Supplement of Hydrol. Earth Syst. Sci., 25, 5517–5534, 2021
https://doi.org/10.5194/hess-25-5517-2021-supplement

Hydrology and Earth System Sciences
Open Access

Supplement of

Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models

Thomas Lees et al.

Correspondence to: Thomas Lees (thomas.lees@chch.ox.ac.uk)

Supplementary Information: Benchmarking Data-Driven Rainfall-Runoff Models in Great Britain: A comparison of LSTM-based models with four lumped conceptual models

Thomas Lees[1], Marcus Buechel[1], Bailey Anderson[1], Louise Slater[1], Steven Reece[2], Gemma Coxon[3], and Simon J. Dadson[1,4]

[1]School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, United Kingdom, OX1 3QY
[2]Department of Engineering, University of Oxford, Oxford, United Kingdom
[3]Geographical Sciences, University of Bristol, Bristol, United Kingdom
[4]UK Centre for Ecology and Hydrology, Maclean Building, Crowmarsh Gifford, Wallingford, United Kingdom, OX10 8BB

Correspondence: Thomas Lees (thomas.lees@chch.ox.ac.uk)

1 LSTM and EA LSTM Model Description

The LSTM captures information that is important over both long and short term time horizons, overcoming a key difficulty with traditional RNNs, which are unable to retain information over longer sequences (Hochreiter, 1991; Bengio et al., 1994). LSTMs do this by maintaining two state vectors, a cell memory vector that captures slowly evolving processes ($\mathbf{C_t}$, Eq. 5) and a more quickly evolving state vector, colloquially named the "hidden" vector ($\mathbf{h_t}$, Eq. 6). The $\mathbf{C_t}$ vector, accounts for longer-term dependencies, and a series of ‘gates’ control the information passing into and out of the memory vector. The $\mathbf{h_t}$ vector evolves more quickly depending on input information and the output of the memory vector (see Fig. S1). The gates include: the forget gate ($\mathbf{f_t}$), which controls the elements of the cell memory vector that are forgotten (i.e. how long water persists in the system, Eq. 1); the input gate ($\mathbf{i_t}$), which controls what information from the new input data at that timestep will be incorporated into the cell memory vector (i.e. what information is stored for future timesteps, Eq. 2); and finally the output gate ($\mathbf{o_t}$), which determines what information from the cell memory will be used to update the hidden state (i.e. what information will impact discharge at the current timestep, Eq. 3). These gates are neural network layers, made up of weights (W_{layer}), biases (b_{layer}) and activation functions. The activation functions allow the LSTM to model nonlinear processes. During training, we seek the values for these weights and biases that best describe observed discharge. The information that passes through the input gate to the cell state ($\mathbf{C_t}$ - see Eq. 5) is itself processed through a neural network layer, producing a series of candidate values that may be used to update the cell state (Eq. 4). Finally, information from the cell state is passed through the output gate ($\mathbf{o_t}$) to produce the hidden output ($\mathbf{h_t}$) at that time-step (Eq. 6). Note that for the LSTM we have explicitly defined the inputs as the concatenation of the dynamic meteorological data and the static catchment attributes, $[\mathbf{X_{t,n}}, \mathbf{A_n}]$. That is, both LSTM models receive the same information. We refer the reader to Kratzert et al. (2018) and Kratzert et al. (2019) for comprehensive descriptions of the LSTM and EA LSTM, and their hydrological interpretation.

$$\mathbf{f_t} = \sigma\left(\mathbf{W}_{\mathrm{f}}\left[\mathbf{X_{t,n}}, \mathbf{A_n}, \boldsymbol{h}_{t-1}\right] + \boldsymbol{b}_{\mathrm{f}}\right) \tag{1}$$

$$\mathbf{i_t} = \sigma\left(\mathbf{W}_{\mathrm{i}}\left[\mathbf{X_{t,n}}, \mathbf{A_n}, \boldsymbol{h}_{t-1}\right] + \boldsymbol{b}_{\mathrm{i}}\right) \tag{2}$$

$$\mathbf{o_t} = \sigma\left(\mathbf{W}_{\mathrm{o}}\left[\mathbf{X_{t,n}}, \mathbf{A_n}, \boldsymbol{h}_{t-1}\right] + \boldsymbol{b}_{\mathrm{o}}\right) \tag{3}$$

$$\tilde{\mathbf{C}}_{\mathbf{t}} = \tanh\left(\mathbf{W}_{\mathrm{C}}\left[\mathbf{X_{t,n}}, \mathbf{A_n}, \boldsymbol{h}_{t-1}\right] + \boldsymbol{b}_{\mathrm{C}}\right) \tag{4}$$

$$\mathbf{C_t} = \mathbf{f_t} * \boldsymbol{C}_{t-1} + i * \tilde{C}_t \tag{5}$$

$$\mathbf{h_t} = \mathbf{o_t} * \tanh(\mathbf{C_t}) \tag{6}$$

The EA LSTM was developed specifically for rainfall-runoff modelling (Kratzert et al., 2019). The key difference between the EA LSTM and the LSTM is that the input gate ($\boldsymbol{i}$) is no longer conditional upon the dynamic (time-varying) data. Instead, the static (time-invariant) catchment attributes ($\mathbf{A_n}$) exclusively influence the input gate (Eq. 2 is replaced with Eq. 8), and all other gates are solely influenced by the dynamic input data (Eq. 7, 9, 10).

