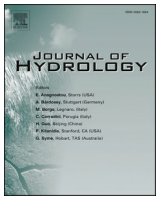
Journal of Hydrology 627 (2023) 130380

Contents lists available at ScienceDirect

Journal of Hydrology

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



Research papers

Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment

Behmard Sabzipour [a,*], Richard Arsenault [a], Magali Troin [a,b], Jean-Luc Martel [a], François Brissette [a], Frédéric Brunet [a], Juliane Mai [c,d,e,f]

[a] Hydrology, Climate and Climate Change Laboratory, École de technologie supérieure, Université du Québec, 1100 Notre-Dame Street West, Montreal, Quebec H3C 1K3, Canada
[b] TVT, Maison du Numérique et de l'Innovation, place Georges Pompidou, 83 000 Toulon, France
[c] Department of Civil and Environmental Engineering, University of Waterloo, 200 University Ave W, Waterloo, Ontario N2L 3G1, Canada
[d] Department of Earth and Environmental Science, University of Waterloo, 200 University Ave W, Waterloo, Ontario N2L 3G1, Canada
[e] Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research - UFZ, Permoser Strasse 15, 04318 Leipzig, Germany
[f] Center for Scalable Data Analytics and Artificial Intelligence - ScaDS.AI, Humboldtstraße 25, 04105 Leipzig, Germany

ARTICLE INFO

This manuscript was handled by Marco Borga, Editor-in-Chief

Keywords:
Long short-term memory (LSTM)
Hydrological forecasting
Data assimilation
Ensemble forecasting
Deep learning

ABSTRACT

Streamflow forecasting is crucial in water planning and management. Physically-based hydrological models have been used for a long time in these fields, but improving forecast quality is still an active area of research. Recently, some artificial neural networks have been found to be effective in simulating and predicting short-term streamflow. In this study, we examine the reliability of Long Short-Term Memory (LSTM) deep learning model in predicting streamflow for lead times of up to ten days over a Canadian catchment. The performance of the LSTM model is compared to that of a process-based distributed hydrological model, with both models using the same weather ensemble forecasts. Furthermore, the LSTM's ability to integrate observed streamflow on the forecast issue date is compared to the data assimilation process required for the hydrological model to reduce initial state biases. Results indicate that the LSTM model forecasted streamflows are more reliable and accurate for lead-times up to 7 and 9 days, respectively. Additionally, it is shown that the LSTM model using recent observed flows as a predictor can forecast flows with smaller errors in the first forecasting days without requiring an explicit data assimilation step, with the LSTM model generating a median value of mean absolute error (MAE) for the first day of lead-time across all forecast issue dates of 25 m^3/s compared to 115 m^3/s for the assimilated hydrological model.

1. Introduction

Accurate, reliable, and easily understandable hydrological forecasts are crucial for a wide range of users in the water-related sectors, such as agriculture, hydropower, and floodplain management (Boucher et al., 2012; Anghileri et al., 2016; Cassagnole et al., 2021). As a result, forecasting streamflow has been the focus of numerous studies since the mid-1970s (Twedt et al., 1977; Day, 1985), and has seen an increasing amount of attention in recent decades (Troin et al., 2021) as demands for water resource management and natural disaster mitigation have risen substantially.

There are two main methods used to forecast streamflow: the first is the use of dynamical (or process-driven) hydrological models, which range from conceptual and lumped to physically-based and distributed models; the second is the use of data-driven statistical models such as machine learning methods (ML) including artificial neural networks (ANNs) and autoregressive models (Rajagopalan et al., 2010). These approaches can provide an ensemble streamflow prediction (ESP) system when ensemble weather forecasts are used as inputs. However, dynamical hydrological models are often limited by the availability of data required for their implementation, such as soil type and depth; or by their simplistic process representations (Damavandi et al., 2019). These issues can be overcome using deep learning approaches, which can lead to reliable simulations of hydrologic systems even when the

* Corresponding author.
E-mail address: behmard.sabzipour.1@ens.etsmtl.ca (B. Sabzipour).

https://doi.org/10.1016/j.jhydrol.2023.130380
Received 7 February 2023; Received in revised form 1 September 2023; Accepted 16 September 2023
Available online 21 October 2023

underlying physical processes are not taken into account (Maier et al., 2010). While they require large volumes of data for training, they do not require certain data types explicitly (such as soil data that can be difficult to obtain). A particular type of deep learning model that has become increasingly popular in the last years in the field of hydrology, is the Long Short-Term Memory (LSTM) network, due to its ability to process sequential data and time series (Zhang et al., 2018; Shen and Lawson, 2021).

