https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


# Do reservoir-influenced gauges need explicit consideration in machine learning models? A case study with Hydra-LSTM

Karan Ruparell[1,2,3], Dai Yamazaki[3], Kieran M. R. Hunt[1,4], Hannah L. Cloke[1,5], Christel Prudhomme[2], Florian Pappenberger[2], and Matthew Chantry[2]

[1]University of Reading, Department of Meteorology, Reading, UK
[2]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK
[3]Institute of Industrial Science, The University of Tokyo, Tokyo, Japan
[4]National Centre for Atmospheric Science (NCAS), Reading, UK
[5]University of Reading, Department of Geography and Environmental Science, Reading, UK

**Correspondence:** Karan Ruparell (k.ruparell2@pgr.reading.ac.uk)

**Abstract.**

Reservoirs fundamentally alter downstream river flow regimes, decoupling discharge from natural meteorological forcing and challenging standard hydrological prediction. While data-driven models, such as Long Short-Term Memory (LSTM) networks, show promise in regulated catchments, it remains unclear how training data composition across natural and regulated rivers influences model generalisability and behaviour. In this study, we investigate how the presence or absence of reservoir-influenced catchments in training data impacts model performance across different flow regimes and alters the physical drivers the models learn to rely on. Using carefully matched subsets of the CAMELS-GB dataset, we trained separate specialist LSTMs (reservoir and non-reservoir), a pooled Full LSTM, and a multi-headed Hydra-LSTM to investigate whether explicit architectural specialisation offers any advantage over pooled training alone. Models were evaluated on held-out test gauges using standard performance metrics and gradient importance analysis to interpret feature reliance. Our results demonstrate that exposure to reservoir-influenced catchments during training is essential. Models trained exclusively on natural catchments consistently overestimate the mean and variance of regulated flows. Conversely, training exclusively on reservoir-influenced data degrades performance on non reservoir-influenced rivers (KGE reduction of $\geq 0.1$) giving importance primarily to anthropogenic static features, such as abstraction rates, at the expense of precipitation drivers. A single Full LSTM trained on combined data matched the performance of both specialist models in their respective domains, implicitly switching its feature reliance between regimes. The Hydra-LSTM performed comparably to the Full LSTM throughout, indicating that the shared body may act as a regulariser limiting over-specialisation, but that explicit architectural specialisation provides no further benefit under these conditions. We conclude that pooling training data across regimes is a highly effective strategy for general-purpose modelling. However, case studies highlight a fundamental limitation: purely meteorological inputs remain insufficient for predicting flows in heavily managed single-purpose reservoirs, where unobserved human operational decisions dominate the hydrograph.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


## 1 Introduction

Reservoirs in the UK serve a wide range of competing purposes. They are essential for stabilising water availability over extended periods, storing water during floods, managing public water supply, generating hydropower, and supporting irrigation and abstraction (Sardo et al., 2023; Carrillo and Frei, 2009). Reservoir operations are designed to balance these demands while minimising adverse ecological impacts, such as algal blooms from prolonged water residence times and downstream ecosystem disruption from rapid large-scale releases. Because operations are typically governed by explicit management rules and procedures (UK Centre for Ecology & Hydrology, n.d.; US Army Corps of Engineers, 2018; of Civil Engineers, 2015), they impose a structured signature on downstream river flow that can propagate substantially beyond the point of intervention (Salwey et al., 2023; Döll et al., 2009; Tebakari et al., 2012). This can decouple river discharge from natural meteorological forcing, making heavily regulated catchments difficult to predict using standard hydrological approaches.

Data-driven models have shown consistent improvements in reservoir-influenced catchments when given access to relevant inputs. Even simple information on upstream reservoir volumes improves lumped model predictions (Payan et al., 2008), and there is growing evidence that machine learning models can learn implicit representations of regulated flow behaviour from static catchment attributes and meteorological inputs alone (DrivenData, 2024; Yoshimi et al.; Li and Razavi, 2024; Fleming et al., 2021). For prediction at ungauged sites, static attributes—such as the number of upstream reservoirs, average annual abstraction, or degree of regulation—are often the only available source of reservoir information, since dynamic operational data such as storage levels and release schedules are frequently not publicly available.

What has not been systematically tested is how training on reservoir-influenced gauges shapes model behaviour more broadly. If models trained on regulated flow learn representations that are specific to reservoir-influenced catchments, two risks follow. First, a model trained without reservoir-influenced gauges may underperform on reservoir-influenced test sites, even if it otherwise generalises well. Second, a model trained exclusively on reservoir-influenced gauges may degrade on non reservoir-influenced catchments by over-applying reservoir flow patterns where they are not appropriate. Whether pooling reservoir-influenced and non reservoir-influenced gauges into a single training set mitigates both risks, leads to better performance by having a larger training set (Kratzert et al., 2024), or introduces its own trade-offs, is also unknown. Isolating these effects requires controlling for dataset size and catchment characteristics that may be correlated with, but not caused by, the presence of upstream reservoirs.

To address these questions, we construct two subsets of CamelsGB (Coxon et al., 2020b) that are closely matched in catchment area, mean precipitation, soil type, and land cover, but differ in the presence or absence of upstream reservoirs. We train standard LSTM models on each subset separately and on both combined. Alongside these, we train a Hydra-LSTM (Ruparell et al., 2025), which combines a shared encoder trained on all data with a specialist head trained only on reservoir-influenced gauges, to assess whether explicit architectural specialisation is necessary to handle competing flow regimes, or if a single combined model is sufficient. This provides a data-driven analogue to the dedicated reservoir routing modules and conceptual frameworks used in traditional hydrology (McCarthy, 1939; Shaw et al., 2010; Moore and Cole, 2022). We then evaluate each model configuration on held-out gauges not seen during training, using the Nash-Sutcliffe Efficiency, components of the

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


Kling-Gupta Efficiency, and bias. We also use individual case studies representing extremes of reservoir influence to identify some of the limitations of meteorological-only forecasting in highly managed catchments. Finally, we use gradient importance analysis to examining whether each model has learned qualitatively different representations of its training domain.

## 2 Data

In this paper, we aim to study the effects of training models exclusively on reservoir-influenced gauges versus non reservoir-influenced gauges. To isolate the effects of reservoir influence, we first create two similar datasets that differ only in the presence of upstream reservoirs. This ensures we can attribute model differences directly to reservoir training rather than confounding factors like geography. We then partitioned these paired gauges into carefully controlled training, validation, and test sets (summarised in Figure 1). The assignment of gauges to each subset was guided by two core principles: maximising conditional independence between the datasets and ensuring symmetry in the difficulty across evaluation stages.

### 2.1 Data Sources and Preprocessing

We use data from CamelsGB and additional variables provided by the extension dataset CamelsGBV2 (Coxon et al., 2020b, 2025b). The daily reanalysis in both Camels datasets come from a number of sources, such as CHESS (Robinson et al., 2020a, b) , HadUK-Grid (Hollis et al., 2019), and CEH-GEAR (Tanguy et al., 2021), providing the models with close-to-truth information about the meteorological conditions and a range of potential values. While we use historical specific river discharge to evaluate our models, we do not include it as an input, as we want our models to be applicable to other, potentially ungauged, locations. For the same reason, we remove all static variables that were computed using observed river discharge, including variables such as climatological mean discharge and streamflow elasticity. We also remove gauge northing and easting, latitude and longitude, and information about the percentage of available data in each region, so as to prevent the models from learning spurious connections. A complete list of static features included in the models can be found in Appendix A. We normalised all variables to have a mean of 0 and variance of 1.

### 2.2 Pairing Reservoir and non reservoir gauges

A common approach to isolating the effect of reservoirs on river flow is to compare paired catchments (Jumani et al., 2020; Gao et al., 2009; Singer, 2007; Brunner, 2021). Following this method, we began by identifying 153 gauges within CamelsGB that have upstream reservoirs, using data from the UK Reservoir Inventory (Durant and Counsell, 2018) and the Scottish Environmental Protection Agency's publicly available controlled reservoirs register.

Before pairing these with non reservoir-influenced gauges, we reduced the pool of candidate non reservoir-influenced gauges to limit other confounding human impacts. Some gauges without upstream reservoirs nevertheless experienced large flow hold and release from upstream fisheries and similar abstractions. As shown in Figure 1(c), while the vast majority of non reservoir-influenced gauges have total abstraction below 0.1 mm/day, a small subset experiences very high abstraction rates. To limit this influence, we filtered the non reservoir-influenced candidate pool to exclude any gauges with a total water abstraction

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


exceeding $0.2\ mm/day$, a threshold corresponding to the 90th percentile of abstraction across all non reservoir-influenced gauges.

After applying this filter, we paired each of the 153 reservoir-influenced gauges with its nearest non reservoir-influenced neighbour, determined by the Euclidean distance between gauge locations. We do this iteratively, removing gauges that have already been paired from the pool to ensure that pairs are distinct. To ensure similarity in the size of data between the reservoir and non-reservoir datasets, we aligned each pair by restricting the time series to dates where observations were available for both gauges. The resulting dates includes for each gauge pair is shown in Figure 1(b).

### 2.3 Partitioning of Evaluation Sets

To create fair training, validation, and test sets, we have a number of constraints. First, to make sure that our reservoir and non-reservoir evaluation sets are similar, we need to make sure that if we assign a reservoir-influenced gauge to a particular evaluation set (training, validation, or test), its pair is also assigned to the same set.

