition, RF models are usually trained in tabular data formats, and do not learn from sequences nor have temporal memory. Unlike LSTMs, antecedent information must be explicitly feature-engineered by the modeler, e.g., accumulation windows, lagged variables, or rolling means.

Thus, the RF architecture and explicit predictor representation have made RFs useful for different tasks relative to the LSTM. Tabular flood datasets, where both flood estimation and inference of controls on flood generation are of interest, can be provided to RF regression models for analysis. For example, Stein et al. (2021) used RF models to predict flood-generating process distributions from catchment attributes, and identify the climatic and landscape controls associated with different mechanisms

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

across the United States. For GB, Zheng et al. (2023) applied RFs to investigate the behavior of rainfall-runoff events, finding that climate exerted the strongest control over spatial variability, while rainfall depth and antecedent soil moisture dominated the event-scale and seasonal variations. Related RF applications have investigated anthropogenic controls on streamflow (Coxon et al., 2024), spatial variations in flood sensitivity to hydroclimatic and catchment characteristics (Slater et al., 2024), and combined RF regression with explainable ML to investigate controls on flood magnitude (Ford et al., 2026).

Importantly, these studies applying LSTMs and RFs for hydrological applications, including flood prediction, vary in evaluation design, predictor representation, hyperparameter tuning, and training sample size, thus making it challenging to draw robust conclusions across works even when applied to the same region. Although the information is supplied to the model explicitly, how that information is used and attributed among predictors across studies remains conditional on its representation and on the fitted model. Therefore, this motivates exploring the robustness of model performance and feature importance to changes in predictor representation, evaluation setting, and architecture across model types. In this context, LSTMs and RFs provide contrasting learning routes to explore this research gap, specifically for flood magnitude estimation.

Here, we adopt these two ML model types for flood magnitude estimation and feature importance analysis across catchments in GB: (i) an event-based RF trained on flood magnitude events; and (ii) a sequence-based LSTM trained on continuous daily data. Both models are evaluated on the same extracted flood magnitude events under multiple robustness experiments, some of which are model-specific, while others are shared across both models. These experiments include evaluation settings, predictor ablations, random seed changes, and antecedent representations. Evaluation settings are designed to test performance on both withheld years and withheld catchments. We pose four research questions, which are addressed using the perturbation experiments outlined above. First, how do event-based and sequence-based ML model types differ in predictive skill for the same flood magnitude events? Second, to what extent do RF and LSTM models rely on the same information when estimating the same held-out flood magnitude events? Third, how is antecedent meteorology represented across timescales in an engineered tabular model in contrast to a sequence-based model that learns lag dependence from daily inputs? Fourth, how stable are model performance and feature importance patterns overall?

2 Data and methods

2.1 RF and LSTM assessment framing

We examine two contrasting ML model types for flood magnitude estimation, as the aim of the study is to assess how predictive skill and feature importance emerge from different representations of the same flood magnitude target. The RF is trained directly on flood magnitude events using engineered predictors of event-day forcing and antecedent conditions, together with static catchment attributes. In contrast, the LSTM is trained to simulate continuous daily discharge from sequences of meteorological forcing and static catchment attributes. The LSTM predictions are subsequently filtered to the same held-out flood magnitude events used to evaluate the RF. Thus, direct comparison of predictive performance evaluates the two model types against a common evaluation target and set of flood magnitude events. We do not assume that these contrasting model

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

types solve equivalent learning problems, and the goal is not to find the best model, but to understand differences in how they learn.

2.2 Data sources and flood magnitude identification

We use the CAMELS-GB version 2 dataset (Coxon et al., 2026), providing daily hydrometeorological time series and static catchment attributes for 671 catchments across GB from 1 October 1970 to 30 September 2022. Hydrometeorological predictors at the daily time-step include precipitation (P, mm d^{-1}), mean air temperature (T, °C), potential evapotranspiration (PET, mm d^{-1}), and daily specific discharge (Q, mm d^{-1}). Static predictors comprised the same 22 catchment attributes used by Lees et al. (2021), obtained from CAMELS-GB version 2 (Coxon et al., 2026). These attributes characterize catchment physiography, soils, land cover, hydrogeology, and long-term climatic conditions. Importantly, the climate attributes are calculated over the 1970 - 2022 record and the land-cover attributes refer to 2015, so they contain information beyond the training period considered in this study (Coxon et al., 2026) (see Sect. 4.6). Full details are provided in Appendix Tables A1 and A2.

Before extracting floods from Q, quality control was applied to exclude catchments that did not have complete coverage of daily hydrometeorological predictors across the analysis period and at least 80 % discharge observations during the training period (1970-2009). After these criteria were applied, 532 of the 671 catchments remained for subsequent analyses.

To identify flood magnitude events, a catchment-specific 99th-percentile specific-discharge threshold was calculated using observations from the training period (1 October 1970 to 31 December 2009), to avoid information leakage from subsequent validation and test periods. This fixed threshold was then applied across the full analysis period (1 October 1970 to 30 September 2022) to identify exceedance days, which were declustered using a 7-day independence window (Brunner and Dougherty, 2022; Ford et al., 2026). The peak-flow day within each declustered exceedance period was retained as an independent *flood magnitude event*, serving as the target variable. The identified flood magnitude event counts used for model evaluation are presented in Table A3. In addition, of the 532 retained catchments, 518 recorded at least one flood magnitude event in the test period. Moreover, 392 catchments were found to have at least 10 events in the test period, and were used to compute within-catchment skill (see Fig. B1).

2.3 Predictor construction

For the RF, the identified flood magnitude events were supplemented with the corresponding static catchment attributes (see Table A2), event-day precipitation, temperature, and PET. In addition, engineered antecedent cumulative precipitation, as well as mean temperature and PET values, were calculated over the preceding 3, 7, 14, and 30 days (including the event day). This formed the flood magnitude event dataset. See full details in Appendix Table A4.

For the LSTM, daily time-steps of precipitation, temperature, and PET were supplied with a sequence length of 30 (i.e., 0-29 days). This sequence length was selected rather than identified through broad hyperparameter tuning in order to match the longest engineered antecedent window supplied to the RF, thereby giving the two models approximately the same maximum antecedent scope. Note that this is a considerably shorter period than used in previous works for streamflow simulation, such as the 365-day sequence length used by Lees et al. (2021), and thus cannot represent seasonal storage (see Sect. 4.6). Sensitivity

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

to sequence lengths of 14, 30, and 60 days is considered in Sect. 2.8. In addition, the same static catchment attributes given to the RF were also supplied to the LSTM.

2.4 Model training

The RF was implemented in *scikit-learn* (Pedregosa et al., 2011) as an event-based regression model trained directly on the flood magnitude day tabular dataset. The LSTM rainfall-runoff model was implemented using the default *NeuralHydrology* framework (Kratzert et al., 2022) with the standard `cudalstm` architecture, where meteorological predictors are provided as daily sequences and static attributes as time-invariant descriptors (see Tables A4 and A5). For the RF, hyperparameters were selected using validation-period flood magnitude events (1 January 2010-31 December 2015) and fixed across all experiments. The LSTM was trained using the standard *NeuralHydrology* workflow for data handling, normalization, batching, sequence construction, optimization, and evaluation. This reference configuration used a hidden size of 128, output dropout of 0.30, a 30-day sequence length, batch size 256, Adam optimization, and Nash-Sutcliffe efficiency (NSE) loss, and was trained for 30 epochs. Training did not use automatic early stopping. Instead, each epoch was evaluated over the temporally held-out validation period, and the epoch with the highest validation NSE was selected. The selected epoch was then used to generate continuous daily discharge predictions over the test period. All subsequent analyses were conducted using this selected epoch for each LSTM model. See Appendix Table A5 for further information on the RF and LSTM design.

2.5 Evaluation settings

We applied three held-out evaluation settings on both the RF and LSTM: (i) temporal; (ii) single-split spatiotemporal; and (iii) five-fold BFI-stratified (Table 1). Within every evaluation setting, the RF and LSTM were evaluated against the same observed flood magnitude events in the corresponding held-out test data (Table 1). Temporal evaluation assesses transfer to withheld years, whereas the single-split spatiotemporal and five-fold evaluation settings additionally examine transfer to withheld catchments. All experiments used fixed, non-overlapping time windows: (i) training (1 October 1970-31 December 2009); (ii) validation (1 January 2010-31 December 2015); and (iii) testing (1 January 2016-30 September 2022). Thus, temporal separation was maintained for all three evaluation settings.

For temporal evaluation, all 532 catchments passing the quality control check were represented in training and validation. For the single-split spatiotemporal evaluation, the 532 catchments were randomly partitioned into 426 training/validation catchments, and 106 held-out test catchments. These catchment partitions were fixed across the RF and LSTM in such a manner that both model types were trained, selected, and evaluated using equivalent spatial partitions. Test performance reflects transfer to catchments withheld during model fitting, as well as to the independent test period.

For the five-fold evaluation setting, the 532 catchments were first divided into four strata according to baseflow-index (BFI) quartiles. Catchments within each BFI quartile were then distributed across the five folds, so that each fold contained approximately equal representation from all four BFI quartiles. As a result, every held-out fold included the full BFI response spectrum, while the complementary training set retained substantial representation of each BFI regime. BFI was selected as the stratification variable because it represents a first-order control on hydrological and flood response in GB (Lees et al., 2021;

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table 1. Model evaluation settings. All RF and LSTM experiments use matching training, validation, and test periods. Performance and feature importance analyses are then computed on the held-out events of each setting (test period).

Evaluation setting	Catchment design	Question posed
Temporal	No spatial withholding. 532 retained catchments after applying quality control, of which 518 contribute events to the test period.	What is the model performance and feature importance when years, but not catchments, are withheld?
Single-split spatiotemporal	Randomly selected spatially held-out test catchments.	Do model performance and feature importance persist when both catchments and years are withheld?
Five-fold BFI-stratified	Five BFI-stratified folds, with each catchment assigned once to a held-out fold.	Are performance and feature importance patterns robust to catchment sampling, while maintaining representation across BFI regimes?

Coxon et al., 2026). Note that neither the single-split nor the five-fold partitions were spatially blocked, so held-out catchments may neighbor training catchments (see Sect. 4.6). Therefore, these settings test transfer to catchments withheld from training that share runoff-response mechanisms, but are not designed as a prediction-in-ungauged-basins test. The five-fold evaluation setting was used for all subsequent sensitivity experiments (Sect. 2.8).

2.6 Model performance on held-out flood magnitude events

Model performance was assessed on aligned held-out flood magnitude events under the three evaluation settings (see Table 1). For the five-fold setting, the performance metrics were calculated separately for each held-out fold and then averaged. The standard deviation across the five folds was additionally calculated. The primary performance diagnostic was the comparison between the magnitudes of the observed and the predicted floods pooled across all catchments in each relevant held-out set (i.e., capturing both inter- and intra-catchment variability). For pooled metrics, we report the coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), and percentage bias (PBIAS). The metric definitions are given in Appendix C.

Pooled R^2 can appear high when a model captures differences between the catchments in typical flood magnitude well, even if it has limited ability to distinguish flood magnitudes within individual catchments. Therefore, we also computed within-catchment performance, using the same metrics listed above, applied to held-out flood magnitude events separately within each catchment. However, this was only calculated when a given catchment had at least 10 held-out flood events, which reduced the catchment count considered from 518 to 392. The within-catchment results are summarized in Appendix Table A9. In addition, to examine errors across the within-catchment flood magnitude distribution (i.e., intra-catchment variability), for catchments with at least 10 held-out flood events, the events were ranked by observed flood magnitude within each catchment, and assigned into quintiles, from Q1 (smallest) to Q5 (largest). Because the groups were defined from ranks independently within each of the catchments, the Q1-Q5 represent relative rather than absolute flood magnitude. This further reduces the influence of between-catchment differences in flood magnitude. The event-level fractional bias, $(\hat{Q} - Q)/Q$, was then evaluated across groups. This was conducted for all three evaluation settings, in addition to input-feature ablations.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

As an additional screening step for the five-fold evaluation setting, within-catchment performance metrics were correlated (i.e., Spearman's rank) against a set of candidate static attributes covering climate, topography, soils, and hydrogeology (the full set of these associations is shown in Appendix Fig. B6). Although BFI was not included as a static attribute predictor in the models, it was selected as a candidate attribute for correlation analysis, due to its established influence on hydrological response in GB (Lees et al., 2021; Coxon et al., 2026; Lane et al., 2019). The candidate attributes were ranked according to their strongest absolute correlation with either the catchment R^2 or PBIAS. Correlations with the magnitude associated metrics (RMSE and MAE) were reported but not used here for ranking, because absolute errors can scale with the magnitude of the flood itself. Therefore, their strong associations with attributes, such as mean precipitation and steepness of the catchment, largely reflect the size of the floods rather than differences in model skill. This screen identified the BFI as the attribute most strongly associated with systematic bias in both model types, motivating the targeted analysis of Sect. 3.4.

2.7 Model feature importance on held-out flood magnitude events

Feature importance was assessed through post-hoc analysis, using methods appropriate to each model type. For the RF, feature importance was quantified using TreeSHAP (Lundberg et al., 2020), with SHAP values computed for every predictor and held-out flood-magnitude day. For the LSTM, feature importance was quantified using Integrated Gradients (IG) (Sundararajan et al., 2017), applied separately to each held-out flood-magnitude prediction. The feature importance analysis was conducted for all three evaluation settings, in addition to the perturbation experiments on the five-fold setting.

In addition to calculating local attribution values for each flood magnitude event, predictor-level attribution was summarized using mean absolute attribution magnitudes across all held-out flood magnitude events per predictor. For the LSTM, due to the temporal lag structure of the dynamic predictors, absolute IG attributions for each dynamic predictor (i.e., precipitation, temperature, and PET) were summed across the input sequence for each event and subsequently averaged across held-out flood magnitude events. The static attributes have only one attribution per event, so were averaged directly across events. For the five-fold evaluation setting experiments, attribution summaries were calculated per fold, and then averaged across the five folds. Further, to allow an approximate comparison of how the two model types distributed attribution share among sources of information when considered as process-groups (see Table 2), absolute attributions were additionally aggregated into three groups: (i) *event-day forcing*, comprising meteorological inputs on the flood-magnitude day; (ii) *antecedent meteorology*, comprising meteorological information preceding the flood-magnitude day; and (iii) *static attributes*. The feature importance value assigned to each information group was expressed as a share of the total absolute attribution across the three groups, such that the group shares summed to one within each fitted model.

Finally, to examine the timescales over which the two different model types use antecedent information, an attribution-weighted mean lag, referred to as a *memory centroid*, was also calculated for each dynamic predictor. For the LSTM, this was calculated across the 30-day lag sequence (lags 0 - 29). For the RF, a measure was calculated by weighting the flood magnitude day and engineered -3, -7, -14, and -30 day predictors (0, 3, 7, 14, and 30 days) by their mean absolute TreeSHAP values. Due to the RF predictors being engineered features, whereas the LSTM implicitly learns from the memory sequence, the resulting centroids are not directly comparable. Instead, the centroids are used to compare changes in antecedent information

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table 2. Process-level groups for comparison of feature importance. The RF and LSTM encode temporal information differently. Note that the RF receives engineered summaries as separate predictors, whereas the LSTM receives daily meteorological sequences. LSTM dynamic feature importance is partitioned using the lag-specific IG distribution rather than by treating individual sequence elements as independent predictors. Static LSTM attributes are attributed with respect to the original static input vector and are not multiplied by sequence length.

Information group	RF representation	LSTM representation	RF n	LSTM n	Interpretation
Event-day forcing	Event-day P, T, and PET	Lag-0 element of each dynamic sequence	3	3	Meteorological conditions on the flood-magnitude day.
Antecedent meteorology	P sums and T/PET means over 3-, 7-, 14-, and 30-day windows	Antecedent elements of each dynamic sequence (lags 1-29)	12	87	Meteorological information preceding the flood event.
Static catchment information	22 CAMELS-GB attributes	The same catchment attributes	22	22	Time-invariant information describing differences among catchments.

across evaluation settings and forcing input ablations within each model type. Details for all feature importance calculations are provided in Appendix C.

2.8 Model performance and feature importance robustness experiments

The five-fold evaluation setting was used as a common framework for a series of matched robustness experiments. These experiments were designed to test whether predictive performance and feature importance persist when different components of the modeling framework are changed, while retaining the same underlying evaluation partitions.