LSTMs, introduced by Hochreiter and Schmidhuber (1997), are an advanced type of recurrent neural networks (RNNs) designed to learn sequence (temporal) data and their long-term dependencies outperforming conventional RNNs (Kratzert et al., 2018). An LSTM consists of a memory cell, which is a neuron with a self-recurrent connection, and three nonlinear gates (i.e., forget, input and output gates) that control the movement of information within and outside the cell. The forget gate, or cell memory, determines what information from previous time steps should be discarded, allowing the LSTM to learn long-term dependencies that other RNNs cannot. The input gate regulates what information should be added to update the cell state, while the output gate determines how much of the cell state should be used to generate the output value. As a result, LSTMs can benefit from a longer memory to learn long-term dependencies by avoiding the exploding or vanishing gradients problematic that strongly affects traditional RNNs structures face (Xu et al., 2020). The capacity of LSTMs to overcome the issues short-term memory is particularly important when it comes to hydrological modelling, since processes such as the accumulation and melting of the snow cover and the evolution of soil moisture conditions are essential to properly model surface runoff. This has also been shown in a comparative study between LSTMs and other machine learning algorithms for hydrological forecasting (Rahimzad et al., 2021).

The potential use and benefits of LSTMs in the field of hydrology have recently begun to be explored (e.g., Hu et al., 2018; Kratzert et al., 2018; Zhang et al., 2018; Sahoo et al., 2019; Hunt et al., 2022; Arsenault et al., 2023). For example, Zhang et al. (2018) evaluated the performance of various RNN architectures in simulating water levels in Norway and found that the LSTM is better suited for multi-step-ahead forecasts than other architectures without cell memory. Kratzert et al. (2018) compared the performance of the LSTM and the SAC-SMA hydrological model (Sorooshian et al., 1993) for simulating long-term streamflow over 241 catchments in North America and found that the LSTM outperformed the SAC-SMA model, highlighting the potential of LSTMs as regional hydrological models. Tounsi et al. (2023) corroborated these findings by showing the LSTM model performance over the MOPEX dataset in the United States. Kratzert et al. (2018) and Le et al. (2021) also noted the possibility of applying LSTMs in regions other than the one used for training. Hunt et al. (2022) evaluated the performance of the LSTM in predicting streamflow in various climate regions of the United States and compared it to the Copernicus Emergency Management Service physics-based Global Flood Awareness System (GloFAS). The authors reported that the LSTM generated skillful forecasts that outperformed the raw and bias-corrected GloFAS forecasts up to a 5-day lead time. In this context, Hu et al. (2018) attributed the better performance of the LSTM compared to conceptual and physical-based hydrological models to the feature of the forget gate. However, other studies have reported the difficulty using LSTMs for predicting streamflow during extreme conditions, as well as for catchments with complex groundwater-river interactions and human abstractions (Kratzert et al., 2018; Lees et al., 2021, Cho and Kim, 2022, Granata et al., 2022). Arsenault et al. (2023) showed that LSTMs systematically outperformed traditional hydrological models in a regionalization experiment performed over 148 North American catchments, with Tang et al. (2023) showing global hydrology models paired with LSTMs could drastically improve streamflow prediction in ungauged basins. Nevo et al. (2022) developed a flood forecasting model using LSTMs and found that it was significantly better than more traditional multiple linear regression models in forecasting river water stages in India and Bangladesh. Other studies have investigated variants of LSTMs for forecasting, including methods that allow forecasting multiple time steps at once rather than generating forecasted time series sequentially. For example, Girihagama et al. (2022) used an attention-based Encoder-Decoder- LSTM to improve streamflow forecasting skill over ten catchments in Canada, and Kao et al. (2020) found Encoder-Decoder-LSTMs to be reliable in converting rainfall sequences to runoff sequences and suitable for forecasting hourly floods. Furthermore, other types of machine learning based forecasts have also shown promising results, such as radial basis function neural networks in Granata and Di Nunno (2023), which performed better than LSTMs for forecasts up to 15 days of lead-time on six UK rivers.