Secondly, we need to make sure that there are no two gauges in the same catchment that we assign to different evaluation sets, as we don't want our models to have information from training of river flow from a gauge upstream or downstream of a test gauge. This is especially important since, unlike other hydrological studies, we cannot split evaluation sets by date. The record period varies greatly across gauges (Figure 1(b)), making a clean temporal split impractical. Thus, we also put all gauges that are part of the same catchment, those that have overlapping catchment polygons, into the same data set. In some catchments there are over 30 gauges, and this heavily influences our split of train, validation, and testing catchments.

Grouping gauges by catchment and pairing reservoirs and non-reservoirs together results in 61 different clusters over the 306 chosen gauges (153 reservoir-influenced gauges and their 153 non reservoir-influenced partners). We assign these clusters to training, validation, and test sets in order of largest cluster to smallest cluster, at each step assigning a cluster to the evaluation set that is furthest away from reaching its designated total number of gauges, using a 70-15-15 split. We show the resulting grouping in Figure 1(a), showing the catchment boundaries and elevation over the UK.

### 2.4 Verification of Dataset Similarity

To verify similarity between our reservoir and non-reservoir evaluation sets, we compare land cover, soil type, mean precipitation, and catchment area. These variables represent the primary controls on natural flow variability in the absence of human regulation (Addor et al., 2017; Coxon et al., 2020b), so matching them between groups helps ensure that any performance differences between models trained on reservoir and non reservoir gauges reflect the influence of regulation rather than underlying differences in catchment hydrology.

Land cover and soil type govern infiltration rates, evapotranspiration, surface runoff generation, and base-flow recession (Bosch and Hewlett, 1982; Beven, 1983), all of which shape the hydrograph in ways that could confound comparisons between model configurations if they were unevenly distributed across groups. Mean precipitation matters because wetter catchments tend to exhibit higher runoff ratios, flashier responses, and different seasonal dynamics (Beven, 2012). Catchment area controls

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


travel time, flow attenuation, and the relative importance of different runoff-generating processes (Beven, 2012), such that area-mismatched datasets may differ in flow dynamics independently of reservoir influence.

The distribution of dominant land cover type for the reservoir-influenced dataset closely resembles that of the non reservoir-influenced dataset, with the greatest difference being a 5% lower occurrence of crop being the major land cover type for reservoir-influenced catchment compared to non reservoir-influenced catchments. Soil type is very similar between these datasets, with no category differing by more than 1%. The spread between training, validation, and test sets are similarly indistinguishable.

The distributions of catchment area and mean precipitation are shown in Figures 1(d) and 1(e) respectively. By running a Kolmogorov-Smirnov test, we confirm that the distributions of area and precipitation in the test set cannot be statistically distinguished from those of the training and validation sets at a 10% significance level.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026

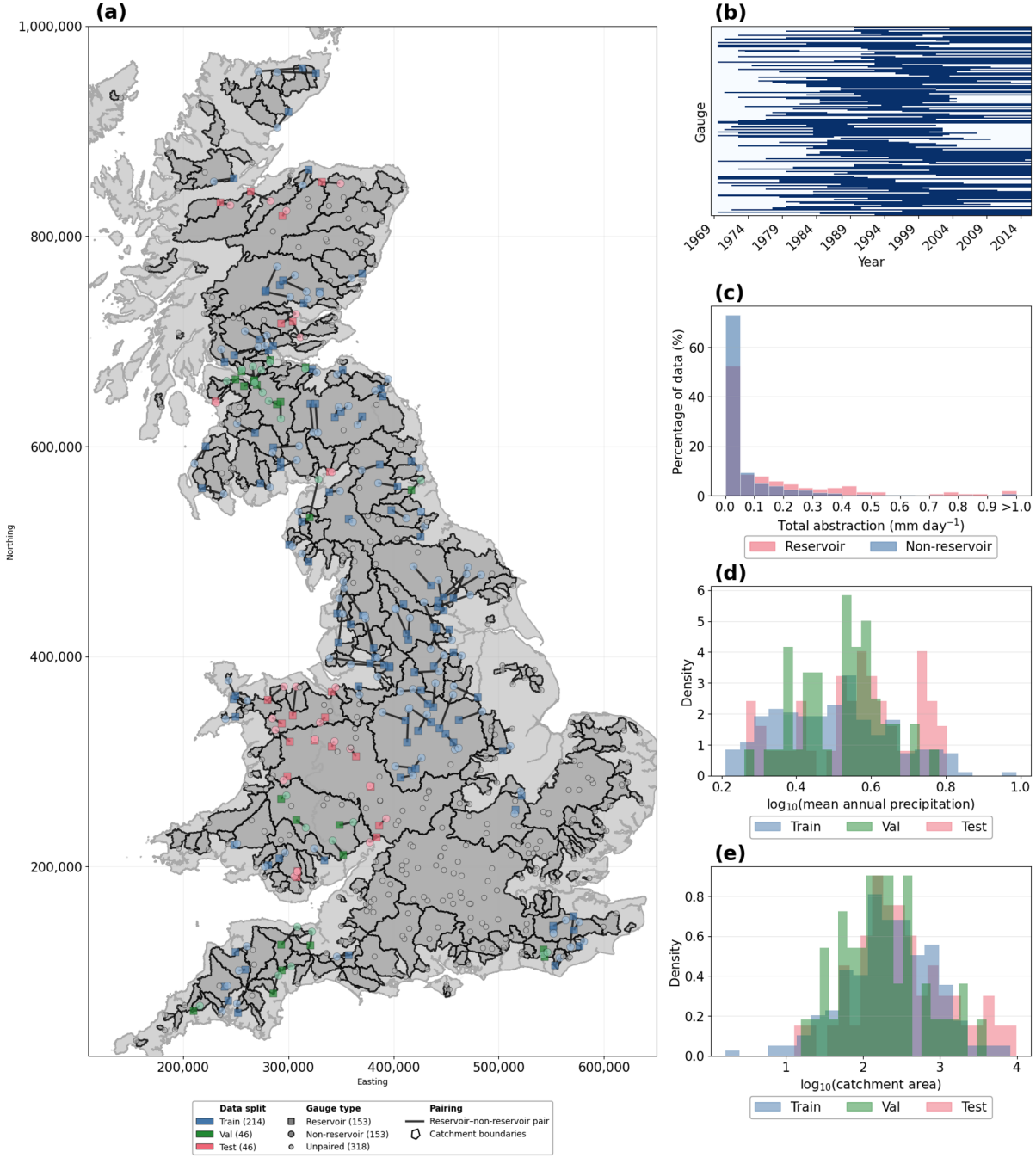




**Figure 1.** Spatial distribution and catchment characteristics of the paired gauge dataset. (a) Location of the reservoir-influenced (squares) and non reservoir-influenced (circles) gauge pairs across Great Britain. Colours denote the evaluation split: training (blue), validation (green), and test (red); solid lines connect each reservoir–non-reservoir pair. Grey circles indicate unpaired non reservoir gauges excluded from the pairing analysis; grey polygons show the corresponding catchment boundaries. (b) Annual data availability for reservoir-influenced gauges over the study period (1969–2015); Blue lines indicate the longest concurrent observation interval per pair, ordered by reservoir ID. (c) Distribution of total annual water abstraction (surface water and groundwater combined) for reservoir-influenced (pink) and non reservoir-influenced (blue) gauges prior to filtering. (d) Mean annual precipitation and (e) upstream catchment area training (pink), validation (blue), and test (green) sets.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


## 3 Methods

### 3.1 Models Trained

To investigate the effect of training machine learning models on reservoir-influenced catchments, we evaluated five distinct architectures built on the Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), summarised in Table 1). LSTMs are a standard benchmark in data-driven hydrology due to their strong performance in river discharge prediction (Hunt et al., 2022; Kratzert et al., 2024). Following the architecture described by (Nearing et al., 2023), all models include a two-layer neural network that projects static catchment attributes into an embedding space before temporal processing.

To isolate the effects of the training data, we defined three standard LSTM models distinguished entirely by their training sets. The Reservoir LSTM and Non-Reservoir LSTM are specialist models trained exclusively on reservoir-influenced gauges and non reservoir-influenced gauges respectively. We hypothesise that each specialist will excel within its own domain but degrade outside it. The Reservoir LSTM should outperform the Non-Reservoir LSTM on reservoir-influenced test gauges, demonstrating that reservoir effects are a learnable signal, while the Non-Reservoir LSTM should outperform the Reservoir LSTM on non reservoir-influenced gauges, demonstrating the risk of over-specialisation. The Full LSTM is a generalist trained on both datasets combined. We hypothesise that pooling will allow the Full LSTM to match the relevant specialist in each domain, since the larger and more varied training set should prevent the over-specialisation of either regime (Kratzert et al., 2024). For the Hydra-LSTM, we hypothesise that the shared Body will act as a regulariser, limiting the over-specialisation that degrades the stand-alone specialist models when applied out-of-distribution. If the Hydra Res-Head outperforms the stand-alone Reservoir LSTM on non reservoir-influenced gauges, that improvement can be attributed to this regularising effect. We further hypothesise that if the Hydra-LSTM Main matches the Full LSTM throughout, explicit architectural specialisation offers no further benefit beyond what pooled training alone achieves, suggesting that the additional complexity of a multi-headed architecture is not warranted. If instead the Full LSTM underperforms the specialists in their respective domains, this would suggest that pooling dilutes regime-specific signals. We also hypothesise that the gradient importance analysis will reflect these performance differences: reservoir-trained models should place greater weight on reservoir-specific static attributes, while models trained without reservoir data should prioritise natural hydrological drivers such as precipitation regime and land cover.