We explored perturbations to three different components of the modeling framework. This includes (1) the predictive information available to the model, (2) stochastic model fitting, and (3) the representation of antecedent information. The first two were applied equivalently to both model types, whereas the third was architecture-specific. Predictor-availability sensitivity was assessed using matched ablation experiments in which precipitation, temperature, PET, and static catchment attributes were independently removed one at a time. Each ablation was repeated across all five folds for both model types, while retaining the corresponding train–test partitions.

Stochastic model fitting sensitivity was then assessed by repeating model fitting with different random seeds. In each fold, the reference seed (193893) was supplemented by four additional seeds. Next, model-specific experiments tested sensitivity to the representation of antecedent hydrometeorological information. For the RF, seven engineered predictor bases were evaluated across the five folds, including the default configuration (Appendix Table A6). The bases vary the number, duration, and overlap of engineered antecedent windows, while retaining a common 30-day maximum horizon. For the LSTM, sensitivity to temporal representation was assessed using sequence lengths of 14, 30, and 60 days, with L=30 defining the reference configuration. The L=14 and L=60 experiments contributed 10 additional model–fold fits across the five folds.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table 3. Robustness experiments for five-fold fitted models. There are 47 directly comparable model fits for the LSTM and RF, including the main evaluation settings, and the four input ablations and training seed changes on the five-fold evaluation setting. The sequence lengths and predictor base changes are LSTM- and RF-specific perturbations, respectively, tested on the five-fold evaluation setting only.

Model	Experiment family	Configurations	Folds	Model fits
RF	Main evaluation settings	Temporal; single-split spatiotemporal; five-fold BFI-stratified	1; 1; 5	7
RF	Input ablations	$-P$; $-T$; −PET; −static attributes	5	20
RF	Training seeds	4 additional seeds	5	20
RF	Predictor bases	7 engineered bases (incl. default)	5	35
LSTM	Main evaluation settings	Temporal; single-split spatiotemporal; five-fold BFI-stratified	1; 1; 5	7
LSTM	Input ablations	$-P$; $-T$; −PET; −static attributes	5	20
LSTM	Training seeds	4 additional seeds	5	20
LSTM	Sequence lengths	L14, L60 (L30 = reference)	5	10
			Total model fits	**139**

Together, these experiments test robustness to catchment sampling, predictor availability, stochastic fitting, and antecedent representation. We use the term *robustness of feature importance* to describe model inference agreement across these perturbations rather than treating their spread as a formal decomposition of hydrological uncertainty. Patterns that persist across perturbations are interpreted as more robust properties of model information use, whereas patterns that vary substantially are considered more conditional on model type or experimental design. We do not claim to have tested for all sources of uncertainty.

3 Results

3.1 Model performance on held-out flood magnitude events

The RF outperformed the LSTM in all 47 directly comparable model fits for R^2, MAE, RMSE, and absolute PBIAS. Pooled R^2 was highest under temporal evaluation (RF 0.86, LSTM 0.68) and decreased when both years and catchments were withheld. Under single-split spatiotemporal evaluation, pooled R^2 was 0.71 for the RF and 0.47 for the LSTM, and 0.77 and 0.56, respectively, under the five-fold setting (Fig. 1a). The single-split spatiotemporal setting rests on one fixed draw of the holdouts and contains substantially fewer aligned flood magnitude events ($n = 1{,}436$ versus 7,147), making its estimate more sensitive to the particular withheld catchments, which were selected randomly. Hence, the five-fold results provide a more robust assessment across alternative catchment partitions.

The LSTM consistently underestimated flood magnitudes, with pooled PBIAS ranging from $-29\,\%$ to $-33\,\%$ across the three evaluation settings, whereas the RF bias remained closer to zero ($-6\,\%$ to $-12\,\%$) (Fig. 1b). Moreover, for both model types, performance calculated after pooling predictions across catchments was substantially higher than performance calculated separately within individual catchments. Within-catchment analyses retained 392 catchments in the temporal and five-fold evaluation settings, and 77 catchments in the single-split spatiotemporal setting. Under the five-fold evaluation setting,

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

pooled R^2 values of 0.77 (RF) and 0.56 (LSTM) contrasted with median within-catchment R^2 values of only 0.13 and −0.73, respectively.

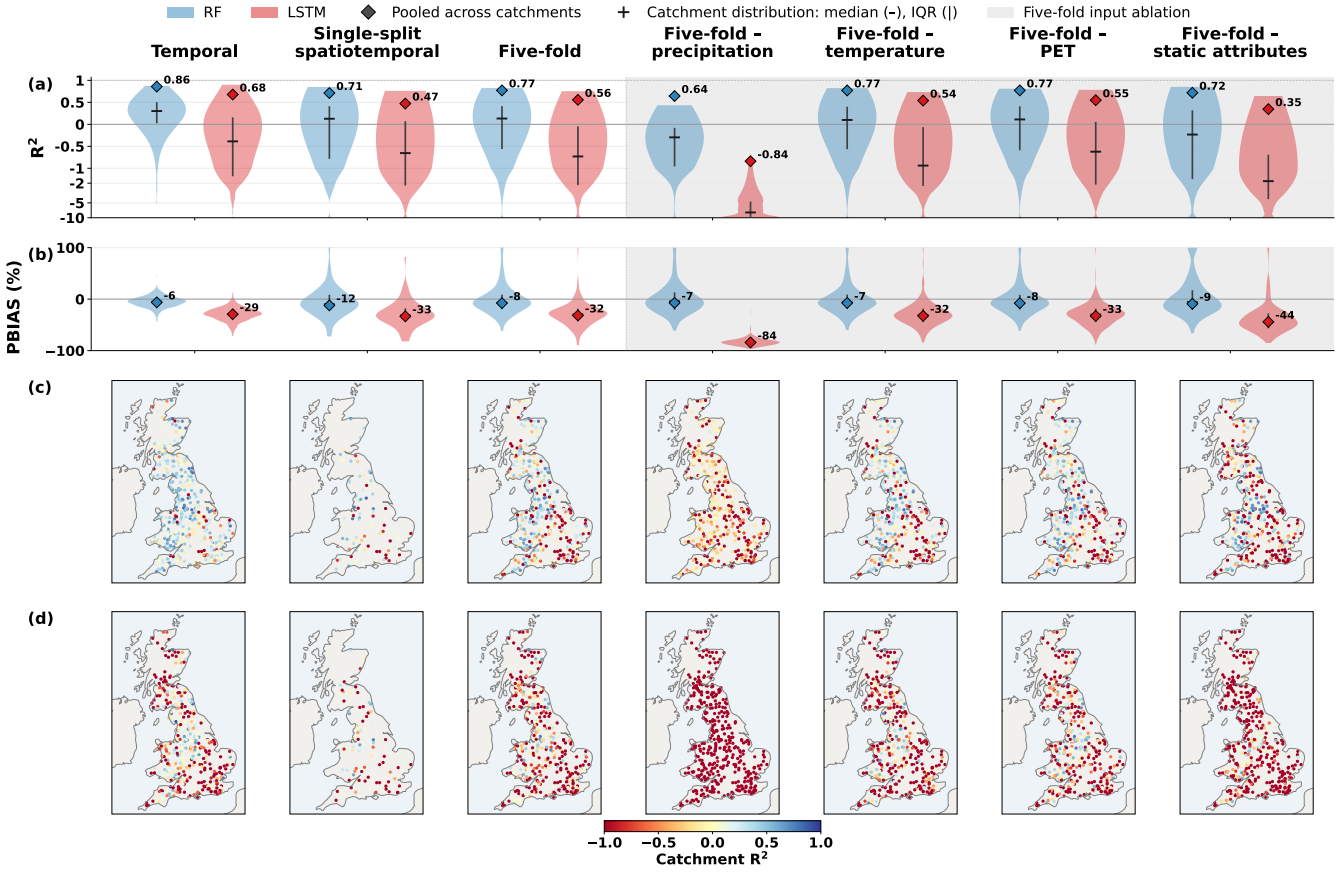

Figure 1. Performance on held-out flood magnitude events across evaluation settings and input-feature ablations. The RF is shown in blue and LSTM in red. Columns show the three evaluation settings, and the input-feature ablations on the five-fold evaluation setting, in which precipitation, temperature, PET, and static attributes were removed in turn. (a) Within-catchment R^2 distributions (violins; horizontal bars mark the catchment median and vertical bars the interquartile range) together with pooled R^2 across held-out flood magnitude events (for the five-fold setting, this is averaged across the five folds). The dashed line marks a maximum R^2 value of 1, and violins are clipped at -10 for visualization purposes. (b) Corresponding pooled and within-catchment PBIAS, where negative values indicate underestimation, and are clipped at ± 100 %. (c, d) Within-catchment R^2 mapped for (c) the RF and (d) the LSTM, with one panel per column setting and map colors capped at ± 1 for visualization. Within-catchment R^2 and PBIAS are computed only for catchments with at least ten held-out flood magnitude events, whereas pooled values use all held-out observations.

285 The within-catchment R^2 was negative in 44 % of eligible RF catchments and 78 % of eligible LSTM catchments (Fig. 1a; Appendix Table A9). The contrast between pooled R^2 and within-catchment R^2 indicates that pooled performance is supported by between-catchment differences in flood magnitude. In contrast, the within-catchment skill is much weaker. This pattern is

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

also evident in the mapped within-catchment R^2 values (Fig. 1c,d), which show widespread negative performance across GB, in particular for the LSTM.

Input-feature ablation experiments within the five-fold framework assessed the sensitivity of predictive skill to removal of predictors (Fig. 1, shaded columns). Removing precipitation reduced RF pooled R^2 from 0.77 to 0.64, while pooled bias remained similar to the reference value. In contrast, removing precipitation caused LSTM performance to collapse, reducing pooled R^2 to -0.84 and increasing PBIAS to $-84\,\%$, while within-catchment R^2 became negative in all 392 eligible catchments. Removing static attributes reduced LSTM pooled R^2 to 0.35 and increased underestimation (PBIAS $= -44\,\%$), whereas the corresponding reduction for the RF was modest (0.77 to 0.72). Removing temperature or PET altered pooled R^2 by less than 0.03 for either model. Finally, additional perturbation experiments (i.e., alternative training seeds, LSTM sequence lengths, and RF predictor bases) generally produced changes in pooled performance that were smaller than the fold-to-fold variation of the reference five-fold models, with the exception of LSTM PBIAS at sequence length $L = 14$ (Appendix Fig. B2; Tables A8 and A10).

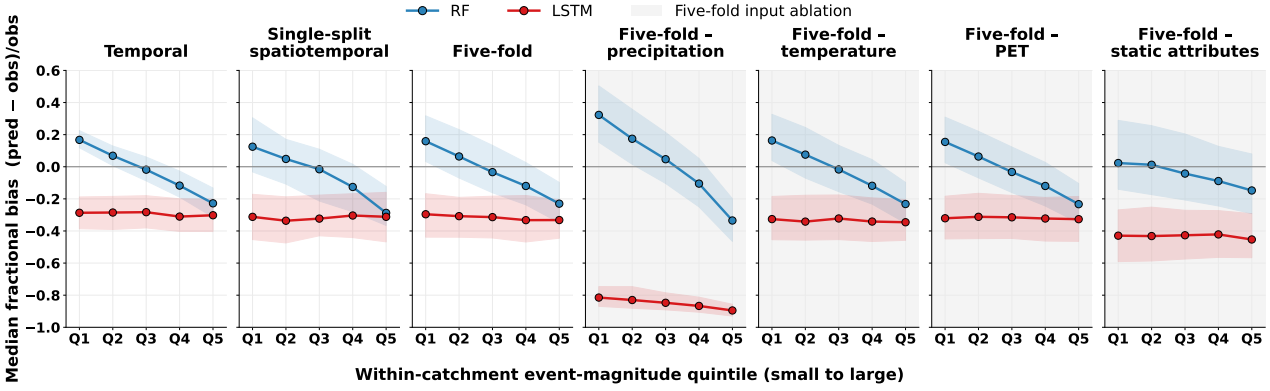

Figure 2. Prediction bias as a function of flood magnitude within each catchment. Within each catchment, held-out flood magnitude events are ranked by observed flood magnitude and grouped into quintiles (Q1, the smallest, to Q5, the largest). Fractional bias is defined here as (prediction − observation)/observation, with zero being unbiased. Negative values indicate underprediction, and positive values indicate overprediction of flood magnitude. For each catchment and quintile, the median fractional bias is calculated, and points show the median of these catchment-level values across catchments. Shaded bands indicate the corresponding interquartile range (RF blue, LSTM red). Panels follow the same columns as Fig. 1, with the three evaluation settings and the input-feature ablations for the five-fold evaluation setting. Catchments with fewer than ten held-out flood magnitude events are excluded from this analysis.

When considering prediction error at the within-catchment scale (Fig. 2), bias varied with flood size. Under the five-fold evaluation setting, the RF overpredicted the smallest floods within each catchment (median fractional bias $\approx +0.15$) and progressively underpredicted larger floods, crossing zero near the median quintile and reaching approximately -0.23 for the largest events. On the other hand, the LSTM underestimated flood magnitudes by an approximately constant fraction (≈ -0.30 to -0.33) across all quintiles. Both models underpredicted the largest floods, and these patterns were consistent across the evaluation settings. Also, removing precipitation steepened the RF bias gradient, and shifted the LSTM toward a nearly uniform

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

fractional bias of approximately -0.85. Removing static attributes deepened the LSTM underestimation (to approximately -0.45) without changing its magnitude independence, while flattening the RF bias gradient. Removing temperature or PET produced little change in either model type.

3.2 Feature importance on held-out flood magnitude events

After outlining predictive skill, we next examine feature importance to understand how the model fits use the available features. Both model types were dominated by precipitation-related information, but differed in how feature importance was distributed across individual features, and whether it was static or dynamic (Fig. 3, which displays the top 12 predictors). Under the five-fold reference setting, the RF's three highest-ranked predictors were long-term mean precipitation (`p_mean`), the frequency of low-precipitation days (`low_prec_freq`), and 3-day accumulated precipitation (`Precipitation_3d`). The first two of the three leading predictors are static attributes, and describe long-term precipitation indices rather than dynamic event-day forcing information, and are computed over the full 1970 - 2022 record. The feature importance of dynamic antecedent precipitation declined approximately monotonically with accumulation window length ($3\,\mathrm{d} > 7\,\mathrm{d} > 14\,\mathrm{d} > 30\,\mathrm{d}$), whereas temperature and PET did not rank among the top 12 predictors. For the LSTM, precipitation also dominated feature importance, approximately four times the feature importance on temperature, and six times that on PET. Feature importance patterns were broadly stable across the three evaluation settings for each of the model types.