$$\mathbf{f_t} = \sigma\left(\mathbf{W}_{\mathrm{f}}\left[X_t, \boldsymbol{h}_{t-1}\right] + \boldsymbol{b}_{\mathrm{f}}\right) \tag{7}$$

$$\boldsymbol{i} = \sigma\left(\mathbf{W}_{\mathrm{i}} A + \boldsymbol{b}_{\mathrm{i}}\right) \tag{8}$$

$$\mathbf{o_t} = \sigma\left(\mathbf{W}_{\mathrm{o}}\left[X_t, \boldsymbol{h}_{t-1}\right] + \boldsymbol{b}_{\mathrm{o}}\right) \tag{9}$$

$$\tilde{\mathbf{C}}_{\mathbf{t}} = \tanh\left(\mathbf{W}_{\mathrm{C}}\left[X_t, \boldsymbol{h}_{t-1}\right] + \boldsymbol{b}_{\mathrm{C}}\right) \tag{10}$$

The EA LSTM is described as “entity-aware” because it explicitly learns how to use $\mathbf{A_n}$ to distinguish between similar dynamic inputs ($\mathbf{X_{t,n}}$) for different catchments ("entities"). For the EA LSTM, $\boldsymbol{i}$ is determined solely by the catchment attributes (Eq. 8). Therefore, each catchment has one unique hs dimensional vector which controls what information should persist in future timesteps. In contrast, the LSTM learns to modify the input gate $\mathbf{i_t}$ based upon the meteorological forcing data ($\mathbf{X_{t,n}}$) *and* the catchment attributes ($\mathbf{A_n}$). The output of the input gate ($\mathbf{i_t}$ or $\boldsymbol{i}$) is a vector of values between 0 and 1, which is learned from data. This vector, also known as an "embedding", translates our catchment attributes into a high-dimensional space that represents catchments in a manner optimised to differentiate between catchment rainfall-runoff behaviours. Kratzert et al. (2019) demonstrated how this embedding represents what the model has learned about our catchments.

For the sake of clarity, it is important to note that both models receive the same information. The LSTM still receives the static catchment attributes. However, rather than affecting only the input gate, the static data can influence all gates, since they are appended to a vector of dynamic inputs ($[\mathbf{X_{t,n}}, \mathbf{A_n}]$) and so the same information is given to the LSTM at each timestep. The static attributes are used by the LSTM in the same way as the dynamic data. This offers extra flexibility for the LSTM compared with the EA LSTM, since the LSTM is able to modify the input gate based on information from time-varying data, whereas the EA LSTM is not. We are using the static nature of the data as a constraint on the EA LSTM to reflect the nature of the input data (separated into static and dynamic inputs - see Fig. S1).