Furthermore, we trained the Hydra-LSTM (Ruparell et al., 2025) to test whether explicit architectural compartmentalisation improves performance. The Hydra-LSTM uses a shared primary encoding Body trained on all available data, but routes final predictions through separate, specialised Heads. We evaluate the outputs of both heads as independent model configurations. The Hydra-LSTM:Main Head is trained on all data to predict both reservoir and non-reservoir flows. Then, the Hydra-LSTM: Res-Head is trained exclusively to predict flow in reservoir-influenced gauges, allowing it to specialise in reservoir-influenced gauges while leveraging the universally trained Body. The Hydra-LSTM: NonRes-Head is trained similarly on exclusively non reservoir-influenced gauges. The weights of the Body are frozen during the training of the Res-Head. This design allows us to test whether performance is improved by combining a universally trained component, the shared Body, with a component strictly specialised for a particular subset of operational conditions, as with the Res-Head.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


**Table 1.** Training data used by each model and model component.

| Model | Reservoir Dataset | Non-Reservoir Dataset |
|---|---|---|
| Reservoir LSTM | ✓ | × |
| Non-Reservoir LSTM | × | ✓ |
| Full LSTM | ✓ | ✓ |
| Hydra-LSTM: Main | ✓ | ✓ |
| Hydra-LSTM: Res-Head | ✓ | Only in the Hydra Body |

### 3.2 Training Methods

All models are trained using a sequence-to-sequence approach over a 90-day window, where the loss is computed across all timesteps in the sequence. During evaluation and operational inference, however, the models operate as sequence-to-one predictors: a 90-day historical lookback is used to forecast only the target day's river discharge. This distinction means the models are exposed to more gradient signal during training, while inference remains computationally efficient and directly comparable across gauges with differing record lengths.

The loss function used during training is the Adapted-NSE (Kratzert et al., 2019), which modifies the standard Nash-Sutcliffe Efficiency by adding a small constant, $\epsilon = 0.1$, to the variance of observed flow in the denominator. This prevents division by zero during low-flow periods and reduces the loss function's sensitivity to errors at very low discharges.

All models are optimised using the Adam optimiser (Kingma and Ba, 2014) with a batch size of 256 and a learning rate that decays by a factor of $\gamma$ after each epoch. To determine the number of training epochs, we use early stopping with a patience of one, ending training when validation loss fails to improve. Hyperparameters are selected via a grid search over the values listed in Table 2, with the configuration yielding the lowest validation loss chosen for each model. The searched parameters include the learning rate, $\gamma$, the hidden and output sizes of the static embedding network, and the hidden size and number of layers of the LSTM.

**Table 2.** Hyperparameter configurations for trained models. Bold values indicate the selected hyperparameters for each model.

| Model | Learning Rate | Gamma | Embedding Hidden Size | Embedding Output Size | LSTM Hidden Size | LSTM Number of Layers |
|---|---|---|---|---|---|---|
| **Reservoir LSTM** | [**1e-3**, 1e-4] | [0.8, **0.95**] | [**64**, 128] | [**16**, 32] | [**128**, 256] | [**1**,2] |
| **Non-Reservoir LSTM** | [**1e-3**, 1e-4] | [0.8, **0.95**] | [64, **128**] | [**16**, 32] | [128, **256**] | [1, **2**] |
| **Full LSTM** | [**1e-3**, 1e-4] | [**0.8**, 0.95] | [**64**, 128] | [**16**, 32] | [128, **256**] | [**1**,2] |
| **Hydra-LSTM** | [**1e-3**, 1e-4] | [**0.8**, 0.95] | **Body:** [64, **128**] | [**16**, 32] | **Body:** [128, **256**] | [**1**,2] |
| *Additional Hydra-LSTM parameters: Body Output Size = [**32**, 64], Head Hidden Size was fixed as that of the Body* | | | | | | |

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


### 3.3 Evaluation Metrics and Explainability

We evaluate model performance using two complementary metrics: the Nash-Sutcliffe Efficiency (NSE) and the Kling-Gupta Efficiency (KGE). The NSE (Nash and Sutcliffe, 1970) is a skill score based on the mean squared error, and is widely used in hydrology as a benchmark against a mean-flow baseline. However, the NSE is known to be sensitive to peak flows and can mask systematic biases in mean flow or flow variability (Gupta et al., 2009). We therefore also report the KGE (Gupta et al., 2009), as it decomposes model error into three interpretable components: the correlation between simulated and observed flow ($r$), the ratio of simulated to observed standard deviation ($\alpha$, capturing variability), and the ratio of simulated to observed mean flow ($\beta$, capturing bias). We report all three components alongside KGE and NSE in our results, as they allow us to diagnose the nature of model errors rather than simply their magnitude. A model with a high KGE but a large positive $\beta$, for example, is systematically overestimating mean flow, which has a different physical interpretation than a model that captures the mean well but misrepresents variability. We also provide absolute bias as a metric, as a familiar metric, and as it is less influenced by low-flow than $\beta$. In addition to evaluating models separately on reservoir-influenced and non reservoir-influenced gauges, we report combined performance across all test gauges, comparing routed specialist strategies, where each specialist is applied only in its own domain, against single generalised models applied uniformly.

To complement the aggregate performance analysis, we examine two individual reservoir-influenced gauges as case studies: Vyrnwy and Gwyfrai at Bontnewydd. These were selected because they represent contrasting levels of regulation intensity, allowing us to identify failure modes in each regime that aggregate metrics such as KGE cannot fully capture. The most direct measure of regulation intensity is reservoir storage capacity relative to mean annual catchment inflow. At Vyrnwy, reservoir capacity is approximately 40% of mean annual precipitation, meaning the reservoir can retain a substantial fraction of incoming water and exert strong control over release timing. At Gwyfrai at Bontnewydd, this figure is only 2%, suggesting a much weaker buffering effect. This contrast is reinforced by the proportion of each catchment drained through a reservoir: 100% at Vyrnwy compared to less than 50% at Gwyfrai at Bontnewydd (Coxon et al., 2025b). Together, these characteristics suggest that Vyrnwy will exhibit far greater flow attenuation and meteorological decoupling than Gwyfrai at Bontnewydd, making them a natural pairing for exploring how the different model configurations respond across a spectrum of reservoir influence.

To understand which input features drive model predictions, we use Integrated Gradients (Sundararajan et al., 2017). Integrated Gradients assigns an importance score to each input feature, by averaging gradients along a straight path from a baseline input to the true input. This averaging makes attributions more stable than single-point gradient methods. We use the zero vector as our baseline, which corresponds to the dataset mean for each feature since all static variables are normalised to zero mean prior to input. This choice means attributions reflect how much each feature's deviation from its typical value shifted the model's prediction.

We apply Integrated Gradients to the static catchment, to see how the importance of static variables, especially reservoir-related variables, changes depending on whether the model is trained on reservoir-influenced flow, and whether the gauge

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


being evaluated is itself reservoir-influenced. We report the top ten features per model for each gauge type in Tables 5a and 5b, allowing direct comparison of which static variables each model prioritises across the two settings.

Formally, the integrated gradient for $i$-th input dimension is:

$$\mathrm{IG}_i(x) = \underbrace{(x_i - x_i')}_{\substack{\text{deviation of feature} \\ \text{from its mean value}}} \times \underbrace{\int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha}_{\substack{\text{prediction sensitivity to that feature,} \\ \text{averaged between mean and true value}}} \tag{1}$$

where $F$ is the model output and $\alpha$ interpolates between the baseline $x'$ and the input $x$. In practice, the integral is discretised and approximated by a Riemann sum over a fixed number of steps.

## 4 Results

Before examining per-regime performance, we note that spatial analysis of model errors revealed no systematic geographic pattern. Prediction difficulty is governed by characteristics of individual gauges rather than regional factors. We therefore focus our analysis on per-regime aggregate performance before examining individual gauges in detail.

It is worth stating the expected behaviour of each model. The Reservoir LSTM and Non-Reservoir LSTM are specialists, each trained on a single flow regime, so we expect strong performance within that regime but degraded performance outside it, since each model has never been exposed to the dynamics it is asked to generalise to. Conversely, since the Full LSTM is trained on both regimes simultaneously, the larger and more varied training set should prevent the over-specialisation of either specialist, though by training on both regimes we may dilute regime-specific signals, weakening performance. The Hydra-LSTM occupies a middle position. Its shared body is trained on all data like the Full LSTM, and an additional reservoir head is fine-tuned exclusively on reservoir-influenced gauges. It is included to test whether explicit architectural specialisation adds anything beyond what the Full LSTM already achieves through pooled training alone. If the Hydra reservoir head outperforms the stand-alone Reservoir LSTM on non reservoir-influenced gauges, that improvement can be attributed to the regularising effect of the shared body. If it matches the Full LSTM throughout, explicit specialisation offers no further benefit.

The Reservoir LSTM, Full LSTM, and Hydra-LSTM perform similarly to one another on reservoir-influenced gauges, and both substantially outperform the Non-Reservoir LSTM, as shown in Table 3 and Figure 2. We run a Kolmogorov-Smirnov test at a 5% significance level to compare model performance. The results reveal that the distribution of the Non-Reservoir LSTM is significantly different from the other models. The other models are statistically indistinguishable from one another under the Kolmogorov-Smirnov test, forming a performance tier above the Non-Reservoir LSTM. The Non-Reservoir LSTM has a KGE of 0.69 compared to 0.81 for the Reservoir LSTM, with the gap most pronounced in the hardest to model catchments at the lower tail of the cumulative distribution function. Even in reservoirs with extreme human control, like Vyrnwy at Vyrnwy, the models trained on reservoir-influenced gauges show less over-prediction than the Non-Reservoir LSTM. This reduction in the lowest KGE is the greatest difference between models. This is mainly due to a systematic positive bias of $0.16 mm/day$ in the

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


Non-Reservoir LSTM, reflecting a tendency to overestimate mean flow in reservoir-influenced catchments. The Hydra-LSTM Main head matches the Reservoir LSTM closely (NSE 0.73, KGE 0.80), suggesting the shared body encodes reservoir-relevant dynamics even without the specialist head.

