Journal of Hydroinformatics

Journal of Hydroinformatics Vol 26 No 4, 835 doi: 10.2166/hydro.2024.268

Leveraging historic streamflow and weather data with deep learning for enhanced streamflow predictions

Christiaan Schutte [a,*], Michael van der Laan [a,b] and Barend van der Merwe [c]

[a] Department of Plant and Soil Sciences, Faculty of Natural and Agricultural Science, University of Pretoria, Pretoria 0002, South Africa
[b] Agricultural Research Council, Pretoria 0002, South Africa
[c] Department of Geography, Geoinformatics and Meteorology, Faculty of Natural and Agricultural Science, University of Pretoria, Pretoria 0002, South Africa
*Corresponding author. E-mail: ceschutte34@gmail.com

CS, 0000-0001-6516-379X; MvdL, 0000-0001-8656-623X; BvdM, 0000-0002-9908-0295

ABSTRACT

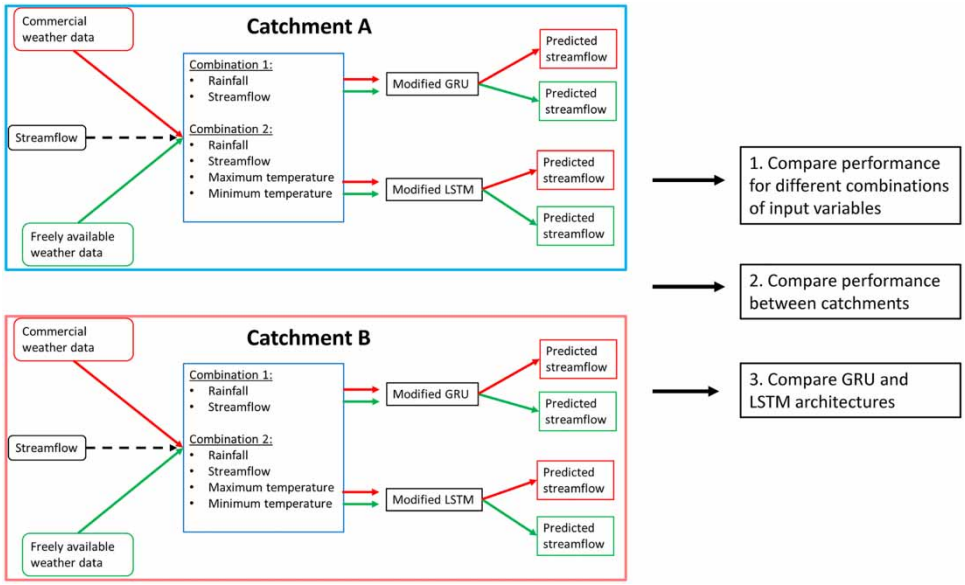
Streamflow information is crucial for effectively managing water resources. The declining number of active gauging stations in many rivers is a global concern, necessitating the need for reliable streamflow estimates. Deep learning techniques offer potential solutions, but their application in southern Africa remains largely underexplored. To fill this gap, this study evaluated the predictive performance of gated recurrent unit (GRU) and long short-term memory (LSTM) networks using two headwater catchments of the Steelpoort River, South Africa, as case studies. The model inputs included rainfall, maximum, and minimum temperature, as well as past streamflow, which was utilized in an autoregressive sense. The inclusion of streamflow in this way allowed for the incorporation of simulated streamflow values into the look-back window for predicting the streamflow of the testing set. Two modifications were required to the GRU and LSTM architectures to ensure physically consistent predictions, including a change in the activation function of the GRU/LSTM cells in the final hidden layer, and a non-negative constraint that was used in the dense layer. Models trained using commercial weather station data produced reliable streamflow estimates, while moderately accurate predictions were obtained using freely available gridded weather data.

Key words: GRU, LSTM, rainfall-runoff modelling

HIGHLIGHTS

- Assessment of Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) networks for streamflow prediction in southern Africa reveals similar predictive performance.
- By incorporating simulated streamflow into the look-back window (LBW) and modifying the GRU and LSTM architectures, this study presents a novel daily streamflow prediction method.
- These models show potential for both long-term data gap filling in streamflow records and for short-term daily forecasts, that could be beneficial for flood risk management and planning.
- LBWs of 10 to 30 days were found to be suitable for achieving accurate predictions with both GRU and LSTM models.

GRAPHICAL ABSTRACT



INTRODUCTION

Developing countries around the world are facing significant water challenges, including limited water resources in low rainfall areas, water pollution, and inadequate infrastructure, all of which are exacerbated by population growth, urbanization, and climate change (Oyebande 2001). Hydrological information is crucial for ensuring the sustainable management of water resources, and understanding the complex dynamics of water availability, flow patterns, and groundwater recharge under a changing climate (Beven 2011). Accurate and reliable hydrological data form the foundation for informed decision-making, enabling stakeholders to gain insights into water system dynamics, identify trends, and make informed choices on water allocation, infrastructure development, and conservation. Streamflow data, a critical component of hydrological information, is vital for monitoring ecosystem health, determining sustainable water abstraction, and assessing flood and drought potential (Beven 2011).

Various governmental and private sector institutions operate streamflow gauging stations to enable informed water resource assessment and planning (Rogers *et al.* 2019). The number of active gauging stations is, however, declining in many catchments around the world (Rogers *et al.* 2019). In South Africa, for instance, the decrease is due to factors such as lack of funding and maintenance, wear and tear, vandalism, and theft (DWS 2021). Moreover, the availability of weather station data has also been declining in South Africa since around 1970 (Engelbrecht *et al.* 2009). This is a major concern as these data are critical inputs to process-based and deep learning (DL) hydrological models, with rainfall being the most important input to water resource studies (Pitman & Bailey 2021). The decreased availability of streamflow and weather data hampers the ability to make informed decisions on water resource management and planning, and this presents a long-term threat to water security (Odendaal 2021).

Where the availability of freshwater is highly variable and where resources required to sustain long-term monitoring programmes are constrained, hydrological models are of particular importance (Hughes 2004). Process-based models used for streamflow prediction face limitations due to the lack of comprehensive information about system properties, including topography, soil characteristics, and vegetation cover. These properties are highly heterogeneous and can change over time, while detailed knowledge of subsurface hydrological processes, where much of hydrology takes place, remains scarce (Kratzert *et al.* 2019a). In addition to data for model parameterization, data for initialization and calibration are also often limited. In contrast, a data-driven approach such as DL, which is based on artificial neural networks (ANNs), offers an alternative approach by predicting streamflow without explicitly defining the underlying physical processes. In addition, the ability of ANNs to learn complex, nonlinear relationships directly from data is particularly important in hydrology, as many hydrological processes exhibit intricate and nonlinear behaviours that process-based models struggle to represent (Kratzert *et al.* 2019a).

Long short-term memory (LSTM) (Hochreiter & Schmidhuber 1997) and gated recurrent unit (GRU) (Cho *et al.* 2014) networks are DL architectures that were specifically designed to analyse sequential data. Streamflow data can be considered sequential, and recent benchmarking studies have illustrated that these LSTM networks can rival and even outperform process-based and conceptual hydrological models in streamflow prediction (Kratzert *et al.* 2019c; Lees *et al.* 2021). Subsequently, GRU networks, as well as hybrid approaches combining LSTM with GRU networks (Muhammad *et al.* 2019) or combining convolutional neural networks (CNN) with either LSTM or GRU networks (Ghimire *et al.* 2021; Anderson & Radić 2022), were also found to be useful for streamflow prediction. Another notable strength of DL approaches lies in its flexibility (Razavi 2021). Since LSTMs and GRUs emerged as streamflow prediction models, modifications to standard LSTM architectures and configurations have also appeared (such as the CNN-LSTM hybrid models). Another example is the ability of DL models to adapt to various forecasting horizons, such as the use of LSTMs for long-lead streamflow prediction (Najafzadeh & Anvari 2023) or cases where LSTMs are used to jointly predict multiple timescales within one mode (Gauch *et al.* 2021), in other words a model that can generate predictions at both hourly and daily timescales.

Despite their advances, DL techniques face challenges such as a lack of interpretability (often termed 'blackbox'), ensuring physical consistency in predictions, and managing multi-dimensional datasets (Reichstein *et al.* 2019). The concept of theory-guided data science emerges in this context, aiming to merge well-established scientific theories with data-driven findings (Karpatne *et al.* 2017). Hybrid approaches, blending physical models' predictability with DL's adaptability, have been advocated in recent discussions (Reichstein *et al.* 2019). In line with this trend, there is a growing interest in physics- and hydrologically informed machine learning (ML)/DL (Nearing *et al.* 2021).

Another major challenge when implementing pure DL or physics and hydrologically informed approaches is the availability of sufficient, good-quality training data, especially in developing countries where data collection and storage can be limited (Gaffoor *et al.* 2022). For example, high-resolution, spatially distributed weather datasets may not be readily or freely available (du Plessis & Kibii 2021), which presents a challenge for streamflow prediction. It should also be considered that a significant portion of existing ML/DL and physics or hydrologically informed research in hydrology has depended on high-quality, spatially distributed datasets such as Catchment Attributes for Large-Sample Studies (CAMELS) (Addor *et al.* 2017). Such datasets are predominantly based in developed countries and are thus not always representative of conditions in less-resourced settings. The performance of LSTM networks has also been shown to be less accurate in drier catchments (Kratzert *et al.* 2019c; Lees *et al.* 2021; Anderson & Radić 2022: 94), a characteristic of much of southern Africa. Such challenges must be considered when evaluating the feasibility of using DL models for streamflow prediction in the hydrological context of semi-arid, data-scarce, and developing countries.