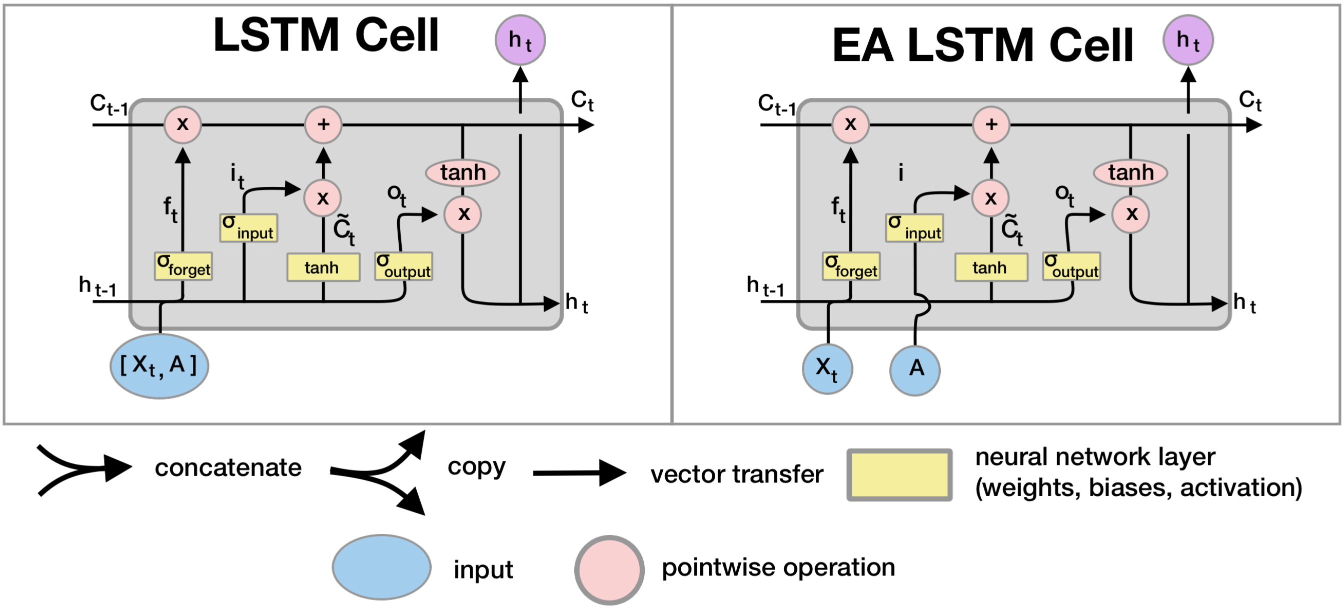


Figure S1. Wiring diagram for the LSTM and EA LSTM recurrent cells, adapted from Olah (2016). These cells are repeated for each input timestep in our sequence length. The key difference between the EA LSTM and the LSTM is the separation of the static data, $\mathbf{A_n}$), from the dynamic data $\mathbf{X_{t,n}}$. In the EA LSTM, the static data is the sole input to the input gate, producing an embedding, $\mathbf{i_t}$. In both LSTM models there is a cell state $\mathbf{C_t}$, that passes from cell to cell, capable of modelling longer-term dependencies. Note that the neural network layers correspond with the weights ($\boldsymbol{W}$), biases (b) and activation functions (σ, $\tanh$). These operations correspond to the yellow layers in the diagram.

Both models have a final layer, a fully connected linear layer, which transforms the $\mathbf{h_t}$ vector into a single discharge prediction, $\mathbf{\hat{y}_{t,n}}$.

2 Comparison of the Train and Test Periods

The calibration (train) period and the evaluation (test) period are similar in terms of their predictability, although the evaluation period was slightly less predictable, as can be seen in the shifting of the two baseline model distributions towards lower NSE values (see Fig. S2). We used two baseline models to test how "predictable" the catchment hydrographs are in these two time periods. Climatology makes a prediction based on the mean discharge for that day of the year. Persistence is equivalent to predicting yesterday's value today, predicting the future will be the same as the past. Fig. S2 shows that the processes are largely stationary, and the period we use for calibration is similar to the period we use for evaluation. Indeed, the period we use for calibration is slightly easier to predict than the test period, since the benchmark models perform better, i.e. the distribution of catchment NSE scores is shifted towards higher NSE scores during the train period. Furthermore, the conditions for precipitation, PET, temperature and specific discharge are very similar between the train and test period. The temperatures have warmed slightly and there are slightly more days with zero precipitation, however, it is unlikely that such small changes have impacted the ability of the DL model to generalize. Discharge has risen slightly in the period of interest, across Great Britain.

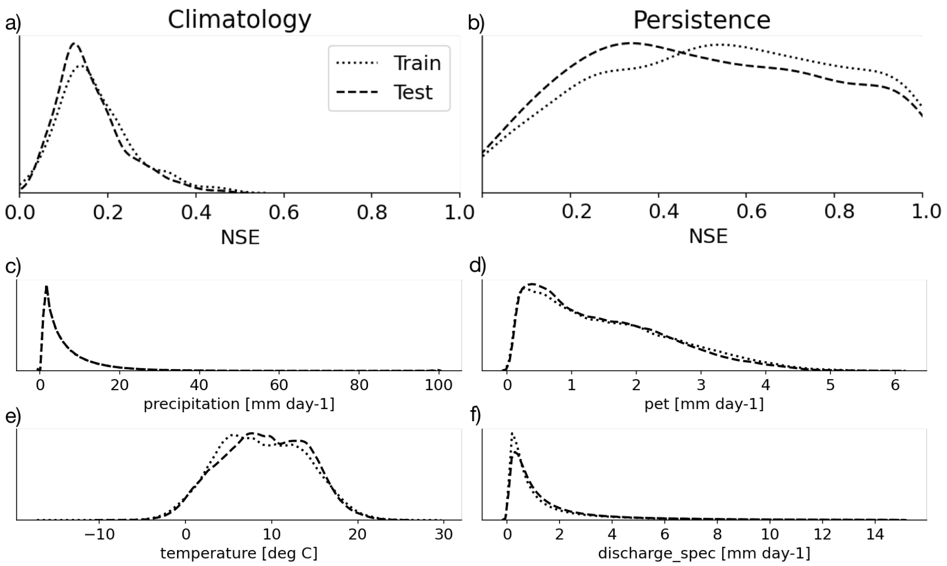


Figure S2. Kernel Density Estimates (KDE) of NSE scores for two baseline models (above), Climatology (a) (calculated as the mean discharge for that day-of-year for each site) and Persistence (b) (calculated as the discharge shifted one day into the future, so yesterdays discharge is a prediction of today). Below, Kernel Density Estimates are provided for hydro-meteorological variables, precipitation (c), potential evaporation (pet) (d), temperature (e) and specific discharge (f) in the training period (1980–1997, dotted line) and the test period (1998–2008, dashed line). Climatology represents the mean conditions for that day of the year. Persistence reflects predicting yesterday's values today, i.e. predicting no change from yesterday. These give an overview of how "predictable" a time period is, since if these baseline models perform well, it will be easier to score at least as well as the baseline.

