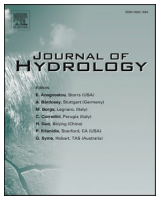
Journal of Hydrology 651 (2025) 132569

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Associations between deep learning runoff predictions and hydrogeological conditions in Australia

Stephanie R. Clark [a,*], Jasmine B.D. Jaffrés [b]

[a] CSIRO, Environment, Sydney, Australia
[b] C&R Consulting, Townsville, Australia

ARTICLE INFO

This manuscript was handled by A. Bardossy

Keywords:
Deep learning
Long short-term memory (LSTM)
Global model
Physics-informed machine learning
Large-sample hydrology (LSH)
Domain knowledge

ABSTRACT

To capture the complexity of hydrological systems across regions, multidimensional domain knowledge (e.g. climate, soils, geology and topography) can be incorporated into deep learning models of streamflow behaviour. Such integration has demonstrated notable improvements in streamflow predictions, thereby enhancing accuracy and offering valuable insights for sustainable water resource management. However, this catchment-specific domain information also holds potential for assessing the suitability of deep learning models for runoff predictions under varied conditions. This study explores the wide-ranging performance of deep learning streamflow predictions across the diverse landscape of Australian catchments through the leveraging of newly-available, comprehensive hydrological and hydrogeological datasets. Data from CAMELS-AUS (the Australian adaptation of CAMELS [Catchment Attributes and MEteorology for Large-sample Studies]) and a nationwide set of hydrogeological catchment attributes are integrated at a continental scale to probe associations between deep learning prediction performance and catchment attributes. The study encompasses three steps: 1) unsupervised learning to identify common patterns of catchment attributes; 2) a continent-wide, deep learning time series model (long short-term memory [LSTM]) incorporating catchment attributes into concurrent predictions across hundreds of basins; and 3) visualising and investigating associations between high (or low) runoff prediction performance and various catchment attributes. The resulting visual analytical tool provides insights into continent-wide differences in performance and also facilitates analysis at the individual catchment level. Key findings reveal a) enhanced LSTM performance in catchments characterised by frequent or variable rainfall, hilly terrain, and low permeability; and b) challenges encountered by the LSTM in flat catchments with slow, infrequent flows, high permeability, and in predicting runoff peaks in regions of substantial summer rainfall. Understanding these performance patterns can help inform the application of global LSTMs in water resource management and hydrological forecasting. Future work may involve assessing how such domain knowledge could improve the extrapolation of predictions to ungauged catchments within each attribute cluster. This multi-catchment study highlights the scalability advantages of machine learning techniques for gaining hydrological insights at a continental scale.

1. Introduction

Accurate prediction of river runoff is essential for sustainable management of water resources. Abundant, detailed predictor data for streamflow estimations are now accessible, primarily due to advancements in remote sensing technologies. These innovations have considerably enhanced the ability to collect comprehensive information about various environmental factors that influence streamflow. However, runoff measurements, which serve as the target variable in streamflow prediction models, are not consistently collected across many river basins. This disparity is especially pronounced in catchments without large population centres, exemplified prominently by Australia's vast, sparsely inhabited interior. The lack of reliable runoff measurements hinders the development and validation of accurate streamflow prediction models, impeding the understanding of hydrological processes in these critical regions. Bridging this data gap is imperative for comprehensive water resource management and ensuring the resilience of communities in remote areas, where the scarcity of streamflow information poses a significant challenge.

Traditional approaches for streamflow estimation, relying on

* Corresponding author.
E-mail address: stephanie.clark@csiro.au (S.R. Clark).

https://doi.org/10.1016/j.jhydrol.2024.132569
Received 19 April 2024; Received in revised form 2 November 2024; Accepted 8 December 2024
Available online 21 December 2024

process-based or conceptual models, require an accurate representation of local rainfall-runoff physics. Detailed information on the physical processes in a basin can be challenging to acquire, with access frequently restricted by economic, geographical or time constraints. This lack of measured data often results in the incorporation of multiple assumptions within the modelling process. Even when physical information is available in specific areas, its inclusion in the modelling framework can limit model transferability to regions where data are less obtainable.

In contrast, the rapidly growing volume of telemetered and remotely-sensed hydrological data and swift development of machine learning techniques have led to the adoption of deep learning methods in runoff modelling. Deep learning techniques excel at capturing nonlinear and complex relationships inherent in hydrological processes, allowing for more accurate modelling of the intricate interactions between climatic, topographic and land-use factors influencing streamflow (Frame et al., 2022; Nearing et al., 2021; Shen et al., 2018). The long short-term memory algorithm (LSTM, Hochreiter & Schmidhuber, 1997), a deep learning neural network model specifically for time series prediction, has been demonstrated to outperform both traditional hydrological models (Arsenault et al., 2023; Kratzert et al., 2018; Lees et al., 2021) and simpler forms of neural networks (Rahimzad et al., 2021) at streamflow estimation tasks.

LSTMs have often been applied on a single-catchment or regional basis because of limited datasets or specific modelling output requirements. However, recent studies have highlighted the benefits to be gained from training over a large set of catchments in what is known as a 'global' model (Kratzert et al., 2024; Lees et al., 2021). Kratzert et al. (2019) first demonstrated the advantages of this method, and in a recent analysis of rainfall-runoff projections under climate change, Wi and Steinschneider (2024) discuss the value of training a deep learning model over a large set of diverse catchments. Benchmark hydrological datasets known as CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) (e.g. Addor et al., 2017) have recently become available for several countries and regions (e.g. United States of America, Great Britain, Chile, Brazil, Australia), providing sets of hundreds of catchments for use in nation-wide or multi-location, large-sample (or 'global') hydrological models.

Although machine learning models are currently providing state-of-the-art streamflow estimates, they do not take the physical realism of predictions into account (Nearing et al., 2019). This raises concerns about the interpretability of results and the lack of insight into – and ability to realistically capture – the underlying physical processes governing streamflow. In order to provide realistic predictions, the incorporation of physical considerations into deep learning models is desirable. To address these concerns, recent research has sought to combine the strengths of both traditional, physically-based models and machine-learning approaches, creating hybrid – or physics-informed – models that leverage the interpretability of traditional models and the predictive power of machine learning (Ng et al., 2023; Tripathy & Mishra, 2023). This integrated approach aims to reconcile the scepticism among traditional modellers by providing a more transparent understanding of the model's decision-making processes and attempting to limit predictions to a physically realistic range. For example, De la Fuente et al. (2023) suggest that architectural modifications to the LSTM structure can enhance hydrological interpretability by ensuring that the neural network's weight patterns are interpretable as hydroclimatic characteristics.

Incorporating domain knowledge about climate, soils, geology and topography is another way to add physical realism to data-driven models that can significantly enhance deep learning modelling (Kratzert et al., 2019; Lees et al., 2021; Wi & Steinschneider, 2024). Climate data – such as precipitation patterns – enrich the model with temporal information critical for understanding hydrological processes. Such processes range from short-term (e.g. extreme weather events) to seasonal and long-term (e.g. multi-year droughts and climate change). Each climate timeframe exerts unique influences on catchment behaviours. Soil types and characteristics determine water retention and infiltration rates, influencing how water moves through the soil profile. Geological data provide insights into subsurface properties that influence groundwater flow, including groundwater / surface water interactions. Topography attributes – like elevation and slope – are fundamental factors influencing surface water flow and runoff patterns. Studies incorporating static catchment attributes such as these in global LSTMs are often known as 'entity-aware' models, a concept that Heudorfer et al. (2024) recently extended from surface water to groundwater modelling.

The consideration of domain knowledge is also important when evaluating the performance of deep learning models for physical systems. It helps ensure that results are physically realistic, provides insights into the model's behaviour, and increases awareness of potential sources of error in real-world applications. Analysing the model errors within a physical context can provide an understanding of physical conditions associated with model performance. For example, model error may be sensitive to critical events in the physical system such as floods or droughts, or to seasonal conditions (e.g. there may be higher error during the wet season than the dry season). Physics-based machine learning frameworks can be used to evaluate a model through domain knowledge and identify physically-relevant error patterns. For example, Anderson and Radić (2022) employed a sensitivity analysis to determine whether a deep learning model's decision making was consistent with the physical system being studied.

Clustering is often a key component of hydrological studies that incorporate physical catchment traits. Kratzert et al. (2019) found that clustering of the attributes allows for the same cell state of the LSTM to be applied to predictions for catchments in each cluster while accounting for interactions between the characteristics. Hashemi et al. (2022) incorporated regime knowledge into individual, global and partitioned LSTMs – clustered on precipitation, runoff and temperature – for 361 catchments in France, finding stronger correlations between LSTM performance and record length in catchments with clear long-term dynamics. To investigate streamflow prediction in ungauged basins with the use of globally-available climate reanalysis data, Wilbrand et al. (2023) clustered CAMELS-US catchments, identifying the characteristics that indicate variations in LSTM performance between models trained on local and global datasets.

Determining links between catchment characteristics and streamflow properties has been the objective of numerous studies (e.g. Janssen & Ameli, 2021; Jehn et al., 2020). Jaffrés et al. (2021) studied the relationship between Australian catchment attributes (such as topography, soil, land use and hydrogeology) and streamflow behaviour to determine which attributes may be relevant to streamflow response in ungauged catchments. Using data from 749 gauging stations, a principal component summary of surface hydrology and baseflow was created based on 20 streamflow descriptors. Then, 2816 catchments were delineated, covering the Australian mainland and Tasmania. Each catchment was assigned a set of 40 attributes. The relationship between these attributes and the hydrological principal components was investigated. It was found that – although climate variables dominate streamflow behaviour – low flows are greatly influenced by topographical features (bifurcation ratio, Horton's drainage density and upstream area), land use and subsurface permeability. In a companion study, Jaffrés et al. (2022) conducted a continent-wide clustering of these 2816 drainage basins into eight clusters, based on the 40 catchment attribute metrics that were reduced to nine orthogonal components via principal component analysis. This clustering process, which included ungauged basins, determined groups of catchments with similar streamflow characteristics. Clusters were described in terms of flow and flood profiles from gauged basins. The results highlight the important impact of catchment attributes (other than climate) on streamflow behaviour and demonstrate the application of a consistent clustering method across all Australian catchments regardless of gauging status. Brunner et al. (2020) considered the relationship

between streamflow and catchment attributes in US basins and attempted to predict the streamflow regime class from static characteristics. Clusters of catchments were created based on mean annual hydrographs. The comparison of twelve attributes found cluster members to also be similar in physical and meteorological characteristics, suggesting that this classification scheme could be of use in determining regime classification of ungauged catchments.

Although it has been established that associations exist between catchment characteristics and streamflow, a question that arises with the use of machine-learning streamflow models is whether certain catchment types might be better suited to representation via deep learning than others. Is there a relationship between catchment attributes and LSTM streamflow prediction performance? When applying global models, this consideration may inform the likelihood of successfully transferring knowledge to ungauged basins of similar characteristics.

The amalgamation of two recent datasets presents an opportunity for developing a comprehensive, continental-scale model capable of assessing the prediction of runoff across various catchment typologies: 1) the extensive rainfall-runoff data from CAMELS-AUS (the Australian version of CAMELS; Fowler et al., 2021), covering 222 catchments, and 2) a curated set of hydrogeological catchment attributes for 2816 subcatchments covering mainland Australia and Tasmania (Jaffrés et al., 2021).

The present study applies clustering and visual analytical techniques to examine the effectiveness of a global, physics-informed LSTM at predicting streamflow in relation to nonlinear combinations of catchment characteristics. This large-sample study is conducted on Australian catchments to determine continent-wide patterns of LSTM performance in relation to the catchment attributes. The physical system of each catchment is represented within the deep learning models through the incorporation of hydrogeological attributes such as soil, topography, geology and climate. Visual analytics are applied to identify relationships between the catchment attributes (and prevalent combinations of attributes) and deep learning prediction performance. Catchments (gauged or ungauged) that were not part of the training set can be analysed to estimate the possible efficacy of LSTMs according to their individual traits. The aim is to understand whether catchments with certain attributes indicate better prediction performance with deep learning and whether this information can enhance predictions in regions where performance is otherwise limited.