The main research question was whether LSTM and GRU networks could be developed in South Africa, a country which is currently experiencing a decline in streamflow measurements, to generate reliable streamflow estimations for predictive and data gap-filling purposes. The increased availability of open-source gridded weather data provides a potential opportunity to address the issue of insufficient weather station data (Reichstein *et al.* 2019). The second research question was, therefore, whether freely available gridded weather data could be used as input to the models to produce reasonably accurate streamflow estimates. Two catchments near Lydenburg, South Africa, were used as a case study to answer the research questions.

MATERIALS AND METHODS

Study area

Situated in the north-east of South Africa, the study area consisted of two catchments in the headwaters of the Steelpoort River located in the Olifants River basin (Figure 1). Catchment A is approximately 100 km^2 and consists of quaternary catchment B42D as defined by the Department of Water and Sanitation (DWS). Catchment B is larger (300 km^2) and consists of quaternary catchments B42A and B42B. The average elevation ranges between 1,336 and 2,263 m above the mean sea level (mamsl). Based on the Köppen–Geiger climate classification system, the study area is classified as Cwb, indicating a warm temperate climate with cool and dry winters, and warm and wet summers. Average daily temperatures range from 6 to 22 °C in winter and from 22 to 32 °C in summer (Herold & Bailey 2016). The study area is in one of the country's higher rainfall areas, with most of the rainfall occurring between October and April. The average annual rainfall is approximately 800 mm yr^{-1}, but highly variable, and the average annual evaporation for an S-Pan is approximately 1,500 mm yr^{-1} (Herold & Bailey 2016). The land cover of Catchment A mainly consists of grasslands, woodlands, and thicket areas. In

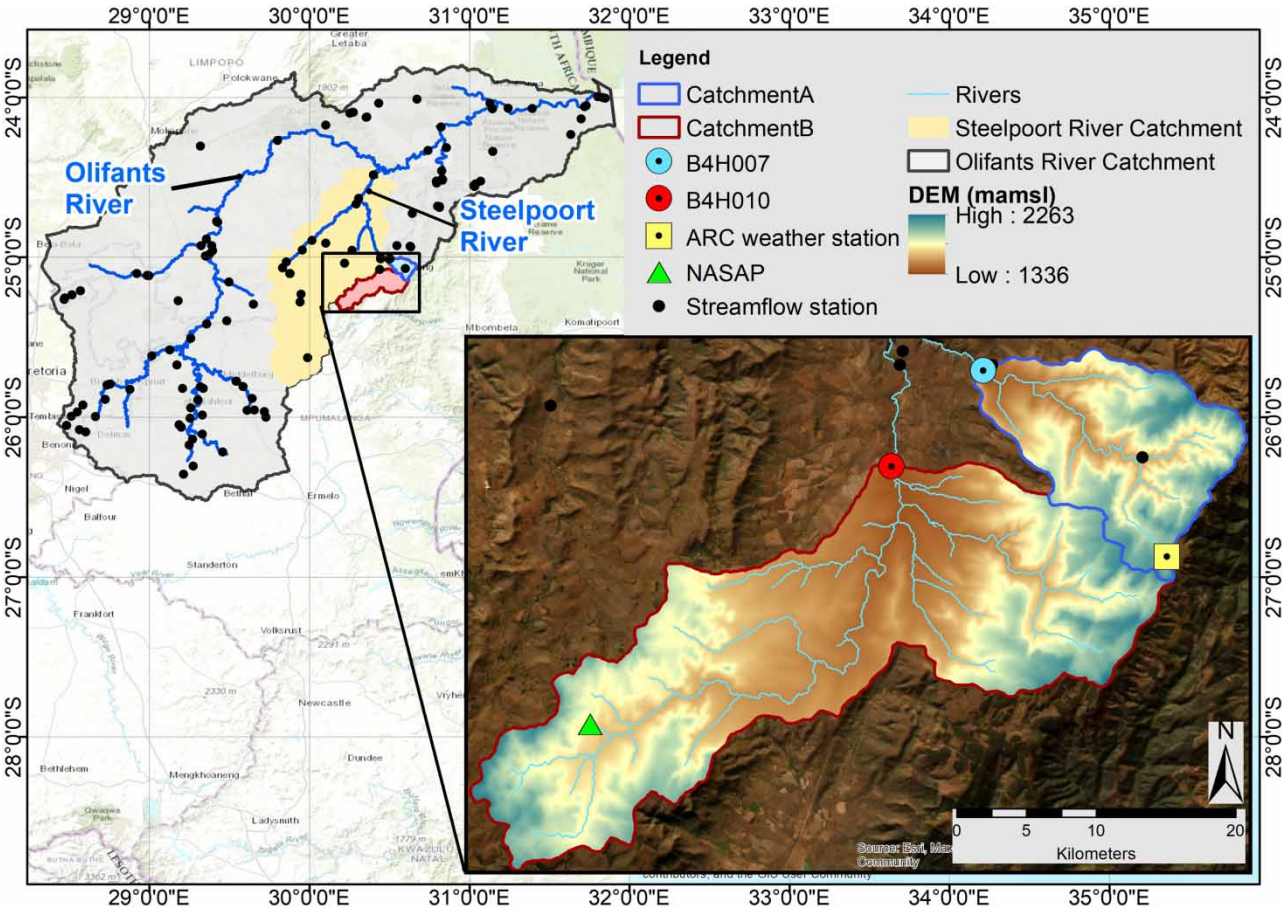


Figure 1 | Locality map of the catchments, streamflow stations, the Agricultural Research Council (ARC) weather station and the National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) grid point.

addition to those land cover types, Catchment B also includes the town of Lydenberg, agricultural activities, and some minor pine plantations.

Data

Daily weather data, consisting of rainfall (mm), minimum and maximum temperature (°C), and streamflow data ($m^3\ s^{-1}$) were used as input variables. The data included both *in situ* measured data and gridded products. Daily weather data were obtained from three sources: (1) a weather station operated by the Agricultural Research Council (ARC), (2) the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) data (Funk *et al.* 2015) (https://www.chc.ucsb.edu/data/chirps), and (3) the National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) project (https://power.larc.nasa.gov/). The ARC weather station had a record spanning from 1979 to 2002. CHIRPS and NASAP data were downloaded for the same period. To download data from CHIRPS, a geojson file for each catchment was uploaded to the CHIRPS portal (https://climateserv.servirglobal.net/). The NASAP data were downloaded from the Water Research Observatory website (https://www.waterresearchobservatory.org/), where the data are available as a grid of points for South Africa. The grid point closest to each catchment was identified, and in this case, it was the same grid point for both catchments. Streamflow data were obtained from the DWS website (https://www.dws.gov.za/Hydrology/Verified/hymain.aspx) for stations B4H007 and B4H001 for Catchment A and Catchment B, respectively.

The streamflow data were combined with each of the weather data sources (ARC, CHIRPS, and NASAP) to create three datasets for each catchment. Each dataset was divided into a training set that was used to derive the optimal network weights and a testing set that was used to assess the prediction accuracy. In hydrological modelling, the first 70–80% of the data are often used to calibrate the model and the last 20–30% to validate the model. In this context, calibration and validation are

analogous to training and testing a DL model. The training period spanned from 1 October 1979 to 30 September 1997, making it 18 years. The testing period was from 1 October 1997 to 22 February 2002 (the date when the ARC weather station was discontinued), which is 4 years, 4 months, and 22 days. The entire dataset covers 22.4 years.

Model development

LSTM networks use a memory cell, which is a long-term memory of the network, and is controlled by three gates: the input gate, the output gate, and the forget gate (Figure 2). These gates control the flow of information into and out of the memory cell, allowing the network to selectively retain or disregard information (Hochreiter & Schmidhuber 1997). The LSTM is described in the following equations:

$$f_t = \sigma(W_{xf}\, x_t + W_{hf} h_{t-1} + W_{cf} c_{-1} + b_f) \tag{1}$$

$$i_t = \sigma(W_{xi}\, x_t + W_{hi} h_{t-1} + W_{ci} c_{-1} + b_i) \tag{2}$$

$$c_t = \sigma\, f_t\, c_{t-1} + i_t \tan h\, (W_{xc}\, x_t + W_{hc} h_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{xo}\, x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \tag{4}$$

$$h_t = o_t \tan h(c_t) \tag{5}$$

$$\hat{y}_t = W_{hy}\, h_y + b_y \tag{6}$$

where σ is the logistic sigmoid function, and f_t, i_t, o_t, and c_t represent the forget gate, input gate, output gate, and cell activation vectors, respectively, all of which are the same length as the hidden vector h. Equation (1) computes the value of the forget gate f_t at time step t, which determines how much of the previous cell state c_{t-1} to keep or forget. Equation (2) computes the value of the input gate i_t at time step t, which determines how much of the new input x_t to store in the current cell state c_t. The vectors f_t and i_t have both entries in the range (0, 1). When values of f_t are close to 0, information is forgotten, when values of f_t are close to 1, information is retained. Similarly, for i_t, information will be added to the cell state when values are close to 1, while information is ignored when values are close to 0. Equation (3) computes the current cell state c_t at time step t, which is a combination of the previous cell state c_{t-1} and the new input x_t that has passed through the input gate and been filtered by the tan h function. Equation (4) computes the value of the output gate o_t at time step t, which determines how much of the current cell state c_t to output as the hidden state h_t. Equation (5) computes the hidden state h_t at time step t by multiplying the output gate o_t with the filtered cell state c_t using the tan h function. Finally,