Incorporating and integrating observational data is important for better performance of hydrological forecasting, as it helps adjust model states such that the model represents actual hydrological conditions as best as possible. This is done by applying methods such as data assimilation (DA) and regressions for state updating (Nearing et al., 2022; Brajard et al., 2020; Fang and Shen, 2020). Nearing et al. (2022) discussed ways of integrating data in LSTMs to represent missing streamflow data for gauged and ungauged basins. They used a regression method and variational DA method. The integration part was added to the cell state of LSTM, i.e., the one which is the recursive state of LSTM, in order to over-train observed streamflow data which are spared. They found DA has advantages over autoregression, since it is able to deal with up to 50 % of missing data. They reported DA worked better when used for catchments including both gauged and ungauged basins. Brajard et al. (2020) reported successful results for combining the ensemble Kalman filter with a surrogate model of a neural network. Feng et al. (2020) showed that it was possible to use data integration of streamflow observations in LSTM models to improve hydrological forecasts at the continental scale, especially in snowmelt dominated catchments. This was also tested with success by Khoshkalam et al. (2023) at the catchment scale, while using simulated streamflow to replace any missing data from the observational record. Fang and Shen (2020) used LSTMs to forecast soil moisture using satellite data using a Data Integration kernel to assimilate and update models with the most recent available observations showing that it was effective to reflect unseen processes in inputs, such as floods.

This study aims to evaluate the potential of LSTM to simulate and forecast daily streamflow over the Lac-Saint-Jean (LSJ) catchment, located in the province of Quebec, Canada. In this region, a large portion of the streamflow comes from snowmelt, making it an ideal environment for testing LSTM forecasts in a wide array of hydrometeorological conditions. As there are only a few applications of LSTM forecasting in Nordic regions, the architecture and reliability of LSTM needs to be investigated to identify its benefits for streamflow forecasting as an operational deployment in this region. The strong snowpack dynamics and the importance of capturing such long-term hydrological processes make it more challenging than in regions with more uniform weather. The only similar study to our knowledge is that of Girihagama et al. (2022), who used an encoder-decoder LSTM with an attention mechanism to provide streamflow forecasts in ten river catchments in the Great Lakes region in Canada. Their study concluded that this variant of LSTM was able to provide excellent forecasting results for up to five days of lead-time.

In particular, this study seeks to address the following research questions:

- How well do LSTM-based models forecast streamflow over a catchment where the hydrologic response is dominated by snowmelt?
- How does LSTM-based model performance compare with the operational forecasting system (a distributed hydrological model combined with a data assimilation scheme) used as a benchmark?

2. Experimental design

2.1. Study area

This study was conducted on the LSJ catchment in the province of Quebec, Canada (Fig. 1). The catchment has an area of 45000 km^2 and is used for hydropower generation by the Rio Tinto corporation for aluminum smelting. The catchment is made up of nine monitored sub-catchments, which drain into the reservoir, as shown in Fig. 1a. This figure also shows the catchment's location (Fig. 1d). The sub-catchment "other tributaries" is made up of a series of small rivers and streams that all flow to the reservoir but are ungauged. The annual average precipitation on the LSJ catchment is 1000 mm, of which 34 % falls as snow, as measured by Rio Tinto's weather monitoring network stations. This provides substantial inflows to the Lac-St-Jean reservoir, a 1000 km^2 reservoir that can store up to 4550 hm^3 of water for the hydropower generating station (Arsenault and Côté, 2019). Spring peak flows are associated with snowmelt events, while high flows in summer and fall are related to precipitation events. Mean annual streamflow is about 900 m^3s^{-1} (Bergeron et al., 2021).

2.2. Datasets

In this study, multiple datasets were used, including streamflow data as well as observed and forecasted weather data. This section describes all datasets and their pre-processed methodology.

2.2.1. Observed hydrometeorological data (used for calibration of hydrologic model)

The observed hydrometeorological data were provided by Rio Tinto, the operator of the hydropower generating station and owner of water rights for the LSJ system. The data covered the period from January 1954 to December 2019 and included daily minimum and maximum temperatures and precipitation from a network containing 16 weather stations distributed throughout the catchment.

In addition to the weather data, Rio Tinto also provided inflows to the main reservoir, which were derived through mass balance calculations by evaluating changes in the reservoir level at various locations along with known outflows from turbines and spillways. This was necessary as multiple rivers and tributaries flow into the reservoir, making a significant portion of the flows ungauged. However, the mass-balance derived inflows can be noisy at times due to wind displacing the water surface, which can bias storage volumes over short periods (Loiselle et al., 2021). To address this, a three-day moving average was applied to smooth out variations in inflows, which did not change the total water balance over longer horizons, or the timing of events in a significant manner. The inflows to the reservoir are the target of the forecasting procedure in this study.