Input-feature ablations produced contrasting patterns of compensation in the two model types. Removing static attributes substantially increased RF feature importance on dynamic precipitation, while reducing pooled R^2 by only 0.05. In the LSTM, removing static attributes increased feature importance of all three dynamic forcings (precipitation $+0.71$, temperature $+0.50$, PET $+0.47$), but with a cost of a 0.21 reduction in pooled R^2. Removing precipitation produced different responses across the model types. The RF redistributed feature importance toward static catchment attributes, with the largest relative increases occurring for land-cover and soil descriptors (`crop_perc` $+68\,\%$, `soil_depth_pelletier` $+71\,\%$), together with an increase in `p_mean` ($+25\,\%$ to 5.18). However, pooled R^2 only decreased by 0.13. It is important to note that `p_mean` is a static attribute, so was not removed when ablating dynamic precipitation. In contrast, when precipitation was removed, the LSTM shifted feature importance toward the remaining meteorological forcings (PET $+1.13$, temperature $+0.79$) and R^2 fell to -0.84. Substitution between the energy-related forcings was asymmetric, as removing temperature approximately tripled feature importance on PET ($+1.07$), whereas removing PET produced no detectable change in temperature feature importance.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

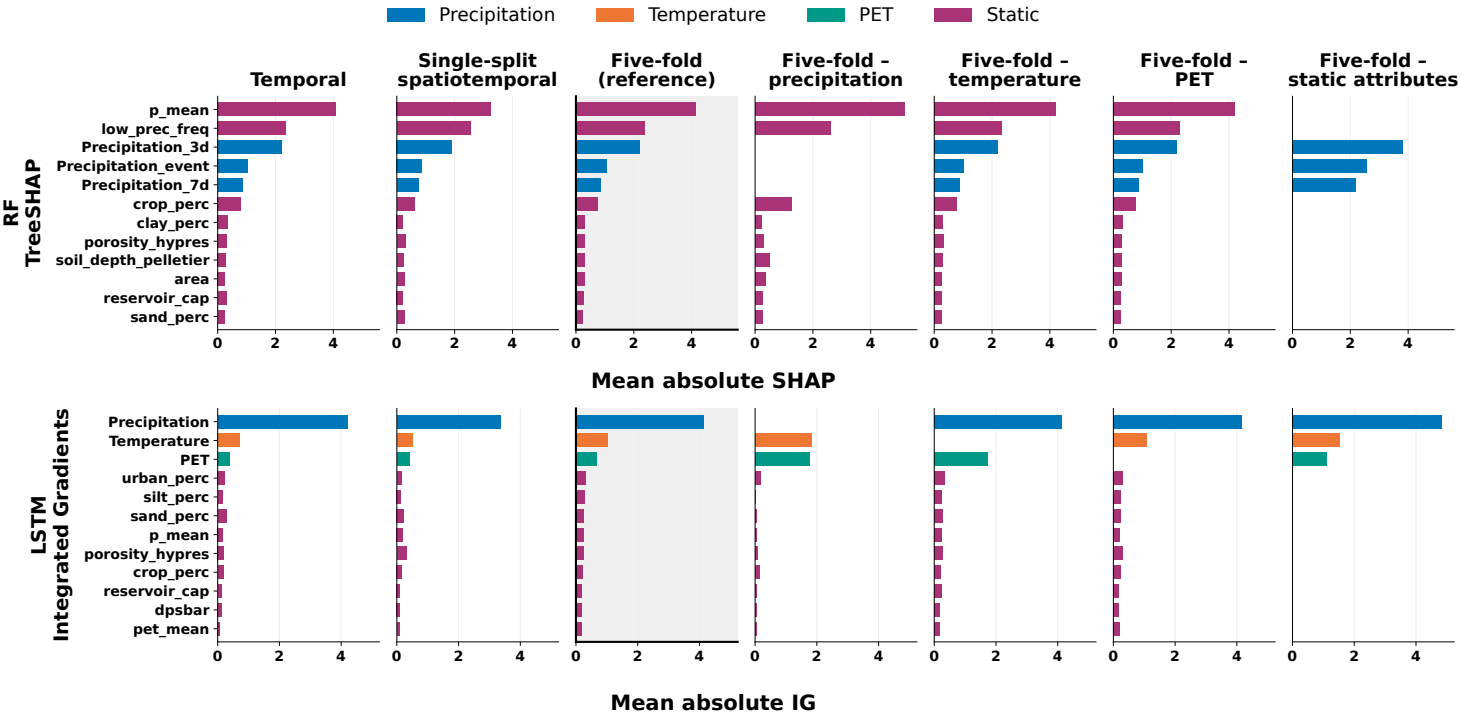

Figure 3. Feature importance on individual predictors across evaluation settings and input ablations. The 12 highest-ranked predictors are shown for each model: (i) RF with TreeSHAP (mean absolute SHAP value per held-out flood magnitude day, mm d^{-1}); and (ii) LSTM with IG (mean absolute attribution per held-out flood magnitude day). Columns show the three evaluation settings followed by the four ablation perturbations performed on the five-fold evaluation setting. Bars are colored by predictor type: dynamic precipitation-related variables (blue), temperature (orange), PET (green), and static attributes (purple). Note that missing bars denote predictors removed in a given ablation. The five-fold setting is the reference (shaded). Values are means across the five folds, except for the temporal and single-split spatiotemporal settings, which are single model fits.

Next, when feature importance was considered in information groups by process-group shares (Fig. 4), event-day forcing accounted for only a small fraction of feature importance in either architecture (0.07 for the RF and 0.15 to 0.19 for the LSTM), with the remainder divided between antecedent meteorology (including precipitation, temperature, and PET occurring before the flood magnitude day itself) and static catchment attributes. Group composition was almost invariant across the three evaluation settings, as no group changed by more than 0.04 despite pooled R^2 differences occurring (see Fig. 1). Removing temperature or PET likewise altered group shares by no more than approximately 0.03. The remaining two ablations produced larger, model-type-specific redistributions. Removing precipitation increased the RF static share from 0.67 to 0.90, whereas the LSTM antecedent share increased from 0.46 to 0.75, indicating that the RF compensated primarily with catchment information while the LSTM relied more heavily on the remaining meteorological sequence. Removing static catchment attributes shifted both models toward dynamic forcing in similar proportions (RF 0.23 event day / 0.77 antecedent; LSTM 0.18 event day / 0.82 antecedent), suggesting that the difference between the two model types lies in the weight assigned to static attribute

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

information rather than in how meteorological information is partitioned between event-day and antecedent conditions. The Appendix displays additional feature importance results, including the additional robustness analyses testing perturbations of training seeds, predictor changes for the RF, and sequence length changes for the LSTM (Figs. B3-B5).

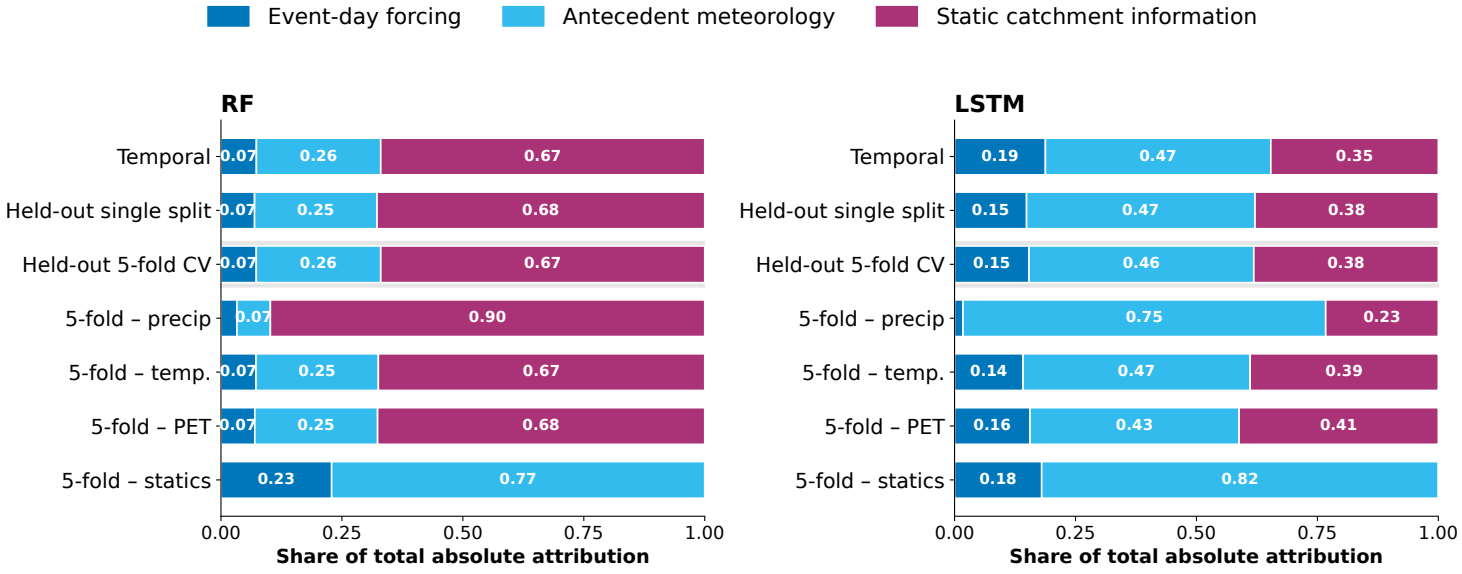

Figure 4. Composition of feature importance by process-group share, and its redistribution under input ablation. Feature importance is aggregated into the three groups defined in Table 2, event-day forcing, antecedent forcing, and static attributes. Results are normalized within each model, so that the three shares sum to one, before averaging across the five folds. RF values are based on TreeSHAP and LSTM values on IG. The LSTM split between event-day and antecedent forcing uses the lag-resolved attribution profile (lag 0 versus lags 1-29). Rows show the three evaluation settings followed by the four matched five-fold ablations. The five-fold setting (shaded) is the reference, and the annotation to the right of each row gives the largest change in group share relative to that reference. Because the two models not only represent antecedent information differently, i.e., 30 sequence steps in the LSTM versus engineered accumulation windows in the RF, but also contain different numbers of predictors per group, absolute group shares are not directly comparable between panels. Changes between rows are comparable because representation is fixed within each model.

3.3 Feature importance on antecedent information across timescales

The two model types concentrated feature importance on antecedent precipitation at short timescales, but with different temporal profiles (Fig. 5 (a) and (b)). Under the five-fold evaluation setting, the RF placed its greatest feature importance on the 3-day accumulated precipitation predictor (2.21), exceeding feature importance on event-day precipitation (1.04), and declined thereafter (7 d 0.86, 14 d 0.22, 30 d 0.18). The LSTM also relied more strongly on precipitation immediately preceding the flood magnitude day than on precipitation occurring on the day itself, peaking one day beforehand (1.51 versus 1.35) and declining by approximately an order of magnitude within three days, with 79 % of total precipitation attribution occurring within the first four lags.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

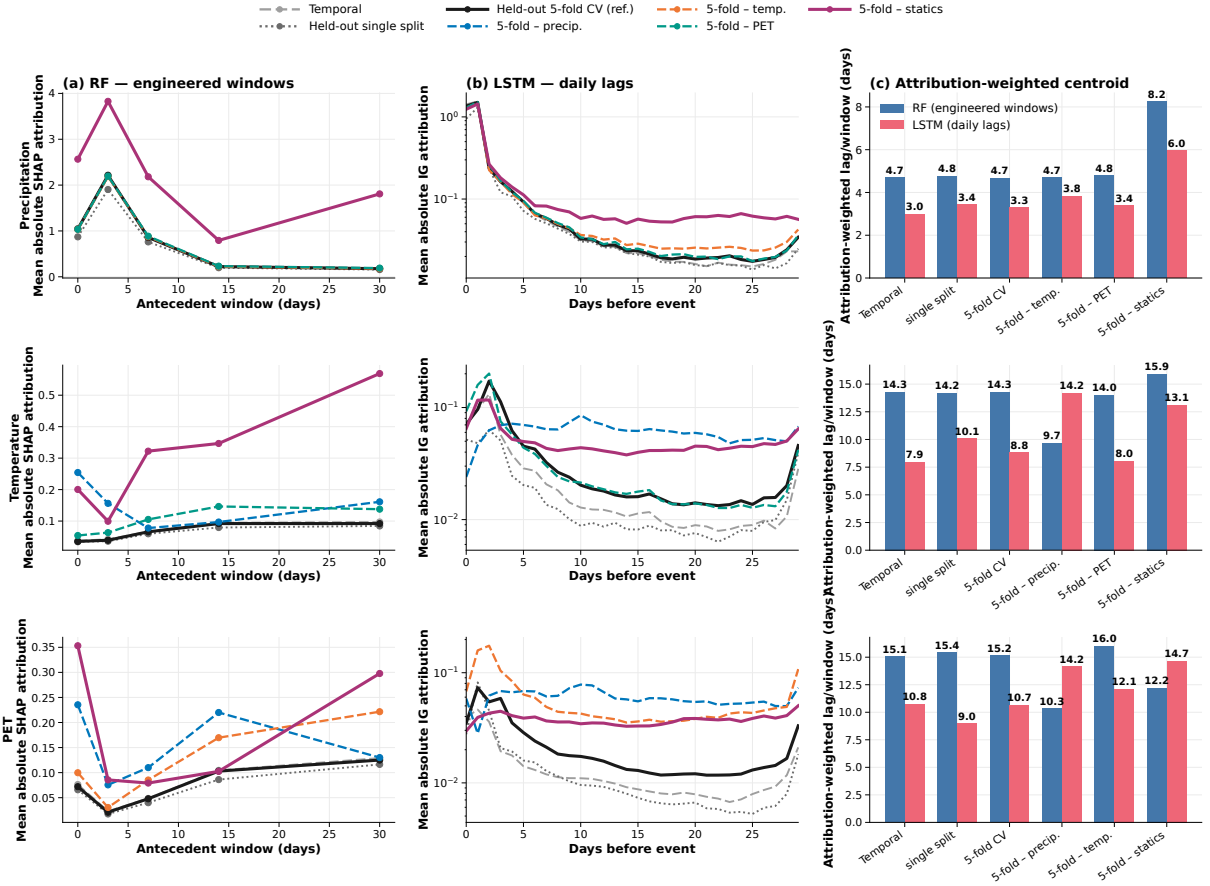

Figure 5. Representation of antecedent meteorological information across timescales. The rows correspond to the three dynamic forcings (precipitation, temperature, and PET). (a) RF feature importance (TreeSHAP; mean absolute SHAP value per held-out flood magnitude day) as a function of engineered antecedent window length, from the flood magnitude day (0) to 30 days beforehand. (b) LSTM feature importance (IG; mean absolute attribution per held-out flood magnitude day) as a function of daily lag, shown on a logarithmic vertical axis because attribution spans approximately two orders of magnitude between the leading lags and the tail. (c) Attribution-weighted mean lag ("memory centroid") for each model and evaluation setting, representing the characteristic timescale over which each model relies on antecedent information. The three evaluation settings are shown in shades of gray (the five-fold evaluation setting, shown in bold black, is the reference), and the ablation experiments are shown in color. A forcing cannot appear in its own ablation, so the corresponding position in its own row is empty. Values for the five-fold settings are means across the five folds. The temporal and single-split spatiotemporal settings are single model fits. LSTM centroids are calculated across the complete 30-day input sequence (lags 0-29). RF and LSTM centroids are not directly comparable in magnitude because RF window importance is assigned at the window's full nominal length, whereas LSTM attributions resolve individual days, so centroids should be compared across settings and forcings within each model.

Despite their different representations of antecedent conditions, both models emphasized precipitation immediately preceding the flood magnitude day over precipitation on the flood magnitude day itself. Temperature and PET were represented over much

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

longer timescales than precipitation. In the RF, feature importance on both these variables increased with antecedent window length rather than declining, resulting in memory centroids approximately three times longer than that of precipitation (14.3 and 15.2 days versus 4.7 days). The LSTM also showed a similar ordering (8.8 and 10.7 days versus 3.3 days). Thus, both model types emphasized precipitation primarily as event-scale forcing, whereas the energy-related variables (temperature and PET) extended over longer antecedent timescales.

Removing static catchment attributes in the five-fold setting lengthened antecedent memory in both model types. The precipitation memory centroid (Fig. 5 (c)) increased from 4.7 to 8.2 days in the RF and from 3.3 to 6.0 days in the LSTM, while the fraction of LSTM precipitation attribution occurring within the first four lags decreased from 0.79 to 0.64 (Fig. 5 (b)). Similar lengthening occurred for temperature (RF 14.3 to 15.9 days; LSTM 8.8 to 13.1 days) and for LSTM PET (10.7 to 14.7 days). The only exception was RF PET, whose memory centroid shortened from 15.2 to 12.2 days. For the RF, the longer precipitation memory following the removal of static attributes did not arise from a uniform shift across accumulation windows but instead from a redistribution toward the longest engineered predictor. Feature importance on the 30-day accumulated precipitation predictor increased approximately tenfold (from 0.18 to 1.81), whereas the 14-day predictor increased by less than fourfold, producing a non-monotonic feature importance profile.

Removing precipitation in the five-fold evaluation setting produced different changes in the representation of antecedent temperature and PET between the model types. For the LSTM, the temperature memory centroid increased from 8.8 to 14.2 days, and the PET from 10.7 to 14.2 days. However, for the RF, the temperature centroid decreased from 14.3 to 9.7 days, and the PET from 15.2 to 10.3 days. Therefore, removing precipitation redistributed feature importance toward longer antecedent timescales in the LSTM, but toward shorter antecedent windows in the RF.

The LSTM precipitation memory centroid was shortest under temporal evaluation, where all catchments were available during training (3.0 days), and slightly longer under the two spatiotemporal evaluation settings (3.3-3.4 days). Removing static attributes represented the strongest expression of this trend. Removing temperature or PET changed precipitation memory by less than 0.6 days in either model type.