On non reservoir-influenced gauges (Table 4 and the right plot of Figure 2), the pattern reverses. Again, we apply a Kolmogorov-Smirnov test. We find that at a 5% significance level the cumulative distribution function of the Reservoir LSTM is different to that of the other models, and at the same significance level the other models are indistinguishable. The Reservoir LSTM degrades to an NSE of 0.58 and KGE of 0.63, introducing a negative bias of $-12mm/day$, while the Non-Reservoir LSTM and Full LSTM both achieve an NSE above 0.72. The Hydra-LSTM Reservoir Head (KGE 0.77) also outperforms the stand-alone Reservoir LSTM (0.63) on these gauges, indicating that the shared body acts as a regulariser that limits over-specialisation. The CDF in Figure 2 shows the Reservoir LSTM lagging across the majority of the distribution, not only at the extremes. A similar pattern is true for the Non-Reservoir LSTM in reservoir-influenced gauges. The aggregate positive bias of $0.16mm/day$ across reservoir-influenced gauges (Table 3) masks substantial variation in modelling difficulty across individual catchments. To examine this, we turn to two contrasting reservoir-influenced gauges that illustrate the boundaries of what meteorological inputs alone can recover.

**Table 3.** Median predictive performance of models on reservoir-influenced gauges. To evaluate seasonal consistency, all metrics were computed over a 90-day rolling window rather than across the entire test period. Within the Kling-Gupta Efficiency (KGE), $\alpha$ represents the ratio of predicted to observed standard deviation, and $\beta$ represents the ratio of predicted to observed mean flow. These capture flow variability and volume bias respectively.

| Model | NSE | KGE | $\alpha$ | $\beta$ | Correlation | Bias($mm/day$) |
|---|---|---|---|---|---|---|
| Reservoir LSTM | **0.74** | **0.81** | 0.94 | 1.02 | 0.92 | 0.04 |
| Hydra-LSTM: Res-Head | 0.73 | 0.80 | **0.98** | 1.01 | **0.93** | 0.03 |
| Non-Reservoir LSTM | 0.57 | 0.69 | 1.06 | 1.10 | 0.90 | 0.16 |
| Hydra-LSTM: Non Res-Head | 0.72 | 0.80 | 1.04 | **1.00** | **0.93** | **-0.01** |
| Full LSTM | **0.74** | 0.78 | 0.94 | 1.03 | 0.93 | 0.05 |
| Hydra-LSTM: Main | 0.73 | 0.80 | **0.98** | **1.00** | **0.93** | **0.00** |

**Table 4.** Median predictive performance of models on non reservoir-influenced gauges. Metrics were calculated just as in Table 3.

| Model | NSE | KGE | Alpha | Beta | Correlation | Bias ($mm/day$) |
|---|---|---|---|---|---|---|
| Reservoir LSTM | 0.58 | 0.63 | **0.99** | 0.91 | 0.89 | -0.12 |
| Hydra-LSTM: Res-Head | 0.70 | **0.73** | **0.99** | 0.97 | **0.93** | **-0.04** |
| Non-Reservoir LSTM | **0.72** | 0.75 | 0.94 | 0.98 | **0.93** | -0.03 |
| Hydra-LSTM: Non Res-Head | 0.71 | **0.77** | 0.97 | **1.00** | **0.93** | **0.00** |
| Full LSTM | 0.72 | 0.75 | 0.95 | 1.03 | **0.93** | -0.03 |
| Hydra-LSTM: Main | 0.71 | **0.77** | 0.96 | **1.00** | **0.93** | **0.00** |

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


**Table 5.** Median predictive performance across all test gauges (combined reservoir-influenced and non reservoir-influenced). To evaluate the benefit of domain-specific modelling, the Specialist models represent a routed evaluation approach: the Reservoir LSTM is evaluated solely on reservoir-influenced gauges, while the Non-Reservoir LSTM is evaluated on non reservoir gauges. The Hydra-LSTM: Specialists follows this same routing approach using its respective domain heads. This compares specialized routing strategies against single, generalized models (Hydra-LSTM: Main and Full LSTM) over the entire dataset.

| Model | NSE | KGE | $\alpha$ | $\beta$ | Correlation | Bias($mm/day$) |
|---|---|---|---|---|---|---|
| Specialist LSTMs | **0.73** | **0.78** | 0.94 | **1.00** | 0.93 | **0.00** |
| Hydra-LSTM: Specialists | 0.72 | **0.78** | **0.97** | 1.01 | **0.93** | 0.01 |
| Full LSTM | 0.73 | 0.76 | 0.94 | 1.03 | **0.93** | 0.04 |
| Hydra-LSTM: Main | 0.72 | **0.78** | **0.97** | **1.00** | **0.93** | **0.00** |

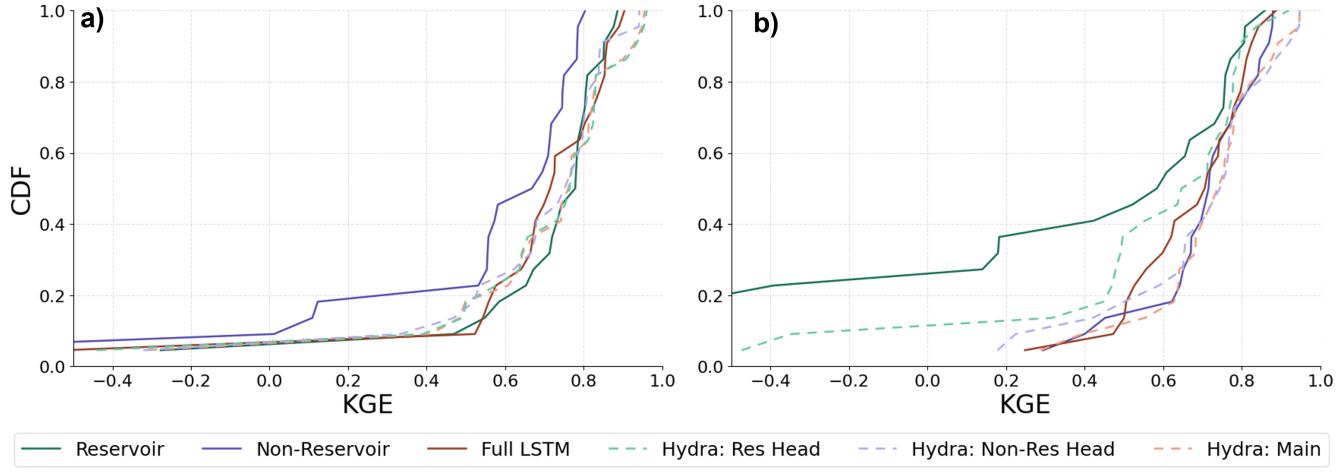



**Figure 2.** Cumulative Distribution Functions (CDF) of Kling-Gupta Efficiency (KGE) scores for reservoir-influenced gauges, a), and non reservoir-influenced gauges, b). Each point in the curves representing the average KGE in a different gauge. In reservoir-influenced gauges, the stand alone Non-Reservoir LSTM consistently underperforms across the majority of the distribution. In non reservoir-influenced gauges, the Reservoir LSTM consistently underperforms. The x-axis has been clipped to -a minimum KGE of -0.5, so as to better see differences between models for greater KGE scores

### 4.1 Case Studies: Vyrnwy reservoir and Gwyfrai at Bontnewydd

Vyrnwy reservoir and Gwyfrai at Bontnewydd illustrate the difference in degree of influence within the reservoir-influenced category, and are discussed together because they represent contrasting conditions.

At Vyrnwy Reservoir, all models fail substantially, returning highly negative 90-day NSE scores (ranging from $-4.4$ to $-15.7$). As shown in Figure 3, the observed hydrograph has near-zero discharge for the majority of the year, with sharp, short flow releases. All models overestimate flow during the low-flow periods, which is consistent with the positive bias across

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


reservoir-influenced gauges (Table 3), though here the effect is extreme. The Reservoir LSTM, Full LSTM, and Hydra-LSTM show lower over-prediction than the Non-Reservoir LSTM, suggesting that even in difficult to model reservoirs, reservoir-trained models have learned some signal. Despite this, no model comes close to capturing the release timing. Vyrnwy is a single, heavily managed reservoir serving multiple competing operational purposes across a 74 $km^2$ catchment, such as energy production, nature conservation, and water resources (Severn Trent, 2017). Its releases reflect decisions that are not recoverable from atmospheric inputs alone.

Gwyfrai at Bontnewydd is much better captured, despite having three upstream reservoirs. Here, all models capture the timing and shape of discharge events well, with NSE scores predominantly above 0.9, as shown in Figure 4. The exception is the Non-Reservoir LSTM, which correctly identifies flow events but consistently overestimates peak magnitudes and overall variance, as is seen in that model across reservoir-influenced gauges (Table 3). The smoother hydrograph at Gwyfrai looks more similar to natural flow, with fewer sharp transitions. The contrast with Vyrnwy suggests that the "reservoir-influenced" label encompasses a wide range of modelling difficulty, and that highly managed reservoirs may require operational data beyond what is available here.