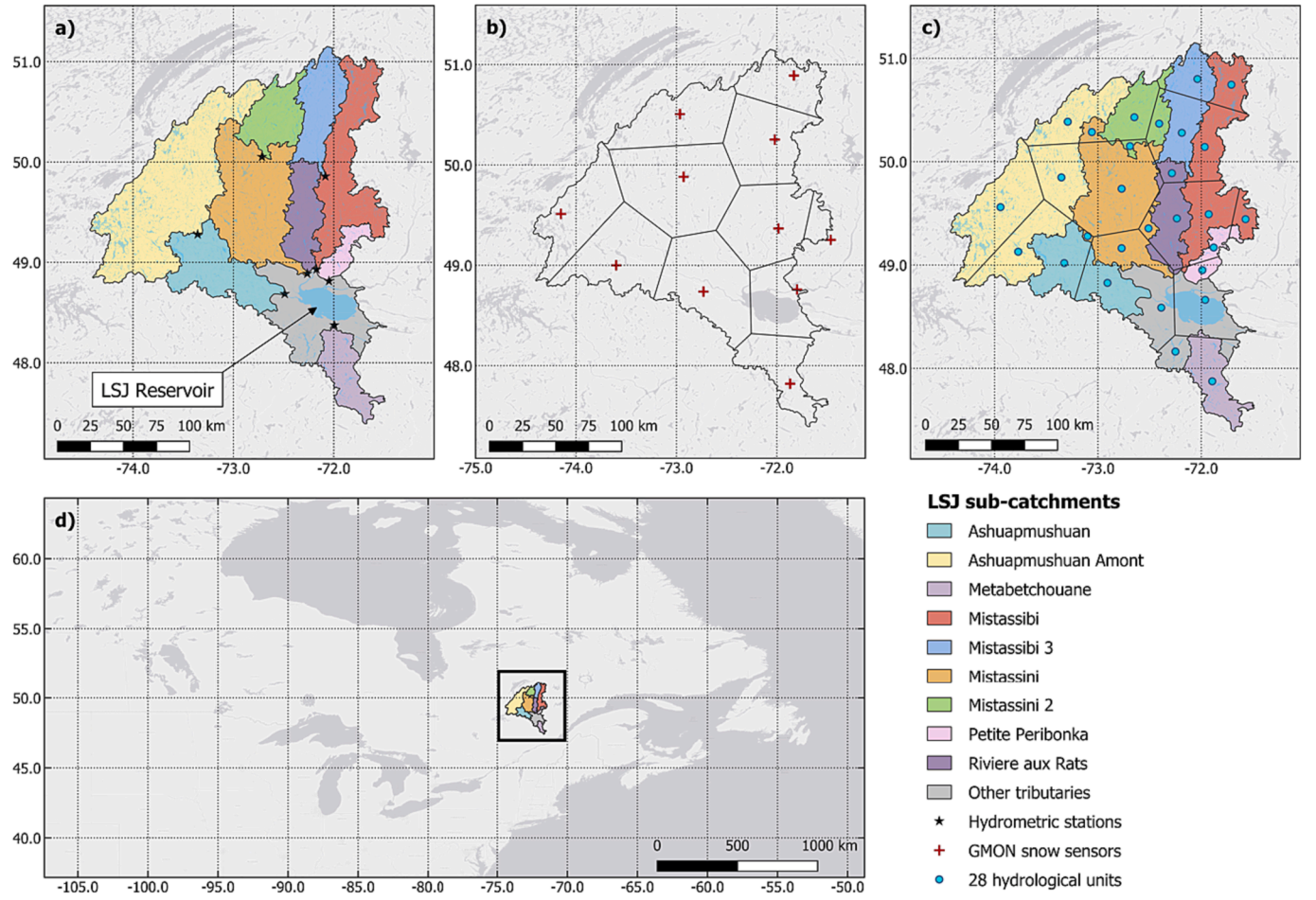


Fig. 1. The Lac St-Jean (LSJ) catchment and its ten contributing sub-catchments, including the "other tributaries" which are the sum of all smaller ungauged rivers flowing into the LSJ reservoir (a). Hydrometric stations are represented by stars (a). The eleven gamma ray monitors (GMON) are represented by red crosses over the Voronoi polygons they generate (b). The intersection of the ten sub-catchments and the eleven Voronoi polygons represent the 28 hydrological response units (HRU) of the CEQUEAU semi-distributed hydrological model used in this study, represented by a blue circle (c). The location of the LSJ catchment in Quebec, Canada, is shown within Eastern North America (d). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2.2. Ensemble weather forecast (used for forecasting with hydrologic and LSTM model)