3 Model Hydrographs

We illustrate the model predictions by showing the hydrographs for three stations from the Thames, the Severn and the Tay, as the largest rivers having at least part of their catchment in England, Wales and Scotland respectively.

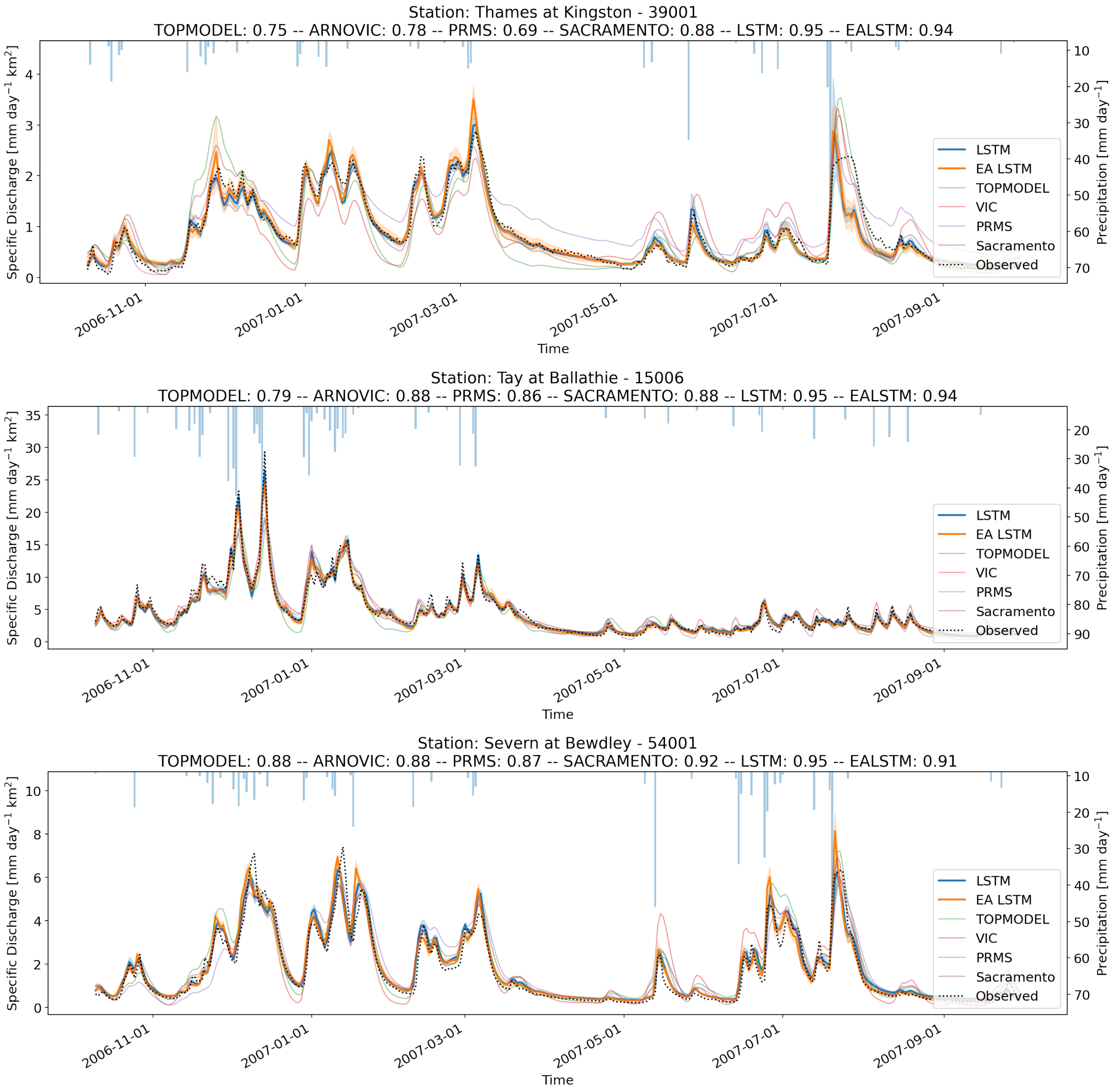


Figure S3. Hydrographs for the Thames at Kingston (Station 39001), the Tay at Ballathie (Station 15006) and the Severn at Bewdley (Station 54001), for the hydrological year from October 2006 – September 2007. The model performances displayed in the header reflect the performance of each model on the entire test period (1998–2008), not just the displayed period. The observed discharge, from (Coxon et al., 2020), is shown as a dotted black line. The bars reflect catchment averaged precipitation with the axis shown on the right side. The LSTM and EA LSTM simulations are shown in blue and orange respectively. Conceptual model simulations for Sacramento (brown), VIC (red), PRMS (purple) and TOPMODEL (green) are taken from published timeseries from Lane et al. (2019).