3.4 Two roles of the static attributes

The static attributes were found to contribute in two different ways to flood magnitude prediction (Fig. 6). First, the static attributes corrected a systematic BFI-related bias, and this pattern was consistent across both model types. Catchment PBIAS was found to be correlated with BFI when static attributes were included (Spearman $\rho = +0.36$ for the RF, and $+0.21$ for the LSTM), and removing the static attributes strengthened these relationships to $\rho = +0.72$ and $+0.61$, respectively (Fig. 6b). Without static attributes, both model types increasingly underpredicted flood magnitude in flashy, low-BFI catchments and overpredicted it in high-BFI, groundwater-dominated catchments. For the LSTM, mean PBIAS without static attributes ranged from -64.5% in catchments with BFI < 0.40 to $+29.0\%$ in catchments with BFI ≥ 0.70, and intermediate BFI bands fell between these extremes (Fig. 6b). The RF showed the same overall pattern, but had a considerably larger positive bias in the highest-BFI band (mean $+142.3\%$ versus $+29.0\%$ for the LSTM). Therefore, the same BFI-related bias structure emerged in both the event-based RF and the sequence-based LSTM. Spatially, LSTM bias without static attributes was predominantly

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

negative across GB, with positive biases confined to a relatively small number of high-BFI catchments in southern and eastern England (Fig. 6a). Because the main five-fold and static-attributes-ablated models were trained and evaluated on identical folds, this comparison is internally controlled with respect to the fold design (Sect. 2.5).

The second role of the static attributes is that they shortened the antecedent period over which both models relied on precipitation. For the LSTM, removing static attributes shifted the mean of the catchment-level precipitation memory centroids (392 catchments) from 3.72 to 6.29 days, corresponding on average to a lengthening of 2.58 days. The increase was quite consistent across the BFI gradient, with band means ranging from $+2.34$ to $+2.79$ days across the four BFI bands. There was nevertheless a weak negative association between BFI and catchment-specific change in centroid (Spearman $\rho = -0.15$, Holm-adjusted $p = 0.009$; Fig. 6d).

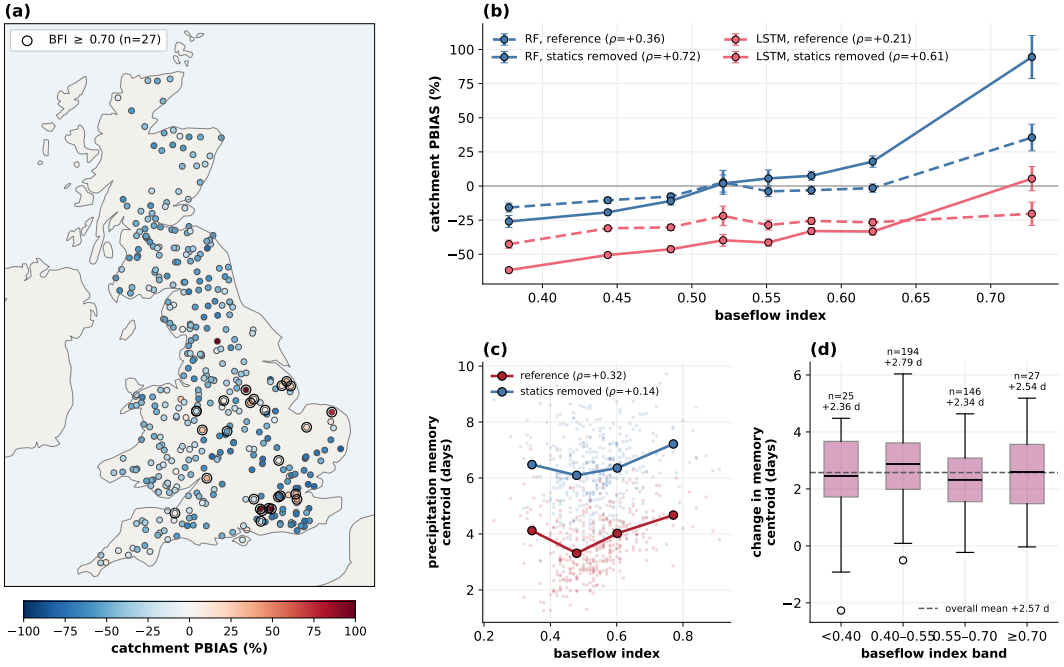

Figure 6. BFI-related bias correction and modulation of antecedent memory usage. (a) Within-catchment PBIAS for the LSTM after removing static attributes under the five-fold evaluation setting. Blue denotes underprediction and red overprediction (catchments with BFI ≥ 0.70 are outlined). Colors are clipped to $\pm 100\,\%$. (b) Catchment PBIAS as a function of BFI for both model types with and without static attributes, and points show means within BFI bands with standard errors. In addition, Spearman correlations are reported in the legend. (c) LSTM precipitation memory centroid as a function of BFI, where points represent individual catchments and lines connect BFI-band means. (d) Increase in the LSTM precipitation memory centroid after removing static attributes, grouped by BFI band. Boxes show the interquartile range, whiskers the non-outlier range, and the dashed line the overall mean. Bands are left-closed (e.g., BFI $\in [0.55, 0.70)$). These reporting bands are distinct from the BFI quartiles used to stratify the folds, and each band contains held-out catchments from all five folds. The memory-centroid analysis is shown only for the LSTM because the RF represents antecedent information using fixed accumulation windows rather than daily lags.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

With static attributes available, the precipitation memory centroid increased modestly with BFI ($\rho = +0.32$). After static attributes were removed, this relationship weakened to $\rho = +0.14$. A permutation test in which the observed per-catchment shifts were randomly re-assigned across catchments (2,000 permutations) reproduced most of this weakening (mean $\rho = +0.21$, 95 % range $+0.13$ to $+0.28$, against $+0.14$ observed). This indicates that the weakening may be largely attributable to BFI-unstructured variability in the size of the shift rather than to a systematically larger lengthening in any catchment type. Thus, static attributes are found to correct a BFI-dependent prediction bias in both model types. In the LSTM, they also shortened precipitation memory by about 2.6 days. The latter effect occurs broadly across the BFI spectrum, although there is a weak tendency for the amount of lengthening to decrease with BFI.

4 Discussion

The central result of this study is that predictive skill and hydrological process interpretation from feature importance analysis are not interchangeable properties of a fitted model. First, we provide large-sample evidence across 139 model fits of two model types that feature importance structure was both nearly invariant to evaluation setting and reorganized when predictor representation was changed (Figs. 3 and 4). Second, both model types achieved useful pooled skill (Fig. 1), but with largely negative within-catchment skill, while maintaining plausible feature importance results. This indicates that predictive skill, and the feature importance patterns supporting it, are largely derived from inter-catchment variability rather than within-catchment variability for flood magnitude estimation. Therefore, feature importance findings should be considered at the scale at which good predictive skill is actually attained, and as conditional on the properties of a fitted model, such as predictor representation, evaluation setting, and training data sample (Breiman, 2001b; Fisher et al., 2019; Rudin, 2019; Slater et al., 2025), rather than as real hydrological process importance. Our findings are consistent with the broader issue of non-unique model representations and interpretations in hydrology (Beven and Freer, 2001; Beven, 2020). These findings do not diminish the value of feature importance methods, but instead suggest that ML-based interpretation of floods is more defensible when feature importance patterns are assessed across robustness experiments, and are interpreted and reported at the scale where skill actually exists.

These findings contribute to the intersection of two areas of recent hydrological ML research. Large-sample rainfall-runoff studies have shown that LSTMs and related DL models achieve strong daily streamflow performance across diverse catchments, including in GB (Kratzert et al., 2018, 2019b; Lees et al., 2021, 2022; Kratzert et al., 2024). At the same time, other studies have argued that strong predictive performance does not necessarily imply reliable extrapolation to extreme events, physically realistic behavior, or interpretable internal representations (Gauch et al., 2021b; Heudorfer et al., 2025; Nearing et al., 2021; Beven, 2020). Our results speak to both findings, and extend both areas to flood magnitude estimation instead of the entire hydrograph for streamflow prediction. We find that the LSTM retained useful pooled skill on held-out flood magnitude events, but systematically underestimated flood magnitudes, including the largest events (Fig. 2), and differed in feature composition across information groups compared to an event-based RF (Fig. 4). The key distinction is not whether DL is useful for flood estimation, but whether predictive performance and feature importance patterns can be interpreted as hydrological evidence

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.



without further testing of their stability under methodological perturbations and across scales of analyses (pooled and within-catchment).

4.1 Predictive skill is largely inter-catchment skill

As anticipated, the RF outperformed the LSTM in every directly comparable model fit and for every performance metric (Fig. 1, Table A9 and Table A8), highlighting that an event-based tabular ML model trained directly on events is suited to estimating flood magnitude. This is because the task of the RF was arguably easier, given that it only had to learn a specific part of the distribution. However, the aim of the study was not a competitive comparison, and this is not evidence that RFs are generally superior to LSTMs. In light of this, our results complement previous demonstrations of strong daily rainfall-runoff performance by LSTMs, by testing their performance and feature importance patterns for flood magnitude events alongside a RF, within a multi-experimental robustness framework. More telling was the scale at which predictive skill emerged for the RF and LSTM. Pooled performance greatly exceeded within-catchment performance, with negative within-catchment R^2 in many RF catchments and in most LSTM catchments (Fig. 1). This divergence indicates that the apparent pooled skill is because both models can predict *where* floods are large much better than how large they may be within a given catchment. However, because within-catchment skill is arguably the problem most relevant to event-scale flood estimation and to process understanding, high pooled performance should not be taken to imply skill at all scales, and the difference between pooled and within-catchment performance should be reported routinely in large-sample studies.

The RF's feature importance structure provides an explanation for this prediction gap between scales. Two of its three leading predictors described static long-term precipitation climatology rather than event-scale dynamic precipitation (Fig. 3). These variables assist RF models in efficiently distinguishing catchments with systematically larger floods, explaining much of the model's pooled performance while contributing relatively little to discriminating flood magnitudes within individual catchments. The small performance cost of removing the static attributes (i.e., 0.05 in pooled R^2), and moderate cost of removing precipitation (i.e., 0.13 in pooled R^2) do not contradict this. This is because not only does feature importance quantify the information a fitted model uses (and not the information it uniquely needs), but also the accompanying redistribution of feature importance shows that the RF reconstructed much of the same between-catchment signal through replacing the static attributes with precipitation, and vice versa (Fig. 3). This is consistent with recent evidence that static and climatological descriptors can function as spatial proxies rather than direct representations of hydrological processes (Li et al., 2022; Heudorfer et al., 2025; Yu et al., 2024; Ford et al., 2026). Crucially, for process inference this flexibility is a cause for concern. Feature importance is revealing which route the fitted model chose, and not which route is necessarily physically correct or a causal explanation.

The two model types also failed differently across the range of flood magnitude event sizes. The RF progressively overpredicted smaller floods and underpredicted larger floods within each catchment (Fig. 2), consistent with regression toward the conditional mean. By contrast, the LSTM underestimated flood magnitudes by an approximately constant fraction across the full range of event sizes, aligning with continuous rainfall-runoff models trained using squared-error objectives (Gupta et al., 2009). Further, the underestimation of the largest peaks has been reported, even where LSTMs have outperformed process-based models for

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

extremes (Frame et al., 2022). This underestimation may be partly explained by peak-timing offsets, because the LSTM was evaluated on the observed flood magnitude day and not tested for its own simulated magnitude, which may differ in timing.

4.2 The two model types rely on similar information, different static weighting, and consistent antecedent timescales

Across both model types, precipitation-related information dominated feature importance, whereas temperature and PET contributed substantially less (Fig. 3). This is consistent with existing knowledge of GB flood generation mechanisms being largely dominated by storm forcing interacting with antecedent catchment wetness, and energy-related variables reflecting broader seasonal conditions (Berghuijs et al., 2019; Griffin et al., 2025). Both models were supplied with dynamic and static precipitation predictor information, and where the two model types differed was in how this information was weighted. For example, the RF concentrated feature importance primarily on precipitation-related static descriptors, whereas the LSTM relied more on temporal sequences of precipitation (Fig. 3). These differences do not necessarily indicate competing hydrological explanations, but reflect that the RF receives antecedent precipitation through predefined engineered accumulation windows, while the LSTM learns temporal dependencies directly from ordered meteorological sequences. In statistical learning, multiple well-performing models may achieve similar predictive accuracy through different predictor relationships, also known as the Rashomon effect (Breiman, 2001b). Hydrology has long recognized a similar principle as equifinality (Beven and Freer, 2001; Beven, 2020), here relating to predictor representation instead of parameters.

The input-feature ablation experiments indicate that the RF and LSTM compensated differently when predictive information was changed (Fig. 3). Removing precipitation reduced predictive skill in both model types, but far more so for the LSTM, collapsing its pooled skill below that of the observed mean. Despite this, both models still provided plausible feature importance patterns (Figs. 3 and 5). In the absence of precipitation, the RF redistributed feature importance toward static catchment attributes and retained good pooled skill, whereas the LSTM shifted focus to temperature and PET (Figs. 3 and 5). When static attributes were removed, both model types partitioned remaining dynamic forcing between event-day and antecedent information in similar proportions. This could suggest that a principal difference between model feature usage lies in the weight assigned to static attributes, rather than how dynamic meteorological information is used. However, it is important to consider that model type and attribution method (TreeSHAP vs IG) are confounded in this comparison, so the feature importance rankings for the same held-out flood events cannot be entirely attributed to architectural differences. Moreover, a small reduction in predictive performance when removing a feature does not imply that the feature is redundant. This is because some features may have overlapping information and be correlated with one another (Strobl et al., 2008). With this in mind, our results suggest that a plausible feature importance pattern does not translate to predictive skill, as the LSTM produced a physically meaningful feature importance pattern while performing poorly.

The agreement between the two model types is informative, because despite representing antecedent information in fundamentally different ways, the RF and LSTM converged on a consistent usage ordering of hydrological timescales. Both models relied more strongly on precipitation immediately preceding the flood event than on precipitation occurring on the flood magnitude event itself, and assigned longer timescales to temperature and PET (Fig. 5). These findings are consistent with previous hydrological studies identifying storm forcing and antecedent wetness as dominant controls on flood response, and energy-

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

related variables providing slower catchment conditioning (Berghuijs et al., 2016, 2019; Brunner and Dougherty, 2022). Recovering the same ordering of timescales despite these contrasting representations suggests that the separation between short precipitation memory and longer energy-related memory reflects the prediction problem more so than the modeling framework. However, in the RF, the 3-day precipitation window includes the event-day precipitation, so event-day and short-term antecedent precipitation are not entirely distinct.

4.3 Feature importance is robust at the process-group level but variable at the predictor level

The robustness experiments distinguish aspects of feature importance that are consistently reproducible, and those that remain more contingent on modeling choices. Across the temporal, single-split spatiotemporal, and five-fold evaluation settings, predictive skill changed substantially, but the broad composition of feature importance changed comparatively little (Figs. 3 and 4). This could be interpreted as a decoupling between feature importance patterns and predictive skill (Sect. 3.2). On the other hand, this stability may be argued to be expected because the three settings resample the same catchment population, so fitted mappings can come from overlapping information, reflecting broad learned patterns.

Alternative training seeds, LSTM sequence lengths, and RF predictor sets produced minor changes in the feature importance patterns (see Figs. B3, B5 and B4). At the level of broad process groups, feature importance therefore proved more stable than might be expected given the known sensitivity of predictive skill to spatial transfer (Kratzert et al., 2019b; Lees et al., 2021) and recent discussion regarding the robustness of explainable ML in hydrology (Slater et al., 2025; Ford et al., 2026).

The limits of model interpretation became apparent at finer, individual predictor-level, scales. Across the engineered RF predictor bases (Fig. B4), feature importance on individual precipitation windows varied substantially, whereas feature importance on the dominant static climatic predictors remained comparatively stable. This may be expected, because overlapping accumulation windows encode much of the same antecedent information, allowing feature importance to redistribute among whichever summaries are supplied by the modeler. Thus, predictor-level rankings within a correlated group can be argued to be a property of the modeler's decision, not that of the catchment behavior.

4.4 Static attributes perform multiple roles

Static attributes are recognized as essential for large-sample hydrological prediction because they encode differences in climate, geology, soils, topography, and storage characteristics that cannot be inferred from meteorological forcings in isolation (Kratzert et al., 2019b; Lees et al., 2021, 2022; Coxon et al., 2026; Ford et al., 2026). Moreover, static attributes have been at the center of recent discourse, regarding their use in ML models and how physically meaningful they may be (Heudorfer et al., 2025, 2026; Li et al., 2022). Our results further refine this interpretation by showing that their contribution cannot be summarized by a single feature importance value (Fig. 6), as our work suggests they may perform at least two distinct functions with different spatial structure.