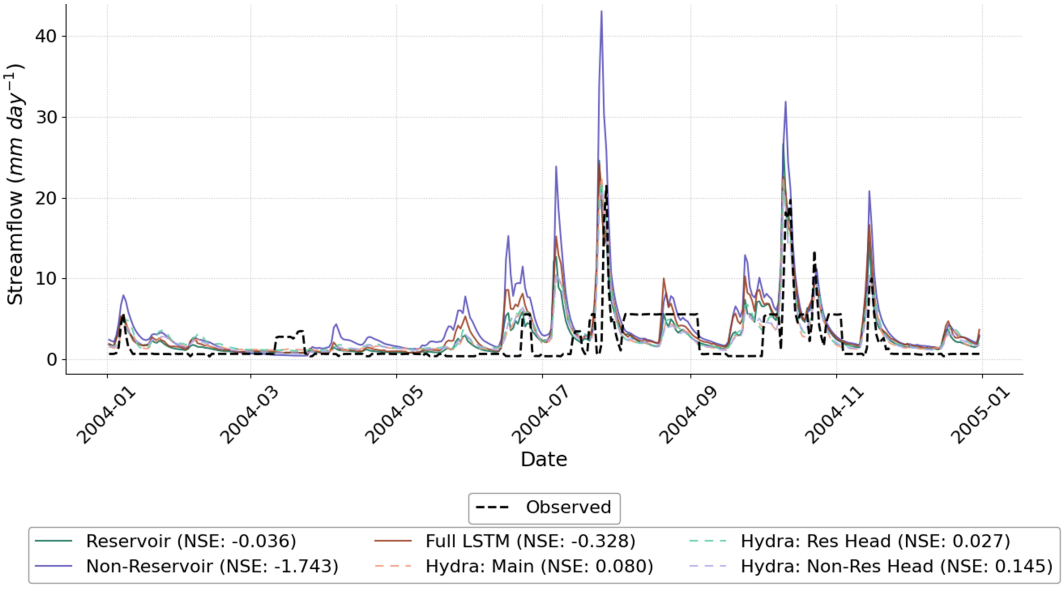
### 4.2 Feature importance analysis via integrated gradients

To better understand what drives model behaviour across the reservoir-influenced and non reservoir-influenced gauges, we examine which static input features each model relies on most heavily. We use integrated gradients to attribute model sensitivity to each input, reporting the ten highest-ranked features per model for both reservoir-influenced and non reservoir-influenced gauges (Tables 5a and 5b). This analysis reveals systematic differences in feature prioritisation, that are consistent with the performance patterns described above.

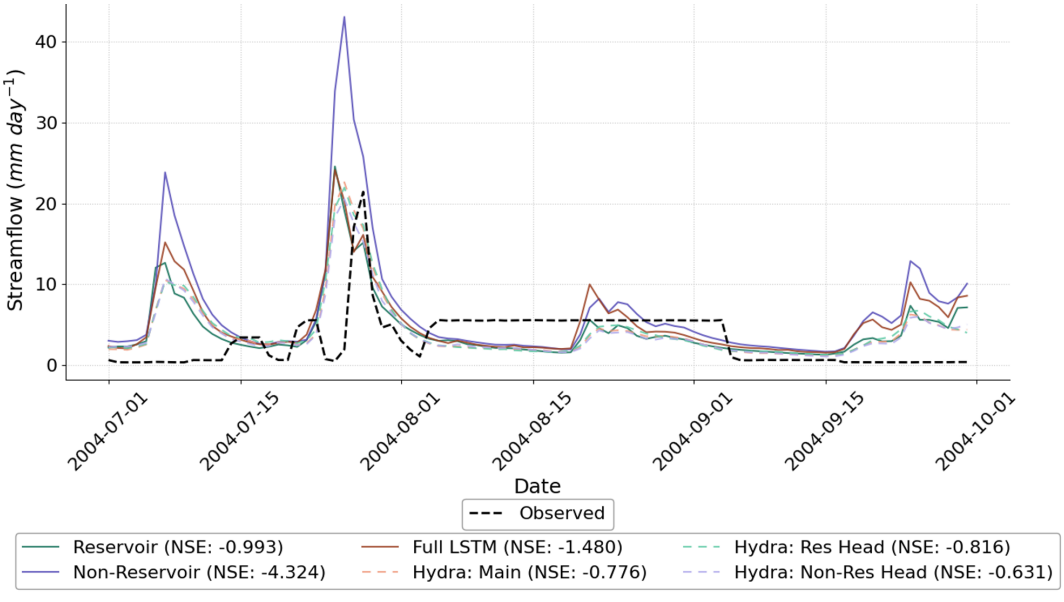
Across reservoir-influenced gauges (Table 5b), the percentage of the upstream catchment area that is inland water ranks first for every model. Beyond this shared signal, reservoir-trained models diverge from the Non-Reservoir LSTM in informative ways. The Reservoir LSTM, Full LSTM, and all Hydra models place reservoir contributing area within their top three features on reservoir gauges. In contrast, the Non-Reservoir LSTM does not have Reservoir contributing area in its top ten static variables at all, and is the only one of the models using soil attributes in its top 10 features. In fact, three of the top ten features used by the Non-Reservoir LSTM are soil attributes. The Non-Res Hydra Head is much more similar to the other Hydra models than the Non-Reservoir LSTM, with the same top 10 features as in the Hydra-Main, with a slight difference in the ordering of those features.

On non reservoir-influenced gauges (Figure 5b) the pattern reverses. The Non-Reservoir LSTM, Full LSTM, and Hydra models are all most sensitive to precipitation regime, with all models but the Reservoir LSTM having mean precipitation as its most important static feature. The Reservoir LSTM is also the only model that does not have precipitation seasonality or root depth as one of its top 10 features, and unlike the other models includes soil depth pelletier. This suggests the Reservoir LSTM is missing the importance of precipitation and other natural features, which is consistent with the degraded performance and negative bias it shows on non reservoir-influenced gauges.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026

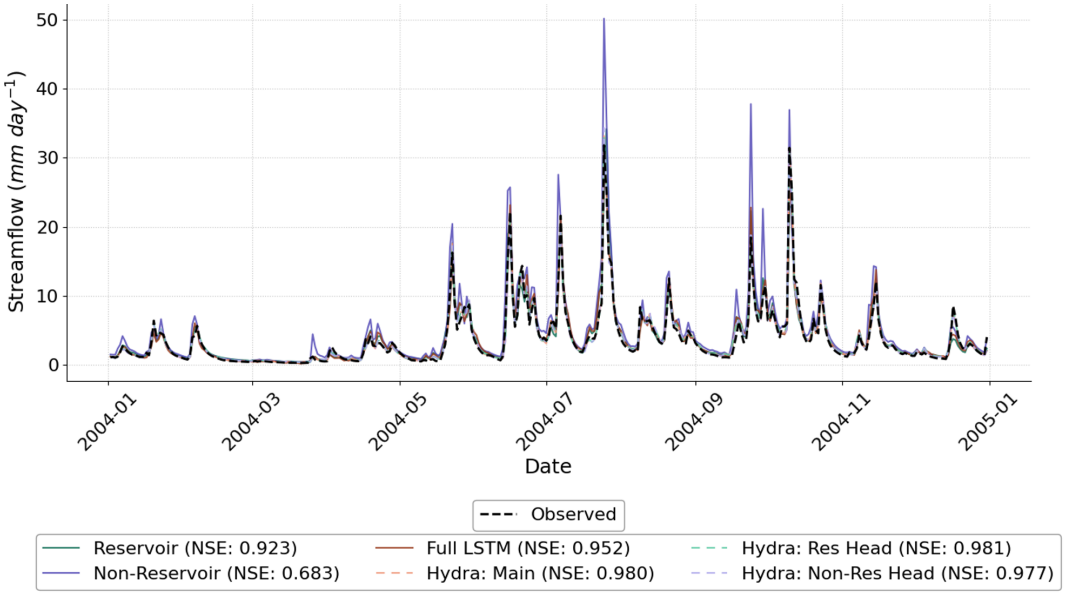

(a) Annual hydrograph (365 days)



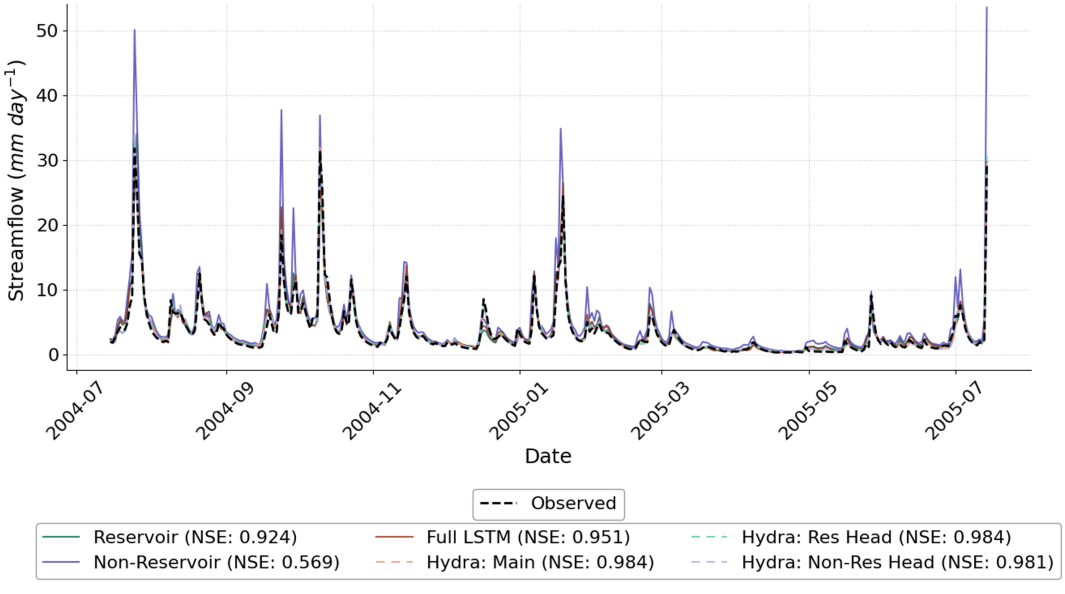

(b) Seasonal hydrograph (92 days)

**Figure 3.** Predicted versus observed river discharge at Vyrnwy Reservoir across annual (a) and seasonal (b) scales. Vyrnwy represents a severe failure mode for all models, with the highest NSE being 0.08. While reservoir-trained models exhibit slightly less over-prediction than the Non-Reservoir LSTM, no model accurately captures the sharp operational releases.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


(a) Annual hydrograph (365 days)




(b) Seasonal hydrograph (92 days)

**Figure 4.** Predicted versus observed river discharge at Gwyfrai at Bontnewydd across annual (a) and seasonal (b) scales. Unlike Vyrnwy, this reservoir-influenced catchment has a smoother, more natural flow regime, allowing most models to achieve NSE scores above 0.90. The Non-Reservoir LSTM is the exception, consistently overestimating peak magnitudes and overall variance.