2. Data sources

Streamflow, rainfall and climate data used in this study are from the open-source CAMELS-AUS dataset (Fowler et al., 2021). This dataset encompasses a set of 222 catchments over the 1950–2014 period, offering 18 streamflow and climate variables and 134 static catchment attributes. The catchments comprising this collection are mainly undeveloped and therefore free of large-scale changes in land use over time. The CAMELS-AUS dataset – along with CAMELS-US, CAMELS-GB and others – is part of Caravan (Kratzert et al., 2023), a large-sample hydrology collection that provides benchmark sets of quality-checked and standardised rainfall-runoff data. These datasets are specifically designed to support large-sample hydrology studies. CAMELS-AUS is particularly suited for the study of arid-zone hydrology.

In this study, daily CAMELS-AUS data for 1980–2014 are used. The study start date of 1980 was chosen because most of the stations were recording consistent daily measurements from that date onwards. Over the study period, there are an average of approximately 17,000 measurements per catchment, ranging from a minimum of 10,229 to a maximum of 23,376. Locations of the CAMELS-AUS catchments are shown on Fig. 1a. The CAMELS-AUS data were normalised to zero mean and unit variance per feature for use in the deep learning model.

Attribute data for 2816 subcatchments covering the Australian continent (shown in Fig. 1b) are sourced from Jaffrés et al. (2021). From this dataset, 21 variables were selected for use in this study based on their low correlation with other variables and their informativeness to cluster structure and pattern identification in exploratory analyses. The chosen attribute variables are listed in Table 1. The subcatchment attribute data were pre-processed before being used in the study. Variables with highly skewed distributions (annual rain days, vertical range, upstream area, elevation standard deviation [SD], slope, hill slope, and longest flow path) were log-transformed. Each variable was then scaled to the range [0,1]. Jaffrés et al. (2022) applied the original subcatchment attribute data to classify the subcatchments into eight clusters with the fuzzy clustering algorithm. These clusters are displayed in Fig. 1b, along with general descriptions of each cluster.

The two datasets are used for different tasks of the study. The complete continental coverage of the subcatchment dataset (Fig. 1b) enables the identification of Australia-wide patterns in catchment attributes. The LSTM model requires rainfall-runoff data pairs for each timestep, and these are provided in the CAMEL-AUS data for specific regions of the continent only (Fig. 1a). By associating these datasets with each other, the LSTM results can be interpreted with respect to the continent-wide

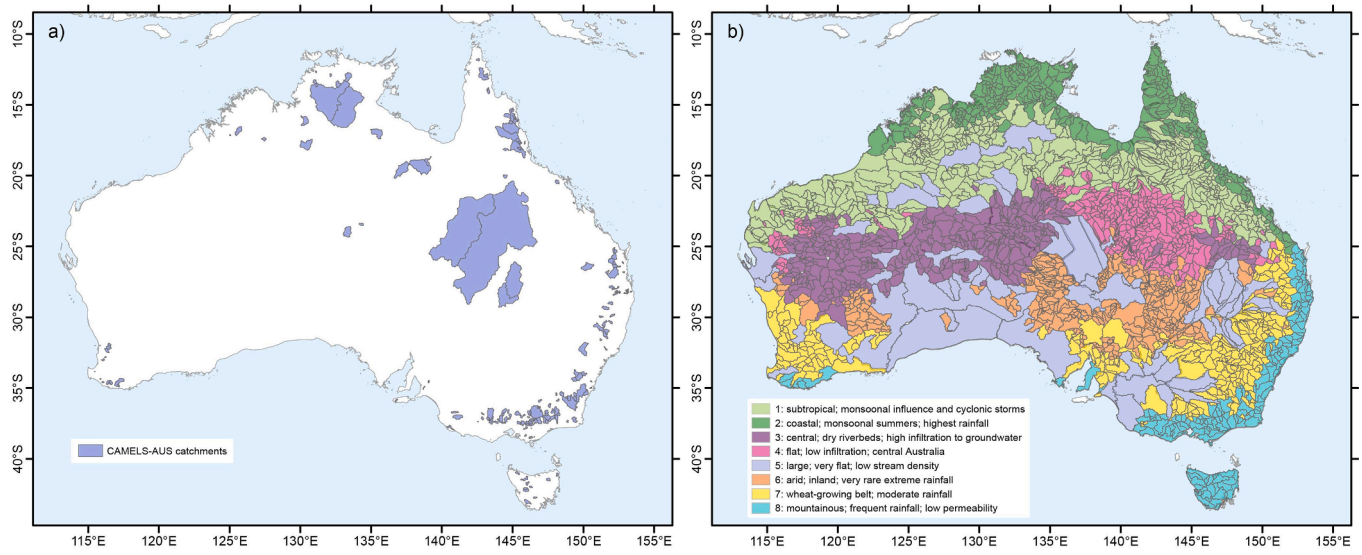


Fig. 1. Location map. a) CAMELS-AUS catchments (n = 222), and b) subcatchments coloured by cluster (n = 2816; recreated from Jaffrés et al., 2022).

Table 1
Subcatchment attributes from Jaffrés et al. (2021) used in study.

Attribute	Unit	Short name (for figures)	Description
Annual rain days	days	Ann rain days	Average (mean) rainy days per year
Summer season (SS) rain	%	Rain % summer	% rain that falls in Nov–Apr
Rain *L*-moment 1 (L-m1)*	mm	Rain mean	Mean daily (rain day) rainfall
Rain L-m2*	mm	Rain SD	Standard deviation (SD) of daily (rain day) rainfall
Rain L-m4*	mm	Rain kurtosis	Kurtosis daily (rain day) rainfall
SS evap L-m1	mm	Evap summer	Mean daily evaporation (Nov–Apr)
SS evap L-m2	mm	Evap summer SD	SD of daily evaporation (Nov–Apr)
Winter season (WS) evap L-m1	mm	Evap winter	Mean daily evaporation (May–Oct)
WS evap L-m2	mm	Evap winter SD	SD of daily evaporation (May–Oct)
SS mean sea level pressure (MSLP) L-m1	hPa	MSLP summer	Mean daily MSLP (Nov–Apr)
SS MSLP L-m2	hPa	MSLP summer SD	SD of daily MSLP (Nov–Apr)
WS MSLP L-m1	hPa	MSLP winter	Mean daily MSLP (May–Oct)
WS MSLP L-m2	hPa	MSLP winter SD	SD of daily MSLP (May–Oct)
Upstream area	m^2	Area upstream	Total area of subcatchment and upstream subcatchments
Slope	°	Slope	Average subcatchment slope
Hill slope	°	Slope hills	Average slope in hilly areas
Longest flow path length	m	Length longest	Length of longest flow path
Vertical range	m	Elev range	Elevation variation in subcatchment
Elevation SD	m	Elev SD	SD of elevation in subcatchment
Mean subsurface permeability	log (m^2)	Subsurf permeability	Measure of ability to transmit water underground
Land surface value (LSV)	–	Land surface value	High value – high runoff; low value – high absorbance

* The rainfall *L*-moment attributes are based on rain days only (i.e. all days with 0 mm were omitted before deriving the variables).

attribute patterns.

3. Methods

This study consists of three steps: 1) unsupervised learning to extract, organise and visualise common patterns of catchment attributes across Australia, followed by 2) a continent-wide, deep learning time series model (LSTM), created to simulate the rainfall-runoff relationship at a large sample of catchments, and, finally, 3) visualising and investigating associations between catchment patterns and the relative success of capturing the rainfall-runoff relationship with deep learning.

3.1. Identification of common catchment attribute patterns

To determine catchment attribute patterns, traits from the 2816 subcatchments are input into the self-organising map (SOM) algorithm (Kohonen, 1990). The SOM is a type of unsupervised neural network that is proficient at nonlinear clustering, pattern extraction and visualisation, making it a popular choice in environmental sciences. It is able to identify nonlinear interactions between variables and produces clusters that are ordered smoothly in a visually-intuitive output. The SOM's resilience to noisy or missing data enhances its applicability for analyses with environmental data.

The SOM consists of an input layer and an output layer (or map), connected to each other by a set of updateable weights, as shown in Fig. 2. Each node on the input layer represents one of the input variables. The nodes of the output map form a regular, two-dimensional grid shape. At the start of training, the grid of map nodes is placed over the input data cloud in the high-dimensional data space. Training progresses as the weights are iteratively updated to more closely align the output node locations with areas of higher input data density. When training is complete, the high-dimensional node locations represent the most prevalent patterns – or combinations of variables – in the dataset. Each node of the trained output map represents a multivariate pattern present in the dataset. Each data observation is then attributed to its most similar (nearest) map node, resulting in ordered clusters of similar data items. Refer to Kohonen (2013) or Clark et al. (2020b) for more detailed information on setup, training and interpretation of a SOM.

In this study, the 2816 observations – each characterised by 21 variables representing catchment attributes – were input to a 36 × 36 SOM (1296 nodes). The grid size for a SOM analysis is based heuristically on a balance between error measures and the required level of information to be extracted (Kohonen, 2013). An iterative process was used to determine the map size, through experimenting with square maps of varying edge lengths (9, 15, 20, 25, 30, 36 and 40). Consideration of the quantisation error (closeness of fit between the map nodes and the data items) and topographic error (ordering of similar data items nearby on the map) metrics led to the selection of a 36 × 36 node map. The 1296 nodes were determined to provide the optimal balance between a suitable representation of the topological structure of the data

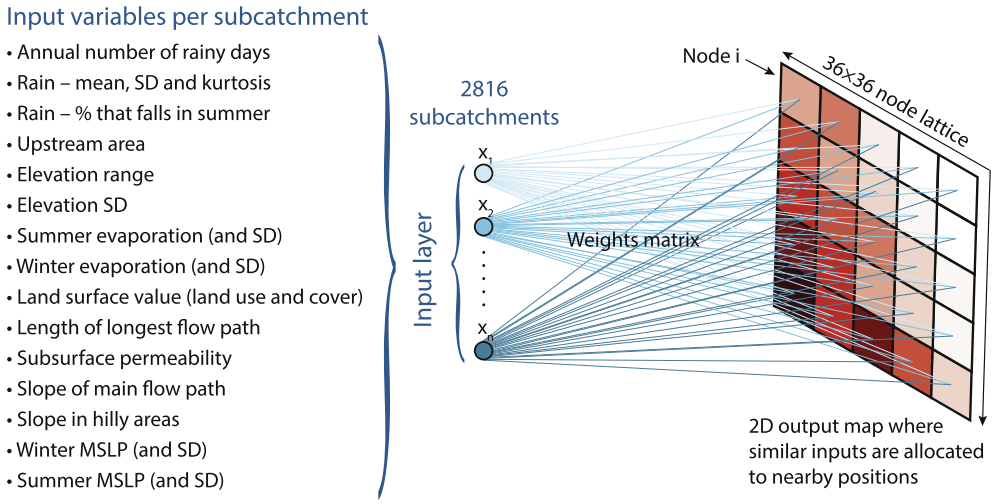


Fig. 2. SOM schematic.

and a manageable resolution of the output map. Prevalent patterns (combinations of characteristics) were determined through SOM training, with nodes on the output map serving as representative indicators. Similar input data items were systematically allocated to proximate positions on the map. The SOM was run with the Kohonen package (Wehrens & Buydens, 2007) in R, on a rectangular grid with a training length of 600 and initial learning rate of 0.05.