4 Model Uncertainty

Uncertainty is present in all rainfall-runoff models. Model uncertainty has three main sources: (i) uncertainties in the observed data used to calibrate (train) hydrological models (McMillan et al., 2010); (ii) uncertainties in model structure (Fenicia et al., 2014; Krueger et al., 2010); and (iii) uncertainties in model parameters (Beven and Freer, 2001; Gupta et al., 2009; Arsenault et al., 2014). Parameter uncertainty can be evaluated by using an uncertainty evaluation framework (Beven and Binley, 2014), often involving a sampling strategy. Model structural uncertainty is often estimated within multi-model frameworks, such as the Modular Modelling System (Leavesley et al., 1996) or the Framework for Understanding Structural Errors (FUSE) (Clark et al., 2008). Uncertainties in observations can be estimated and accounted for by using multiple forcing products (Kratzert et al., 2021) or by resampling the input data. This study addresses predictive uncertainty in the LSTM-based models by using an ensemble of 8 LSTM models trained with different random seeds, representing different starting conditions for the training process.

The results in the main text, unless otherwise specified, show diagnostic scores given the ensemble mean discharge. Here, we discuss the ensemble range and the uncertainty that this represents. The ensemble is produced by different random seeds, and therefore different starting parameters used during the training process. The mean catchment ensemble variability is 0.16 mm^3 day^{-1}. The median is 0.12 mm^3 day^{-1}. However, model uncertainties and their relationship with catchment attributes are in accordance with our hydrological intuition. For example, we see increasing uncertainty at increased streamflow (Fig. S4). Furthermore, by normalising for mean catchment discharge we can calculate ensemble standard deviation as a ratio of total discharge. This coefficient of variability is greatest in the South East of England (Fig. S5). A more principled treatment of uncertainty, which benchmarks various methods for using DL models to directly simulate a distribution can be found in Klotz et al. (2020).

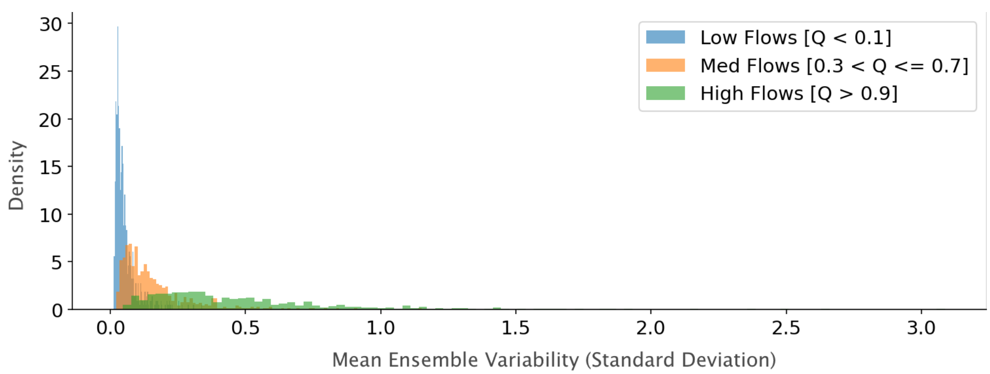


Figure S4. Histogram of raw station averaged variability (standard deviation) across ensemble members. The blue histogram reflects the variability in simulations where observed discharge is lower than the 10th percentile ($y\_true_{t,n}Q_n^{0.1}$). The green histogram shows variability for only those times where observed discharge is greater than the 90th percentile ($y\_true_{t,n}Q_n^{0.9}$). The orange histogram shows variability for all times when the observed discharge is between the 30th and 70th percentile ($Q_n^{0.3}y\_true_{t,n}Q_n^{0.7}$.)