## 5 Discussion

The results demonstrate that LSTM-based hydrological models can implicitly learn representations of reservoir operations from static catchment attributes and temporal meteorological inputs alone, and that this learning generalizes to unseen gauges.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


| Rank | Reservoir LSTM | Hydra: Res-Head | Non Reservoir LSTM | Hydra: Non Res-Head | Full LSTM | Hydra: Main |
|---|---|---|---|---|---|---|
| 1 | inwater % | inwater % | inwater % | inwater % | inwater % | inwater % |
| 2 | reservoir water resources | reservoir contributing area | precipitation mean | reservoir contributing area | reservoir hydroelectricity | reservoir contributing area |
| 3 | reservoir contributing area | precipitation mean | area | precipitation mean | reservoir contributing area | precipitation mean |
| 4 | frac snow | area | reservoir hydroelectricity | shrub % | precipitation mean | shrub % |
| 5 | bare soil % | reservoir hydroelectricity | saturated hydraulic conductivity 95 | reservoir hydroelectricity | area | area |
| 6 | reservoir hydroelectricity | shrub % | precipitation seasonality | precipitation seasonality | reservoir normalised upstream capacity | reservoir hydroelectricity |
| 7 | high precipitation duration | elevation 10% | total available water content 50% | bare soil % | precipitation seasonality | bare soil % |
| 8 | elevation 10% | bare soil % | bare soil % | aridity | bare soil % | elevation 10% |
| 9 | precipitation seasonality | precipitation seasonality | reservoir navigation | area | reservoir water resources | precipitation seasonality |
| 10 | precipitation mean | reservoir normalised upstream capacity | root depth | elevation 10 | shrub % | aridity |

Human Influence | Climate Indices | Hydrogeology | Soil | Land Cover | Topography

(a) Top 10 static input features ranked by integrated gradient importance for models evaluated on reservoir-influenced gauges. While all models prioritise the percentage of inland water, models exposed to reservoir training data correctly identify reservoir attributes such as reservoir contributing area and the percentage of reservoir storage used for hydroelectricity. The Non-Reservoir LSTM defaults to natural drivers like precipitation and catchment area.

| Rank | Reservoir LSTM | Hydra: Res-Head | Non Reservoir LSTM | Hydra: Non Res-Head | Full LSTM | Hydra: Main |
|---|---|---|---|---|---|---|
| 1 | frac snow | precipitation mean | precipitation mean | precipitation mean | precipitation mean | precipitation mean |
| 2 | aridity | shrub % | no groundwater % | shrub % | deciduous wood % | shrub % |
| 3 | elevation 10 | aridity | precipitation seasonality | dpsbar | precipitation seasonality | dpsbar |
| 4 | precipitation mean | dpsbar | dpsbar | aridity | aridity | aridity |
| 5 | elevation mean | no groundwater % | high precipitation duration | no groundwater | root depth | no groundwater |
| 6 | grass % | elevation 10 | aridity | root depth | low productivity flow through fractures | root depth |
| 7 | soil depth pelletier 50 | root depth | low productivity flow through fractures | precipitation seasonality | grass % | grass % |
| 8 | low productivity flow through fractures | grass % | root depth | grass % | elevation 50 | precipitation seasonality |
| 9 | soil depth pelletier 95 | deciduous wood % | bare soil % | low productivity flow through fractures | dpsbar | elevation 10% |
| 10 | bare soil % | precipitation seasonality | bulk dens | deciduous wood % | inwater % | low productivity flow through fractures |

Human Influence | Climate Indices | Hydrogeology | Soil | Land Cover | Topography

(b) Top 10 static input features ranked by integrated gradient importance for models evaluated on non reservoir-influenced gauges. The Non-Reservoir, Full, and Hydra LSTMs appropriately shift their focus to natural hydrological drivers such as precipitation duration and land cover. In contrast, the Reservoir LSTM incorrectly prioritises human-management proxies (e.g., industrial abstraction).

**Figure 5.** Integrated gradient importance rankings for static input features evaluated on reservoir-influenced and non reservoir-influenced gauges.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


However, our analysis also reveals a trade-off: while specialized training improves performance within a specific flow regime, it introduces a performance cost when the model is applied out-of-distribution to non reservoir-influenced rivers. Together, the results establish that reservoir-influenced flow is a learnable but domain-specific signal, and can be learned as long as there are a sufficient number of reservoir-influenced gauges in the training dataset. They also show that it is not necessary to train the model only on reservoir-influenced flow, LSTMs are capable of performing well in both reservoir-influenced and non reservoir-influenced gauges, as long as both are provided in training.

### 5.1 Learning from reservoir-influenced areas generalises to new reservoir-influenced gauges

Models trained on reservoir-influenced gauges substantially outperform the Non-Reservoir LSTM when applied to held-out reservoir-influenced test catchments, with a KGE gap of $0.12$ compared to the Reservoir LSTM and a persistent positive bias of $0.16 mm/day$ in the Non-Reservoir LSTM (Table 3).

This benefit is greatest in the hardest to model gauges, as seen in the lower tail of the CDF of per-catchment KGE scores in Figure 2, but is also present across gauges. Because we constructed the reservoir and non-reservoir training sets to be otherwise very similar, we attribute this difference to exposure to upstream reservoir dynamics during training, rather than to any other incidental dataset properties. This is further supported by gradient importance analysis, which shows that reservoir-trained models place greater weight on the features of reservoir contributing area compared to the Non-Reservoir LSTM, consistent with having learned an implicit representation of reservoir-influenced flow behaviour.

### 5.2 Training on reservoir-influenced gauges alone hinders performance on non reservoir-influenced gayges

The Reservoir LSTM degrades on non reservoir-influenced gauges (NSE 0.58, KGE 0.63) compared to reservoir-influenced gauges. The Reservoir LSTM over-applies reservoir-influenced-flow patterns where they are not appropriate. It is highly sensitive to features such as depth to bedrock, while being less sensitive to precipitation seasonality or root depth. This means that, even in catchments with little abstraction and human influence, the Reservoir LSTM prioritises changes underweights seasonal variation. In non reservoir-influenced gauges, all other models include at three precipitations seasonality and root depth in their top ten most important variables, while the Reservoir LSTM does not. Because the reservoir and non-reservoir training sets were otherwise matched, this gap is due to exposure to reservoirs, rather than other differences between the datasets. This is supported by gradient analysis showing qualitatively different feature prioritisation in each specialist.

That specialist models excel in their own domain but degrade in the other confirms that reservoir influence is a learnable signal, but also that exposure to non reservoir-influenced catchments during training is necessary for a model to retain skill in uninfluenced gauges. A model trained exclusively on regulated flow has not been exposed to the underlying hydrological dynamics that govern natural catchments, and cannot recover that understanding at inference time. Because the reservoir and non-reservoir training sets were otherwise matched, this gap is due to exposure to reservoirs, rather than other differences between the datasets.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


### 5.3 A single combined model achieves specialist-level performance across both flow regimes

A key practical question is whether training on pooled data, rather than maintaining separate specialist models for each regime, sacrifices performance in either domain. We address this in two stages. First we compare the specialist models against the Full LSTM, then we ask whether the Hydra-LSTM's explicit architectural specialisation adds anything beyond what pooled training alone achieves.

The Full LSTM and Hydra: Main, both trained on the full combined dataset, consistently match or exceed the relevant specialist in each domain (Tables 3 and 4). Kolmogorov-Smirnov tests confirm that the performance distributions of the generalist and specialist models are statistically indistinguishable at the 10% level. That specialist models trained on smaller but domain-relevant datasets match the combined models suggests that data relevance, rather than dataset size alone, drives these gains. Gradient importance analysis is consistent with this interpretation. The Full LSTM prioritises reservoir-specific attributes, such as reservoir contributing area and hydroelectric fraction — on reservoir-influenced gauges, and shifts toward precipitation-regime and land-cover features on non reservoir-influenced gauges. This implicit feature switching mirrors the behaviour of the respective specialists, and suggests that the Full LSTM has learned to partition flow regimes internally without architectural guidance.