3.2. Continent-wide deep learning time series model (LSTM)

The LSTM is a deep-learning, recurrent neural network (RNN) specifically designed for time series analyses. An updated version of the RNN, the LSTM overcomes a common issue known as the vanishing/exploding gradient problem encountered in RNNs when applied over more than a few time steps. The LSTM makes predictions at the current timestep based on all previous conditions. It is recurrent in that it cycles over successive timesteps of input data before producing an output. The model architecture consists of an input layer, one or more hidden layers, and an output layer. 'Gates' are included in the form of additional nodes that regulate which information from the past is retained in a memory cell over long time periods, and which non-essential information is forgotten or discarded. Another gate regulates the flow of information from the memory cell to the hidden layer nodes, and eventually to the output nodes. At each iteration, the state of the memory cell is updated based on the previous system state and the current input, regulated by the gates. For a more detailed description of the LSTM training process, see Goodfellow et al. (2016) or Clark et al. (2020a).

The LSTM is gaining popularity in hydrological studies due to its ability to learn both short-term (response to meteorological events) and long-term (response to multi-year climatic variations) hydrological processes concurrently. Lagged input/output relationships can be found between, for example, rainfall, evapotranspiration, temperature and river flow without a requirement to pre-specify the length of the lagged response.

The LSTM is employed in this study to model the input/output relationships for predictors of streamflow as depicted in Fig. 3. Dynamic (n = 4) and static (n = 45) input variables are used as predictors for the target variable of streamflow (mm/day). The dynamic variables are input as daily time series of precipitation, actual evapotranspiration, minimum temperature and maximum temperature. The static input variables cover precipitation statistics (e.g. mean, seasonality, frequency), runoff characteristics, climate, soil, land cover and geological traits. All data for the LSTM are sourced from the CAMELS-AUS dataset, although the dynamic input variables originate from SILO (Scientific Information for Land Owners) daily climate time-series products (Jeffrey et al., 2001). A full list of the predictors used in the model is provided in Table A1.

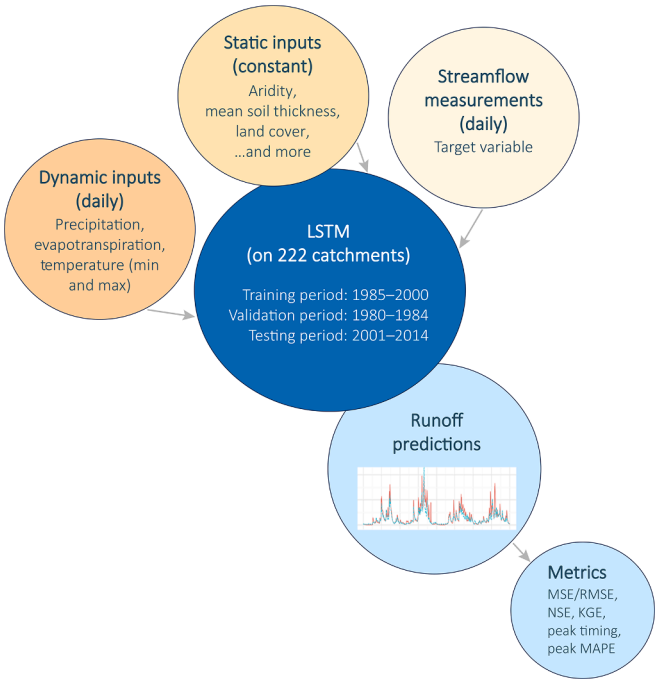


Fig. 3. LSTM flowchart.

The NeuralHydrology model is applied in this study to create a global LSTM to predict runoff for all CAMELS-AUS catchments concurrently. NeuralHydrology (Kratzert et al., 2022) – a large-scale, hydrologically-focussed deep learning platform – allows the standardised creation of global models on benchmark datasets, incorporating hundreds of catchments simultaneously. NeuralHydrology is an open-source Python LSTM package based on PyTorch, which has been specifically designed for hydrological applications and for use with CAMELS datasets. The package allows comparison of different models or different datasets, prediction in ungauged basins, uncertainty estimation and internal inspection of the LSTM cells for interpretability. It has been used in pioneering machine learning large-sample hydrology studies (Gauch et al., 2021; Lees et al., 2021) and is the current basis for Google FloodHub (Nearing et al., 2024).

The LSTM model uses the Nash–Sutcliffe efficiency (NSE; Hiscock & Bense, 2021) – the most common hydrological model performance metric – for training loss. In NeuralHydrology, the NSE loss is averaged over all the study basins. At each timestep, the static attribute variables are concatenated with the dynamic variables in the input layer. The model can then incorporate catchment-specific conditions into its predictions throughout the time series. For more information on model training, see the NeuralHydrology documentation (NeuralHydrology Team, 2024).

The dataset was split into training (01/01/1985 – 30/12/2000), validation (01/01/1980 – 30/12/1984) and testing (01/01/2001 – 30/12/2014) time periods for use in the global LSTM. The LSTM is trained on the training set, and the validation set is applied to monitor for errors during the selection of hyperparameters. The testing set is kept unseen until training is complete. This selection of training, validation and testing dates allows for the model to be trained on data obtained during both non-drought and drought conditions, and tested under similar settings.

Hyperparameters were chosen with the use of TensorBoard (Abadi et al., 2015), based on the minimisation of errors in the validation set and mitigating signs of overfitting. The range of values tested for each hyperparameter is listed here, with the chosen value in bold: number of hidden nodes [36, 64, **128**, 192, 256, 320], number of training epochs [up to 50, **30**], batch size [64, **128**, 256, 512], learning rate (steady [0.001], **declining [0.001–0.0001]**, declining [0.005–0.0005]), dropout proportion [0.2, **0.4**, 0.5], and lookback length [180, 270, **365** days]. In particular, a network with 128 nodes resulted in the most consistently low errors over the final training epochs (larger networks tended towards error spikes). The model showed signs of overfitting after 30 epochs (increased error in the validation set even though the training set error continued to decline). Further, a declining learning rate reliably provided the lowest loss on the validation data, and 365 days of lookback resulted in the highest median NSE across all basins. The NeuralHydrology model was run on Google Colab (Bisong, 2019).

The LSTM predictions on the testing dataset are compared with observed streamflow data. A range of metrics are recorded for each catchment. The metrics used to assess runoff prediction performance in this study are: 1) NSE, measuring the overall match between the observed and measured streamflow, 2) peak timing, quantifying the accuracy of predicting when a high-flow event occurs, and 3) peak MAPE (mean absolute percentage error), evaluating the prediction of high-event magnitudes.

3.3. Visualising associations between catchment attributes and LSTM performance

The CAMELS-AUS catchments incorporated in the global LSTM for modelling runoff are spatially correlated with the subcatchments used in the SOM for determining catchment attribute patterns. This process involved matching subcatchments to their respective CAMELS-AUS catchments using open-source QGIS software (v3.32.3-Lima). Each catchment in CAMELS-AUS was assigned corresponding subcatchments based on geographical relationships, including intersection, overlap, containment or equality. Following the initial matching procedure in QGIS, all combinations underwent manual verification to identify and rectify any flawed pairings. Removed pairings predominantly related to subcatchments from adjacent basins that marginally overlapped with CAMELS-AUS because of differences in catchment delineation (i.e. differences in the digital elevation models). Additionally, downstream subcatchments with only minor, partial overlaps were omitted, if the corresponding CAMELS-AUS catchment covered several upstream subcatchments.

4. Results

4.1. Investigation of continent-wide attribute patterns

Prevalent patterns of catchment attributes in the subcatchment dataset were identified with the 36 × 36 SOM created from the 21-dimensional attribute data. A visualisation of the results is shown in Fig. 4, on which the 1296 (36 × 36) node grid is repeated 21 times. The nodes are coloured according to the magnitude of a single input variable on each replicate. High values are depicted in purple and low values in white. Each of the 1296 SOM nodes represents a unique combination of the 21 variables, consisting of the value of each variable at that node on each of the replicate grids. For example, a node in the lower-left corner represents a pattern of low subsurface permeability, and low summer and winter mean sea level pressure (MSLP; mean and SD), as indicated by the white colour of these nodes on the respective replicate grids. In contrast, the same lower-left node on other grids is mid-hued (e.g. mean rainfall) or dark in colour (e.g. percentage of rain that falls in summer), describing the unique weather and land surface characteristics represented in this region of the map. Thus, the 1296 nodes reflect 1296 patterns, which vary gradually along the coordinates of the grid, with similar (diverse) combinations located close to (far from) each other in map space. A reciprocal figure that highlights the map regions corresponding to low values of each variable is shown in Fig. A1 of Appendix A.

The 2816 subcatchments are then matched to the most similar node on the SOM based on the values of their 21 attribute variables. Because the nodes of the SOM are constrained in a mesh formation (see Clark et al., 2020b), some nodes of the trained map will fall in regions of data space where no subcatchments match the specific combination of attributes. On this grid, approximately 12% of the cells do not have subcatchments assigned. These cells are coloured by the values of data space they occupy regardless of whether subcatchments are assigned to them because the patterns of attributes continue smoothly across the map. The same node on each grid of Fig. 4 represents the same set of subcatchments. The subcatchments allocated to nodes in the lower left corner of the grid will be those with most rainfall during the summer months, low subsurface permeability, and low MSLP in both summer and winter. Distributions of variables for subcatchments allocated to this corner of the map are shown in the boxplots of Fig. 5a. Characteristics of subcatchments attributed to the other corners of the grid are shown in the remaining three panels, with the variables of highest and lowest magnitude highlighted.

The SOM map can be used to investigate the patterns of variables associated with the clusters of Jaffrés et al. (2022). On Fig. 6, the colour tint of the SOM nodes relates to the number of subcatchments in each cluster that are allocated to them. For visualisation purposes, a separate grid is shown for each cluster. The distinctness and cohesiveness of the clusters are evident – except for Cluster 5, which is distributed sparsely over the grid. As discussed in Jaffrés et al. (2022), the subcatchments of Cluster 5 have a disparate spatial distribution and are generally characterised as being large with very low stream density. The seven other clusters are organised in distinct areas of the SOM.

With this information, it is possible to investigate in detail the properties of the subcatchments that make up each cluster through the intersection of Fig. 4 and Fig. 6. As shown in Fig. 7, individual grids from the SOM output can be overlaid on the cluster structure grids, indicating for example that Cluster 1 has a high percentage of its rainfall in the summer months, whereas Cluster 7 does not. The other variables and clusters can be compared in the same way, as shown in Fig. A3. This

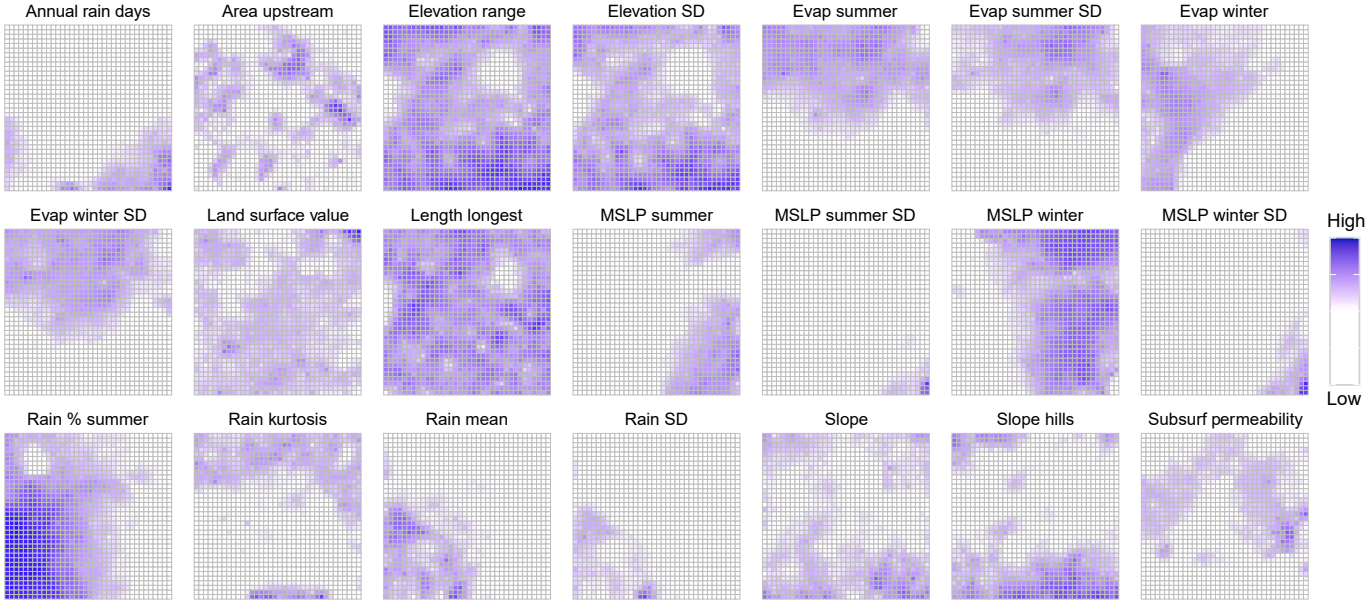


Fig. 4. SOM output. The grid is repeated for each of the 21 input variables, with dark colouring indicating high values of each variable. A portion of the dataset is represented by the same square on each of the grids (i.e. node (2,2) represents the same subset of data on the 'elevation SD' grid as on the 'rain mean' grid).