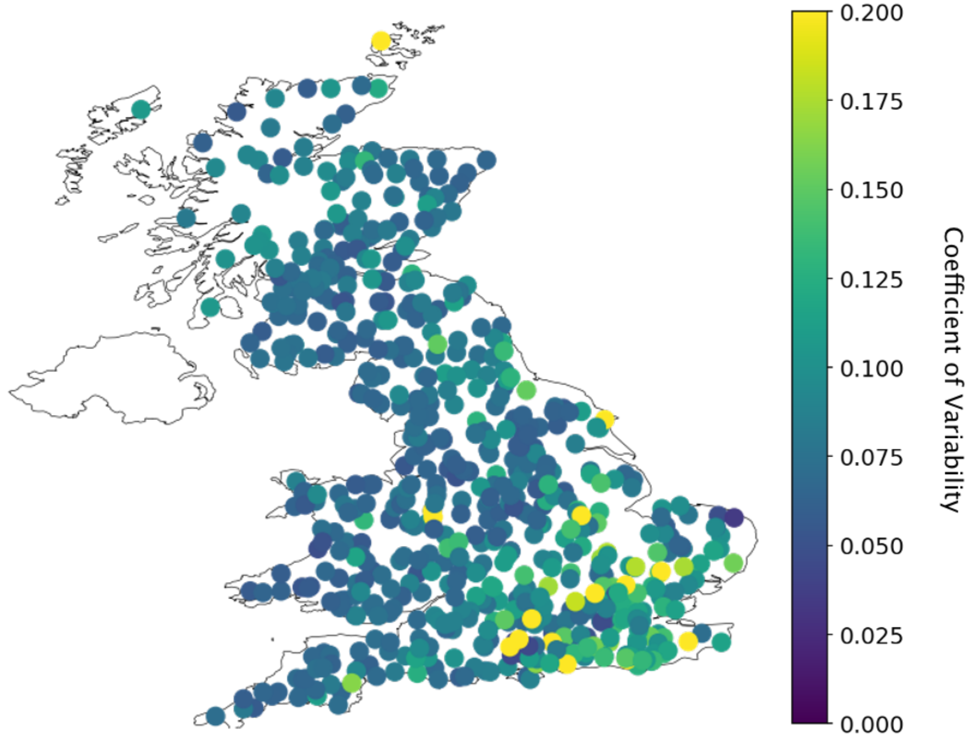
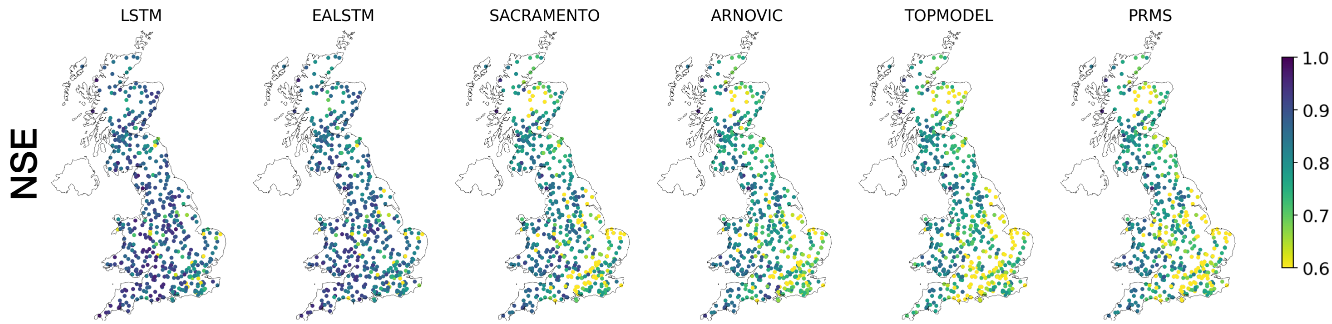
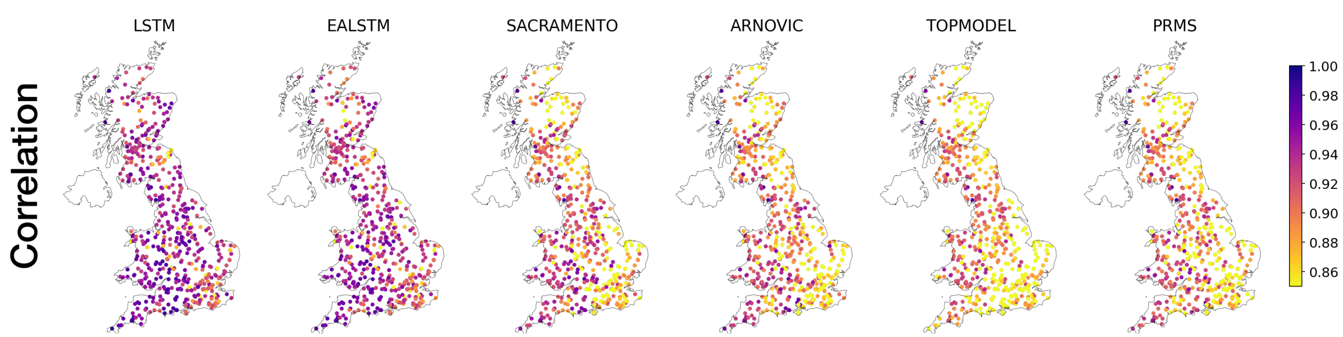
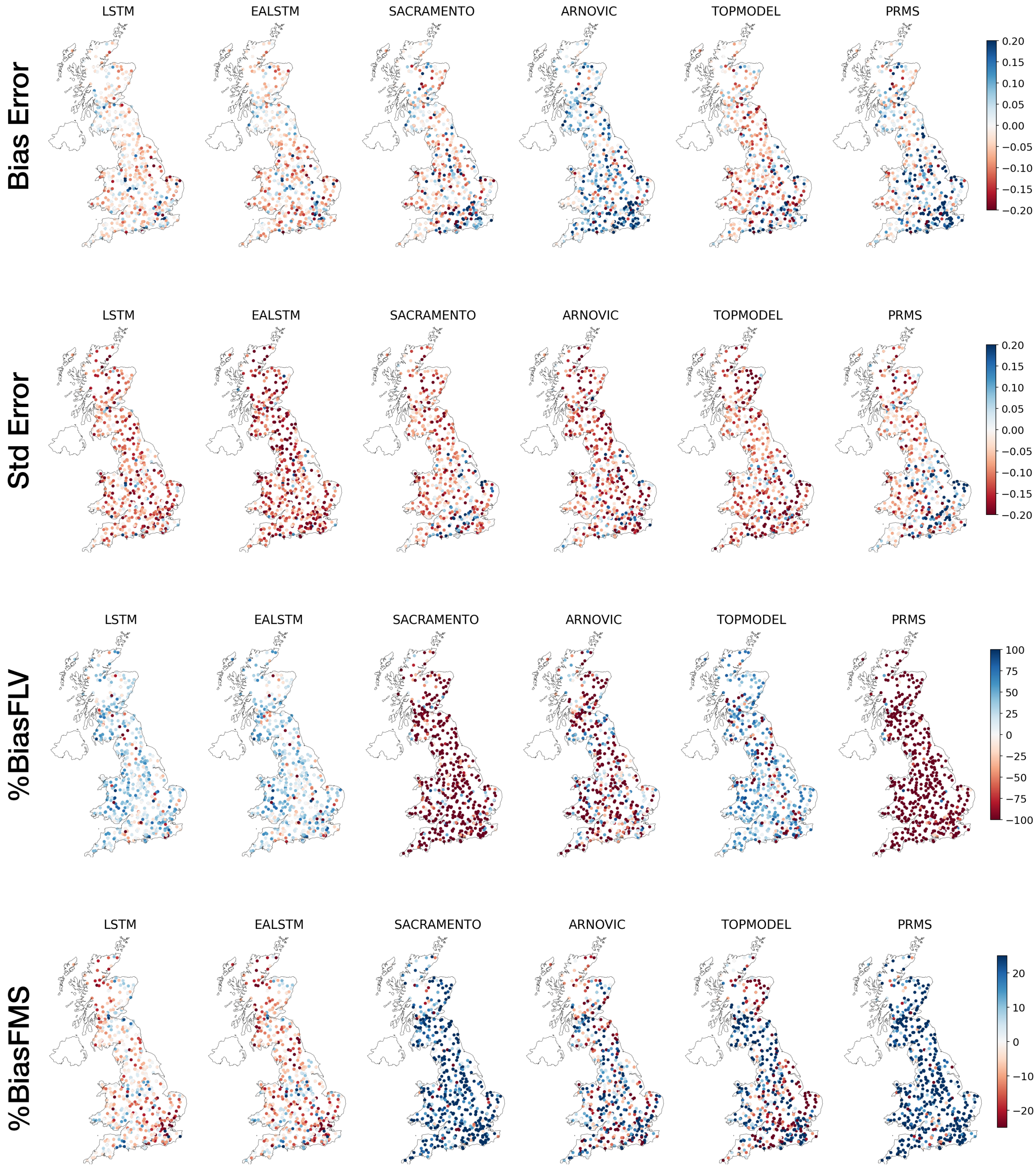