The first role is a bias correction organized along the BFI gradient (Fig. 6). Removing static attributes strengthened the relationship between prediction bias and BFI in both model types, approximately doubling the correlation for the RF and tripling it for the LSTM. This related to systematic underprediction in flashy catchments and overprediction in groundwater-

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

dominated catchments (Fig. 6). This could be ~~reasoned due to~~ without static attributes, the models may learn a single pooled average mapping from forcing to streamflow. Therefore, for the same rainfall, steep and low-permeability catchments produce much larger peak flows than high-BFI catchments, with the average mapping underpredicting the former and overpredicting the latter. Static attributes can be used as proxies to allow the model to scale responses. Because a similar bias correction pattern emerged in both the event-based and sequence-based models, this finding is more attributable to the availability of static attributes than to modeling architecture. It is also arguably physically meaningful, being consistent with the role of catchment storage and flow attenuation in modulating flood response (Lees et al., 2021; Lane et al., 2019). However, this could also be interpreted as spatial-proxy behavior.

The identified bias structure is not attributable to the BFI-stratified five-fold evaluation setting design. This is because the quartiles used to stratify the folds are distinct from the fixed-threshold bands used for reporting, and each band contains held-out catchments from all five folds. Critically, any optimism introduced by the fold design applies equally to the reference five-fold and static-attribute-ablated models. Therefore, if anything, the stratification is conservative with respect to this result because every held-out high-BFI catchment had other high-BFI catchments available in training. Despite this, the bias continued to emerge. This can be explained because a model without static attributes is unable to identify which held-out catchments those training examples may resemble.

Further, static attributes shortened the antecedent period over which both models relied on precipitation (Sect. 3.3). For the LSTM where this response can be understood at the catchment level, little spatial organization was observed. Antecedent memory lengthened by a similar amount across the full BFI gradient rather than preferentially in catchments where longer hydrological memory might be expected (Fig. 6). Static attributes supported a modest BFI grading of the memory timescale, with longer antecedent integration in higher-baseflow catchments. Their removal degraded that gradient chiefly by adding a lengthening whose magnitude was largely indifferent to catchment type. This suggests that the static attributes assist the model in identifying the variability in different catchments long-term wetness. Without the static attributes, this can be inferred only from dynamic forcing alone, and longer meteorological totals are the best available indicator for this. The lengthening being largely BFI unstructured, provides consistency with a catchment identification role, rather than a storage memory one.

Our results suggest that static attributes can both be physically meaningful and act as spatial proxies simultaneously. The bias correction is spatially structured and physically interpretable, whereas the change in antecedent memory appears to be a more global property of the fitted sequence model. Therefore, these findings reinforce the broader conclusion that feature importance should not be interpreted as a direct measure of process importance. A feature importance value for the static predictor group conflates multiple functions operating at different spatial scales and through different model mechanisms. Separating these roles in future work provides a more informative interpretation of how static information contributes to flood-magnitude prediction.

4.5 Implications for flood estimation

Together, these results support the careful interpretation of feature importance in large-sample hydrology ML studies in layers, which we regard as the study's main epistemic contribution. We emphasize the need to complement inference methods with robustness testing. This is because hydrological studies may use a variety of different predictor data sources and information

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

groups. More generally, for UK flood estimation, the Flood Estimation Handbook (FEH) (Reed, 1999) remains a foundation of design flood estimation as it provides standardized, auditable procedures across gauged and ungauged catchments.

The event-based RF estimated magnitudes on identified flood days with useful skill and interpretable feature importance, but addresses neither flood timing nor extrapolation to design standards. Here, its usefulness is conditional on the event already being identified, and is being applied to understand what the model is learning about flood magnitude. Applications requiring continuous simulation and hydrograph evolution, retain clear advantages for an LSTM (Kratzert et al., 2018; Lees et al., 2021), despite its lower event-magnitude accuracy (Fig. 1). The choice between model types should follow modeling purpose rather than average predictive performance (Yilmaz et al., 2008; Wagener et al., 2001; Gupta et al., 2009; Nearing et al., 2026).

4.6 Limitations and future work

The LSTM was evaluated in a standard continuous-simulation configuration, with a squared-error (NSE) loss function, and a 30-day sequence length. Extreme-specific losses, probabilistic objectives, considerably longer memory horizons, event-aware sampling, hybrid constraints, or larger training samples could potentially reduce the systematic underestimation and result in different feature importance patterns (Frame et al., 2022; Klotz et al., 2022). The 30-day sequence length is relatively short compared to the 365 days used in previous works for streamflow simulation (Kratzert et al., 2018; Lees et al., 2021). Although sensitivity analysis was run for sequence lengths of 14 and 60 days, future work could test sensitivity over longer periods. This could be useful to further interrogate the underestimation and weaker skill in high-BFI catchments.

The two model types also required different feature importance methods, TreeSHAP for tree ensembles and lag-resolved gradient attribution for sequence models. Therefore, cross-model comparisons are restricted to relative patterns and aggregated process-group shares. Furthermore, the CAMELS-GB climate attributes are computed over 1970 - 2022, so they contain information from the validation and test periods. Because `p_mean` is the leading RF predictor, this must be considered, but the effect is likely small because a 52-year mean is dominated by training-period years. Future work could involve recomputing indices over the training period. Moreover, the catchment system is only partially observable in the daily data used from Coxon et al. (2026), i.e., the models receive no explicit information on storm structure, sub-daily precipitation, soil moisture, or groundwater. High-flow observations are uncertain, and rating curves are often extrapolated beyond the gauged range (McMillan et al., 2012; Coxon et al., 2015), which may influence performance and feature importance patterns.

Future work should therefore extend the framework of this study to include a sub-daily comparison, utilizing higher resolution precipitation and streamflow data. In addition, research could also further test what the ML models learn, using spatially blocked holdout designs. Also, probabilistic model types, such as quantile RFs could separate conditional-mean error from predictive uncertainty. Boosted tabular learners, and mass-conserving recurrent architectures, could test whether the feature importance conclusions and the interpretation of static attributes generalize beyond this study's limits. It is also advised to extend this work beyond flood magnitude to flood frequency, event clustering, hydrograph shape, and compound events.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

5 Conclusions

This study assessed predictive skill and feature importance for large-sample ML flood estimation across two model types, to investigate what ML models learn about floods. An event-based RF and a continuous sequence-based LSTM were evaluated on identical held-out flood magnitude events across GB catchments, using 139 model fits spanning three evaluation settings, input-feature ablations, training seeds, and alternative representations of antecedent information. The central finding of this work is that predictive skill and feature importance are not interchangeable properties of a fitted model, and what the model learns should be investigated alongside optimizing for good predictive skill. Feature importance structure was nearly invariant to evaluation setting, even as pooled model skill changed considerably. Further, pooled performance substantially exceeded within-catchment performance across both model types, as much of the apparent skill emerged from the models' ability to capture inter-catchment rather than intra-catchment variability. Thus, the corresponding feature importance patterns describe how each model used the between-catchment signal, and should be interpreted at the scale at which good predictive skill was attained (pooled).

The model types also failed differently on the largest floods. The RF had a magnitude-dependent bias, while the LSTM displayed an approximately constant fractional underestimation. Despite the contrasting model type architectures, the two models use broadly similar hydrological information. Both were dominated by precipitation, relied much more on antecedent conditions than on event days, and represented temperature and PET over longer timescales than precipitation. Their main difference was found in the weight assigned to static attributes. When static attributes were removed, both models converged onto dynamic forcing. The RF's greater feature importance on static long-term precipitation climatology also explains its strong between-catchment performance and comparatively weak discrimination among events within catchments. Finally, static catchment attributes were found to provide two functions: (i) they suppressed a systematic bias organized along the BFI gradient; and (ii) they shortened the antecedent period over which both models relied on precipitation.

These findings support practices for large-sample ML studies of floods. Pooled and within-catchment skill should be reported together, and feature importance should be tested across evaluation settings and predictor representations before process conclusions are inferred.

Appendix A: Supplementary tables

Table A1. Dynamic hydrometeorological variables used in this study, from CAMELS-GB (Coxon et al., 2026).

Variable	Source	Units	Role in this study
Precipitation (P)	HadUK-Grid	mm d^{-1}	Daily dynamic input to the LSTM and event-day/antecedent predictor for the RF.
Mean air temperature (T)	HadUK-Grid	°C	Daily dynamic input to the LSTM and event-day/antecedent predictor for the RF.
PET	Hydro-PE	mm d^{-1}	Daily dynamic input to the LSTM and event-day/antecedent predictor for the RF.
Specific discharge (Q)	UK National River Flow Archive	mm d^{-1}	Observed target variable used to identify flood magnitude events and evaluate model predictions.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table A2. Static attributes used in this study. These are from CAMELS-GB (Coxon et al., 2026), and are supplied to both the RF and LSTM models. The same 22 static attribute types were used in the LSTM configuration of Lees et al. (2021). The static attribute BFI was used only to define the five folds and was not included as a model predictor.

Attribute group	Predictors	Hydrological interpretation
Physiography	`area`, `elev_mean`, `dpsbar`	Describes catchment size, elevation, relief, and topographic controls on drainage response and runoff routing.
Soil properties	`sand_perc`, `silt_perc`, `clay_perc`, `porosity_hypres`, `conductivity_hypres`, `soil_depth_pelletier`	Describes soil texture, storage capacity, infiltration potential, hydraulic conductivity, and soil depth.
Snow and land cover	`frac_snow`, `dwood_perc_2015`, `ewood_perc_2015`, `crop_perc_2015`, `urban_perc_2015`	Describes snow occurrence, vegetation cover, agricultural land, impervious surfaces, interception, evapotranspiration, and runoff generation.
Water storage and regulation	`reservoir_cap`	Represents the influence of upstream reservoir storage and flow regulation.
Long-term hydroclimatic context	`p_mean`, `pet_mean`, `p_seasonality`, `high_prec_freq`, `low_prec_freq`, `high_prec_dur`, `low_prec_dur`	Describes long-term water and energy availability, precipitation seasonality, and the frequency and duration of wet and dry conditions.

Table A3. Number of held-out flood magnitude events used for model evaluation. RF and LSTM were evaluated on the same held-out events within each evaluation setting. All perturbation experiments used the same held-out events as the corresponding reference five folds, and predictive skill metrics are calculated per fold and then averaged across the five folds per model fit.

Evaluation setting	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Total
Temporal	-	-	-	-	-	7,147
Single-split spatiotemporal	-	-	-	-	-	1,436
Five-fold BFI stratified	1,399	1,426	1,507	1,365	1,450	7,147
Five-fold BFI stratified perturbations	1,399	1,426	1,507	1,365	1,450	7,147

Table A4. The default RF dynamic and static predictors used in the temporal, single-split spatiotemporal, and five-fold evaluation settings. These do not represent the perturbation experiment predictor sets for the RF, which are provided in Table A6.

Predictor type	Variables	Hydrological interpretation
Event-day forcing	P, T, and PET on the flood magnitude day	Event day forcing and hydroclimatic conditions.
Short antecedent precipitation	3- and 7-day accumulated precipitation	Recent catchment wetness and short-term storm accumulation.
Medium antecedent precipitation	14- and 30-day accumulated precipitation	Antecedent catchment wetness and hydrologic preconditioning.
Antecedent temperature	3-, 7-, 14-, and 30-day mean temperature	Antecedent energy-related and seasonal conditions.
Antecedent PET	3-, 7-, 14-, and 30-day mean PET	Antecedent evaporative demand and seasonal energy availability.
Static catchment attributes	22 CAMELS-GB predictors	Spatially varying catchment characteristics and long-term hydroclimatic context.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table A5. Final RF and LSTM model configurations. RF hyperparameters were selected using the validation period then held fixed across all experiments. LSTM hyperparameters were selected to match the default settings of *NeuralHydrology*. For the LSTM, validation period metrics were used for final checkpoint selection.

Model	Configuration item	Value or description
RF	Software	*scikit-learn 1.9.1*, regression model.
RF	Number of trees	1000.
RF	Maximum tree depth	None.
RF	Minimum samples per leaf	5.
RF	Maximum features	0.5 of the available predictors considered at each split.
RF	Random seed	193893 for the main experiments, with additional matched seeds evaluated as sensitivity analyses.
RF	Hyperparameter-selection data	validation events from 1 January 2010 to 31 December 2015.
RF	Selection criterion	Highest validation R^2
LSTM	Software and architecture	*NeuralHydrology* version 1.12.0, standard `cudalstm`.
LSTM	Hidden size	128.
LSTM	Sequence length	30 days, including the target day.
LSTM	Dynamic inputs	Daily precipitation, PET, and mean air temperature.
LSTM	Static inputs	The 22 attributes listed in Table A2.
LSTM	Target	Daily specific discharge, `discharge_spec`.
LSTM	Loss function	Nash-Sutcliffe efficiency loss.
LSTM	Optimizer	Adam.
LSTM	Learning-rate schedule	0.001 from epoch 1 and 0.0005 from epoch 20.
LSTM	Batch size	256.
LSTM	Training epochs	30 fixed epochs.
LSTM	Output dropout	0.3.
LSTM	Initial forget bias	3.
LSTM	Gradient clipping	Maximum gradient norm of 1.
LSTM	Random seed	193893 for the main experiments, with additional matched seeds evaluated as sensitivity analyses.
LSTM	Model selection	Weights saved after every epoch, and validation period metrics used for final checkpoint selection.
LSTM	Hyperparameter tuning	Settings were fixed by design, with sequence-length and random-seed effects evaluated separately as sensitivity analyses.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table A6. RF predictor-set perturbations used to evaluate the sensitivity of feature importance to antecedent-window representation. All models used identical static predictors, training and validation data, hyperparameter-selection procedures, and TreeSHAP attribution. Only the dynamic predictor set was modified. All antecedent summaries were restricted to the same 30-day horizon used for the LSTM look-back window.

Predictor set	Dynamic predictors	Hydrologic information	Purpose
Default	Event-day P, T, and PET; 3-, 7-, 14-, and 30-day accumulated P; 3-, 7-, 14-, and 30-day mean T and PET	Event forcing and short- to medium-term antecedent wetness	Reference RF configuration used throughout the primary analysis.
Short-window	Event-day P, T, and PET; 1-, 3-, and 7-day accumulated P; 1-, 3-, and 7-day mean T and PET	Near-event forcing and short-term antecedent wetness	Evaluates whether feature importance remains concentrated near the event day after removing longer antecedent windows.
Medium-window	Event-day P, T, and PET; 7-, 14-, and 30-day accumulated P; 7-, 14-, and 30-day mean T and PET	Medium-term antecedent wetness and catchment preconditioning	Evaluates whether feature importance shifts toward longer antecedent conditions when short windows are excluded.
Nested-window	Event-day P, T, and PET; 1-, 3-, 7-, 14-, and 30-day accumulated P; 1-, 3-, 7-, 14-, and 30-day mean T and PET	Nested antecedent-window representation over 30 days	Evaluates whether correlated overlapping windows redistribute feature importance.
Non-overlapping-window	Event-day P, T, and PET; accumulated P over days 1-3, 4-7, 8-14, and 15-30 before the event; corresponding mean T and PET	Independent antecedent periods within the 30-day horizon	Evaluates whether feature importance is assigned to distinct antecedent periods when overlapping windows are removed.
Precipitation-only	Event-day P; 3-, 7-, 14-, and 30-day accumulated P; static predictors retained	Precipitation forcing and spatial catchment controls	Evaluates whether the primary feature importance structure persists without dynamic temperature and PET inputs.
Reduced	Event-day P, T, and PET; 7- and 30-day accumulated P; 7- and 30-day mean T and PET	Simplified event and antecedent representation	Evaluates whether process-group feature importance is preserved with a reduced predictor set.

Table A7. Robustness and sensitivity analyses used to evaluate predictive performance and model-feature importance interpretation.