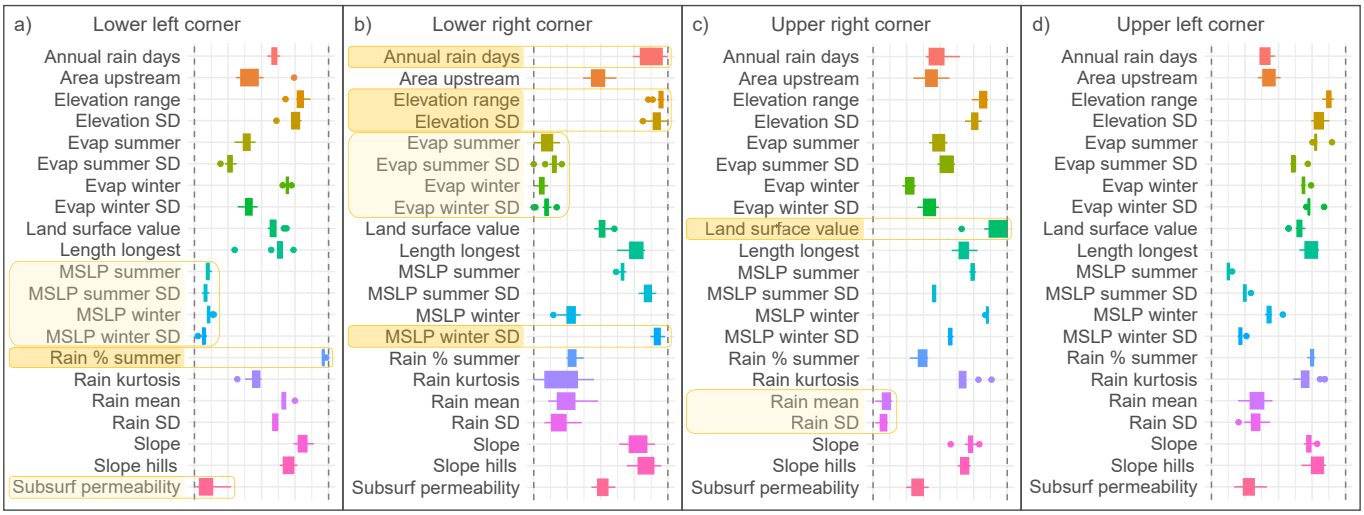


Fig. 5. Boxplots of characteristics for subcatchments attributed to each 2 × 2 set of corner nodes on Fig. 4. The variables with high values in each map corner are highlighted in dark yellow, and those with low values in light yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 6. The SOM grid coloured by the cluster membership of the subcatchments allocated to each node. The colours correspond to the map on the right (from Fig. 1b).

process provides a detailed view into the variations of attributes that comprise each cluster.

4.2. Deep learning streamflow predictions (global LSTM)

A global LSTM is trained on predictor and streamflow data from the 222 catchments of the CAMELS-AUS dataset (as outlined on Fig. 1a). This model encompasses all 222 rainfall-runoff time series in a single LSTM that can produce predictions on an individual-catchment basis. Metrics of prediction performance are determined for each station on the testing portion of the dataset.

Runoff predictions for a four-year segment of the testing data are shown in Fig. 8 for a varied subset of the catchments. The model predicts unique runoff patterns (blue) at each station, matching relatively consistently with the respective observations (orange). In some cases, the predicted runoff does not reach the observed peaks and in other instances it exceeds them. In Fig. 8a, station 215002 (Shoalhaven River at Warri) experienced drought conditions for the first two years of this four-year plot. The model accurately captures the state of drought as well as the response to renewed rainfall events as the drought eased in the latter years. For station A5040517 (First Creek at Waterfall Gully; Fig. 8f), the model produces unique streamflow predictions where there are four consecutive years of missing measurements.

4.3. Visual analysis of relationship between deep learning predictions and catchment attributes

The LSTM metrics were transferred onto the SOM grid to investigate model performance in the map space. To facilitate this, the CAMELS-AUS catchments were matched with subcatchments defined by Jaffrés et al. (2021) in a one-to-one, one-to-many or many-to-one configuration. The LSTM streamflow metrics of CAMELS-AUS catchments were then attributed to the corresponding subcatchments.

This subset of subcatchments is represented in Fig. 9, with each data point positioned on the SOM grid based on the subcatchment attribute patterns. The data points are coloured by LSTM prediction performance metrics obtained from the previous step. The performance metrics represent: the overall match of the predictions to the observed time

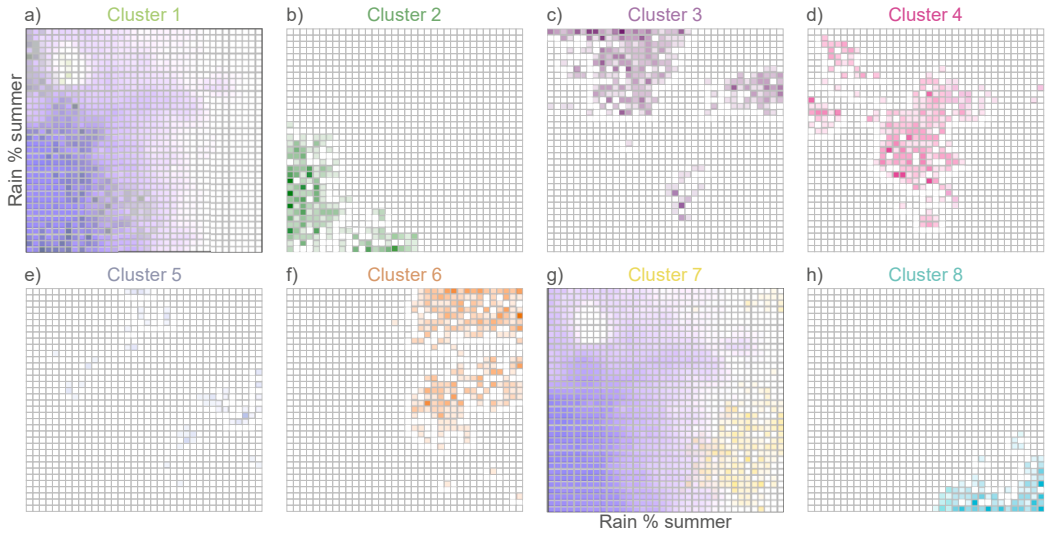


Fig. 7. SOM grid for the 'percentage of rainfall that falls in summer' variable (from Fig. 4) overlaid on cluster information (from Fig. 6). As examples, Cluster 1 falls within the 'high percentage of rainfall in summer' portion of the SOM, and Cluster 7 falls outside of it. Detailed information on the distribution and combinations of variables related to each cluster can be abstracted from this visualisation.

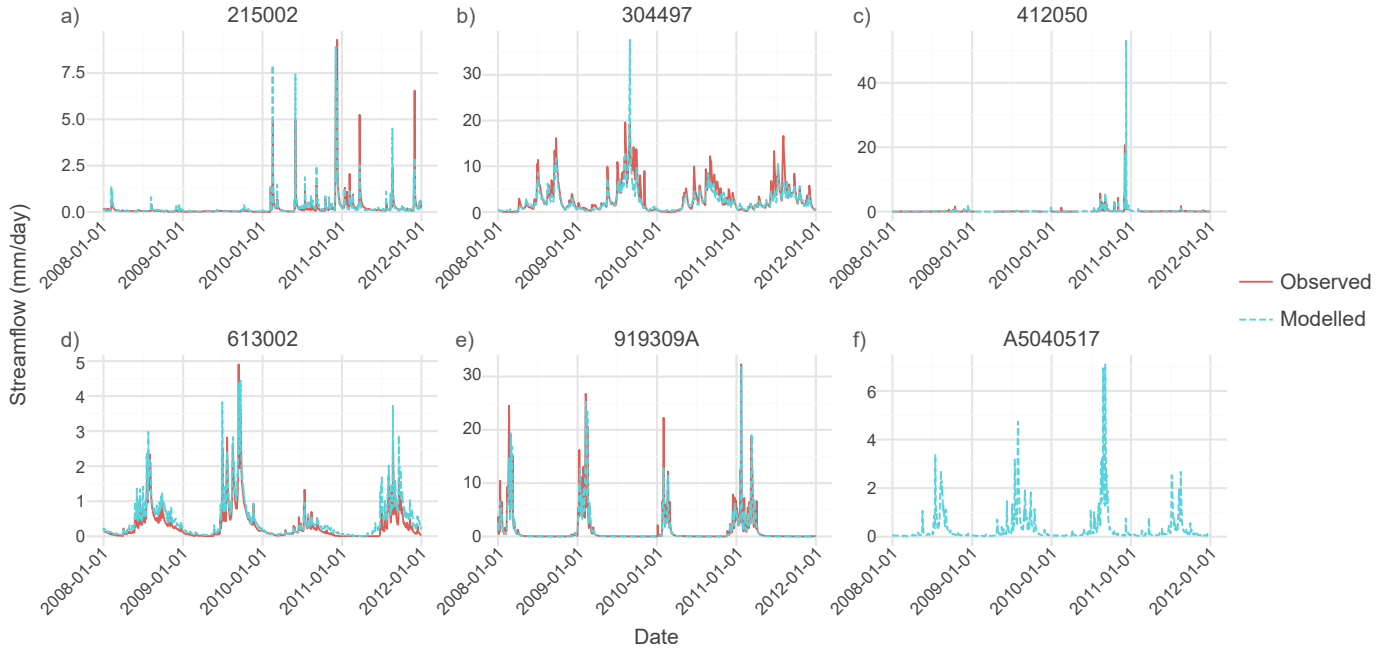


Fig. 8. LSTM prediction results (blue) compared to observed streamflow (orange) for a selection of six of the 222 CAMELS catchments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

series (NSE; Fig. 9a), the accuracy of predicting the timing of high-runoff events (peak timing; Fig. 9b), and the magnitude of high-runoff events (peak MAPE; Fig. 9c). Turquoise colouring indicates subcatchments with better predictions, and orange signals subcatchments with weaker, or suboptimal, performance. For NSE, the split between *good* and *suboptimal* predictions is made at NSE = 0.65. For the peak timing and peak MAPE metrics, the split is at the respective median values (0.37 days and 44.4%).

Based on this colour scheme, the lower left and right edges of the map tend to exhibit better NSE values than the centre region. Whereas the centre of the map uniformly indicates *suboptimal* prediction performance, the results along the lower edges are more variable. Consequently, subcatchments in this lower region need to be assessed with care. Considering the overall match of predictions to observations (NSE; Fig. 9a), a high proportion of the subcatchments along the lower edge obtain *good* predictions. However, subcatchments in the lower left corner of Fig. 9b appear to perform more poorly when assessed in terms of the timing of high-runoff events than when assessed by NSE. The LSTM also has difficulty predicting the height of runoff peaks in many subcatchments in the lower left and right of the grid (Fig. 9c).