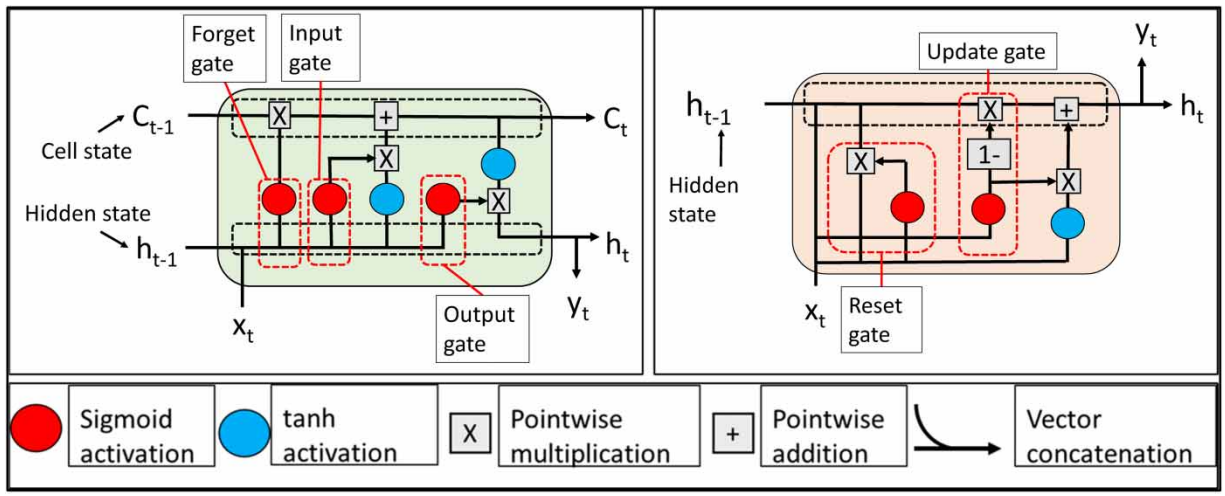


Figure 2 | A comparison of the internal architecture of a standard long short-term memory (LSTM) cell (left) and a gated recurrent unit (GRU) cell (right). The LSTM cell is characterized by its input, output, and forget gates as well as its cell state. The GRU cell uses a more streamlined architecture with update and reset gates but without a separate cell state.

Equation (6) computes the final output $\hat{y}_t$ at time step t by multiplying the hidden state h_t with a weight matrix W_{hy} and adding a bias term b_y.

GRUs do not have a separate cell state, only a hidden state h_t (Figure 2) (Cho *et al.* 2014). While LSTMs utilize three gates to manage the flow of information, GRUs use only two gates. The update gate that determines the extent to which the previous hidden state is updated, and the reset gate that determines the extent to which the previous hidden state is reset. The GRU is described in the following equations:

$$z_t = \sigma(W_{xz}\, x_t + W_{hz} h_{t-1} + b_z) \tag{7}$$

$$r_t = \sigma(W_{xr}\, x_t + W_{hr} h_{t-1} + b_r) \tag{8}$$

$$h_t = (1 - z_t) h_{t-1} + z_t \tan h(W_{xh}^T x_t + \ W_{hh}^T (r_t\, h_{t-1}) + b_h) \tag{9}$$

$$\hat{y}_t = \sigma(W_o^T h_t + b_o) \tag{10}$$

Equation (7) computes the update gate, z_t, which determines the extent to which the previous hidden state, h_{t-1}, is updated based on the current input x_t. Similarly, Equation (8) computes the reset gate, r_t, which determines the extent to which the previous hidden state, h_{t-1}, is reset based on the current input, x_t. Using the update and reset gates, Equation (9) computes the new hidden state, h_t, by combining the previous hidden state, h_{t-1}, with the current input, x_t, and the reset gate, r_t. The output of the tan h function is multiplied by the update gate, z_t, and added to the previous hidden state, $(1 - z_t)\, h_{t-1}$, to obtain the new hidden state, h_t. Finally, the output prediction, $\hat{y}_t$, is computed using a traditional dense layer with weights, W_o, and biases, b_o, applied to the hidden state, h_t. The output prediction is then passed through the sigmoid activation function to obtain the final output of the network.

In Recurrent Neural Network (RNN) architectures, the term 'rolled-out' diagram refers to an expanded visual representation of the network's operations across multiple time steps (Goodfellow *et al.* 2016). Instead of showing a single compact cell, the diagram is 'unrolled' to depict how the network processes sequential data over a given time period. The rolled-out diagrams for a GRU cell and an LSTM cell are used to display their operations over a 30-day look-back window (LBW) (Figure 3). These cells ingest sequences consisting of 30 days' worth of historical weather and hydrological data. For each time step within this window, the LSTM employs both the previous hidden and cell states to produce an individual streamflow prediction. Similarly, the GRU uses the preceding hidden state to output a streamflow prediction at each corresponding time step.

Hyperparameters were selected based on a combination of experimentation and previous literature (Kratzert *et al.* 2018; Kratzert *et al.* 2019b; Fan *et al.* 2020; Nifa *et al.* 2023). An LBW of 30 days was selected, which has been proven effective in a semi-arid catchment (Nifa *et al.* 2023) that has a climate more similar to the catchments in this study compared to those in snow-dominated or tropical basins (see Appendix I for additional information on the choice of LBW). The model architecture consisted of five hidden layers with 25 GRU or LSTM cells per layer and a dense layer for the final streamflow prediction (Figure 4). Each of the 25 GRU or LSTM cells in the first hidden layer processes four input features: rainfall, streamflow, maximum temperature, and minimum temperature (Figure 3). From the second hidden layer onwards, each GRU or LSTM cell receives a 25-dimensional input, being the hidden state output from each cell in the previous layer. The dense layer then receives the outputs from all 25 GRU or LSTM cells in the fifth hidden layer to produce the final streamflow prediction. No activation function was used for the dense layer, which was linear.

A challenge in DL models is the generation of physically inconsistent predictions (Reichstein *et al.* 2019). In many instances, during the dry seasons, when streamflow was close to 0 $m^3\ s^{-1}$, the models would start simulating slightly negative streamflow values and then return to normal in the wet seasons (Appendix I). To prevent the models from predicting these negative streamflow values, two modifications to the GRU and LSTM network architectures were required. Firstly, the hyperbolic tan h activation function (Goodfellow *et al.* 2016) of each of the 25 GRU or LSTM cells, contained in the fifth hidden layer of the network, was changed to the rectified linear unit (ReLU) activation function (Krizhevsky *et al.* 2017) (refer to Figure 2 for the position of the tan h within a standard GRU or LSTM cell). Secondly, a non-negative constraint (Abadi *et al.* 2016) was used in the dense layer. Neither the ReLU activation function nor non-negative constraint on its own prevented negative streamflow predictions, and both modifications were necessary and complemented each other effectively. The ReLU activation function in the fifth layer could control the internal representations and mitigate the occurrence of

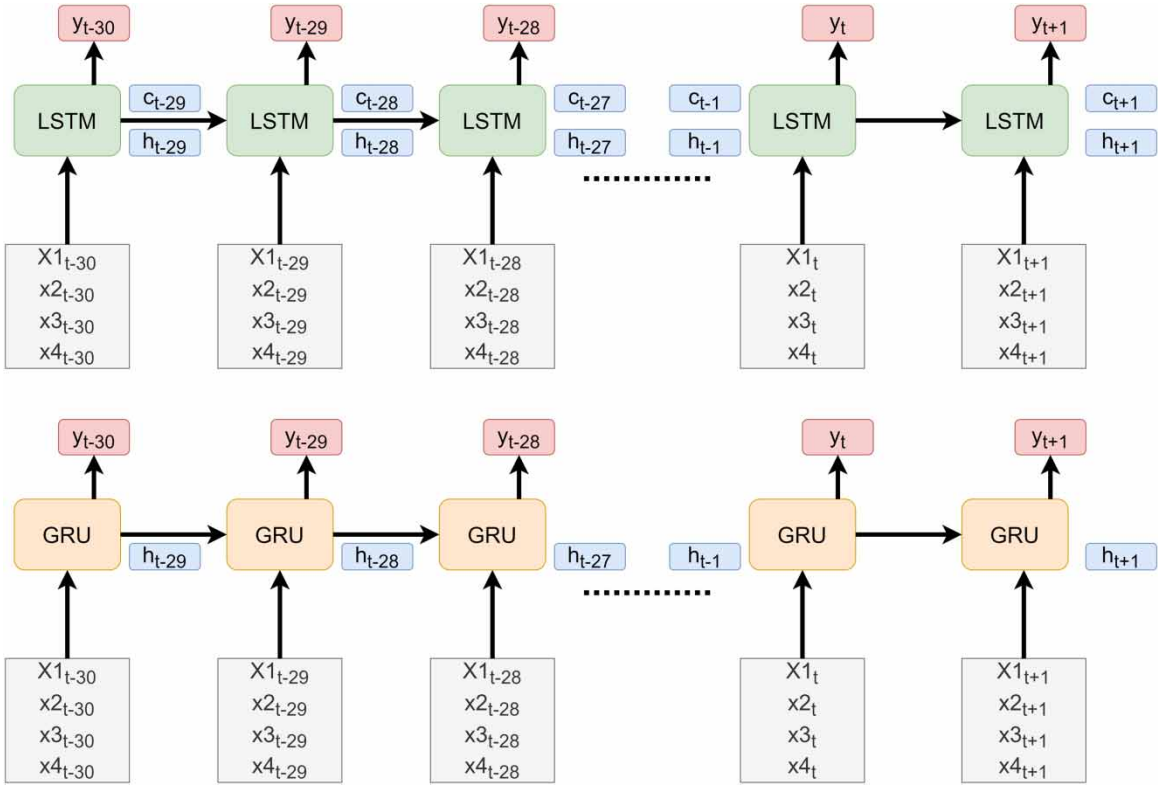


Figure 3 | Rolled-out diagram of a single long short-term memory (LSTM) (above) and gated recurrent unit (GRU) cell (below) configured with a 30-day LBW and four inputs. $(\chi_{n_{t-30}})$, $(\chi_{n_{t-29}})$, ..., (χ_{n_t}), and $(\chi_{n_{t+1}})$ denote the previous, current, and next time step inputs, respectively. At each time step, the LSTM cell receives the inputs and both the hidden state (h_{t-1}), and cell state (c_{t-1}) from the previous time step, whereas the GRU cell receives only the inputs and hidden state (h_{t-1}) from the previous time step. Both architectures produce a prediction ($\hat{y}_t$) at each time step.

negative values within the model layers, while the non-negative constraint in the dense layer served as an additional measure to enforce non-negativity in the final predictions.