Across all test gauges (Table 5), the Specialist LSTMs, Hydra: Main, and Hydra Specialists all achieve a combined KGE of $0.78$, with the Full LSTM marginally behind at 0.76. All models perform similarly on the $\beta$ and correlation components of KGE. The one exception is flow variability: the Hydra models achieve an $\alpha$ of 0.97, compared to 0.94 for the specialist and Full LSTM configurations. This likely reflects the Hydra architecture's static embedding in both body and head, rather than only at initialisation, which may give it a richer representation of catchment characteristics. The difference is slight and does not reach significance at the 10% level, but is worth noting for applications where capturing flow variability is a priority.

Using the Hydra-LSTM to finetune to reservoir-influenced or non reservoir-influenced flow seems unnecessary. The Hydra Res-Head performs comparably to the Full LSTM in both domains rather than surpassing it. It does have an advantage over the specialist models, it provides stronger performance on non reservoir-influenced gauges compared to the Reservoir LSTM, with a KGE 0.77 versus 0.63. This may be attributable to the regularising effect of the shared body rather than to specialisation itself. Howeverm pooled training alone is sufficient for implicit domain partitioning. Whether this holds when reservoir-influenced gauges are substantially fewer, or when catchment-specific dynamic variables could be routed through specialist heads, remains an open question for future work.

### 5.4 Limits in highly managed catchments

The case studies at Vyrnwy Reservoir and Gwyfrai at Bontnewydd illustrate that the reservoir-influenced label encompasses a wide range of overall reservoir influence, and point to a structural limitation of purely meteorological approaches. At Gwyfrai, a much smaller fraction of the upstream area is drained through reservoirs, and all reservoir-trained models achieve NSE scores above $0.90$, suggesting that where reservoir operations produce a flow regime that remains broadly shaped by meteorological forcing, LSTM models can learn a reasonable approximation of that behaviour from static attributes and atmospheric inputs

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


alone. The reservoir here modulates the natural signal, but doesn't completely remove it. Vyrnwy represents the opposite extreme. The whole upstream area is drained through the large reservoir, and the observed hydrograph is dominated by sharp operational releases against a near-zero baseline, a pattern that is unrecoverable from precipitation and temperature inputs alone. No model approaches acceptable performance, but reservoir-training does improve performance. The reservoir-trained models, which show less over prediction than the Non-Reservoir LSTM, fail to capture release timing. Spatial analysis of model errors revealed no systematic geographic patterns, suggesting that prediction difficulty is governed by characteristics of indivdual gauges rather than regional factors. Reservoir releases at Vyrnwy reflect real-time operational decisions driven by energy demand, conservation targets, and water resource management (Severn Trent, 2017), none of which are encoded in the available timeseries data. For heavily managed single-purpose reservoirs, dramatically improving predictions will likely require dynamic operational inputs, such as storage levels or demand schedules, rather than architectural refinements.

## 6 Conclusions

In this paper, we investigated whether LSTM-based hydrological models can represent the effects of upstream reservoir operation, and whether this learning generalises to unseen gauges. Answering this required us to create matched training sets, apply gradient importance analysis to investigate what each model had learned, and use statistical tests to distinguish genuine performance differences from noise. By training and evaluating five model configurations including specialized, combined, and multi-headed architectures across reservoir-influenced and non reservoir-influenced test gauges in the UK, we identified three main findings.

First, exposure to reservoir-influenced gauges during training is necessary for good performance at held-out reservoir-influenced test gauges: the Non-Reservoir LSTM consistently overestimated the mean and variance of flow in these gauges. Second, the inverse is also true. A model trained exclusively on reservoir-influenced gauges degrades when applied to non reservoir-influenced gauges, potentially over-relying on its understanding of reservoir-influenced-flow in non reservoir-influenced areas. In non reservoir-influenced gauges the Reservoir LSTM had a weaker performance, with a KGE of $0.63$, at least $0.1$ lower than the KGE of all other models. Gradient importance analysis supports both of these findings: reservoir-trained and non-reservoir-trained models prioritise qualitatively different static input features, consistent with each having learned a qualitatively distinct implicit representation of its respective flow regime. Finally, the Full LSTM and Hydra-LSTM both perform comparably to the relevant specialist in each domain, and overall (Table 5). Differences among these models are small and should be seen as of equivalent performance, indicating that pooling training data across regimes is sufficient to recover specialist-level skill without domain-specific architectures.

These results suggest that the distinction between reservoir-influenced and non reservoir-influenced flow is a learnable signal that data-driven models can exploit across catchment types. It also suggests that pooling training data across both regimes is a viable and efficient strategy for operational deployment when datasets are of similar size. However, case studies at Vyrnwy Reservoir and Gwyfrai at Bontnewydd demonstrate that while prediction in multi-reservoir catchments is tractable, highly managed releases remain a fundamental challenge for purely meteorological data-driven approaches.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


The clearest limitation is that this learnable signal has a ceiling, regardless of architecture. To better predict reservoir operation, we may need more temporal data that influence reservoir operation. Future work should study this through three main avenues. First, a more rigorous explainability analysis such as eigenvector decompositions of gradients across a large number of gauges could reveal whether models have learned physically interpretable proxies for reservoir operation or just statistical regularities. Second, while this study leveraged the clean CAMELS-GB dataset, replicating this analysis across other countries and regulatory contexts is necessary to assess generalisability. This will require careful attention to the quality of reservoir-influence labels, as exceptionally mislabelled gauges can meaningfully damage model training. Finally, where operational data (e.g., storage levels, electricity demands), or earth observation data is available (Mason and Dance, 2026), future research should assess whether routing these dynamic, catchment-specific variables through specialist heads like the Hydra-LSTM can close the predictability gap in highly managed catchments like Vyrnwy.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


*Acknowledgements.* This work was supported by the Advanced Frontiers for Earth System Prediction Doctoral Training Programme, funded by the University of Reading. It was also supported by the Japan Society for the Promotion of Science. Finally, we thank Ieuan Higgs for his advice on using integrated gradients.

*Code and data availability.* Code for training, evaluation, and gradient importance analysis is available at https://github.com/KarRups/UK_Forecasting_With_HydraLSTM. Data is available from the Centre of Ecology and Hydrology (Coxon et al., 2020a, 2025a).

*Author contributions.* KR was the lead researcher, developing the code and research question. DY, KH, HC, CP, FP, and MC provided supervision, review and editing. All authors have read and agreed to the current version of the paper.

*Competing interests.* The authors declare that they have no conflict of interest.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


## References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, 2017.

Beven, K.: Surface water hydrology—runoff generation and basin structure, Reviews of Geophysics, 21, 721–730, 1983.

Beven, K. J.: Rainfall-runoff modelling: the primer, John Wiley & Sons, 2012.

Bosch, J. M. and Hewlett, J.: A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration, Journal of hydrology, 55, 3–23, 1982.

Brunner, M. I.: Reservoir regulation affects droughts and floods at local and regional scales, Environmental Research Letters, 16, 124 016, 2021.

Carrillo, A. M. R. and Frei, C.: Water: A key resource in energy production, Energy policy, 37, 4303–4312, 2009.

Coxon, G., Addor, N., Bloomfield, J., Freer, J., Fry, M., Hannaford, J., Howden, N., Lane, R., Lewis, M., Robinson, E., Wagener, T., and Woods, R.: Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain (CAMELS-GB), https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9, 2020a.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J., Lane, R., Lewis, M., Robinson, E. L., et al.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth System Science Data, 12, 2459–2483, 2020b.

Coxon, G., Zheng, Y., Barbedo, R., Cooper, H., Fileni, F., Fowler, H., Fry, M., Green, A., Gribbin, T., Harfoot, H., Lewis, E., Qiu, X., Salwey, S., Wendt, D., and Ribeiro Neto, G.: Catchment boundaries, daily and sub-daily hydrometeorological time series, groundwater level time series and attributes for 671 catchments in Great Britain (CAMELS-GB v2), https://doi.org/10.5285/9a46d428-958f-4ac1-86eb-94eee70c0955, 2025a.

Coxon, G., Zheng, Y., Barbedo, R., Cooper, H., Fileni, F., Fowler, H. J., Fry, M., Green, A., Gribbin, T., Harfoot, H., et al.: CAMELS-GB v2: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth System Science Data Discussions, 2025, 1–44, 2025b.

Döll, P., Fiedler, K., and Zhang, J.: Global-scale analysis of river flow alterations due to water withdrawals and reservoirs, Hydrology and Earth System Sciences, 13, 2413–2432, 2009.

DrivenData: Reclamation Water Supply Forecast Challenge, https://www.drivendata.org/competitions/254/reclamation-water-supply-forecast-dev/, accessed: 2024-09-04, 2024.

Durant, M. and Counsell, C.: Inventory of reservoirs amounting to 90

Fleming, S. W., Garen, D. C., Goodbody, A. G., McCarthy, C. S., and Landers, L. C.: Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence, Journal of Hydrology, 602, 126 782, 2021.

Gao, Y., Vogel, R. M., Kroll, C. N., Poff, N. L., and Olden, J. D.: Development of representative indicators of hydrologic alteration, Journal of Hydrology, 374, 136–147, 2009.

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, 2009.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, 9, 1735–1780, 1997.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


Hollis, D., McCarthy, M., Kendon, M., Legg, T., and Simpson, I.: HadUK-Grid—A new UK dataset of gridded climate observations, Geoscience data journal, 6, 151–159, 2019.

Hunt, K. M., Matthews, G. R., Pappenberger, F., and Prudhomme, C.: Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States, Hydrology and Earth System Sciences, 26, 5449–5472, 2022.