Analysis	Perturbation or diagnostic	What it tests
Spatial partitioning	Five-fold evaluation setting	Dependence of spatiotemporal skill and feature importance on the held-out catchment partition.
Input ablation (RF and LSTM)	Removal, in turn, of P, T, PET, or all static attributes across all five folds	Predictive contribution of each information source and the associated reliance redistribution.
Training-seed sensitivity (RF and LSTM)	Four additional matched seeds per fold	Stability of performance and feature importance to stochastic variation in model fitting.
RF predictor-basis sensitivity	Six alternative engineered antecedent-window predictor sets	Stability of RF feature importance to feature-engineering choices.
LSTM sequence-length sensitivity	Look-back lengths of 14, 30, and 60 days	Dependence of memory-timescale feature importance on the input horizon.
Observational uncertainty	High-flow discharge and precipitation uncertainty considered in interpretation	Recognizes that model error and feature importance reflect imperfect observations.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table A8. Pooled performance across held-out flood magnitude events. Values for the five-fold runs are the fold mean ± standard deviation, n is the total number of held-out events, and "default" is the reference RF predictor basis, identical to the 5-fold evaluation setting.

Setting	Model	n	R^2	MAE	RMSE	PBIAS (%)
Temporal	RF	7147	0.86	3.45	6.13	−6.2
	LSTM	7147	0.68	5.92	9.25	−29.3
Spatiotemporal (single split)	RF	1436	0.71	4.10	6.45	−12.1
	LSTM	1436	0.47	6.12	8.72	−33.3
5-fold	RF	7147	0.77 ± 0.02	4.51 ± 0.44	7.65 ± 1.57	-7.8 ± 2.1
	LSTM	7147	0.56 ± 0.05	6.90 ± 0.76	10.69 ± 2.39	-31.7 ± 2.5
−precipitation	RF	7147	0.64 ± 0.03	5.59 ± 0.51	9.55 ± 1.88	-7.2 ± 2.8
	LSTM	7147	-0.84 ± 0.15	15.95 ± 1.25	21.54 ± 2.75	-84.1 ± 0.8
−temperature	RF	7147	0.77 ± 0.02	4.54 ± 0.45	7.66 ± 1.54	-7.4 ± 2.2
	LSTM	7147	0.54 ± 0.10	7.05 ± 1.14	10.93 ± 3.02	-32.5 ± 4.2
−PET	RF	7147	0.77 ± 0.02	4.52 ± 0.46	7.65 ± 1.56	-7.9 ± 2.2
	LSTM	7147	0.55 ± 0.06	6.95 ± 1.01	10.81 ± 2.68	-32.6 ± 3.8
−static attributes	RF	7147	0.72 ± 0.03	5.46 ± 0.53	8.51 ± 1.76	-9.1 ± 1.8
	LSTM	7147	0.35 ± 0.04	8.91 ± 1.07	12.96 ± 2.55	-44.1 ± 3.0
seed 42	RF	7147	0.77 ± 0.02	4.52 ± 0.45	7.67 ± 1.58	-7.7 ± 2.1
	LSTM	7147	0.56 ± 0.06	7.04 ± 0.97	10.68 ± 2.43	-33.3 ± 3.6
seed 123	RF	7147	0.77 ± 0.02	4.52 ± 0.45	7.68 ± 1.59	-7.7 ± 2.2
	LSTM	7147	0.55 ± 0.08	6.97 ± 0.95	10.86 ± 2.92	-33.0 ± 3.4
seed 2024	RF	7147	0.77 ± 0.02	4.52 ± 0.45	7.66 ± 1.58	-7.8 ± 2.2
	LSTM	7147	0.54 ± 0.07	7.06 ± 0.89	10.92 ± 2.50	-33.1 ± 3.4
seed 8675309	RF	7147	0.77 ± 0.02	4.51 ± 0.44	7.67 ± 1.58	-7.7 ± 2.1
	LSTM	7147	0.57 ± 0.07	6.74 ± 0.87	10.57 ± 2.72	-31.3 ± 3.0
default (RF basis)	RF	7147	0.77 ± 0.02	4.51 ± 0.44	7.65 ± 1.57	-7.8 ± 2.1
short	RF	7147	0.77 ± 0.02	4.57 ± 0.46	7.70 ± 1.56	-7.5 ± 2.1
medium	RF	7147	0.76 ± 0.02	4.70 ± 0.46	7.92 ± 1.57	-7.1 ± 2.4
nested	RF	7147	0.77 ± 0.02	4.52 ± 0.46	7.66 ± 1.57	-7.6 ± 2.1
non overlap	RF	7147	0.76 ± 0.02	4.70 ± 0.41	7.91 ± 1.56	-6.9 ± 2.3
precip-only	RF	7147	0.77 ± 0.02	4.59 ± 0.44	7.70 ± 1.46	-7.4 ± 2.1
reduced	RF	7147	0.76 ± 0.02	4.69 ± 0.46	7.90 ± 1.57	-7.1 ± 2.4
$L = 14$	LSTM	7147	0.52 ± 0.06	7.37 ± 0.92	11.15 ± 2.41	-35.0 ± 3.0
$L = 60$	LSTM	7147	0.53 ± 0.06	7.04 ± 0.75	11.06 ± 2.52	-32.9 ± 1.2

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table A9. Within-catchment performance on held-out flood magnitude events (catchments with ≥ 10 events). R^2 and PBIAS are given as median [interquartile range]. $R^2 < 0$ is the percentage of catchments with negative R^2. Temporal and five-fold evaluation settings retain 392 eligible catchments, while the single spatiotemporal split retains 77 catchments, after requiring at least ten held-out flood magnitude events per catchment.

Setting	Model	R^2 median [IQR]	$R^2 < 0$ (%)	PBIAS median [IQR]
Temporal	RF	0.30 [0.04, 0.49]	22	−5 [−12, 2]
	LSTM	−0.39 [−1.40, 0.14]	67	−29 [−36, −21]
Spatiotemporal	RF	0.13 [−0.77, 0.39]	42	−10 [−20, 7]
	LSTM	−0.66 [−2.15, 0.05]	71	−32 [−42, −20]
5-fold	RF	0.13 [−0.54, 0.39]	44	−7 [−18, 7]
	LSTM	−0.73 [−2.11, −0.06]	78	−30 [−41, −20]
−precipitation	RF	−0.30 [−0.94, −0.10]	93	−4 [−19, 11]
	LSTM	−7.82 [−11.98, −4.87]	100	−84 [−88, −80]
−temperature	RF	0.10 [−0.55, 0.38]	45	−6 [−17, 7]
	LSTM	−0.94 [−2.19, −0.08]	79	−31 [−43, −20]
−PET	RF	0.11 [−0.57, 0.39]	43	−7 [−18, 7]
	LSTM	−0.62 [−2.06, 0.04]	73	−30 [−42, −20]
−static attributes	RF	−0.23 [−1.59, 0.30]	61	−6 [−19, 16]
	LSTM	−1.82 [−4.04, −0.71]	91	−44 [−55, −29]

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table A10. Within-catchment performance for the perturbation runs, all within the five-fold evaluation setting (catchments with ≥ 10 events). R^2 and PBIAS are median [IQR]. $R^2 < 0$ is the percentage of catchments with negative R^2. "default" is the reference RF predictor basis (identical to the 5-fold setting in Table A9).

Run	Model	R^2 median [IQR]	$R^2 < 0$ (%)	PBIAS median [IQR]
Training seeds				
seed 42	RF	0.11 [−0.54, 0.40]	45	−7 [−18, 7]
	LSTM	−0.79 [−2.31, −0.09]	80	−33 [−43, −22]
seed 123	RF	0.12 [−0.55, 0.39]	44	−7 [−18, 7]
	LSTM	−0.73 [−2.18, −0.04]	77	−31 [−43, −20]
seed 2024	RF	0.13 [−0.54, 0.39]	44	−7 [−18, 7]
	LSTM	−0.81 [−2.29, −0.05]	77	−32 [−42, −20]
seed 8675309	RF	0.12 [−0.54, 0.39]	43	−7 [−18, 7]
	LSTM	−0.69 [−2.09, −0.03]	76	−31 [−41, −19]
RF predictor bases				
default	RF	0.13 [−0.54, 0.39]	44	−7 [−18, 7]
short	RF	0.05 [−0.60, 0.35]	48	−6 [−17, 8]
medium	RF	0.01 [−0.60, 0.26]	50	−7 [−18, 7]
nested	RF	0.13 [−0.55, 0.38]	43	−7 [−18, 7]
non overlap	RF	0.02 [−0.63, 0.28]	49	−6 [−18, 8]
precip-only	RF	0.04 [−0.57, 0.34]	47	−6 [−18, 6]
reduced	RF	0.02 [−0.60, 0.26]	48	−7 [−17, 8]
LSTM sequence lengths				
L = 14	LSTM	−1.02 [−2.51, −0.15]	82	−35 [−45, −21]
L = 60	LSTM	−0.73 [−2.03, −0.03]	76	−31 [−41, −19]

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table A11. Paired comparison of within-catchment R^2 between the RF and LSTM (Wilcoxon signed-rank test). "RF > LSTM" is the fraction of paired catchments in which the RF has the higher R^2. p is Holm-adjusted across the 11 comparisons. Catchments with fewer than ten held-out events are excluded.

Setting	n	W	RF > LSTM	p
Temporal	392	5512	0.89	< 0.001
Spatiotemporal	77	323	0.87	< 0.001
5-fold	392	12924	0.81	< 0.001
−precipitation	392	2709	0.96	< 0.001
−temperature	392	11137	0.85	< 0.001
−PET	392	13512	0.82	< 0.001
−static attributes	392	19074	0.80	< 0.001
seed 42	392	12526	0.84	< 0.001
seed 123	392	13404	0.82	< 0.001
seed 2024	392	12640	0.82	< 0.001
seed 8675309	392	13640	0.79	< 0.001

Appendix B: Supplementary figures

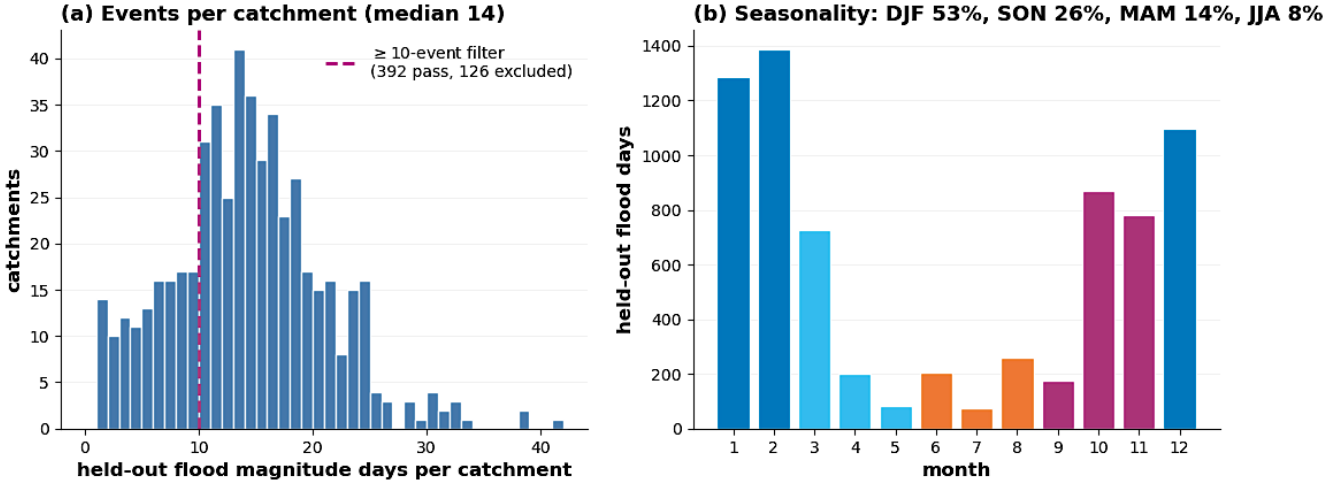

Figure B1. Within-catchment retained flood magnitude events in the evaluation periods In the evaluation period for each model, catchments with at least 10 events were retained for within-catchment predictive skill computations. This resulted in 392 catchments from the 518. The climatology of these flood magnitude events, including counts, frequency, and seasonal distribution are displayed.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

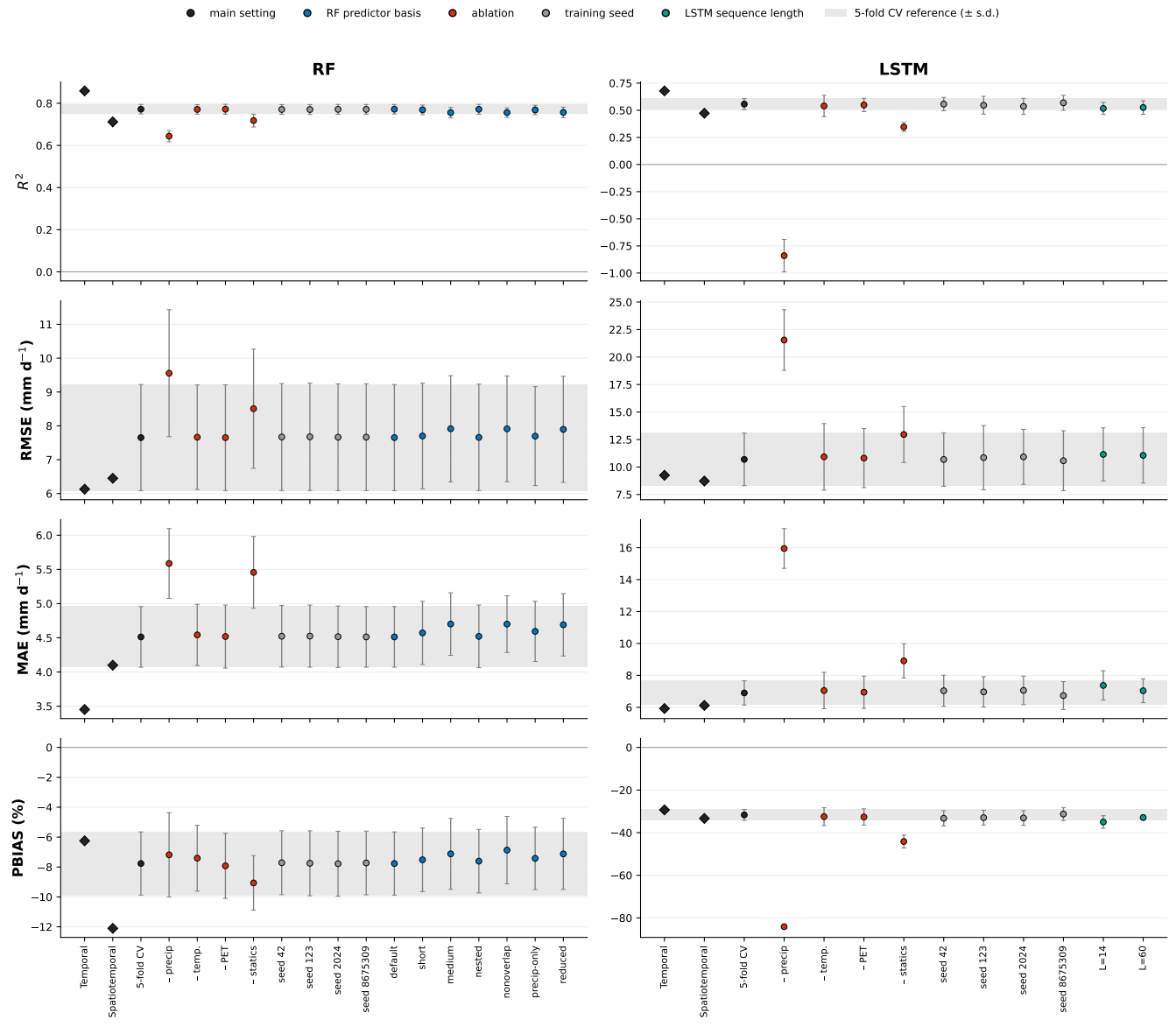

Figure B2. Pooled performance across all 139 fitted model runs. Each marker is the pooled (inter-catchment) metric across five folds; RF (left) and LSTM (right), with rows showing R^2, RMSE, MAE, and PBIAS. Runs are ordered along the x-axis by experiment, the three evaluation settings, the four five-fold ablations, four training seeds, and the RF predictor bases (RF) or sequence lengths $L = 14, 60$ (LSTM). Five-fold runs show the fold mean with error bars spanning ± 1 standard deviation across folds. The temporal and single-split settings are single model fits (diamonds). The gray band marks the 5-fold reference run (mean $\pm$ s.d.), and point color denotes experiment type. Training-seed, predictor-basis, and sequence-length perturbations broadly fall within the reference band, whereas precipitation and static attribute removal move metrics beyond it.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

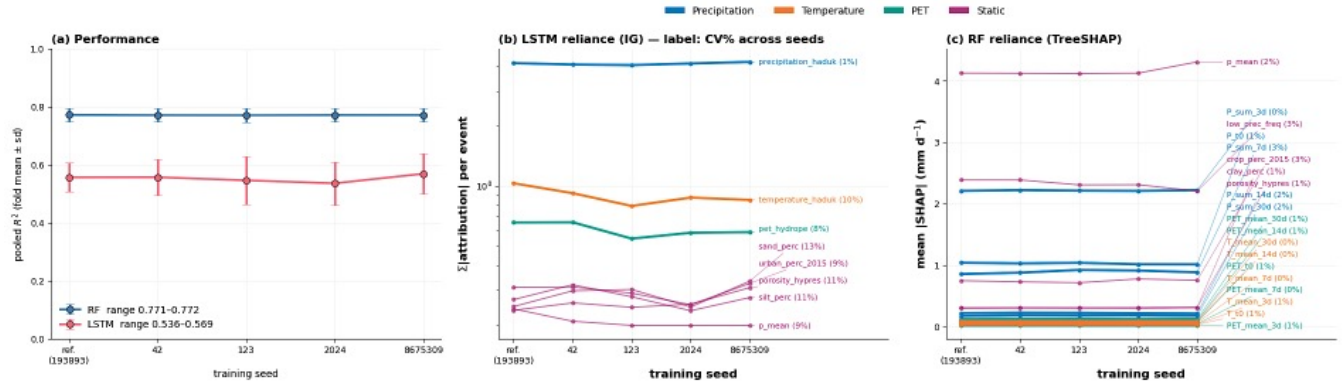

Figure B3. Sensitivity of predictive performance and feature importance to training seed. Results are shown for the five-fold cross-validated experiments for the reference seed and four additional training seeds. (a) Pooled R^2 for the RF and LSTM, where points show the mean across the five held-out folds and error bars show ± 1 s.d. (b) LSTM predictor-level feature importance quantified using IG (mean absolute attribution per held-out flood magnitude event). (c) RF predictor-level feature importance quantified using TreeSHAP (mean absolute SHAP value per held-out flood magnitude event, mm d^{-1}). Labels report the coefficient of variation (CV) across training seeds.