Before training, the MinMaxScaler function of the Scikit Learn Library (Pedregosa *et al.* 2011) was used to scale the training set and the testing set to a range between 0 and 1. The models were trained for 55 epochs (number of iterations) with a batch size of 256 (number of samples extracted from the training set at a time) (Kratzert *et al.* 2019b). To prevent overfitting, a dropout rate of 0.1 was used for regularization. The mean-squared error (Bishop & Nasrabadi 2006) was used as the loss function and the Adam optimizer (Kingma & Ba 2014) was used for gradient descent. The weather variables of the testing set were used as input to the trained GRU and LSTM models to predict a time series consisting of 1 605 days of streamflow values. An essential aspect to note is that the models used the preceding 30 days of weather and streamflow values to predict the next day of streamflow (Figure 5). As only the measured weather variables (not the measured streamflow) contained in the testing set were used, the models had to incorporate the predicted streamflow values into the LBW to predict the entire time series. To do this, the model used the preceding 30 days of measured weather and predicted streamflow values to predict streamflow on the next day. The model then moved one time step forward and incorporated the predicted streamflow value and the measured weather values of that day into the LBW. This process continued for the entire testing set and was achieved with a for loop (Shirzadi 2023).

Ensemble predictions refer to the practice of combining the predictions of multiple individual models to obtain a more accurate and robust prediction (Goodfellow *et al.* 2016). For each catchment and architecture, 20 individual models were trained and a streamflow time series was predicted with each model. The ensemble predictions were created by computing the average across a set of 20 models. This approach allowed accounting for the variability in model performance due to random initialization and different training runs.

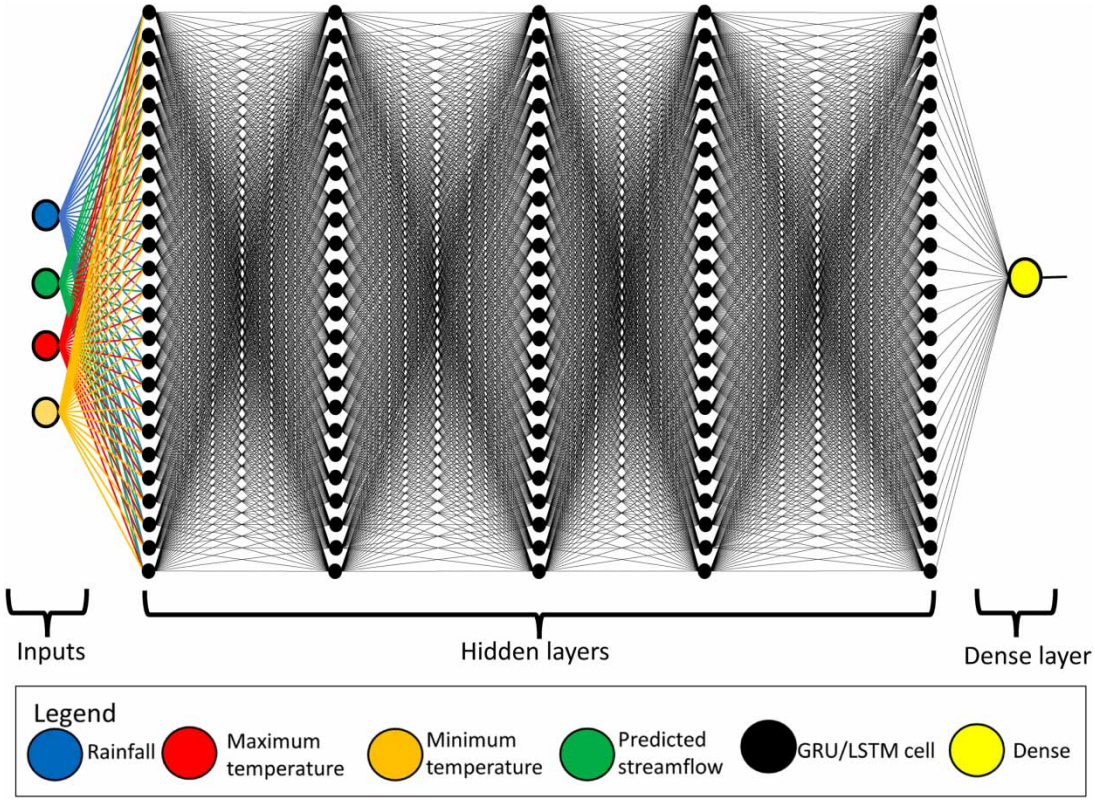


Figure 4 | Schematic representation of the deep learning architecture with four input nodes, five hidden layers, each comprising 25 gated recurrent unit (GRU) or 25 long short-term memory (LSTM) cells, where each of these cells utilizes a look-back window of 30 days. A dense layer follows the hidden layers, leading to the final prediction.

As the streamflow predictions were in a range between 0 and 1, they were scaled back to the original range with the inverse function of the Scikit Learn Library, to compare with a measured streamflow contained in the original unscaled testing set. The Nash–Sutcliffe efficiency (NSE) (Nash & Sutcliffe 1970) and Kling–Gupta efficiency (KGE) (Kling *et al.* 2012) were used for the evaluation of model performance and were carried out in two stages. The first part of model evaluation involved assessing variation in prediction accuracy for each set of 20 GRU or LSTM models, while in the second stage, the evaluation focused on a comparative analysis between the best-performing models of each set of 20 models and the corresponding ensemble predictions within each set.

Both the NSE and KGE range from negative infinity to 1, with 1 indicating perfect model fit and values closer to 0 indicating poorer model performance. In hydrology, an acceptable NSE/KGE value often depends on the nature of the specific application; however, an NSE/KGE value of 0.5 or higher is generally considered an acceptable model, while a value of 0.8 or higher is considered a very good model (Moriasi *et al.* 2007). The Hydrostats package (Roberts *et al.* 2018) was used to compute the NSE/KGE values.

Experiment: testing different sources and combinations of weather input data

Three different sources of weather data and two different combinations of weather input variables were tested as input for the GRU and LSTM models to explore the effect on model performance (Table 1). The first combination (Combination 1) consisted only of rainfall and streamflow as input variables, with observed streamflow used during training and model-predicted streamflow during testing. Only rainfall was obtained for CHIRPS that was only included under Combination 1, while ARC and NASAP were used in both combinations. The second combination (Combination 2) included minimum and maximum temperature in addition to rainfall and streamflow. For each catchment and each combination, 20 GRU and 20 LSTM networks were trained, and ensemble predictions were calculated for each set of 20 models.

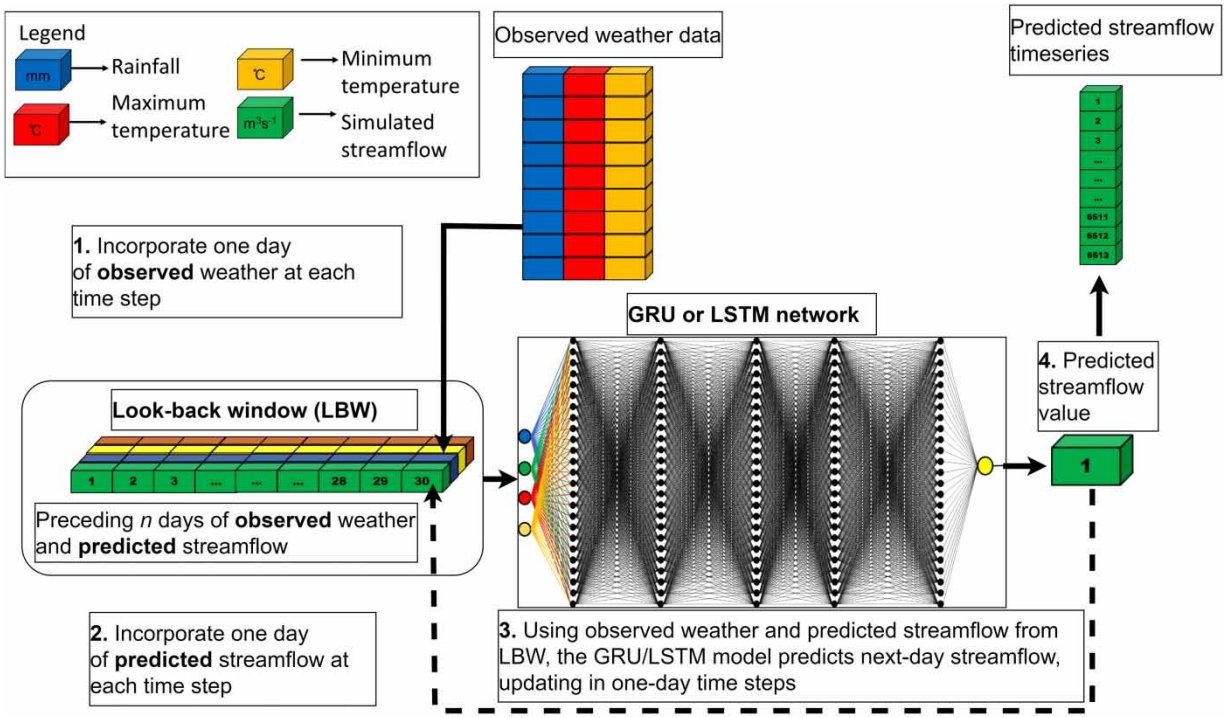


Figure 5 | Based on the preceding 30 days of past weather and streamflow values in the look-back window (LBW), the network predicts a streamflow value for the next day. The model then moves one time step forward, ingesting the predicted streamflow value, along with observed weather values for that day, into the LBW to make the next prediction.