To assess the performance of the distributed hydrological model and the LSTM model in forecasting mode, the 50-member operational ensemble weather forecast data was obtained from the European Center for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) using the Meteorological Archival and Retrieval System (MARS) archive (https://www.ecmwf.int/en/forecasts/dataset/operational-archive; Grawinkel et al., 2015). The variables of total precipitation, maximum and minimum temperature were available at a spatial resolution of 0.2° x 0.2° on a 6-hourly time step twice per day (0Z and 12Z). For this study, only the 0Z-emitted forecasts were utilized and the four 6-hour timesteps were aggregated to a daily time step for each day. The forecast lead-times from one to ten days were downloaded, but the last day was truncated due to the five-to-six-hour time zone offset that made the tenth day unavailable for the entire period for the study location, resulting in an effective nine-day ensemble weather forecast. The forecasts were downloaded for the period of January 2015 to December 2019. Data prior to 2015 were not included due to a major update of the ECMWF integrated forecasting system in 2015, which caused a change in weather forecast statistics that would not be representative of the more recent model versions. The ensemble forecasts of precipitation and temperature were then spatially aggregated to the scale of the 28 distributed hydrological model sub-regions for the hydrological model and over the entire catchment as input to the LSTM model.

2.2.3. Reanalysis data (used for training of LSTM model)

In this study, a second set of pseudo-observed meteorological data, known as reanalysis data, was used to train the LSTM model. This data is from the ECMWF fifth generation reanalysis (ERA5; Hersbach et al., 2020). The reanalysis data used to provide observations on the historical period, so that the LSTM model could be used to forecast data with the same variables in the forecasting period. The station-based observations provided by Rio Tinto (Section 2.2.1) did not contain the desired spatial and temporal coverage of wind, solar radiation, and pressure variables for this training, thus ERA5 data was used instead. The variables used from ERA5 are the same as for the forecast data, but at an hourly time step on a 0.25° x 0.25° resolution. The ERA5 data was aggregated at the daily scale and was spatially aggregated over the entire catchment, to maintain consistency with the spatial and temporal scales of the ensemble weather forecast data. The data is available from January 1979 to the present day with a latency of approximately five days, which was deemed sufficient for training the LSTM model in this study.

3. Methods

This study aims to evaluate the ability of the LSTM neural network model to forecast streamflows on a large, snowmelt-dominated catchment. A traditional semi-distributed hydrological model is also used as a comparison over the same catchment and time periods. The methods used to achieve the study objectives are described in detail in this section, including the forecast evaluation metrics (Section 3.1), the traditional hydrological modeling and forecasting (Section 3.2), and the LSTM model training and forecast testing (Section 3.3).

3.1. Performance evaluation criteria

The performance of the different streamflow forecasts is evaluated using the Kling-Gupta Efficiency (KGE; Gupta et al., 2009), the Continuous Ranked Probability Score (CRPS; Hersbach, 2000) and the Mean Absolute Error (MAE; Mather and Johnson, 2016) at different lead times (from days one to nine) during the forecast.

The KGE, which is unit less, is defined as:

$$KGE = 1 - \sqrt[2]{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \quad (1)$$

where r is the Pearson correlation coefficient; β is the bias ratio ($\mu_{q_{sim}}/\mu_{q_{obs}}$); and γ is the variability ratio ($\sigma_{q_{sim}}/\sigma_{q_{obs}}$) where $\mu_{q_{sim}}$ and $\sigma_{q_{sim}}$ are the mean and standard deviation of the forecasted (simulated) streamflow and $\mu_{q_{obs}}$ and $\sigma_{q_{obs}}$ are the equivalent for the observed streamflow. A perfect forecast would have a KGE of 1, while suboptimal forecasts show KGE values lower than 1.

The CRPS is a suitable metric for probabilistic forecasts. It measures the average distance between the observed probability density function ($F(q_{obs})$), and the ensemble forecast probability density function ($F(q_{sim})$), given by:

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\left[F(q_{sim_i}) - F(q_{obs})\right]^2 dq \quad (2)$$

where N is the size of the forecast ensemble. The CRPS ranges from 0 to + ∞, with zero being a perfect forecast. For streamflow, the units of CRPS is m^3s^{-1}. CRPS has the same dimension as q (dq).

MAE measures the difference between the observed and forecasted results but in a deterministic (rather than ensemble) setting. The MAE is the equivalent of the CRPS when a single member is used, and therefore the CRPS can be seen as an extension of MAE to ensembles. MAE is an average of the absolute error, as:

$$MAE_{(t)} = \frac{\sum_{i=1}^{N}\left|q_{sim_i}(t) - q_{obs}(t)\right|}{N} \quad (3)$$

where q_{obs} is the observed streamflow, q_{sim_i} is the simulated streamflow (which also means pseudo-forecasted as well) for the i-th member of the ensemble of size N at lead-time t. The optimal value for MAE is 0. In this study, MAE values presented are the average MAE over all forecasts $MAE_{(t)}$.