Jumani, S., Deitch, M. J., Kaplan, D., Anderson, E. P., Krishnaswamy, J., Lecours, V., and Whiles, M. R.: River fragmentation and flow alteration metrics: a review of methods and directions for future research, Environmental Research Letters, 15, 123 009, 2020.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resources Research, 55, 11 344–11 354, 2019.

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, 2024.

Li, K. and Razavi, S.: What controls hydrology? An assessment across the contiguous United States through an interpretable machine learning approach, Journal of Hydrology, p. 131835, 2024.

Mason, D. and Dance, S.: Improved urban flood detection using Sentinel-1 by effective combination of elevation data with image intensity and interferometric coherence, Journal of Applied Remote Sensing, 2026.

McCarthy, G. T.: The unit hydrograph and flood routing, Army Engineer District, Providence, 1939.

Moore, R. J. and Cole, S. J.: IMPRESS: approaches to IMProve flood and drought forecasting and warning in catchments influenced by REServoirS, 2022.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, 1970.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T., Weitzner, D., and Yoss, M.: AI Increases Global Access to Reliable Flood Forecasts, arXiv preprint arXiv:2307.16104, 2023.

of Civil Engineers, I.: Floods and Reservoir Safety, ICE Publishing, ISBN 978-0-7277-6006-7, https://doi.org/10.1680/frs.60067, 2015.

Payan, J.-L., Perrin, C., Andréassian, V., and Michel, C.: How can man-made water reservoirs be accounted for in a lumped rainfall-runoff model?, Water Resources Research, 44, 2008.

Robinson, E., Blyth, E., Clark, D., Comyn-Platt, E., and Rudd, A.: Climate hydrology and ecology research support system meteorology dataset for Great Britain (1961-2017)[CHESS-met], 2020a.

Robinson, E., Blyth, E., Clark, D., Comyn-Platt, E., and Rudd, A.: Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961-2017)[CHESS-PE], 2020b.

Ruparell, K., Marks, R. J., Wood, A., Hunt, K. M., Cloke, H. L., Prudhomme, C., Pappenberger, F., and Chantry, M.: Hydra-LSTM: A semi-shared Machine Learning architecture for prediction across Watersheds, Artificial Intelligence for the Earth Systems, 4, 240 103, 2025.

Salwey, S., Coxon, G., Pianosi, F., Singer, M. B., and Hutton, C.: National-scale detection of reservoir impacts through hydrological signatures, Water Resources Research, 59, e2022WR033 893, 2023.

Sardo, M., Epifani, I., D'Odorico, P., Galli, N., and Rulli, M. C.: Exploring the water–food nexus reveals the interlinkages with urban human conflicts in Central America, Nature Water, 1, 348–358, 2023.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


Severn Trent: Severn Trent becomes first UK company to be inducted into the Hydro Hall of Fame, https://www.stwater.co.uk/news/news-releases/severntrentbecomesfirstukcompanytobeinductedintothehydrohalloffame/, accessed: 2026-04-14, 2017.

Shaw, E. M., Beven, K. J., Chappell, N. A., and Lamb, R.: Hydrology in Practice, CRC Press, London, 4th edn., ISBN 978-0415370424, 2010.

Singer, M. B.: The influence of major dams on hydrology through the drainage network of the Sacramento River basin, California, River Research and Applications, 23, 55–72, 2007.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic attribution for deep networks, in: International conference on machine learning, pp. 3319–3328, PMLR, 2017.

Tanguy, M., Dixon, H., Prosdocimi, I., Morris, D., and Keller, V.: Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2019)[CEH-GEAR], Environmental Information Data Centre, 10, 2021.

Tebakari, T., Yoshitani, J., and Suvanpimol, P.: Impact of large-scale reservoir operation on flow regime in the Chao Phraya River basin, Thailand, Hydrological Processes, 26, 2411–2420, 2012.

UK Centre for Ecology & Hydrology: Reservoir management, https://www.ceh.ac.uk/solutions/industries/water/reservoir-management, accessed: 2026-03-16, n.d.

US Army Corps of Engineers: Hydrologic Engineering Requirements for Reservoirs, Engineer Manual EM 1110-2-1420, Department of the Army, U.S. Army Corps of Engineers, chapter 4, 2018.

Yoshimi, K., Hascoet, T., Dossa, R., Takashima, R., Takiguchi, T., and Oishi, S.: Optimizing Japanese dam reservoir inflow forecast for efficient operation.

## Appendix A: Static Catchment Attributes

This appendix details the static catchment attributes used as predictor variables. The initial candidate set was drawn from CAMELS-GB and CAMELS-GBV2 (Coxon et al., 2020b, 2025b), then filtered to ensure all predictors were strictly numerical, independent of discharge, and representative of physical or static human characteristics.

Three filtering decisions were applied. Hydrological signatures (e.g., `q mean`, `baseflow index`) were excluded to prevent data leakage. Spatial and temporal identifiers, coordinates, categorical labels, and date markers, were removed to prevent the model from memorising specific locations or arbitrary historical periods. Land cover percentages were retained across multiple years to capture structural changes in land use over time.

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


**Table A1.** Topography and morphology attributes.

| Attribute Name | Description | Unit |
|---|---|---|
| area | Catchment area | $km^2$ |
| gauge_elev | Gauge elevation | m a.s.l. |
| dpsbar | Mean drainage path slope | m $km^{-1}$ |
| elev_mean, elev_min, elev_10, elev_50, elev_90, elev_max | Elevation distribution (mean, min, and percentiles) | m a.s.l. |

**Table A2.** Climatic index attributes.

| Attribute Name | Description | Unit |
|---|---|---|
| p_mean | Mean daily precipitation | mm $d^{-1}$ |
| pet_mean | Mean daily potential evapotranspiration | mm $d^{-1}$ |
| aridity | Aridity index (PET / P) | – |
| p_seasonality | Seasonality and timing of precipitation | – |
| frac_snow | Fraction of precipitation falling as snow | – |
| high_prec_freq, high_prec_dur | Frequency (d $yr^{-1}$) and mean duration (d) of high-precipitation events ($\geq 5\times$ mean daily) | – |
| low_prec_freq, low_prec_dur | Frequency (d $yr^{-1}$) and mean duration (d) of dry periods ($<1$ mm $d^{-1}$) | – |

**Table A3.** Land cover attributes. Each variable is provided with a [YEAR] suffix for 1990, 2015, and 2017–2022.

| Attribute Name | Description | Unit |
|---|---|---|
| dwood_perc_[YEAR] | Deciduous woodland cover | % |
| ewood_perc_[YEAR] | Evergreen woodland cover | % |
| grass_perc_[YEAR] | Grassland cover | % |
| shrub_perc_[YEAR] | Shrubland cover | % |
| crop_perc_[YEAR] | Cropland cover | % |
| urban_perc_[YEAR] | Urban cover | % |
| inwater_perc_[YEAR] | Inland water cover | % |
| bares_perc_[YEAR] | Bare soil cover | % |

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


**Table A4.** Soil attributes. All physical properties also include `_missing`, `_5`, `_50`, and `_95` percentile variants across the catchment.

| Attribute Name | Description | Unit |
|---|---|---|
| `sand_perc`, `silt_perc`, `clay_perc`, `organic_perc` | Particle size fractions and organic content | % |
| `bulkdens` | Bulk density | g $cm^{-3}$ |
| `tawc` | Total available water content | mm |
| `porosity_cosby`, `porosity_hypres` | Volumetric porosity (two pedotransfer estimates) | – |
| `conductivity_cosby`, `conductivity_hypres` | Saturated hydraulic conductivity (two pedotransfer estimates) | cm $h^{-1}$ |
| `root_depth` | Depth available for roots | m |
| `soil_depth_pelletier` | Depth to bedrock | m |

**Table A5.** Hydrogeology attributes.

| Attribute Name | Description | Unit |
|---|---|---|
| `inter_high_perc`, `inter_mod_perc`, `inter_low_perc` | Intergranular flow productivity (high, moderate, low) | % |
| `frac_high_perc`, `frac_mod_perc`, `frac_low_perc` | Fracture flow productivity (high, moderate, low) | % |
| `no_gw_perc`, `low_nsig_perc`, `nsig_low_perc` | Negligible groundwater / low-productivity aquifer fractions | % |

**Table A6.** Human influence attributes.

| Attribute Name | Unit | Description |
|---|---|---|
| `surfacewater_abs`, `groundwater_abs` | mm $d^{-1}$ | Mean surface water and groundwater abstraction |
| `discharges` | mm $d^{-1}$ | Mean daily discharges into watercourses |
| `abs_[sector]_perc` | % | Abstraction share by sector: agriculture, amenities, energy, environment, industry, water supply |
| `abs_[loss]_perc` | % | Abstraction share by loss class: high, medium, low, very low |
| `num_reservoir`, `reservoir_cap` | –, ML | Number of reservoirs; total storage capacity |
| `reservoir_contributing_area` | % | Catchment area drained through reservoirs |
| `reservoir_normalised_upstream_capacity` | – | Reservoir capacity relative to mean annual precipitation volume |
| `reservoir_[use]` | % | Storage allocation by use: hydroelectricity, navigation, drainage, water resources, flood storage, environment |

https://doi.org/10.5194/egusphere-2026-2909
Preprint. Discussion started: 16 June 2026


**Table A7.** Hydrometry and well attributes.

| Attribute Name | Description | Unit |
|---|---|---|
| `gw_well_depth` | Groundwater well depth | m |
| `aquifer` | Aquifer attributed to well water-level variations | – |
| `bankfull_flow` | Flow at which river overtops banks | $m^3\ s^{-1}$ |
| `structurefull_flow` | Flow at which river reaches structure wingwalls | $m^3\ s^{-1}$ |