Figure S5. Spatial Patterns of normalised catchment averaged variability (standard deviation) of ensemble predictions. Brighter colours reflect greater variability across members of the ensemble of LSTMs.

5 Spatial Performances of Error Metrics







Bias Error
Std Error
%BiasFLV
%BiasFMS
LSTM
EALSTM
SACRAMENTO
ARNOVIC
TOPMODEL
PRMS
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
100
75
50
25
0
−25
−50
−75
−100
20
10
0
−10
−20

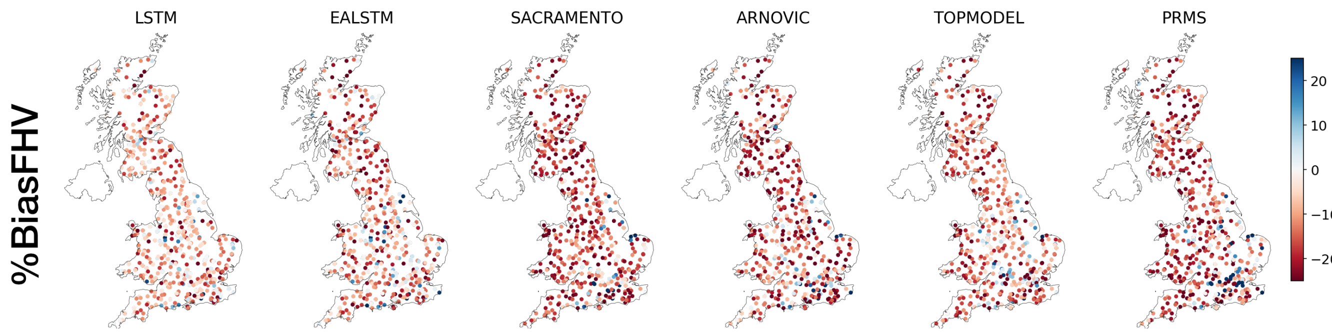


Figure S6. Spatial Patterns of different performance metrics. Each point is a single station-gauge, and the point is coloured according to the performance metric. For performance metrics with a diverging score (above and below an optimum, e.g. Bias Error) more intense colours represent worse performance. Red represents an under-prediction, blue an over-prediction. For scores which are increasing (e.g. NSE, Correlation), darker colours reflect improved performance.