Table 1 | Combinations of weather sources and input variables for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and weather station (ARC) data

Dataset	Combination 1	Combination 2
CHIRPS	Rainfall Streamflow	–
NASAP	Rainfall Streamflow	Rainfall Streamflow Maximum temperature Minimum temperature
ARC	Rainfall Streamflow	Rainfall Streamflow Maximum temperature Minimum temperature

In conventional ML paradigms, manual feature engineering is often a prerequisite to effectively model complex systems (Zheng & Casari 2018). This frequently involves the calculation of lagged variables, especially in hydrological applications. However, one of the unique strengths of DL models lies in their capacity to automatically learn and identify relevant features from the data (Goodfellow *et al.* 2016). Leveraging these capabilities, the present study opted not to incorporate lagged variables for runoff prediction. Instead, we employed deep recurrent neural networks, which are inherently designed to capture essential temporal relationships and have been demonstrated to be effective in streamflow prediction (Kratzert *et al.* 2019c).

Combination 1 predictions for each catchment, architecture and weather data source, were compared to the corresponding Combination 2, to determine if adding minimum and maximum temperature improved prediction accuracy. Next, the three weather data sources for each catchment and architecture were compared to each other to determine which weather data source performed the best overall. The Shapiro–Wilk test (Shapiro & Wilk 1965) confirmed that the NSE values violated

assumptions of normality; therefore, the need for non-parametric tests. The non-parametric Mann–Whitney U test (Mann & Whitney 1947) was used to test for significant differences between each of the three weather data sources and combinations.

RESULTS

Experiment: testing input variables and weather data sources

The NSE values varied from −9.4 (very poor) to 0.74 (very good), while the KGE values varied from −4.4 (very poor) to 0.77 (very good) (Figure 6). This range in the NSE and KGE values indicated a wide variability in model performance. The models also showed different performance spreads in cumulative frequency distribution (CDF) curves, and specifically, the ARC models (both GRU and LSTM) tended to cluster around higher NSE and KGE values (their CDF curves shifted towards the right), implying better performance. A steeper slope, as observed for ARC data-driven models, implied that a significant portion of those models had closely clustered performance values. In contrast, the CHIRPS and NASAP models displayed a flatter curve, indicating a wider spread in performances. While the CDF plots for both catchments exhibited some differences, a consistent observation was the relative superiority of ARC data-driven models. Both in terms of NSE and KGE values, ARC models consistently outperformed those based on CHIRPS or NASAP data. There were notable differences between the CDF plots of Catchment A and Catchment B. While ARC remained superior in both cases, the degree of superiority and the relative performance of CHIRPS and NASAP models differed.

The CHIRPS and NASAP data-driven models were capable of moderately accurate predictions in a few instances with maximum NSE and KGE values greater than 0.5. The ARC weather station data-driven models achieved considerably higher NSE and KGE values of 0.74 and 0.77, respectively, and the predictions were considered satisfactory. The Mann–

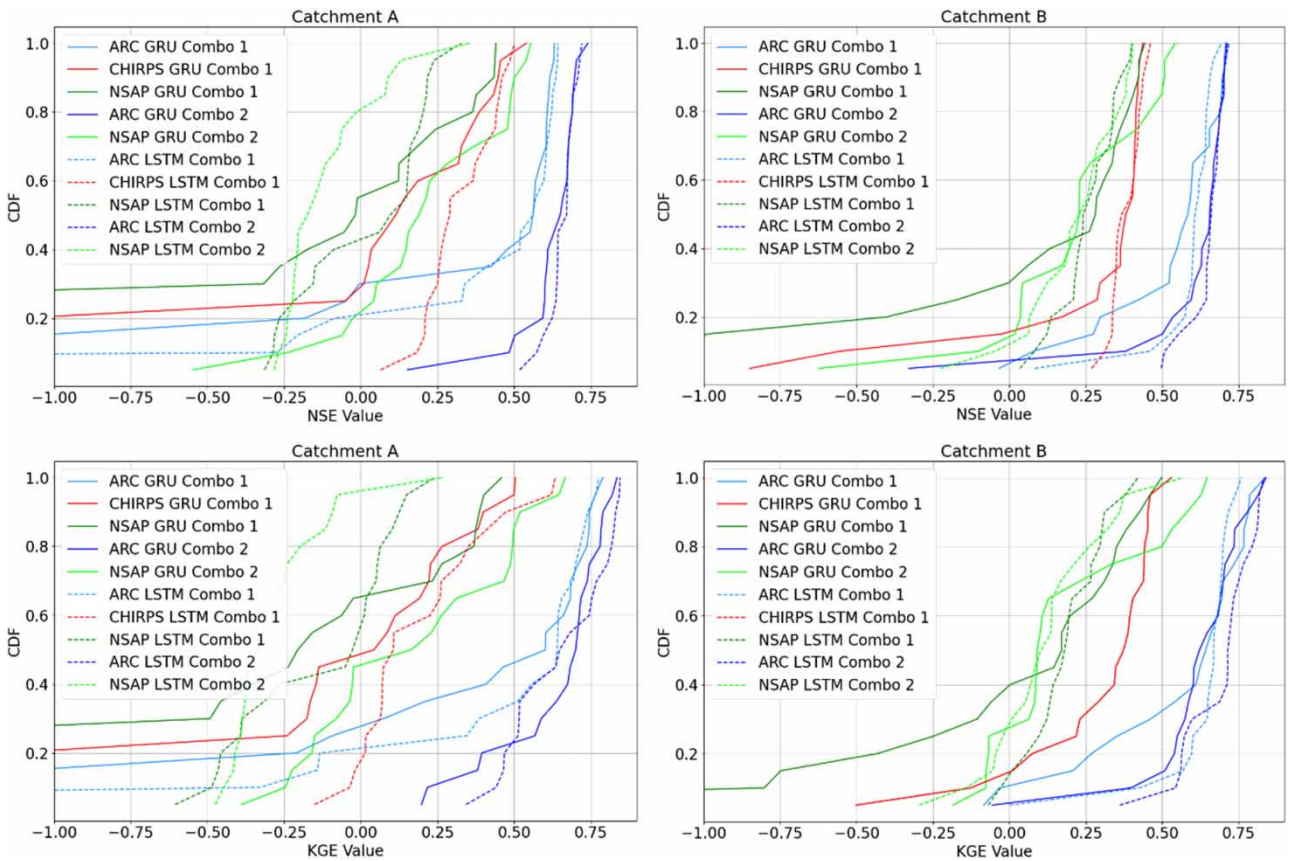


Figure 6 | Cumulative frequency distribution (CDF) for Nash–Sutcliffe efficiency (NSE) (top) and Kling–Gupta efficiency (KGE) values (bottom) for gated recurrent unit (GRU) and long short-term memory (LSTM) models for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP), and Agricultural Research Council weather station data (ARC). Values smaller than −1 are not shown, with −10.9 being the lowest.

Whitney U test revealed significant differences in both NSE and KGE values between the ARC models and both the CHIRPS and NASAP models for Combination 1 and for Combination 2 in both catchments (Table 2). The test also revealed significant differences in both NSE and KGE values between CHIRPS and NASAP, for both the LSTMs for Catchment A, and for both the GRUs and LSTMs for Catchment B (Table 2). There were no significant differences in NSE or KGE values between the CHIRPS and NASAP GRU of Catchment A. While significant differences exist between CHIRPS and NASAP in some scenarios, it is challenging to decisively point out which data source leads to better performance consistently across all configurations and catchments. This suggests that while they may differ, neither CHIRPS nor NASAP consistently outperforms the other.

The test revealed significant differences in NSE values between Combinations 1 and 2 for the NASAP GRUs and LSTMs of Catchment A, while no differences in NSE were observed for the GRUs or LSTMs of Catchment B (Table 3). The test revealed significant differences in KGE values between Combinations 1 and 2 for both the NASAP GRUs and LSTMs of Catchment A, while no differences in KGE values were observed for the GRUs and LSTMs of Catchment B (Table 3). The test also revealed significant differences in NSE values between Combinations 1 and 2 for the ARC GRUs and LSTMs for Catchment A (Table 3), as well as for the LSTM models of Catchment B (Table 3). No differences in NSE values were observed between Combinations 1 and 2 for the GRUs of Catchment B. No significant differences in KGE values were observed between Combinations 1 and 2 for the ARC data-driven models. While the addition of temperature in Combination 2 led to increased prediction accuracy in certain scenarios, particularly in Catchment A, it did not consistently enhance performance across all model configurations and catchments.

No statistical significance tests were conducted for NSE (Table 4) and KGE (Table 5) values between the best-performing models and the corresponding ensemble predictions, due to the fact that only one ensemble was produced for each combination. The performance of the ensemble predictions varied across different data sources and combinations. The ensemble predictions for the CHIRPS data-driven models did not exceed NSE or KGE values of 0.5; however, the best CHIRPS GRU model for Combination 1 achieved an NSE value of 0.54 for Catchment A. The best CHIRPS GRU models for Combination 1 achieved KGE values of 0.58 and 0.57 for Catchment A and B, respectively.