3.2. The CEQUEAU hydrological model

The CEQUEAU hydrological model is used as a benchmark for evaluating model performance statistics in this study. CEQUEAU is a semi-distributed model that incorporates three conceptual reservoirs to simulate key hydrologic processes: a surface reservoir to model land surface processes such as evapotranspiration and snowmelt), a soil reservoir that describes the soil infiltration and interflow processes, and a deeper reservoir that simulates groundwater and base flows (Morin and Paquet, 2007). Evapotranspiration is estimated using the Oudin method (Oudin et al., 2005) and snow processes are modeled with the degree-day CEMANEIGE model (Valéry, 2010).

The catchment is divided into 28 hydrologic response units, which corresponds to the intersection of the influence area of GMON snow monitoring stations and the 10 hydrologic sub-catchments (Fig. 1a-1c). The simulation of vertical fluxes is independently performed on each of the 28 sub-regions, as described in Mai et al. (2020). Unit hydrographs are then employed to route flows to the downstream sub-catchments until they reach the outlet. CEQUEAU only requires daily precipitation and average daily temperature as meteorological inputs. The model was manually adjusted and calibrated by Rio Tinto, which provided their operational streamflow forecasting model for this study. The model was set up using data from January 1954 to December 2014 and has been their primary hydrological forecasting tool since.

CEQUEAU was then employed to generate streamflow forecasts for the period of January 2015 to December 2019, both in open-loop mode (i.e., without data assimilation) and using an implementation of the Ensemble Kalman Filter (EnKF; Evensen, 2003) data assimilation procedure. The EnKF used in this study was optimized for forecasting performance on the LSJ catchment and was implemented according to the procedure described in Sabzipour et al. (2023), which is summarized in the supplementary material (Section S1). CEQUEAU was forced with

observed meteorological variables (Section 2.2.1) until the forecast date using the EnKF to maintain robust state variables up to the day the forecast was issued. At this point, the hydrological model was run using the ECMWF ensemble weather forecasts (Section 2.2.2) as meteorological input to generate the ensemble streamflow forecasts. Data assimilation was performed every three days to increase computation efficiency and to prevent the model's initial states from drifting too far away from the observations. Streamflow forecasts generated by CEQUEAU (with and without data assimilation) were only generated on the dates that data assimilation was performed to ensure the best possible initial states while allowing a small temporal gap between streamflow forecasts to prevent strong autocorrelations between successive forecast dates.

3.3. Long short-term memory (LSTM) network

In this study, a single-layer LSTM model with a variable number of units was chosen as the model structure. The setup details and hyperparameters of the LSTM are summarized in Table 1. Different hyperparameters and model structures were considered for each lead-time during the model training and were tested to obtain the best results as measured by the KGE metric. The dropout rate, number of epochs, batch size, and number of LSTM units hyperparameters were modified and optimized for each lead-time through trial-and-error to improve performance on the validation period. The dropout rate was included to randomly turn off a certain percentage of LSTM units during training, in order to reduce overfitting and improve modeling performance during forecasting. The number of epochs, or the number of times the model sees the full dataset during the training, and batch size, or the size of the dataset sub-sample used to estimate the gradient descent, were adjusted to speed up the training and ensure its convergence. This led to a larger number of epochs for longer lead-times.

The LSTM models were trained and tested using multiple meteorological variables from the ERA5 reanalysis dataset, including precipitation, temperature, wind, surface pressure, net solar radiation and dewpoint temperature, (Section 2.2.3) and daily observed streamflow provided by Rio Tinto (Section 2.2.1) for the period from January 1979 to December 2014. The LSTM models were trained on sequences of [365-k] days of hydrometeorological data prior to a single observed streamflow target, where k is the lead-time for which the model will be used to forecast flows (see Fig. 2a). Additionally, streamflow data up to the forecast date were used as inputs to the LSTM model, allowing the network to access recent information about the current hydrological state at the time of forecast, similar to how CEQUEAU has access to assimilated initial states. However, this also means that for each forecast lead-time, the LSTM model needs to be retrained with streamflow observations being lagged by the same number of days as the forecast lead-time. For example, for a 3-day lead-time forecast, the LSTM inputs would be a combination of 362 days of observed weather data prior to the forecast day as well as 362 days of streamflow observations lagged by 3 days (see Fig. 2a for an example of data and periods used to train a 1-day and a 3-day lead-time forecasting model). This ensures that no observed streamflow from the forecast period is used during model training. To train these models, observed data (both weather and hydrometric) are used for all time steps; however, in forecasting, the forecast lead-time days are replaced with actual forecast data and the observed streamflow time series ends on the day of the forecast issue (see Fig. 2b for an example of application of a 1-day and a 3-day forecasting model). Nine different LSTM models were thus trained, corresponding to each lead-time (one to nine days) of the forecast. It is important to note that in this study, ERA5 reanalysis data was used as the study is performed in hindcasting mode. However, in a real-world application, ERA5 data would not be available for the 3–4 more recent days as the ERA5 data are released with a few days of lag. These data would thus need to be taken elsewhere in an operational setting.