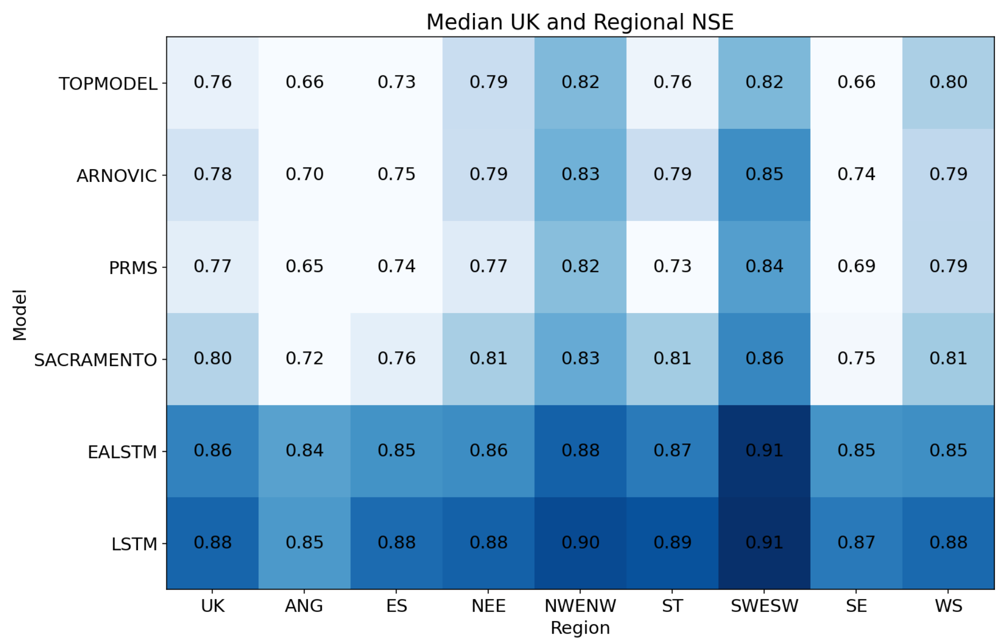


Figure S7. Median NSE scores for eight Great Britain river basin regions. The regions are based on the UKCP09 river basins (mur) aggregated from 21 river basin districts to eight regions. The leftmost column is the median score for all GB catchments, which is the same as in Table 3 in the main text. It is included here for reference.

References

Arsenault, R., Poulin, A., Côté, P., and Brissette, F.: Comparison of stochastic optimization algorithms in hydrological model calibration, Journal of Hydrologic Engineering, 19, 1374–1384, 2014.

Bengio, Y., Simard, P., and Frasconi, P.: Learning Long-Term Dependencies with Gradient Descent is Difficult, IEEE Transactions on Neural Networks, 5, 157–166, 1994.

Beven, K. and Binley, A.: GLUE: 20 years on, Hydrological processes, 28, 5897–5918, 2014.

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, 2001.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, 2008.

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, https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9, 2020.

Fenicia, F., Kavetski, D., Savenije, H. H., Clark, M. P., Schoups, G., Pfister, L., and Freer, J.: Catchment properties, function, and conceptual model representation: is there a correspondence?, Hydrological Processes, 28, 2451–2467, 2014.

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 (Amsterdam), 377, 80–91, 2009.

Hochreiter, S.: Untersuchungen zu dynamischen neuronalen Netzen, Diploma, Technische Universität München, 91, 1991.

Klotz, D., Kratzert, F., Gauch, M., Sampson, A. K., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty Estimation with Deep Learning for Rainfall-Runoff Modelling, arXiv preprint arXiv:2012.14295, 2020.

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, https://hess.copernicus.org/articles/22/6005/2018/, 2018.

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, https://hess.copernicus.org/articles/23/5089/2019/, 2019.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 25, 2685–2703, 2021.

Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J., Bilotta, G. S., Brazier, R. E., Butler, P., and Haygarth, P. M.: Ensemble evaluation of hydrological model hypotheses, Water Resources Research, 46, 2010.

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., Greene, S., Macleod, C. J., 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, 2019.

Leavesley, G., Markstrom, S., Brewer, M., and Viger, R.: The modular modeling system (MMS)—The physical process modeling component of a database-centered decision support system for water and power management, Water, Air, & Soil Pollution, 90, 303–311, 1996.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M.: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, Hydrological Processes: An International Journal, 24, 1270–1284, 2010.

Olah, C.: Understanding LSTM Networks – colah's blog, http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 2016.