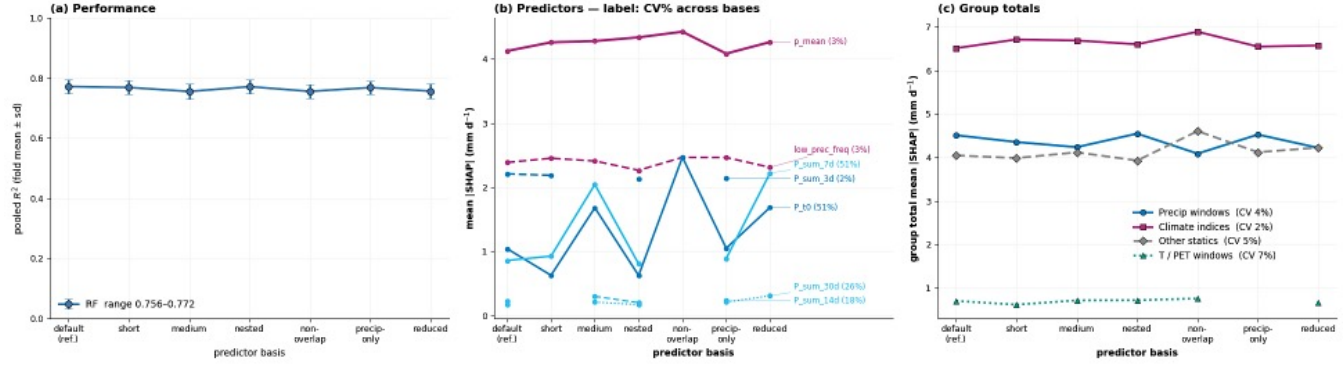

Figure B4. Sensitivity of RF predictive performance and feature importance to the representation of antecedent predictors. Results are shown for the seven engineered predictor bases evaluated within the five-fold BFI-stratified framework. (a) Pooled R^2; points show the mean across the five held-out folds and error bars show ± 1 s.d. (b) Predictor-level feature importance quantified using TreeSHAP (mean absolute SHAP value per held-out flood-magnitude event, mm d^{-1}). Missing points indicate predictors not included in a given predictor basis; labels report the coefficient of variation (CV) across bases. (c) Absolute SHAP attribution aggregated by predictor group, showing precipitation windows, long-term climate attributes, other static attributes, and temperature/PET windows. Labels report the CV across predictor bases.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

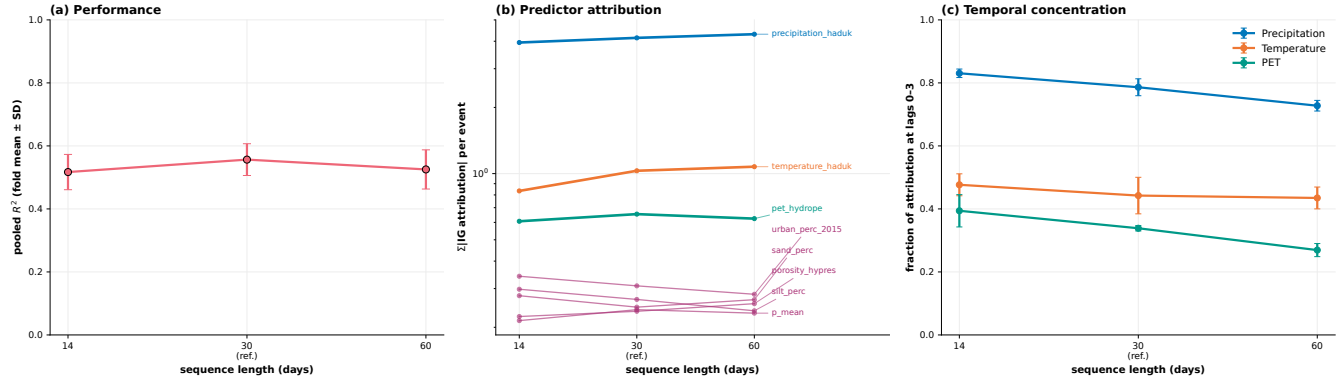

Figure B5. Sensitivity of LSTM predictive performance and feature importance to input sequence length. Results are shown for sequence lengths of 14, 30, and 60 days within the five-fold evaluation setting framework, with the 30-day model representing the reference configuration. (a) Pooled R^2; points show the mean across the five held-out folds and error bars show ± 1 s.d. (b) Predictor-level feature importance quantified using IG (mean absolute attribution per held-out flood magnitude event). (c) Temporal concentration of dynamic-predictor attribution, expressed as the fraction of absolute IG attribution occurring at lags 0-3 days. Points show the mean across the five held-out folds and error bars show ± 1 s.d.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

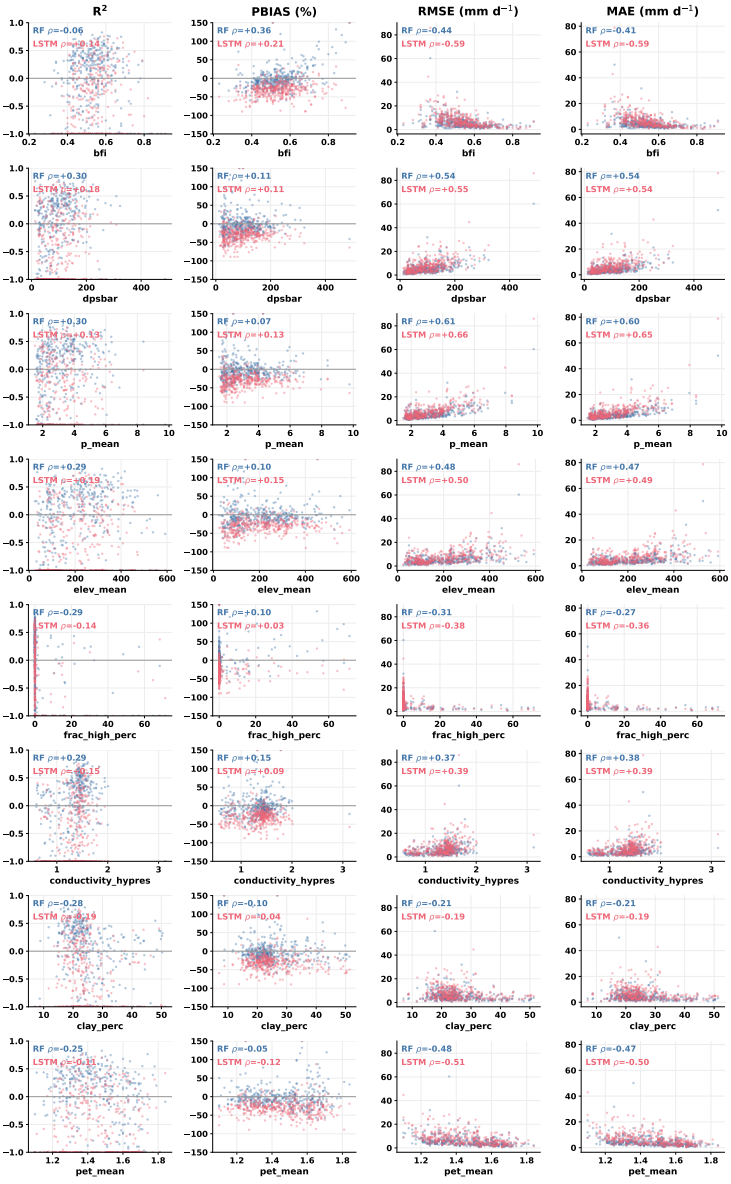

Figure B6. Attribute screen relating within-catchment performance to static catchment attributes under BFI-stratified five-fold evaluation setting. Rows show the eight attributes with the strongest absolute Spearman correlation with catchment R^2 or PBIAS. Columns show R^2, PBIAS, RMSE, and MAE. Each point represents one catchment, where blue denotes the RF and red the LSTM, with Spearman correlations annotated in each panel. Catchments with fewer than ten held-out flood magnitude events are excluded, giving 392 catchments for each model type. R^2 and PBIAS are clipped for display. Correlations of RMSE and MAE with attributes such as mean precipitation and catchment steepness largely reflect the scaling of absolute errors with flood magnitude and were therefore not used for ranking. PBIAS correlations with BFI for the reference models ($\rho = +0.36$ for the RF and $\rho = +0.21$ for the LSTM) are identical to the reference values in Fig. 6b because they represent the same quantity. This screen motivated the targeted analysis in Sect. 3.4.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix C: Methods and calculations

C1 Performance metrics

Here, the performance metric calculations are detailed. In terms of notation, let y_i denote observed flood magnitude, $\hat{y}_i$ predicted flood magnitude, $\bar{y}$ the mean observed flood magnitude over the evaluation subset, and n the number of held-out flood magnitude events. Higher R^2 indicates greater explained variance, lower MAE and RMSE indicate lower error, and PBIAS indicates mean signed bias, with negative values corresponding to underestimation.

The coefficient of determination (R^2), MAE, and RMSE, were calculated following Chicco et al. (2021):

R^2 as,

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{C1}$$

MAE as,

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{C2}$$

and RMSE as,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{C3}$$

Percent bias (PBIAS) was calculated following Towler et al. (2023) as,

$$\mathrm{PBIAS} = 100 \times \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)}{\sum_{i=1}^{n} y_i}. \tag{C4}$$

C2 Feature attribution

Feature attributions were computed at three different levels. (1) local attribution, (2) predictor level attribution magnitude, and (3) process group attribution share. For the LSTM, the temporal structure of the dynamic attributions was also retained to examine antecedent lag contribution decomposition over the sequence length (0-29 days). Unless otherwise stated, absolute attribution magnitudes were not normalized across predictors, and normalization was specifically applied for the process-group share analysis.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

C2.1 RF TreeSHAP local, predictor-level, and process-group comparison

SHapley Additive exPlanations (SHAP) provide local feature attribution values for individual model predictions (Lundberg and Lee, 2017). For the RF model, SHAP values were computed using TreeSHAP. This is designed for tree-based algorithms to enable efficient computation of SHAP values (Lundberg et al., 2020).

For the RF, TreeSHAP was used to obtain a local signed attribution ϕ_{ij} for predictor j and held-out flood-magnitude event i. For a prediction $f(x_i)$, the SHAP decomposition is:

$$f(x_i) = \phi_0 + \sum_{j=1}^{p} \phi_{ij}, \tag{C5}$$

where ϕ_0 is the expected model prediction, ϕ_{ij} is the signed attribution assigned to predictor j for event i, and p is the number of RF predictors provided to the model.

To summarize local attributions at the predictor level, local SHAP values were converted to absolute attribution magnitudes and averaged across the n held-out flood-magnitude events:

$$I_j^{\mathrm{RF}} = \frac{1}{n} \sum_{i=1}^{n} |\phi_{ij}|, \tag{C6}$$

Thus, I_j^{RF} represents the mean absolute SHAP attribution assigned to predictor j. These predictor-level attribution magnitudes were used for the RF predictor-importance analysis and were not normalized to sum to one across predictors.

For the process-group comparison, absolute SHAP attributions were additionally aggregated according to the three groups. Then, the group totals were normalized by their combined attribution magnitude so that the resulting process-group shares summed to one.

C2.2 LSTM IG

For the LSTM, IG were used to quantify feature attribution, a method introduced by Sundararajan et al. (2017) and previously applied to hydrological LSTMs by Kratzert et al. (2019a). IG is a gradient-based attribution method that integrates the model gradient along a linear path from a specified reference baseline and the observed input (Sundararajan et al., 2017). Thus, this method provides model sensitivity along the path.

The LSTM receives both dynamic inputs that vary over the input sequence, and static catchment attributes. For each flood magnitude i, dynamic predictor j, and sequence lag ℓ, IG was calculated following Sundararajan et al. (2017) as:

$$\mathrm{IG}_{ij\ell} = \left(x_{ij\ell} - x'_{ij\ell}\right) \int_0^1 \frac{\partial \hat{y}_i(x' + \alpha\,(x - x'))}{\partial x_{ij\ell}}\, d\alpha, \tag{C7}$$

where x' is the reference baseline, and $\hat{y}_i$ is the corresponding LSTM prediction. Lag $\ell = 0$ denotes the flood magnitude event, with increasing values of ℓ representing the progression of antecedent days.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

The same IG formulation was applied to the static inputs. Here, static predictor is denoted q for event i by x_{iq}:

$$\mathrm{IG}_{iq} = \left(x_{iq} - x'_{iq}\right) \int_0^1 \frac{\partial \hat{y}_i \left(x' + \alpha(x - x')\right)}{\partial x_{iq}} \, d\alpha. \tag{C8}$$

For both dynamic and static predictors, the reference baseline x' was defined as an all-zero input in normalized feature space. Because the predictors were standardized using the training data, zero thus corresponds to the training-period mean of each predictor. The path integral was approximated using a 32-step right-endpoint Riemann sum, following the numerical formulation of Sundararajan et al. (2017). IG satisfies the completeness property described by Sundararajan et al. (2017), whereby the signed attributions over all input dimensions sum to the difference between the model output at the observed input and at the reference baseline, up to numerical integration error.

In this study, the complete input comprises both the dynamic predictor-lag combinations and the static catchment predictors:

$$\sum_j \sum_{\ell=0}^{L-1} \mathrm{IG}_{ij\ell} + \sum_q \mathrm{IG}_{iq} \approx \hat{y}_i(x) - \hat{y}_i(x'), \tag{C9}$$

where the approximation reflects numerical integration of the IG path integral.