The network weights were optimized to maximize model performance over all training sequences. The training set was composed of 70 % of the years selected at random from the training period, 15 % of the years were kept as the validation dataset to prevent the LSTM from overfitting during training, and the remaining 15 % of the dataset is used as testing data which are used to evaluate the LSTM's robustness on an independent period prior to forecasting.

Data normalization was applied to all hydrometeorological data using a scaling model calibrated on the training period data only, to ensure no contamination of the training dataset was accidentally introduced. Training and validation were performed using the Tensorflow and Keras libraries in Python, with the Adam optimizer and the KGE metric as the objective (loss) function. The model was trained by generating one streamflow value for each [365-k]-day window of training data, and repeated for all days in the dataset, generating a hydrograph one day at a time in multiple batches. The trained model was then used to perform forecasts for the period from January 2015 to December 2019 by combining historical (ERA5 reanalysis dataset) and forecast (ECMWF) data into sequences of [365-k] days, ending on the desired lead-time's meteorological data. One LSTM model was trained for each forecast lead-time due to lagged streamflow observations. The LSTM forecasts were also generated every three days and on the same dates as the CEQUEAU-DA data assimilation and forecasts were performed.

4. Results

4.1. Performance of the LSTM model in simulation

Table 1 presents the statistics of the LSTM performance over the training, validation, and testing periods. Overall, the LSTM model performs well in simulating daily streamflow as suggested with KGE values above 0.90 for all lead-times over the training period and above 0.89 for the testing period.

Although the LSTM model captures the lowest peak flows in the daily hydrograph of Fig. 3a, a slight underestimation of the highest peak flows is observed, ranging from 5 to 10 %. The timing of all peak flows is successfully captured during the training period. As for the testing period, the KGE values are slightly better than those obtained over the training period for day one, and the performance is comparable until day three (Table 1). However, for lead-times beyond day four, the LSTM

Table 1
Description of the LSTM models hyperparameters used for each of the forecast lead-times and training results.

Lead times (days)	Nb. of LSTM units	Dropout rate	Nb. of training epochs	Batch size	Training KGE	Validation KGE	Testing KGE
1	128	0.2	100	128	0.97	0.99	0.98
2	128	0.2	200	128	0.98	0.99	0.98
3	128	0.2	200	128	0.95	0.95	0.95
4	128	0.2	250	128	0.95	0.91	0.94
5	128	0.2	250	128	0.92	0.94	0.91
6	128	0.2	300	128	0.93	0.91	0.92
7	64	0.1	1000	64	0.94	0.94	0.93
8	64	0.1	1000	64	0.95	0.93	0.92
9	64	0.1	1000	64	0.90	0.92	0.88