The metrics grids can be overlaid on the SOM individual variable maps for a visual comparison of distributions. In Fig. 10, the LSTM metrics from Fig. 9a are draped over grids from Fig. 4 to investigate the relationship between each catchment attribute and the LSTM prediction performance (in terms of high NSE). Fig. 10 shows this for a selection of the grids only. A corresponding figure displaying the relationship between the NSE metrics grid and all attribute variables is provided in Fig. A2 of Appendix A.

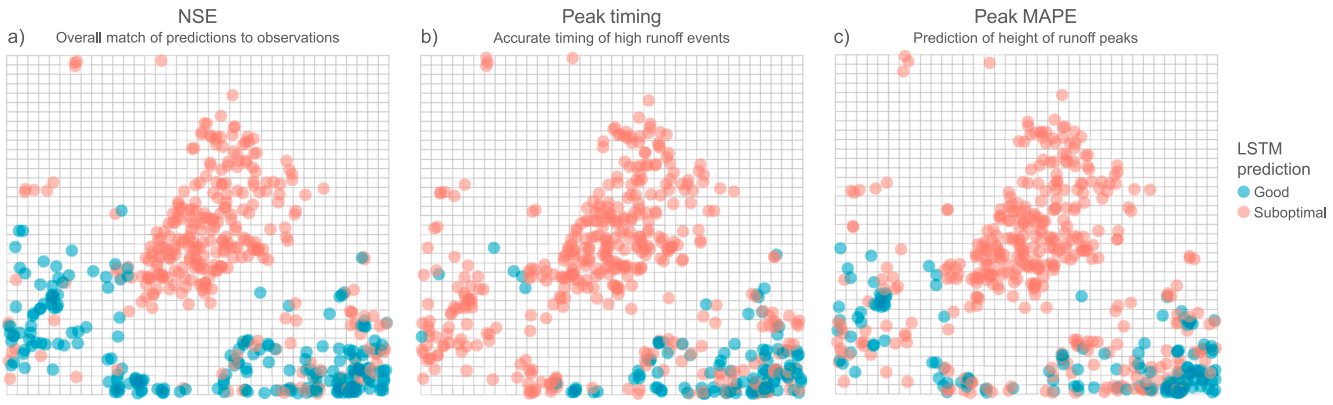


Fig. 9. LSTM metrics on the SOM grid: Nash-Sutcliffe efficiency, accuracy of peak runoff timing, and peak magnitude (MAPE). Each data point on the jitter plot represents a subcatchment, coloured by the prediction performance of the LSTM on the associated CAMELS-AUS catchment. Predictions with NSE $>$ 0.65 – and below-median peak timing error ($\leq$ 0.37) or peak MAPE ($<$ 44.4%) – are considered '*good*'. Only subcatchments with associated LSTM metrics are shown.

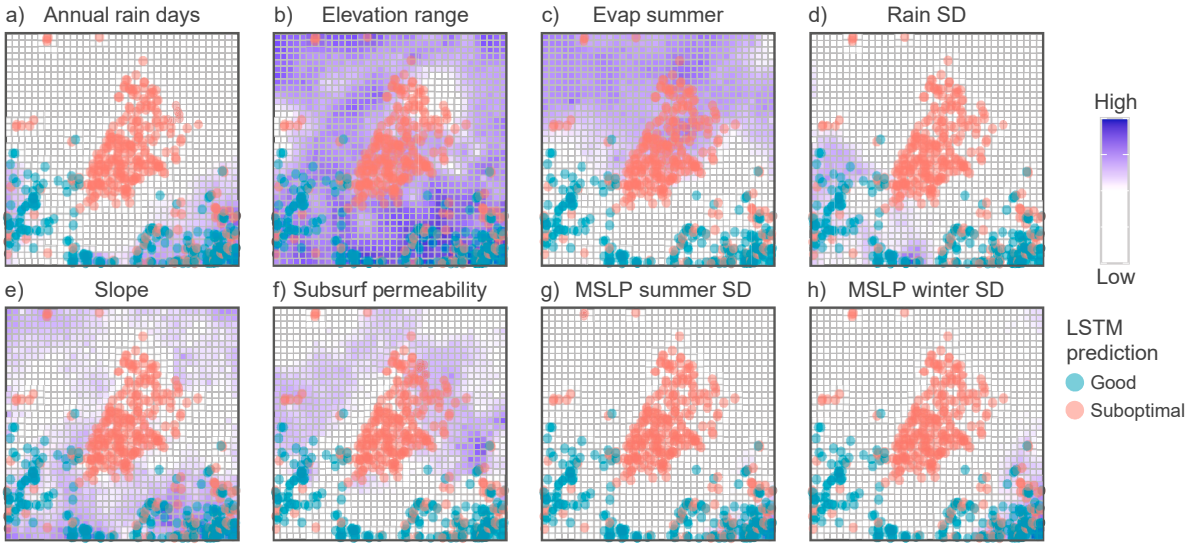


Fig. 10. Overlay of LSTM prediction performance metrics (NSE) on the SOM output for eight catchment attributes.

The distribution of high values (purple) of annual rain days in Fig. 10a matches well with the distribution of *good* LSTM prediction performance (turquoise) on the superimposed metrics grid. Following this method, in addition to frequent rainfall, better overall prediction performance could also be expected at catchments with a large elevation range and steeper slopes, highly variable rainfall and MSLP, low summer evaporation, and low subsurface permeability, as shown in the additional grids on Fig. 10 (refer to Fig. A1 for distributions of low values of each variable). This indicates that the model was proficient at capturing dynamic weather patterns, storms and precipitation, and the resulting impact on streamflow, especially in hilly or mountainous areas where the topography likely induces a relatively quick response.

The region of weaker LSTM prediction performance corresponds with flatter catchments, higher summer evaporation and high subsurface permeability. This suggests that the model was challenged by predicting streamflow responses in catchments with slower responses and where infiltration to groundwater tends to affect the amount of surface runoff. If the metrics grid from Fig. 9b were used instead, the resulting comparison would indicate the model facing challenges in accurately predicting the timing of peak runoff events in areas of high summer rainfall.

Any subcatchment with some available attribute information can be located on the SOM grid based on its characteristics, regardless of whether the catchment was part of the training dataset. The SOM algorithm is resilient to large amounts of missing data and, consequently, information for a subset of the 21 variables is sufficient for this task. The performance of an LSTM for predicting runoff at this subcatchment could then be estimated. In Fig. 11, the scaled attribute data for two subcatchments are shown, along with the location that each subcatchment would be ascribed to on the SOM grid based on these characteristics. From the metrics grid, it is evident that one of the subcatchments falls within a region typified by *suboptimal* prediction performance (centre of the map) and the other in a region associated with *good* prediction performance (lower-right edge).

The cluster structure can also be compared to the LSTM metrics grid, as shown in Fig. 12. Clusters that correspond to the higher NSE region of the grid (from Fig. 9a) are Cluster 2 (northern, coastal, monsoonal summers) and Cluster 8 (mountainous, southern, frequent rain). A proportion of members of Cluster 1 (subtropical, monsoonal influence) and Cluster 7 (moderate rainfall, wheatbelt) are also within this region. Subcatchments in Cluster 4 (flat, central, low infiltration) are located in the region of the grid typified by low NSE values. In terms of predicting the timing and magnitude of peak runoff (if Fig. 9b and Fig. 9c were used for this task instead of Fig. 9a), the performance of the LSTM on the catchments in Cluster 2 would appear lower than when assessed by NSE (Fig. 9a).

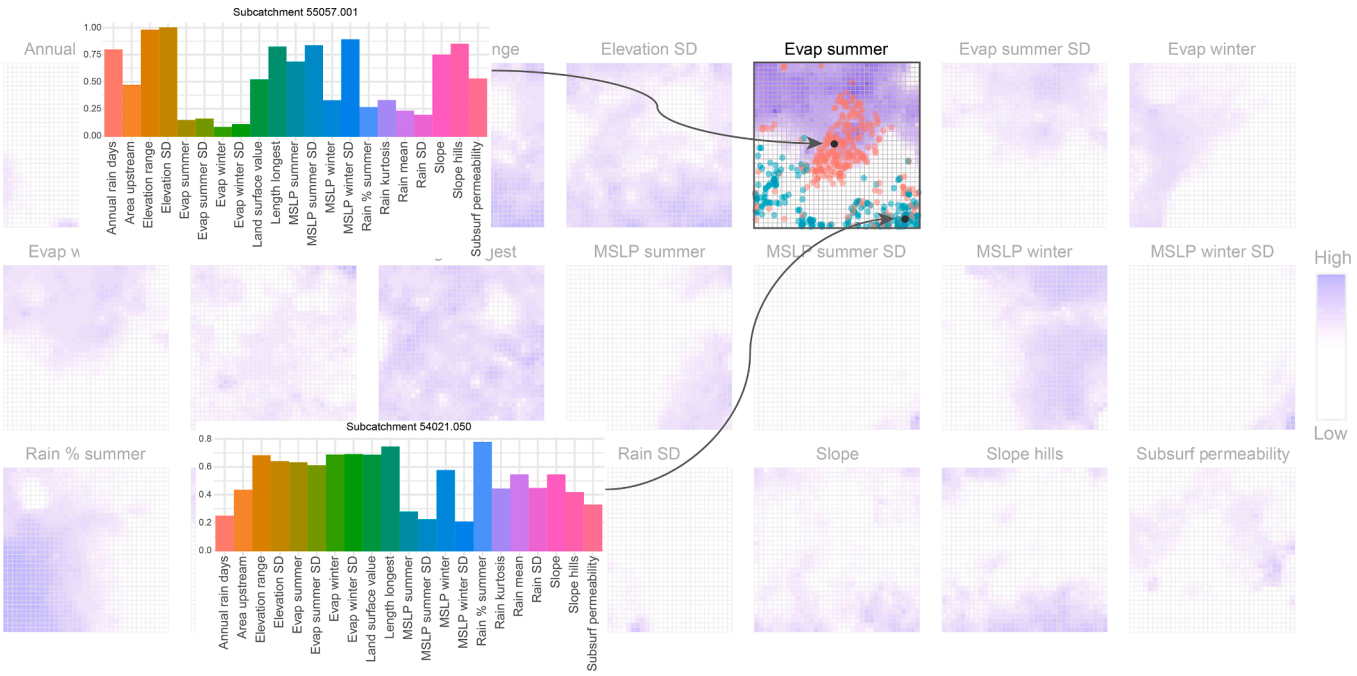


Fig. 11. Subcatchments not used in SOM training can be assigned to a node on the map based on their attributes. LSTM metrics associated with that region can then be checked.

Fig. 12. LSTM metrics (NSE results overlaid on Cluster 2) and cluster structure.

5. Discussion

This study integrates contemporary advances in large-sample machine learning hydrology to perform a continental-scale investigation into the relationship between deep learning prediction performance and Australian catchment characteristics. A SOM has been created to establish a mapping of the 2816 subcatchments onto 1296 ordered nodes. This mapping becomes the basis for visualisations throughout the study. Building on the fuzzy clustering of the subcatchments in Jaffrés et al. (2022), each node of the SOM can be considered a sub-cluster, representing small groups of subcatchments with very similar attributes. The groups are arranged in a two-dimensional space, with similar groups closer together. Values of the catchment attributes flow smoothly across the grid, providing an intuitive visual analysis of the combinations of attributes in each region of the map. Exhibiting the structure of the fuzzy clusters on the grid enables the intervariable combinations characterising each cluster – and the variations within the clusters – to be explored. Individual catchments are positioned on the map based on their attribute values and can be compared with other catchments by map location. Finally, a global LSTM incorporating static variables allows unique predictions to be made at each catchment, informed by the specific local attributes. LSTM performance is assessed with respect to catchment attributes and cluster membership. This multi-catchment study demonstrates the scalability advantages of machine learning techniques for deriving hydrological insights at a continental scale.