For Combination 1, the best NASAP GRU models achieved NSE values of 0.44 in Catchment A and B. For the NASAP LSTM models, the NSE values were 0.48 and 0.46 for Catchment A and B, respectively. The ensemble predictions achieved

Table 2 | *P*-values for differences in Nash–Sutcliffe Efficiency (NSE) and Kling–Gupta Efficiency (KGE) values between the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP), and Agricultural Research Council (ARC) weather station data for gated recurrent unit (GRU) and long short-term memory (LSTM)

RNN	ARC vs CHIRPS	ARC vs NASAP	CHIRPS vs NASAP	ARC vs NASAP
		Combination 1		Combination 2
		Catchment A		
		NSE		
GRU	1.1×10^{-2}*	2.3×10^{-3}*	0.2	1.1×10^{-6}*
LSTM	9.8×10^{-3}*	2.5×10^{-4}*	6.7×10^{-6}*	6.8×10^{-8}*
KGE				
GRU	7.7×10^{-3}*	1.7×10^{-3}*	0.38	1.6×10^{-5}*
LSTM	3.3×10^{-3}*	1.3×10^{-4}*	5.6×10^{-4}*	1.1×10^{-6}*
		Catchment B		
		NSE		
GRU	3.4×10^{-4}*	5.9×10^{-5}*	4.7×10^{-2}*	4.0×10^{-6}*
LSTM	1.4×10^{-6}*	9.1×10^{-7}*	2.9×10^{-5}*	6.8×10^{-8}*
		KGE		
GRU	1.0×10^{-3}*	4.1×10^{-5}*	1.9×10^{-2}*	1.8×10^{-5}*
LSTM	2.4×10^{-6}*	9.1×10^{-8}*	7.7×10^{-3}*	1.7×10^{-7}*

*Significantly different at values of $p \leq 0.05$.

Table 3 | *P*-values for differences in Nash–Sutcliffe Efficiency (NSE) and Kling–Gupta Efficiency (KGE) values between Combination 1 and 2 for National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council (ARC) weather station data for gated recurrent unit (GRU) and long short-term memory (LSTM)

Catchment A			Catchment B	
GRU	**LSTM**		**GRU**	**LSTM**
		NSE		
ARC				
4.2×10^{-4}*	2.0×10^{-5}*		0.12	3.3×10^{-3}*
		NASAP		
2.6×10^{-2}*	7.2×10^{-2}*		0.61	0.58
		KGE		
		ARC		
0.06	0.11		0.82	0.06
		NASAP		
1.7×10^{-2}*	0.09		0.62	0.19

*Significantly different at values of $p \leq 0.05$.

Table 4 | Nash–Sutcliffe efficiency values for the best long short-term memory (LSTM) and gated recurrent unit (GRU) and ensemble predictions for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP), and Agricultural Research Council (ARC) weather station data

		Catchment A		Catchment B	
Dataset	**RNN**	**Best model**	**Ensemble**	**Best model**	**Ensemble**
		Combination 1			
CHIRPS	GRU	0.54	0.38	0.44	0.43
	LSTM	0.48	0.42	0.46	0.41
NASAP	GRU	0.44	0.19	0.44	0.40
	LSTM	0.24	0.27	0.40	0.31
ARC	GRU	0.63	0.51	0.71	0.66
	LSTM	0.64	0.52	0.69	0.65
		Combination 2			
NASAP	GRU	0.55	0.39	0.54	0.38
	LSTM	0.35	−0.09	0.40	0.26
ARC	GRU	0.74	0.73	0.71	0.72
	LSTM	0.72	0.72	0.72	0.72

NSE values of 0.19 and 0.27 for GRU and LSTM models in Catchment A, while for Catchment B, they were both 0.40. When observing the KGE values, the best NASAP GRU models achieved values of 0.51 in Catchment A and 0.45 in Catchment B. The LSTM models had KGE values of 0.50 and 0.46 for Catchment A and B, respectively. The ensemble predictions for KGE values were noticeably lower at −0.15 and 0.22 for the GRU and LSTM models in Catchment A and 0.20 and 0.30 for Catchment B, respectively.

The best NASAP models for Combination 2 achieved higher NSE (Table 4) and KGE (Table 5) values than Combination 1. The best NASAP models for Catchment A achieved NSE values of 0.55 and 0.35 for the GRU and LSTM models, respectively. The corresponding NSE values for the ensemble predictions were 0.39 and −0.09, respectively. The best NASAP models for Catchment A had maximum KGE values of 0.61 and 0.39 for the GRU and LSTM models, respectively. The corresponding KGE values for the ensemble predictions were 0.45 and 0.06, respectively. The best NASAP GRU and LSTM models of Catchment B achieved maximum NSE values of 0.54 and 0.40, respectively, and 0.38 and 0.26 for the corresponding

Table 5 | Kling–Gupta efficiency values for the best long short-term memory (LSTM) and gated recurrent unit (GRU) and ensemble predictions for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP), and Agricultural Research Council (ARC) weather station data

		Catchment A		Catchment B	
Dataset	**RNN**	**Best model**	**Ensemble**	**Best model**	**Ensemble**
		Combination 1			
CHIRPS	GRU	0.58	0.22	0.57	0.38
	LSTM	0.50	0.47	0.46	0.49
NASAP	GRU	0.51	−0.15	0.47	0.20
	LSTM	0.14	0.21	0.35	0.31
ARC	GRU	0.72	0.25	0.81	0.62
	LSTM	0.74	0.35	0.76	0.68
		Combination 2			
NASAP	GRU	0.61	0.45	0.59	0.45
	LSTM	0.39	0.06	0.44	0.36
ARC	GRU	0.77	0.86	0.73	0.79
	LSTM	0.76	0.79	0.77	0.83

ensemble predictions. The best GRU and LSTM models for Catchment B achieved maximum KGE values of 0.59 and 0.44, respectively, and 0.45 and 0.36 for the corresponding ensemble predictions.

For Combination 1, the best ARC models for Catchment A had maximum NSE values of 0.64 and 0.63 for the LSTM and GRU models, respectively. The corresponding NSE values for the ensemble predictions were 0.52 and 0.51, respectively. The best ARC models for Catchment A had maximum KGE values of 0.72 and 0.74 for the GRU and LSTM models, respectively. The corresponding KGE values for the ensemble predictions were 0.25 and 0.35, respectively. The best ARC GRU and LSTM models of Catchment B achieved maximum NSE values of 0.71 and 0.69, respectively, and 0.66 and 0.65 for the corresponding ensemble predictions, respectively. The best GRU and LSTM models for Catchment B achieved maximum KGE values of 0.81 and 0.76, respectively, and 0.62 and 0.68 for the corresponding ensemble predictions, respectively.

The best ARC models of Combination 2 for Catchment A achieved maximum NSE values of 0.74 and 0.72 for the GRU and LSTM models, respectively, and NSE values of 0.73 and 0.72, respectively, for the corresponding ensemble predictions. The increase in ensemble prediction NSE values for Catchment B was smaller than for Catchment A. The best ARC models of Combination 2 for Catchment B achieved NSE values of 0.72 and 0.71 for the LSTM and GRU models, respectively, and NSE values of 0.72 for both ensemble predictions. The best ARC models of Combination 2 for Catchment A achieved KGE values of 0.77 and 0.76 for the GRU and LSTM models, respectively, and KGE values of 0.86 and 0.79 for the corresponding ensemble predictions. The increase in ensemble prediction KGE values for Catchment B was smaller than for Catchment A, with the best ARC models of Combination 2 for Catchment B achieving KGE values of 0.73 and 0.77 for the GRU and LSTM models, respectively, and KGE values of 0.79 and 0.83 for the corresponding ensemble predictions.

For all Combination 1 models, the NSE (Table 4) and KGE (Table 5) values of the ensemble predictions were lower than the NSE and KGE values of any single best model in all cases. For Combination 2 (only NASAP and ARC data-driven models), this was still the case for the NASAP models; however, for the ARC models, the NSE values of the ensemble predictions were equal to the best models, and the KGE values of the ensemble predictions were better than the best models in all cases.

Hydrographs for 2 years from the testing set for Catchment A are provided as an illustration (Figure 7). The model predictions generally showed a good fit with the observed streamflow, although they struggled to predict the very high levels of extreme peak flow events during 2000. Devastating floods in Mozambique, Zimbabwe, and South Africa (including the study area) during February to March 2000 brought about by Tropical Cyclone Eline in late February and a tropical depression early in March (Reason & Keibel 2004). The extreme conditions of the cyclone provided a valuable opportunity to assess the ability of the models to extrapolate to unobserved conditions (such extreme flood events were not present in the training set), a crucial aspect in the context of climate change.