C2.3 LSTM predictor-level attribution

For the predictor-level analysis, the signed local IG values were converted to absolute attribution magnitudes. Dynamic and static predictors were aggregated differently because each dynamic predictor is represented at multiple sequence positions, whereas each static predictor is represented by a single input dimension. For each dynamic predictor, an event-level absolute attribution magnitude was calculated by summing across the complete input sequence:

$$A_{ij}^{\mathrm{dyn}} = \sum_{\ell=0}^{L-1} \left|\mathrm{IG}_{ij\ell}\right|, \tag{C10}$$

where L is the sequence length. Predictor-level attribution magnitude was then calculated by averaging across the n held-out flood-magnitude events:

$$I_j^{\mathrm{LSTM,dyn}} = \frac{1}{n} \sum_{i=1}^{n} A_{ij}^{\mathrm{dyn}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=0}^{L-1} \left|\mathrm{IG}_{ij\ell}\right|. \tag{C11}$$

Absolute attributions were summed rather than averaged across sequence positions so that the attribution magnitude assigned to a dynamic predictor represented its attribution across the complete input sequence. Averaging over lags would scale each dynamic predictor by the number of sequence positions, whereas static predictors are represented by a single input dimension. For static predictor q, predictor-level attribution magnitude was calculated directly across events:

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

$$I_q^{\mathrm{LSTM,stat}} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{IG}_{iq}| . \tag{C12}$$

These predictor-level attribution magnitudes were used for the LSTM predictor-level barplots. For the five-fold evaluation setting experiments, attribution magnitudes were calculated separately within each held-out fold and then averaged across the five folds.

C2.4 LSTM temporal attribution

The antecedent lag profile analysis used the same absolute IG values but retained the lag dimension rather than summing across the sequence. For dynamic predictor j, mean absolute attribution at lag ℓ was calculated as:

$$A_j(\ell) = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{IG}_{ij\ell}| . \tag{C13}$$

The resulting lag profiles therefore represent mean absolute IG attribution magnitude at each sequence position, with $\ell = 0$ corresponding to the flood magnitude event. These lag-specific attribution magnitudes were not normalized across predictors or across lags. Again, for the five-fold evaluation setting experiments, lag-specific attribution magnitudes and derived temporal summaries were calculated separately within each held-out fold and subsequently averaged across the five folds.

C2.5 LSTM process-group attribution shares

Normalization was applied specifically for the process-group comparison. For the LSTM, event-day forcing comprised the absolute attribution of precipitation, temperature, and PET at lag $\ell = 0$. Antecedent meteorology comprised the absolute attribution of these dynamic predictors over lags $\ell = 1, \ldots, L-1$, and static catchment information comprised the absolute attributions of the 22 static catchment predictors.

For model m, process group g, and evaluation setting s, let $I_g^{m,s}$ denote the total absolute attribution magnitude assigned to that information group.

The normalized process-group attribution share was then calculated as

$$R_g^{m,s} = \frac{I_g^{m,s}}{\sum_h I_h^{m,s}}, \tag{C14}$$

such that $\sum_g R_g^{m,s} = 1$.

Code availability. The code used to produce the analyses and figures in this study is available at https://github.com/emma-michelle.

Data availability. All datasets used in this study are publicly available through the CAMELS-GB version 2 data release described by Coxon et al. (2026).

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Author contributions.

EF designed the experiments and conducted the analysis under the supervision of HC, MB, and LSl. MB, HC, and LSl revised and edited the manuscript. SM, EL, and LSh provided feedback on the manuscript, including revisions and editing.

Competing interests.

At least one of the (co-)authors is a member of the editorial board of Hydrology and Earth System Sciences.

Acknowledgements.

EF would like to thank the UKRI Natural Environmental Research Council, for funding the work conducted in this paper. The award number is NE/S007474/1. HC was funded through the Leverhulme Trust Research Leadership Award 'SUQCES'. LSl is supported by UKRI (UKRI2054). EF was also funded by the Natural Environment Research Council CANARI project (NE/W004984/1), and the Co-Centre for Climate + Biodiversity + Water programme (NE/Y006496/1) funded by Science Foundation Ireland (SFI), Northern Ireland's Department of Agriculture, Environment and Rural Affairs (DAERA) and UK Research and Innovation (UKRI).

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

References

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., and Wilby, R. L.: Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting, Progress in Physical Geography, 36, 480–513, https://doi.org/10.1177/0309133312444943, 2012.

Bayati, A., Ameli, A. A., and Razavi, S.: Evaluating the Functional Realism of Deep Learning Rainfall-Runoff Models Using Catchment Hydrology Principles, Water Resources Research, 62, e2025WR040 076, https://doi.org/10.1029/2025WR040076, 2026.

Berghuijs, W. R., Woods, R. A., Hutton, C. J., and Sivapalan, M.: Dominant flood generating mechanisms across the United States, Geophysical Research Letters, 43, 4382–4390, https://doi.org/10.1002/2016GL068070, 2016.

Berghuijs, W. R., Harrigan, S., Molnar, P., Slater, L. J., and Kirchner, J. W.: The Relative Importance of Different Flood-Generating Mechanisms Across Europe, Water Resources Research, 55, 4582–4593, https://doi.org/10.1029/2019WR024841, 2019.

Beven, K.: Deep learning, hydrological processes and the uniqueness of place, Hydrological Processes, 34, 3608–3613, https://doi.org/10.1002/hyp.13805, 2020.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of Hydrology, 249, 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., Boháč, M., Bonacci, O., Borga, M., Čanjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Frolova, N., Ganora, D., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Salinas, J. L., Sauquet, E., Šraj, M., Szolgay, J., Volpi, E., Wilson, D., Zaimi, K., and Živković, N.: Changing climate both increases and decreases European river floods, Nature, 573, 108–111, https://doi.org/10.1038/s41586-019-1495-6, 2019.

Breiman, L.: Random forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001a.

Breiman, L.: Statistical Modeling: The Two Cultures, Statistical Science, 16, 199–231, https://doi.org/10.1214/ss/1009213726, 2001b.

Brunner, M. I. and Dougherty, E. M.: Varying Importance of Storm Types and Antecedent Conditions for Local and Regional Floods, Water Resources Research, 58, https://doi.org/10.1029/2022WR033249, 2022.

Brunner, M. I., Slater, L., Tallaksen, L. M., and Clark, M.: Challenges in modeling and predicting floods and droughts: A review, Wiley Interdisciplinary Reviews: Water, 8, https://doi.org/10.1002/wat2.1520, 2021.

Chicco, D., Warrens, M. J., and Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Computer Science, 7, e623, https://doi.org/10.7717/peerj-cs.623, 2021.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M.: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resources Research, 51, 2498–2514, https://doi.org/10.1002/2015WR017198, 2015.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., and Smith, P. J.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resources Research, 51, 5531–5546, https://doi.org/10.1002/2014WR016532, 2015.

Coxon, G., McMillan, H., Bloomfield, J. P., Bolotin, L., Dean, J. F., Kelleher, C., Slater, L., and Zheng, Y.: Wastewater discharges and urban land cover dominate urban hydrology signals across England and Wales, Environmental Research Letters, 19, 084 016, https://doi.org/10.1088/1748-9326/AD5BF2, 2024.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.

Coxon, G., Zheng, Y., Barbedo, R., Cooper, H., Fileni, F., Fowler, H. J., Fry, M., Green, A., Gribbin, T., Harfoot, H., Lewis, E., Gondim, G., Neto, R., Qiu, X., Salwey, S., and Wendt, D. E.: CAMELS-GB v2: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth Syst. Sci. Data, 18, 4345–4371, https://doi.org/10.5194/essd-18-4345-2026, 2026.

Dawson, C. W. and Wilby, R. L.: Hydrological modelling using artificial neural networks, Progress in Physical Geography, 25, 80–108, https://doi.org/10.1177/030913330102500104, 2001.

De la Fuente, L., Ehsani, M., Gupta, H., and Condon, L.: Toward interpretable LSTM-based modeling of hydrological systems, Hydrol. Earth Syst. Sci., 28, 945–971, https://doi.org/10.5194/hess-28-945-2024, 2024.

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., Downer, C. W., Camporese, M., Davison, J. H., Ebel, B., Jones, N., Kim, J., Mascaro, G., Niswonger, R., Restrepo, P., Rigon, R., Shen, C., Sulis, M., and Tarboton, D.: An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, Journal of Hydrology, 537, 45–60, https://doi.org/10.1016/J.JHYDROL.2016.03.026, 2016.

Fisher, A., Rudin, C., and Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously, Journal of machine learning research, 20, 177, https://jmlr.org/papers/v20/18-760.html, 2019.

Ford, E., Brunner, M. I., Christensen, H., and Slater, L.: Interpretable feature incorporation machine-learning framework for flood magnitude estimation, Hydrology and Earth System Sciences, 30, 2135–2160, https://doi.org/10.5194/HESS-30-2135-2026, 2026.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall-runoff predictions of extreme events, Hydrol. Earth Syst. Sci., 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.

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

Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: How limited training data affects streamflow prediction, Environmental Modelling & Software, 135, 104 926, https://doi.org/10.1016/j.envsoft.2020.104926, 2021b.

Griffin, A., Vesuviano, G., Wilson, D., Sefton, C., Turner, S., Armitage, R., and Suman, G.: Putting the English Flooding of 2019–2021 in the Context of Antecedent Conditions, Journal of Flood Risk Management, 18, e70 016, https://doi.org/10.1111/JFR3.70016, 2025.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Heudorfer, B., Gupta, H. V., and Loritz, R.: Are Deep Learning Models in Hydrology Entity Aware?, Geophysical Research Letters, 52, e2024GL113 036, https://doi.org/10.1029/2024GL113036, 2025.

Heudorfer, B., Gupta, H., Dolich, A., and Loritz, R.: Better data or better architecture? Improving deep-learning-based prediction in ungauged basins, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2026-1965, 2026.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/NECO.1997.9.8.1735, 1997.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, https://doi.org/10.1029/2005WR004362, 2006.

Klotz, D., Kratzert, F., Gauch, M., Sampson, A. K., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall-runoff modeling, Hydrol. Earth Syst. Sci., 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.



Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., and Klambauer, G.: NeuralHydrology – Interpreting LSTMs in Hydrology, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11700 LNCS, 347–362, https://doi.org/10.1007/978-3-030-28954-6_19, 2019a.

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

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology — A Python library for Deep Learning research in hydrology, Journal of Open Source Software, 7, 4050, https://doi.org/10.21105/JOSS.04050, 2022.

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

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., Greene, S., Macleod, C. J. A., and Reaney, S. M.: Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain, Hydrology and Earth System Sciences, 23, 4011–4032, https://doi.org/10.5194/hess-23-4011-2019, 2019.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall-runoff models in Great Britain: A comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, Hydrology and Earth System Sciences, 25, 5517–5534, https://doi.org/10.5194/hess-25-5517-2021, 2021.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Sahu, R. K., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci., 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J., Duffy, C., Steinbach, M., and Kumar, V.: Regionalization in a Global Hydrologic Deep Learning Model: From Physical Descriptors to Random Vectors, Water Resources Research, 58, e2021WR031 794, https://doi.org/10.1029/2021WR031794, 2022.

Lundberg, S. and Lee, S.: A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, 30, 4765–4774, https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. I.: From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence, 2, 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

McMillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality, Hydrological Processes, 26, 4078–4111, https://doi.org/10.1002/hyp.9384, 2012.

Nearing, G., Cohen, D., Kratzert, F., Gauch, M., Gilon, O., Ginnetti, J., Hassidim, A., Harrigan, S., Klotz, D., Matias, Y., Metzger, A., Nevo, S., Pappenberger, F., and Prudhomme, C.: Comment on "When are AI models ready for deployment? reassessing Google's global AI flood forecasting system through the lens of responsible modelling", https://doi.org/10.5281/ZENODO.19097819, 2026.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, https://doi.org/10.1029/2020WR028091, 2021.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.



Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, https://jmlr.org/papers/v12/pedregosa11a.html, 2011.

Reed, D. W.: Overview, Volume 1 of the Flood Estimation Handbook, Centre for Ecology and Hydrology, Wallingford, UK, 1999.

Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, 1, 206–215, https://doi.org/10.1038/s42256-019-0048-x, 2019.

Scornet, E. and Hooker, G.: Theory of Random Forests, Annual Review of Statistics and Its Application, 13, 99–121, https://doi.org/10.1146/annurev-statistics-112723-034707, 2026.

Shen, M. and Chui, T. F. M.: Quantifying the relative contributions of different flood generating mechanisms to floods across CONUS, Journal of Hydrology, 626, 130 255, https://doi.org/10.1016/j.jhydrol.2023.130255, 2023.

Slater, L., Coxon, G., Brunner, M., McMillan, H., Yu, L., Zheng, Y., Khouakhi, A., Moulds, S., and Berghuijs, W.: Spatial Sensitivity of River Flooding to Changes in Climate and Land Cover Through Explainable AI, Earth's Future, 12, https://doi.org/10.1029/2023EF004035, 2024.

Slater, L., Blougouras, G., Deng, L., Deng, Q., Ford, E., Hoek van Dijke, A., Huang, F., Jiang, S., Liu, Y., Moulds, S., Schepen, A., Yin, J., and Zhang, B.: Challenges and opportunities of ML and explainable AI in large-sample hydrology, Philosophical Transactions of the Royal Society A, 383, https://doi.org/10.1098/rsta.2024.0287, 2025.

Slater, L. J., Arnal, L., Boucher, M.-A., Y-Y Chang, A., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R. L., Wood, A., and Zappa, M.: Hybrid forecasting: blending climate predictions with AI models, Hydrol. Earth Syst. Sci., 27, 1865–1889, https://doi.org/10.5194/hess-27-1865-2023, 2023.

Stein, L., Clark, M. P., Knoben, W. J., Pianosi, F., and Woods, R. A.: How Do Climate and Catchment Attributes Influence Flood Generating Processes? A Large-Sample Study for 671 Catchments Across the Contiguous USA, Water Resources Research, 57, e2020WR028 300, https://doi.org/10.1029/2020WR028300, 2021.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., and Zeileis, A.: Conditional variable importance for random forests, BMC Bioinformatics, 9, 307, https://doi.org/10.1186/1471-2105-9-307, 2008.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, Proceedings of the 34th International Conference on Machine Learning, 70, 3319–3328, https://proceedings.mlr.press/v70/sundararajan17a.html, 2017.

Tarasova, L., Lun, D., Merz, R., Blöschl, G., Basso, S., Bertola, M., Miniussi, A., Rakovec, O., Samaniego, L., Thober, S., and Kumar, R.: Shifts in flood generation processes exacerbate regional flood anomalies in Europe, Communications Earth & Environment, 4, 49, https://doi.org/10.1038/s43247-023-00714-8, 2023.

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., Viger, R. J., and Zhang, Y.: Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States, Hydrology and Earth System Sciences, 27, 1809–1825, https://doi.org/10.5194/HESS-27-1809-2023, 2023.

Tramblay, Y., Thirel, G., Strohmenger, L., Evin, G., Corre, L., Heraut, L., and Sauquet, E.: Evolution of flood generating processes under climate change in France, Hydrology and Earth System Sciences, 29, 7023–7039, https://doi.org/10.5194/hess-29-7023-2025, 2025.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrology and Earth System Sciences, 5, 13–26, https://doi.org/10.5194/HESS-5-13-2001, 2001.

Wang, Z., Tarasova, L., and Merz, R.: Event-Type-Based Multi-Dimensional Diagnostics of Process Limitations in Hydrological Models, Water Resources Research, 62, e2025WR040 264, https://doi.org/10.1029/2025WR040264, 2026.

https://doi.org/10.5194/egusphere-2026-5890
Preprint. Discussion started: 19 October 2026
© Author(s) 2026. CC BY 4.0 License.



Wu, Y. and Li, N.: Nonlinear control of climate, hydrology, and topography on streamflow response through the use of interpretable machine learning across the contiguous United States, Journal of Water and Climate Change, 14, 4084–4098, https://doi.org/10.2166/WCC.2023.279, 2023.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, W09 417, https://doi.org/10.1029/2007WR006716, 2008.

Yokoo, K., Ishida, K., Ercan, A., Tu, T., Nagasato, T., Kiyama, M., and Amagasaki, M.: Capabilities of deep learning models on learning physical relationships: Case of rainfall-runoff modeling with LSTM, Science of The Total Environment, 802, 149 876, https://doi.org/10.1016/J.SCITOTENV.2021.149876, 2022.

Yu, Q., Jiang, L., Schneider, R., Zheng, Y., and Liu, J.: Deciphering the Mechanism of Better Predictions of Regional LSTM Models in Ungauged Basins, Water Resources Research, 60, e2023WR035 876, https://doi.org/10.1029/2023WR035876, 2024.

Zheng, Y., Coxon, G., Woods, R., Li, J., and Feng, P.: Controls on the Spatial and Temporal Patterns of Rainfall-Runoff Event Characteristics—A Large Sample of Catchments Across Great Britain, Water Resources Research, 59, e2022WR033 226, https://doi.org/10.1029/2022WR033226, 2023.