5.1. Relationship between catchment attributes and LSTM performance

The global LSTM performed well in catchments characterised by frequent, variable or high rainfall, corresponding to the areas of the SOM grid covered by Cluster 2 and Cluster 8. Jaffrés et al. (2022) showed that subcatchments in Cluster 8 (situated in the lower-right corner of the SOM) experienced much more frequent rainfall (117.3 days/year on average) than any of the other groups. However, rainfall in that cluster

also has a relatively low mean and variability (see supplementary Table A.3 in Jaffrés et al., 2022), suggesting that precipitation was generally of low intensity. In conjunction with relatively high MSLP throughout the year (i.e. limited and fast-moving cold fronts and low-pressure systems) and steep terrain, rare high-intensity rainfall events are expected to produce short-lived periods of elevated runoff in Cluster 8 catchments. In comparison, Cluster 2 subcatchments (lower-left corner of SOM) are typified by the highest rainfall mean and variability but only a moderate number of rain days. Combined with low MSLP throughout the year, these properties are representative of regions affected by intense rainfall events such as those produced by atmospheric lows (including tropical cyclones).

The global LSTM was less accurate in capturing the rainfall-runoff relationship in catchments with slow, infrequent flows, higher subsurface permeability and extensive summer evaporation. This finding is consistent with knowledge about the data-generating system. In the flat, dry, inland catchments of Cluster 4, a large proportion of the daily observations will have no precipitation or flow, reducing the number of non-zero rainfall-runoff pairs available for training the model. The high subsurface permeability suggests a strong interaction between surface water and groundwater, indicating a complicated relationship that cannot be recognised by the model because no groundwater component is included. Further, soil permeability is strongly associated with evaporation and rainfall processes, although their relationship is complex and nonlinear. Fully saturated soils inhibit uptake of rainfall, which is thus nearly entirely converted to runoff. Conversely, a decrease in soil moisture enhances water infiltration, thus reducing runoff. The infiltration potential is typically greatest when antecedent soil moisture is low. However, long dry spells often produce soil crusting which diminishes surface permeability (Jaffrés et al., 2021).

5.2. Comparison of results with previous studies

An important consideration – beyond identifying catchment types well-suited to global LSTMs – is whether these differ from regions that are well-represented by other hydrological models. The results of the global LSTM modelling can be compared with results from Clark et al. (2024), in which 496 diverse catchments across the Australian continent were modelled with individual LSTMs and traditional conceptual models (WAPABA). The individual-basin scale was used for this study as this is the most common scale for conceptual hydrological models. The two sources (i.e. the CAMELS-AUS dataset and Clark et al., 2024) have 126 catchments in common, enabling a rough comparison between results from these individual catchments. This comparison is shown in Fig. 13, with results coloured by cluster. Note that this is for rough comparison purposes only because a difference in the dates of the testing period makes the results not straightforwardly comparable. The catchments most commonly available for comparison are in Cluster 7 and Cluster 8, with many points surrounding or above the 1:1 line on Fig. 13. This indicates that the information shared across the catchments in the current global LSTM has led to a similar or better representation of the rainfall-runoff relationship at many of these catchments when compared to the individual LSTMs (Fig. 13a) and traditional, conceptual models (Fig. 13b).

The findings of Mathevet et al. (2020) and Yao et al. (2023) also suggest that hydrological model predictions are influenced by specific catchment characteristics and hydrometeorological conditions. Mathevet et al. (2020) investigated whether the prediction performance of two conceptual models differed depending on catchment characteristics or hydrometeorological processes for over 2000 basins worldwide. They found that performance was dependent on hydroclimatic conditions, with both model types performing less effectively under arid or dry conditions. Yao et al. (2023) trained an LSTM on 671 US catchments, including static catchment attributes, to enable predictions in the Tibetan plateau. Their study indicates that attributes such as soil and geology had less influence on model accuracy than climate data. However, the influence of static properties is expected to become more prominent during periods of low flow, when the relative contribution of groundwater to streamflow tends to be greater.

5.3. LSTM performance assessment notes

It is important to note that larger CAMELS-AUS catchments encompass several subcatchments, as shown in Fig. A4. For example, CAMELS-AUS contains catchments that almost completely cover two very large, central Australian basins: Diamantina River (NSE = 0.46) and Cooper Creek (NSE = 0.61). These two CAMELS-AUS catchments overlap with 94 and 114 subcatchments respectively. Consequently, LSTM model performance for each of these two catchments (and other CAMELS catchments) is represented by multiple points in each subplot of Fig. 9. The subcatchments of these two basins mostly belong to Cluster 4 and Cluster 6, situated towards the centre of the grid (Fig. 6d,f). However, the points are quite dispersed, reflecting the differing attributes of each subcatchment. Further, the NSE threshold of 0.65 is relatively stringent. For example, model performances of $0.50 < \text{NSE} \leq 0.65$ could be classed as *satisfactory* rather than *suboptimal*. This NSE range would reclassify

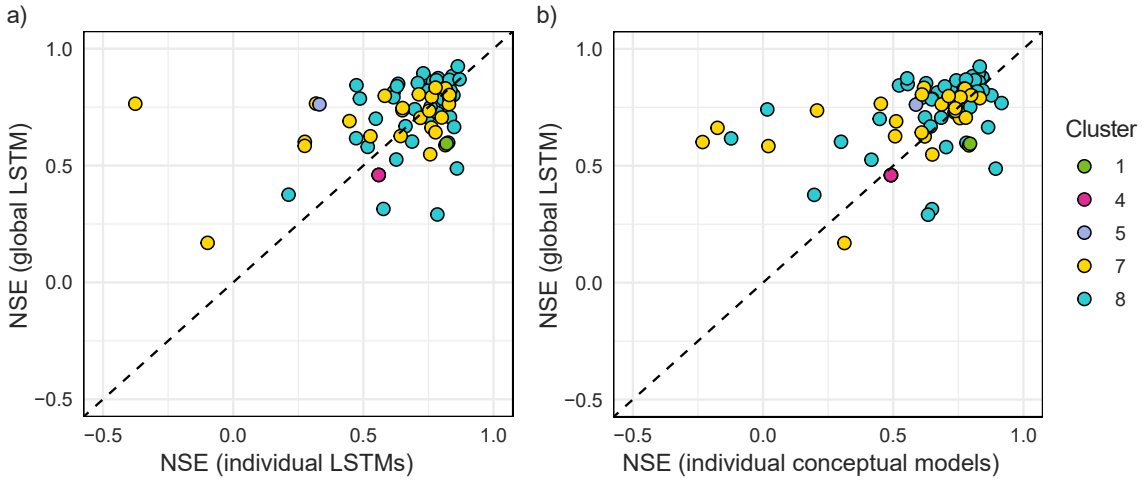


Fig. 13. Comparison of global LSTM results to a) individual-catchment LSTMs and b) conceptual models from Clark et al. (2024). Each point represents a catchment, coloured by its cluster membership from Fig. 1.

approximately 2/3 of the suboptimally performing CAMELS catchments as *satisfactory*. Therefore, the higher proportion of *suboptimal* prediction results in the centre of Fig. 9 is influenced by both the applied thresholds and the larger area of some CAMELS catchments.

5.4. Non-stationarity of hydrological conditions

The non-stationarity of hydrological regimes complicates the development and application of all hydrological models. Australia has encountered shifting regimes as a result of the Millennium Drought (1997–2010), the most protracted drought on record in southeastern Australia. During this period, the region experienced a profound alteration in rainfall-runoff relationships within particular catchments, connected to a shift in the climate regime (Verdon-Kidd et al., 2014). Notably, some catchments are yet to revert to their pre-drought hydrological patterns, with enduring consequences (Fowler et al., 2022). This highlights the lasting impact of the Millennium Drought on the region's hydrological dynamics. Hydrological models, if trained on pre-drought data, may no longer accurately represent the rainfall-runoff relationships post-drought. Fig. 14 shows the observed and predicted hydrographs for one such gauging station. At this station, the rainfall-runoff regime had altered post-2010, and the model was less accurate in capturing the new hydrological processes when making streamflow predictions. Historical data collected before the drought may thus no longer adequately capture the full range of hydrological processes present under the changed regime. However, datasets containing only post-drought data will be relatively small and may have limited examples of extreme events or coverage of under-represented geographical areas. A difficulty therefore arises when attempting to use the historically available data to make predictions in the new regime. A dataset must be split sensitively into training, validation and testing sets – to enable these shifting relationships to be captured whilst not solely training on data before the drought and testing on the post-drought regime. Although this is an issue encountered in the deep learning modelling in this study, it is also an issue that affects traditional hydrological modelling.

5.5. Limitations

The data, model setup and evaluation in this study come with certain limitations. Streamflow data in CAMELS-AUS are available for a finite selection of basins and only extend to 2014. In this study, the training and testing datasets are split during the Millennium Drought, ensuring that both datasets partially overlap with drought conditions. When more recent post-drought data become available, further investigation may lead to a more practical split. Expanding the dataset with more recent rainfall-runoff data – including from additional catchments – would provide a richer foundation for model training.

Using the same metrics across all basins assumes uniform modelling priorities, which may not hold true. Clark et al. (2021) caution that the NSE can be skewed by a small number of data points, suggesting that more research is needed to refine the use and interpretation of hydrological metrics across large samples of catchments. Metrics designed to capture temporal shifts or broader trends beyond point-in-time accuracy might offer valuable insights. With the NSE, for instance, a shifted peak flow incurs greater penalties than a peak that is missed altogether (Magyar & Sambridge, 2022).

Further, the lack of groundwater information as an input limits the model's ability to account for surface water and groundwater interactions, which vary significantly across the continent. Previous studies (e.g. Rassam et al., 2013) have illustrated that incorporating these interactions improves model performance, particularly during periods of low flows. This is especially relevant for catchments typified by infrequent flow events (e.g. central Australia) and high subsurface permeability, where the inclusion of groundwater data may enhance predictive accuracy.

5.6. Future work

Future research could involve expanding this study to include rainfall-runoff paired data for as many catchments as possible. In this project, the homogeneity of the standardised CAMELS-AUS dataset was prioritised. However, there is potential to incorporate many more gauging stations across Australia. This would increase the coverage of metrics on the SOM, filling in more of the white space in Fig. 9. Including recent data (post-2014) may further strengthen results by enlarging the training dataset and providing more information on post-drought rainfall-runoff conditions. Integrating groundwater data could enhance the model's understanding of surface and subsurface water interactions, leading to better predictions in areas of high subsurface permeability. Investigating the evolution of cluster structures – especially whether they differ before and after the Millennium Drought – could offer

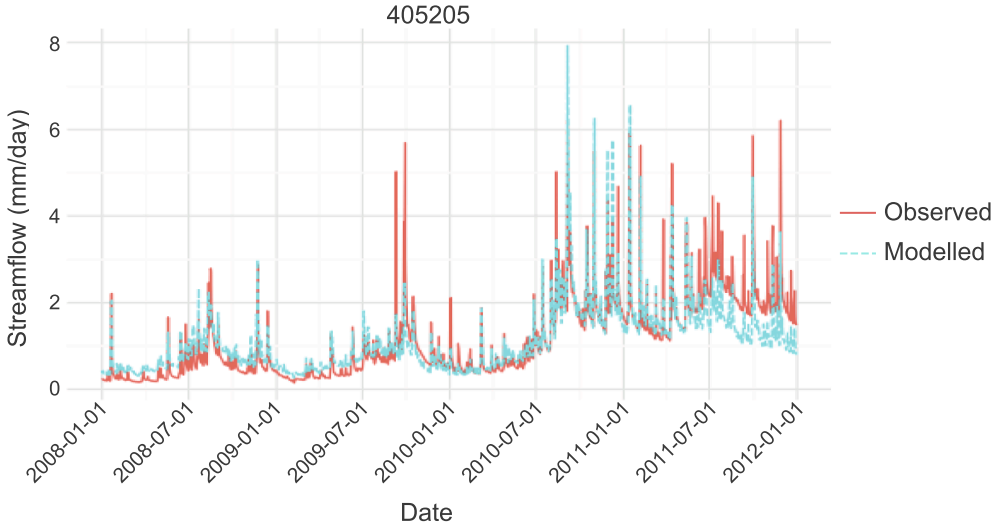


Fig. 14. Shifting runoff regime after the Millennium Drought.

additional insights. Finally, applying the results towards runoff predictions in basins with limited or no discharge measurements would be a valuable step toward sustainable water management in critical regions of Australia with limited historical monitoring records.