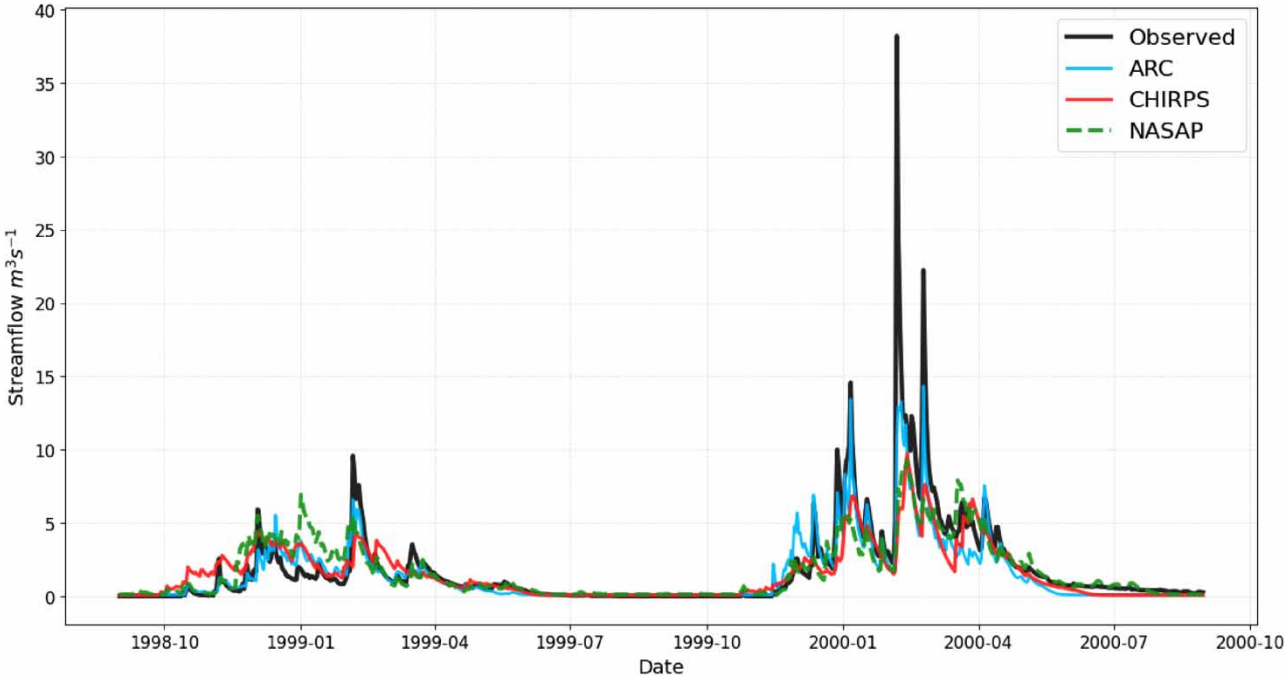


Figure 7 | Best model of the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP), and Agricultural Research Council (ARC) weather station data plotted against observed streamflow for 2 years of the testing set.

DISCUSSION

The method used in this study to predict streamflow involved two key aspects that distinguish it from previously published approaches. Firstly, the incorporation of simulated streamflow values into the LBW for predicting the streamflow of the testing set was a novel approach that has not, to the best knowledge of the authors, been explored in the context of streamflow prediction. Secondly, the required modifications to the GRU and LSTM architectures, specifically the change in activation function of the cells in the final hidden layer of the network and the non-negative constraint used in the dense layer, played a critical role in making this method effective. The unique combination of these aspects has contributed to the advancement of DL techniques in hydrological modelling.

One notable observation is that researchers using DL in hydrology often lack details in their reporting methodologies (Sit *et al.* 2020), making it difficult to determine the exact methods used. While the use of calculated lagged streamflow variables has been observed in numerous articles, the specific approach adopted in this study, to incorporate predicted streamflow into the LBW, appears to be distinct. Similar approaches may, however, have been used for hourly data (Sit *et al.* 2021). Contrary to the approach used in this study, Sit *et al.*'s (2021) model also utilized streamflow data from gauging stations situated in the upstream river network, which might be a limitation when such information is not available.

The approach of constraining the architectures is in line with the principles of theory-guided data science, specifically the concept of theory-guided design of model architecture (Karpatne *et al.* 2017). The targeted modifications to the architectures ensured physically consistent predictions and were aimed at ensuring that the models adhered to certain physical principles, albeit in a simplified form, and proved effective in mitigating physically unrealistic predictions (negative streamflow values), allowing for the prediction of long-term streamflow time series. However, the 'black box' nature of DL models remains a challenge for the hydrological community (Anderson & Radić 2022). The fact that it is not completely understood exactly how the two modifications prevented negative streamflow predictions is recommended for future work.

Wang & Karimi (2022) recently utilized the CAMELS dataset to analyse the effects of catchment mean rainfall and spatially distributed rainfall data on LSTM networks and found that including spatial distribution information of rainfall could improve performance. However, as in this case, high-quality, spatially distributed catchment weather datasets are not always available in less-developed countries. The method developed here addresses the limitation by incorporating

past streamflow data into the LBW of the model. This served as an implicit but effective proxy for hydrological processes that go unaccounted for and a lack of more spatially representative weather data. Specifically, the inclusion of historical streamflow captured some of the aggregated impacts of spatial and temporal variations within the catchment area, albeit indirectly.

The best models trained using weather station data could be used to generate reliable streamflow estimates with both GRU and LSTM networks. The best models trained using data from CHIRPS and NASAP were only moderately accurate. However, given the decline in weather station data availability, it is still encouraging that moderately accurate predictions were achieved using free data sources that are available for most of the surface of the earth. Future work could also explore combining CHIRPS and NASAP as input data to take advantage of the strengths of each data product while minimizing limitations (Kratzert *et al.* 2021). For example, CHIRPS may provide better accuracy for precipitation data, while NASAP could provide additional variables, such as temperature.

The results showed that rainfall together with streamflow was sufficient to obtain accurate streamflow predictions. Including minimum and maximum temperatures in the models did not consistently increase the accuracy, contrasting with the study by Fan *et al.* (2020). The reason for this might be attributed to the fact that streamflow predictions are primarily driven by rainfall and its subsequent runoff, with temperature playing a secondary role in influencing evapotranspiration rates and soil moisture dynamics.

The ARC ensemble predictions for Combination 2 outperformed the best single models in terms of KGE for both catchments. For the NASAP dataset, which relied on remote sensing products, the ensemble did improve in some instances, but not as much as the ARC-based ensemble, and decreased in others. This discrepancy is considered to be related to the quality of the input data, specifically, the noise inherent in remote sensing products like NASAP, which leads to more varied model outcomes and, consequently, lower average accuracy in the ensemble. It is also essential to consider the nature of the catchments. Catchment characteristics, such as topography, soil type, and land use, can play a significant role in how different models perform (Beven 2011). It seems that Catchment A has some specific features that make the ARC model ensemble, especially when temperature data are incorporated, perform exceptionally well, while the models with the NASAP dataset struggled.

Rainfall and temperature only were used in this study to test the applicability of this method for data-scarce regions, but further research with higher numbers of input variables is recommended. Future work could also investigate (1) the integration of other relevant data, such as evapotranspiration and normalized difference vegetation index, and (2) feature design (such as lagged variables), to further enhance the predictive capabilities of the hydrological models.

Although accurate streamflow predictions were achieved, there is room for improvement in capturing extreme peak flows, particularly in the context of climate change with an expected increase in extreme flood events (Allan 2021). Future work could explore if combining this method with process-based models in a hybrid approach (Senent-Aparicio *et al.* 2019), or through informing the DL networks with physical principles (Reichstein *et al.* 2019), could increase accuracy in extreme peak flow prediction. The fact that these models were able to exploit patterns in historic streamflow and weather data only to predict streamflow accurately, while not requiring information of geophysical properties of the catchments, is both an advantage and a limitation. This is because these models cannot be used to account for or assess the impact of land cover changes. Future work should explore the addition of land cover/land use change data, especially in more impacted catchments, where these changes can have significant impacts on streamflow (Beven 2011). Finally, the data of many catchments could be combined together in an attempt to develop a model to predict streamflow in ungauged basins (Kratzert *et al.* 2019b).

CONCLUSION

In this study, the research questions were as follows: (1) if LSTM and GRU networks could be successfully developed in semi-arid South African catchments to generate reliable streamflow estimations, and (2) whether freely available gridded weather data could be used to produce reasonably accurate streamflow estimates.

A challenge was obtaining spatially distributed weather data for the catchments. To compensate, 30 days' worth of past streamflow was used alongside rainfall and temperature to train the models to predict the next day's streamflow. Predicted streamflow was incorporated into the model input data during model testing to simulate long-term time series. This resulted in negative streamflow predictions, mainly during the dry seasons. The GRU and LSTM were constrained in two ways to

eliminate these predictions. This allowed for the successful training of GRU and LSTM networks using historic streamflow and weather data that could be used to generate satisfactory streamflow estimates based on two statistical criteria.

While the approach does not fully integrate hydrological or physical theories in the way that recent advances in physics or hydrologically informed ML/DL have, it does employ domain-specific adjustments to improve the physical plausibility of model outputs, thereby offering a robust solution especially well-suited for regions where spatially comprehensive meteorological data are not readily available.

Both GRU and LSTM networks could be used as a fast and efficient technique to generate streamflow information. These techniques showed great potential and could be used, for example, to generate missing streamflow data for streamflow stations that were monitored in the past but are no longer monitored, or to fill gaps in records – a crucial need for long-term climate and hydrological studies. There is also potential to further develop and apply these models for short-term (daily time step) streamflow forecasts, which could have important implications for flood risk management and water resource planning. Testing this method in more catchments under varying characteristics, where sufficient historic streamflow and weather data are available, is recommended to further validate this method.

This study showed that 17 years of training data was sufficient to successfully train the models, although longer training periods could increase performance. The DWS website (https://www.dws.gov.za/Hydrology/Verified/hymain.aspx) contains thousands of measured streamflow records, and although monitoring at many stations ended decades ago, there are hundreds of stations with at least 20 years of data and many stations with longer records. These old records could be potentially used to train low-cost GRU and LSTM streamflow prediction models with the method developed here to obtain estimates of streamflow where observations are currently unavailable due to a lack of monitoring.

Given that LSTM and GRU networks are inherently designed for sequential data, the method, though initially developed for streamflow prediction, holds promise for a broad range of time series applications within hydrological modelling. These could include forecasting variables such as evapotranspiration (ET), soil moisture, and groundwater levels.