6. Conclusion

This study has produced a visual exploration of the link between catchment attributes and the success of a deep learning model (the global LSTM) at predicting streamflow for a diverse set of catchments across continental Australia.

Using an unsupervised clustering and visualisation algorithm (the SOM), patterns of catchment attributes have been mapped for 2816 subcatchments dispersed across the country. These attributes include factors such as terrain, rainfall patterns and subsurface permeability. Ordered, two-dimensional SOM grids provide information on the unique combinations of factors present across the country. The smooth transition amid patterns on the map allows a comparison between subcatchments in terms of their multi-variable characteristics. This initial step offers a comprehensive understanding of the varied combinations of environmental conditions present across the different catchment areas – and a visual tool for analysis of the prevalent catchment traits and comparisons between catchments.

A global LSTM was developed to predict streamflow at 222 catchments. By creating a single model for use at all catchments, the study builds on recent research supporting the integration of multiple hydrological time series to share information on common hydrological responses across catchments. Prediction results of the LSTM were evaluated at each catchment using metrics based on errors of mean, peak magnitude and peak timing estimation. The metrics have been mapped to the catchment attribute patterns of the SOM for a visual analysis of LSTM performance compared to prevalent Australian catchment characteristics.

The study found that the global LSTM performed best in certain types of catchments characterised by specific attributes. These include:

- **Frequent, variable and/or high rainfall:** Catchments experiencing frequent, variable or generally elevated rainfall events showed better predictive performance. This suggests that the model was effective in capturing the dynamics of high-intensity and seasonal precipitation and its impact on streamflow.
- **Variable MSLP:** Areas with high MSLP variations may experience more dynamic or rapidly changing weather patterns. Fluctuations in MSLP are associated with weather systems such as tropical cyclones or weather fronts.
- **Hilly or mountainous terrain:** Areas with rugged terrain were also associated with better prediction performance. The complex topography likely induces relatively quick flow responses to precipitation, which the LSTM is able to capture.
- **Low subsurface permeability:** Catchments with lower subsurface permeability – implying less infiltration and more surface runoff – were better predicted by the LSTM model.

Despite its overall effectiveness, the LSTM model encountered challenges in certain catchment types:

- **Flat catchments with infrequent, slow flows:** The rainfall-runoff relationship in these catchments was not well captured by the model. In these arid, inland catchments, there are fewer examples of the flow response to precipitation events (non-zero rainfall-runoff paired data) to train the model on. This impacts the model's ability to predict flow for given levels of precipitation. These catchments may therefore have unique hydrological characteristics that are not captured by the model.
- **High subsurface permeability:** Catchments with higher subsurface permeability – indicating faster infiltration and potentially less surface runoff as flow is transferred into the groundwater regime – presented challenges for accurate streamflow prediction.
- **High summer evaporation:** Evaporation is intrinsically linked to surface permeability and antecedent soil moisture. A decrease in soil moisture can enhance rainfall infiltration, thus reducing runoff. However, extended dry periods often promote soil crusting which diminishes surface permeability. The model may have difficulty capturing these complex dynamics.
- **Peak runoff in areas of high summer rainfall:** The model encountered difficulties in accurately predicting peak runoff in regions characterised by high summer rainfall. This could be due to the complexity of hydrological processes associated with the magnitude of peak flows in these areas.

In summary, although the global LSTM showed promising performance in predicting streamflow across diverse catchment types in Australia, its effectiveness varied depending on the specific environmental attributes of each catchment. The visual analytical tool provided by the SOM output aided in disentangling these continent-wide differences in performance and also allowed analysis at an individual catchment level. Understanding these patterns of performance can help inform the development and application of accurate deep-learning hydrological forecasting models that transfer knowledge amongst numerous catchments.

CRediT authorship contribution statement

Stephanie R. Clark: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jasmine B.D. Jaffrés:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

SRC would like to thank the CSIRO Digital Water and Landscapes initiative for their support of this project. Suncorp Insurance is acknowledged for making the Australian subcatchment dataset available for use in this project.

Appendix A

Table A1
LSTM input features. Refer to Fowler et al. (2021) for a description of each.

Type		Predictor name (as described in CAMELS documentation)
Dynamic	Climate	Precipitation, actual evapotranspiration, minimum temperature, maximum temperature
Static	Climate	Aridity, p_mean, pet_mean, p_seasonality, high_prec_freq, high_prec_dur, low_prec_freq, low_prec_dur
	Soil	Mean soil thickness, saturated hydraulic conductivity, erosivity (power of rainfall to cause soil erosion)
	Hydrology	Runoff ratio, baseflow index
	Geology	Unconsoldted, igneous, silicsed, carbnatesed, othersed, metamorph, sedvolc, oldrock, claya, sanda
	Land cover and vegetation (proportion of catchment occupied by land cover categories within the Dynamic Land Cover Dataset [DLCD])	lc01_extracti, lc03_waterbo, lc04_saltlak, lc05_irrcrop, lc06_irrpast, lc07_irrsuga, lc08_rfcropp, lc09_rfpastu, lc10_rfsugar, lc11_wetlands, lc14_tussclo, lc15_alpineg, lc16_openhum, lc18_opentus, lc19_shrbsca, lc24_shrbden, lc25_shrbope, lc31_forclos, lc32_foropen, lc33_woodope, lc34_woodspa, lc35_urbanar

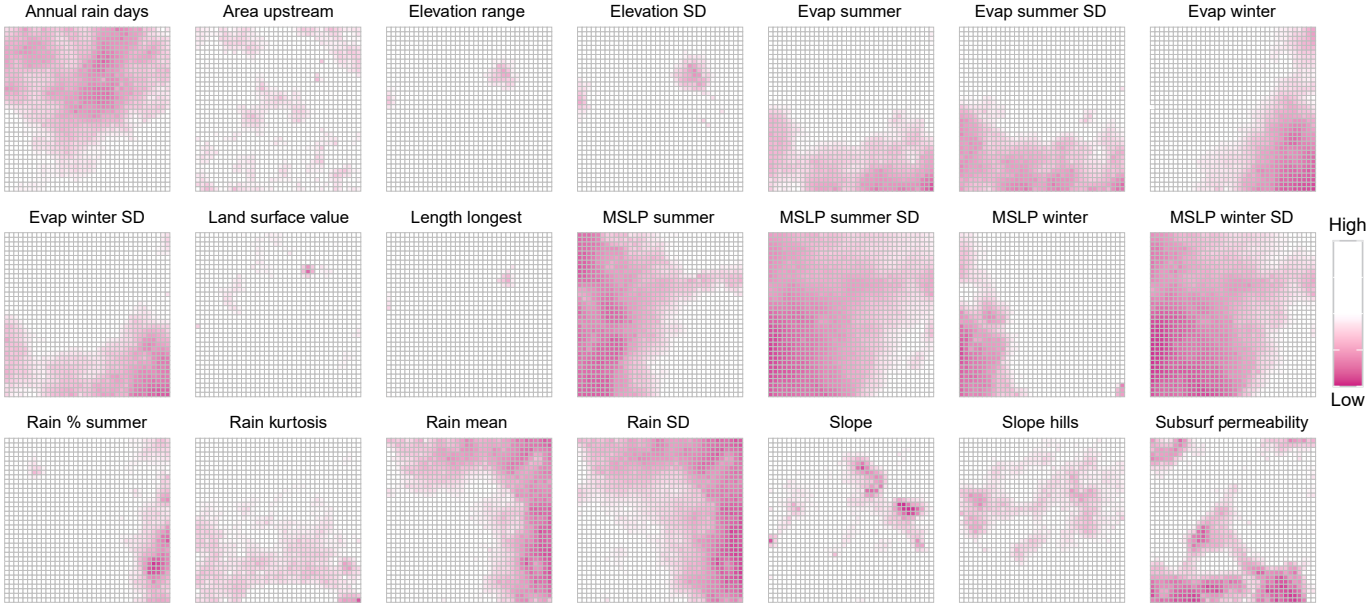


Fig. A1. SOM output grids coloured to highlight the low values of each variable. This figure is the reciprocal to Fig. 4, indicating the map regions representing the lower end of the scale for each characteristic.

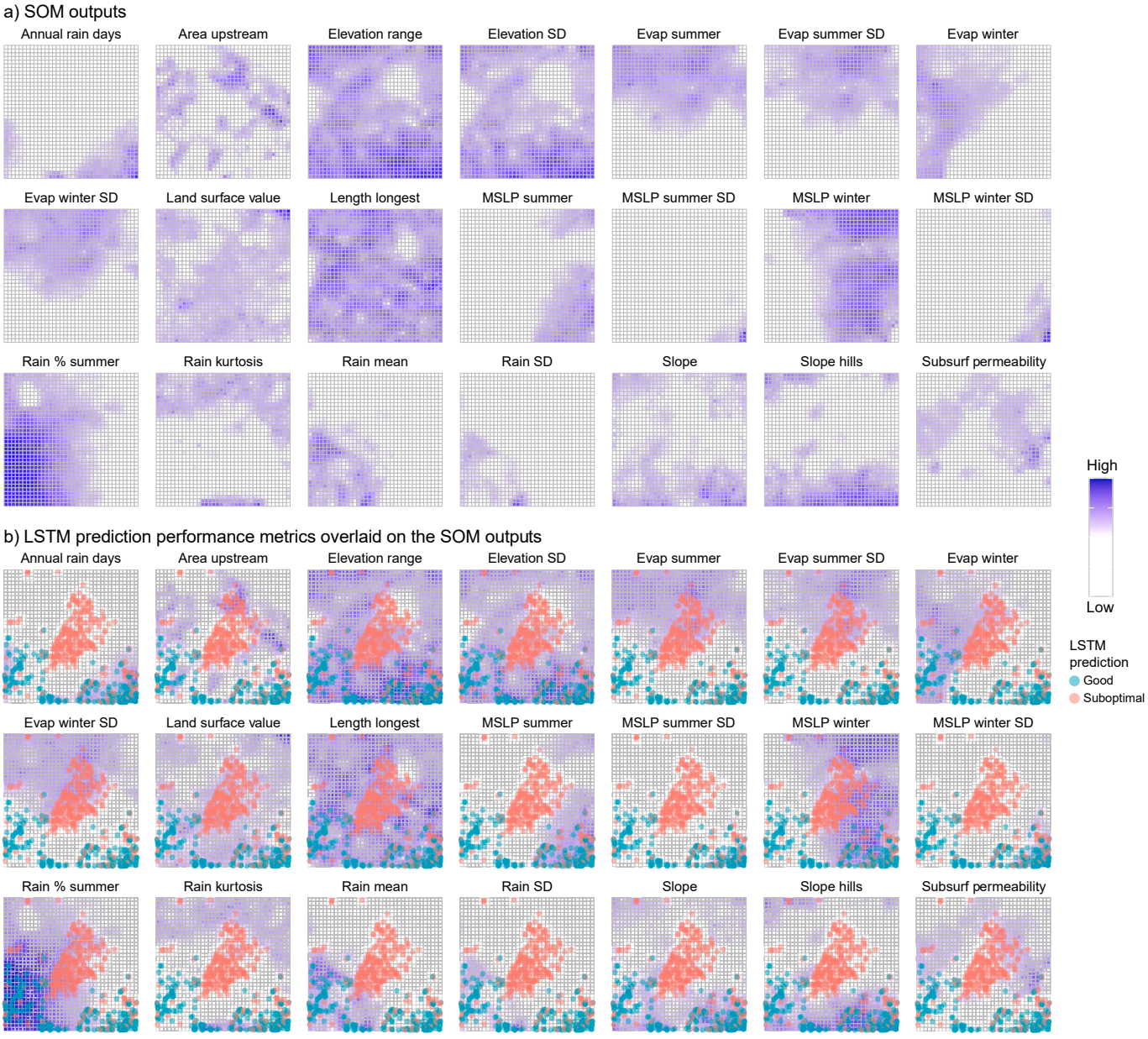


Fig. A2. Additional material for Fig. 10, showing LSTM metrics (NSE) on all attribute maps.

Fig. A3. Additional material for Fig. 12, showing LSTM metrics (NSE) on all clusters.

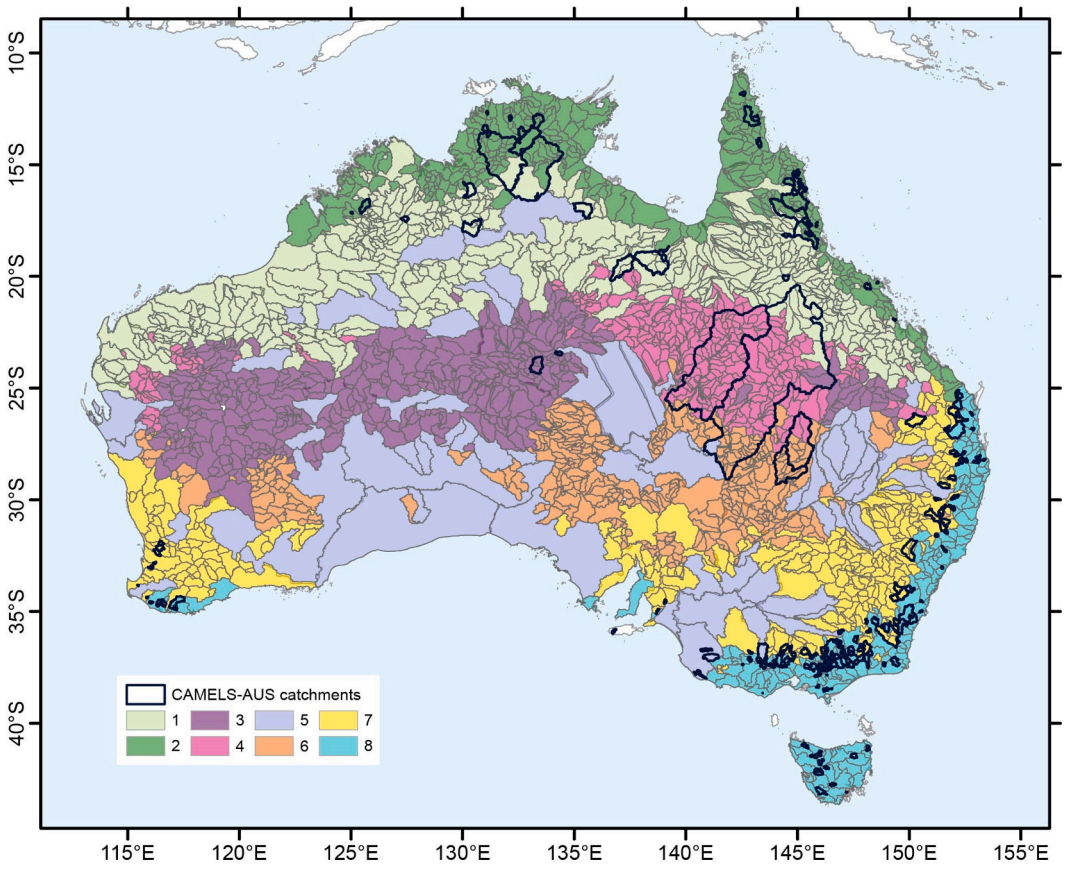


Fig. A4. CAMELS-AUS catchments (n = 222) overlying the subcatchments (n = 2816).

Data availability

Data will be made available on request.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Mountain View. Tensorflow, CA.

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci. 21 (10), 5293–5313.

Anderson, S., Radić, V., 2022. Interpreting deep machine learning for streamflow modeling across glacial, nival, and pluvial regimes in southwestern Canada. Front. Water 4, 934709.

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., Mai, J., 2023. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. Hydrol. Earth Syst. Sci. 27 (1), 139–157.

Bisong, E., 2019. Google colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Apress, pp. 59–64. https://doi.org/10.1007/978-1-4842-4470-8_7.

Brunner, M.I., Melsen, L.A., Newman, A.J., Wood, A.W., Clark, M.P., 2020. Future streamflow regime changes in the United States: assessment using functional classification. Hydrol. Earth Syst. Sci. 24 (8), 3951–3966.

Clark, S., Hyndman, R.J., Pagendam, D., Ryan, L.M., 2020a. Modern strategies for time series regression. Int. Stat. Rev. 88, S179–S204.

Clark, S.R., Lerat, J., Perraud, J.M., Fitch, P., 2024. Deep learning for monthly rainfall–runoff modelling: a large-sample comparison with conceptual models across Australia. Hydrol. Earth Syst. Sci. 28 (5), 1191–1213. https://doi.org/10.5194/hess-28-1191-2024.

Clark, S., Sisson, S.A., Sharma, A., 2020b. Tools for enhancing the application of self-organizing maps in water resources research and engineering. Adv. Water Resour. 143, 103676.

Clark, M.P., Vogel, R.M., Lamontagne, J.R., Mizukami, N., Knoben, W.J., Tang, G., Gharari, S., Freer, J.E., Whitfield, P.H., Shook, K.R., 2021. The abuse of popular performance metrics in hydrologic modeling. Water Resour. Res. 57 (9), e2020WR029001.

De la Fuente, L.A., Ehsani, M.R., Gupta, H.V., Condon, L.E., 2023. Towards interpretable LSTM-based modelling of hydrological systems. Hydrol. Earth Syst. Sci. Discuss. 2023, 1–36.

Fowler, K.J., Acharya, S.C., Addor, N., Chou, C., Peel, M.C., 2021. CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia. Earth Syst. Sci. Data 13 (8), 3847–3867.

Fowler, K., Peel, M., Saft, M., Peterson, T.J., Western, A., Band, L., Petheram, C., Dharmadi, S., Tan, K.S., Zhang, L., 2022. Explaining changes in rainfall–runoff relationships during and after Australia's Millennium Drought: a community perspective. Hydrol. Earth Syst. Sci. 26 (23), 6073–6120.

Frame, J.M., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L.M., Gupta, H.V., Nearing, G.S., 2022. Deep learning rainfall–runoff predictions of extreme events. Hydrol. Earth Syst. Sci. 26 (13), 3377–3392.

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

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, Massachusetts.

Hashemi, R., Brigode, P., Garambois, P.-A., Javelle, P., 2022. How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models? Hydrol. Earth Syst. Sci. 26 (22), 5793–5816.

Heudorfer, B., Liesch, T., Broda, S., 2024. On the challenges of global entity-aware deep learning models for groundwater level prediction. Hydrol. Earth Syst. Sci. 28 (3), 525–543.

Hiscock, K.M., Bense, V.F., 2021. Hydrogeology: Principles and Practice. John Wiley & Sons.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Jaffrés, J.B., Cuff, B., Cuff, C., Faichney, I., Knott, M., Rasmussen, C., 2021. Hydrological characteristics of Australia: relationship between surface flow, climate and intrinsic catchment properties. J. Hydrol. 603, 126911.

Jaffrés, J.B., Cuff, B., Cuff, C., Knott, M., Rasmussen, C., 2022. Hydrological characteristics of Australia: national catchment classification and regional relationships. J. Hydrol. 612, 127969.

Janssen, J., Ameli, A.A., 2021. A hydrologic functional approach for improving large-sample hydrology performance in poorly gauged regions. Water Resour. Res. 57 (9), e2021WR030263. https://doi.org/10.1029/2021WR030263.

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ. Model. Softw. 16 (4), 309–330. https://doi.org/10.1016/S1364-8152(01)00008-1.

Jehn, F.U., Bestian, K., Breuer, L., Kraft, P., Houska, T., 2020. Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. Hydrol. Earth Syst. Sci. 24 (3), 1081–1100. https://doi.org/10.5194/hess-24-1081-2020.

Kohonen, T., 1990. The self-organizing map. Proc. IEEE 78 (9), 1464–1480.

Kohonen, T., 2013. Essentials of the self-organizing map. Neural networks 37, 52–65.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22 (11), 6005–6022.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23 (12), 5089–5110.

Kratzert, F., Gauch, M., Nearing, G., Klotz, D., 2022. Neural Hydrology–-A Python library for Deep Learning research in hydrology. J. Open Source Soft. 7 (71), 4050.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., Matias, Y., 2023. Caravan - A global community dataset for large-sample hydrology. Sci. Data 10 (1), 61. https://doi.org/10.1038/s41597-023-01975-w.

Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS Opinions: Never train an LSTM on a single basin. Hydrol. Earth Syst. Sci. Discuss. 2024, 1–19.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., Dadson, S.J., 2021. Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. Hydrol. Earth Syst. Sci. 25 (10), 5517–5534.

Magyar, J.C., Sambridge, M.S., 2022. The Wasserstein distance as a hydrological objective function. Egusphere 2022, 1–32.

Mathevet, T., Gupta, H., Perrin, C., Andréassian, V., Le Moine, N., 2020. Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds. J. Hydrol. 585, 124698.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., 2024. Global prediction of extreme floods in ungauged watersheds. Nature 627 (8004), 559–563.

Nearing, G.S., Pelissier, C.S., Kratzert, F., Klotz, D., Gupta, H.V., Frame, J.M., Sampson, A.K., 2019. Physically informed machine learning for hydrological modeling under climate nonstationarity. UMBC Faculty Collection.

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H.V., 2021. What role does hydrological science play in the age of machine learning? Water Resour. Res. 57 (3), e2020WR028091.

NeuralHydrology Team (2024). *NeuralHydrology documentation.* Accessed October 22, 2024. https://neuralhydrology.readthedocs.io/en/latest/.

Ng, K., Huang, Y., Koo, C., Chong, K., El-Shafie, A., Ahmed, A.N., 2023. A review of hybrid deep learning applications for streamflow forecasting. J. Hydrol. 130141.

Rahimzad, M., Moghaddam Nia, A., Zolfonoon, H., Soltani, J., Danandeh Mehr, A., Kwon, H.-H., 2021. Performance comparison of an LSTM-based deep learning model versus conventional machine learning algorithms for streamflow forecasting. Water Resour. Manag. 35 (12), 4167–4187.

Rassam, D.W., Peeters, L., Pickett, T., Jolly, I., Holz, L., 2013. Accounting for surface–groundwater interactions and their uncertainty in river and groundwater models: A case study in the Namoi River, Australia. Environ. Model. Softw. 50, 108–119. https://doi.org/10.1016/j.envsoft.2013.09.004.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., 2018. HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. Hydrol. Earth Syst. Sci. 22 (11), 5639–5656.

Tripathy, K.P., Mishra, A.K., 2023. Deep learning in hydrology and water resources disciplines: Concepts, methods, applications, and research directions. J. Hydrol. 130458.

Verdon-Kidd, D., Kiem, A., Moran, R., 2014. Links between the Big Dry in Australia and hemispheric multi-decadal climate variability–implications for water resource management. Hydrol. Earth Syst. Sci. 18 (6), 2235–2256.

Wehrens, R., Buydens, L.M., 2007. Self-and super-organizing maps in R: the Kohonen package. J. Stat. Softw. 21, 1–19.

Wi, S., Steinschneider, S., 2024. On the need for physical constraints in deep learning rainfall–runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration. Hydrol. Earth Syst. Sci. 28 (3), 479–503.

Wilbrand, K., Taormina, R., ten Veldhuis, M.-C., Visser, M., Hrachowitz, M., Nuttall, J., Dahm, R., 2023. Predicting streamflow with LSTM networks using global datasets. Front. Water 5, 1166124.

Yao, Y., Zhao, Y., Li, X., Feng, D., Shen, C., Liu, C., Kuang, X., Zheng, C., 2023. Can transfer learning improve hydrological predictions in the alpine regions? J. Hydrol. 625, 130038.