In conclusion, this study contributes to the growing body of literature affirming the capabilities of DL in hydrological modelling. It also provides valuable insights into the specific challenges and solutions associated with applying these models to streamflow prediction.

ACKNOWLEDGEMENTS

The Water Research Commission of South Africa provided financial support for the conduct of the research under project number: C2020/2021-00440. The weather station data were provided by the Agricultural Research Council (https://www.agroclimate.agric.za/WP/WP/).

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories: https://data.waterresearchobservatory.org/metadata-form/deep-learning-for-streamflow-prediction-project-data.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J. & Devin, M. 2016 Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv*, 1. https://doi.org/10.48550/arXiv.1603.04467.

Addor, N., Newman, A. J., Mizukami, N. & Clark, M. P. 2017 The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences* **21**, 5293–5313. https://doi.org/10.5194/hess-21-5293-2017.

Allan, R. P. 2021 *Climate Change 2021: The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/9781009157896.

Anderson, S. & Radić, V. 2022 Evaluation and interpretation of convolutional long short-term memory networks for regional hydrological modelling. *Hydrology and Earth System Sciences* **26**, 795–825. https://doi.org/10.5194/hess-26-795-2022.

Beven, K. J. 2011 *Rainfall-runoff Modelling: The Primer*. John Wiley & Sons, Sussex, UK. ISBN: 1119951011.

Bishop, C. M. & Nasrabadi, N. M. 2006 *Pattern Recognition and Machine Learning*. Springer, New York, NY. ISBN: 978-0-387-31073-2.

Cho, K., van Merriënboer, B., Bahdanau, D. & Bengio, Y. 2014 On the properties of neural machine translation: Encoder-decoder approaches. *arXiv*. https://doi.org/10.48550/arXiv.1409.1259.

du Plessis, J. & Kibii, J. 2021 Applicability of CHIRPS-based satellite rainfall estimates for South Africa. *Journal of the South African Institution of Civil Engineering* **63**, 43–54. http://dx.doi.org/10.17159/2309-8775/2021/v63n3a4.

Dws, D. O. W. A. S. 2021 *National State of Water Report for South Africa – Hydrological Year 2019/20*. Department of Water and Sanitation, Pretoria.

Engelbrecht, F., Mcgregor, J. & Engelbrecht, C. 2009 Dynamics of the Conformal-Cubic Atmospheric Model projected climate-change signal over southern Africa. *International Journal of Climatology: A Journal of the Royal Meteorological Society* **29**, 1013–1033. https://doi.org/10.1002/joc.1742.

Fan, H., Jiang, M., Xu, L., Zhu, H., Cheng, J. & Jiang, J. 2020 Comparison of long short term memory networks and the hydrological model in runoff simulation. *Water* **12**, 175. https://doi.org/10.3390/w12010175.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L. & Hoell, A. 2015 The climate hazards infrared precipitation with stations – A new environmental record for monitoring extremes. *Scientific Data* **2**, 1–21. https://doi.org/10.1038/sdata.2015.66.

Gaffoor, Z., Pietersen, K., Jovanovic, N., Bagula, A., Kanyerere, T., Ajayi, O. & Wanangwa, G. 2022 A comparison of ensemble and deep learning algorithms to model groundwater levels in a data-scarce aquifer of Southern Africa. *Hydrology* **9**. https://doi.org/10.3390/hydrology9070125.

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

Ghimire, S., Yaseen, Z. M., Farooque, A. A., Deo, R. C., Zhang, J. & Tao, X. 2021 Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. *Scientiffic Reports* **11**, 17497. https://doi.org/10.1038/s41598-021-96751-4.

Goodfellow, I., Bengio, Y. & Courville, A. 2016 *Deep Learning*. MIT Press, Cambridge, MA. ISBN: 0262337371.

Herold, C. & Bailey, A. 2016 *Water Resources of South Africa, 2012 Study (WR2012)*. Water Research Commission, Pretoria, South Africa.

Hochreiter, S. & Schmidhuber, J. 1997 Long short-term memory. *Neural Computation* **9**, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Hughes, D. 2004 Three decades of hydrological modelling research in South Africa. *South African Journal of Science* **100**, 638–642. Available from: https://hdl.handle.net/10520/EJC96172.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. & Kumar, V. 2017 Theory-guided data science: A new paradigm for scientific discovery from data. *Institute of Electrical and Electronics Engineers Transactions on Knowledge and Data Engineering* **29**, 2318–2331. doi:10.1109/TKDE.2017.2720168.

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

Kling, H., Fuchs, M. & Paulin, M. 2012 Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology* **424**, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. 2018 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.

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S. & Klambauer, G. 2019a Neuralhydrology–interpreting LSTMs in hydrology. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* 347–362. https://doi.org/10.1007/978-3-030-28954-6_19.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S. & Nearing, G. S. 2019b Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research* **55**, 11344–11354. https://doi.org/10.1029/2019WR026065.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. & Nearing, G. 2019c 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.

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

Krizhevsky, A., Sutskever, I. & Hinton, G. E. 2017 Imagenet classification with deep convolutional neural networks. *Communications of the Association for Computing Machinery* **60**, 84–90. https://doi.org/10.1145/3065386.

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 LSTM-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-25-5517-2021.

Mann, H. B. & Whitney, D. R. 1947 On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* **18**, 50–60.

Moriasi, D. N., Arnold, J. G., van Liew, M. W., Bingner, R. L., Harmel, R. D. & Veith, T. L. 2007 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the American Society of Agricultural and Biological Engineers* **50**, 885–900. doi:10.13031/2013.23153.

Muhammad, A. U., Li, X. & Feng, J. 2019 Using LSTM GRU and hybrid models for streamflow forecasting. In: *Machine Learning and Intelligent Communications: 4th International Conference, MLICOM 2019*, August 24–25, 2019, Nanjing, China. Proceedings 4. Springer, pp. 510–524. https://doi.org/10.1007/978-3-030-32388-2_44.

Najafzadeh, M. & Anvari, S. 2023 Long-lead streamflow forecasting using computational intelligence methods while considering uncertainty issue. *Environmental Science and Pollution Research* **30**, 84474–84490. https://doi.org/10.1007/s11356-023-28236-y.

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

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 Resources Research* **57**. https://doi.org/10.1029/2020WR028091.

Nifa, K., Boudhar, A., Ouatiki, H., Elyoussfi, H., Bargam, B. & Chehbouni, A. 2023 Deep learning approach with LSTM for daily streamflow prediction in a semi-arid area: A case study of Oum Er-Rbia River Basin, Morocco. *Water* **15**, 262. https://doi.org/10.3390/w15020262.

Odendaal, N. 2021 *Govt Needs to Focus on Securing Reliable Hydrological Information to Ensure Water Security*. Available from: https://www.engineeringnews.co.za/article/govt-needs-to-focus-on-securing-reliable-hydrological-information-to-ensure-water-security-2021-02-24/rep_id:4136.

Oyebande, L. 2001 Water problems in Africa – How can the sciences help? *Hydrological Sciences Journal* **46**, 947–962. https://doi.org/10.1080/02626660109492888.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011 Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **12**, 2825–2830.

Pitman, W. V. & Bailey, A. K. 2021 Can CHIRPS fill the gap left by the decline in the availability of rainfall stations in southern Africa? *Water SA* **47**. https://doi.org/10.17159/wsa/2021.v47.i2.10912.

Razavi S. 2021 Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *Environmental Modelling & Software* **144**, 105159. https://doi.org/10.1016/j.envsoft.2021.105159.

Reason, C. & Keibel, A. 2004 Tropical cyclone Eline and its unusual penetration and impacts over the southern African mainland. *Weather and Forecasting* **19**, 789–805. https://doi.org/10.1175/1520-0434(2004)019 < 0789:TCEAIU > 2.0.CO;2.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J. & Carvalhais, N. 2019 Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Roberts, W., Williams, G. P., Jackson, E., Nelson, E. J. & Ames, D. P. 2018 Hydrostats: A Python package for characterizing errors between observed and predicted time series. *Hydrology* **5**, 66. https://doi.org/10.3390/hydrology5040066.

Rogers, D. P., Tsirkunov, V. V., Kootval, H., Soares, A., Kull, D., Bogdanova, A.-M. & Suwa, M. 2019 *Weathering the Change: How to Improve Hydromet Services in Developing Countries?* World Bank, Washington, DC.

Senent-Aparicio, J., Jimeno-Sáez, P., Bueno-Crespo, A., Pérez-Sánchez, J. & Pulido-Velázquez, D. 2019 Coupling machine-learning techniques with SWAT model for instantaneous peak flow prediction. *Biosystems Engineering* **177**, 67–77. https://doi.org/10.1016/j.biosystemseng.2018.04.022.

Shapiro, S. S. & Wilk, M. B. 1965 An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611.

Shirzadi, N. 2023 *Time Series Analysis and Forecasting with Python*. Available from: https://www.udemy.com/course/time-series-analysis-and-forecasting-with-python/ (Accessed 6 June 2022).

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y. & Demir, I. 2020 A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology* **82**, 2635–2670. https://doi.org/10.2166/wst.2020.369.

Sit, M., Demiray, B. & Demir, I. 2021 Short-term hourly streamflow prediction with graph convolutional gru networks. *arXiv*. https://doi.org/10.48550/arXiv.2107.07039.

Wang, Y. & Karimi, H. A. 2022 Impact of spatial distribution information of rainfall in runoff simulation using deep learning method. *Hydrology and Earth System Sciences* **26**, 2387–2403. https://doi.org/10.5194/hess-26-2387-2022.

Zheng, A. & Casari, A. 2018 *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc, Sebastopol. ISBN: 9781491953242.

First received 17 November 2023; accepted in revised form 16 February 2024. Available online 28 February 2024