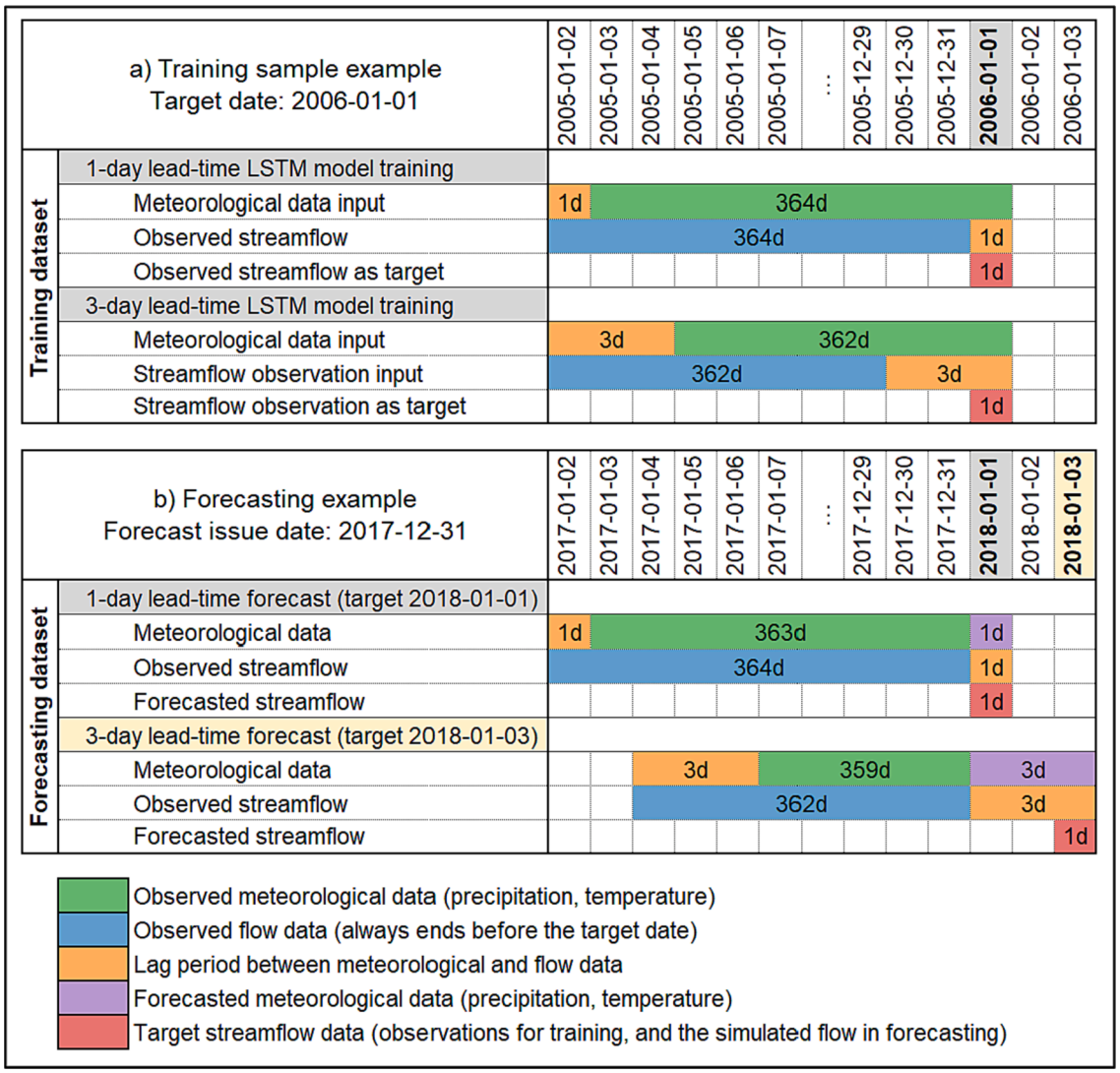


Fig. 2. Example implementation of a training sample for a 1-day and a 3-day lead-time forecasting model (a) and an example implementation of such forecasting models for 1-day and 3-day lead-time streamflow forecasts (b). Panel (a) presents a single sample, but the process is repeated on the entire dataset available for the training period and the same process is repeated on the validation and testing periods. Panel (b) shows an example for a single issue date, but the process is repeated for other forecast issue dates, combining the outputs of the 1-day to 9-day forecasts for each forecast issue date.

performance over the testing period is slightly lower, even though the KGE values remain very good (KGE $\geq$ 0.88).

In the testing period, the LSTM model slightly underestimates the highest peak flows while the timing of peak flows is generally well simulated. These results can be seen in Fig. 3b for the 1-day lead-time LSTM model. Results for longer lead-times (from 2- to 9-day) show similar skill levels.

Fig. 4 shows a quantile–quantile plot of the observed and simulated streamflows for each of the nine lead-times during the testing period. A general underestimation of peak flows is seen in most lead-times. Additionally, a small overestimation of the lowest flows can be noted for most lead-times, which can be seen in more detail in the supplementary materials (Figure S2).

4.2. LSTM and CEQUEAU comparison in forecasting

The main focus of this section is to compare the performance of the LSTM and the CEQUEAU models with (CEQUEAU-DA) and without (open-loop, CEQUEAU-OL) data assimilation when simulating streamflow over the LSJ catchment at the annual and seasonal scales during the testing period.

The skill of ensemble forecasts for one to nine days of lead time is analyzed in Figs. 5 to 7. Fig. 5 presents the CRPS and MAE scores using data from all years and all seasons as a first step to assess overall model performance. Figs. 5 and 6 present the forecast skill metrics for forecasts issued for each season (i.e., winter: December to March - DJFM; spring: April and May - AM; summer: June, to August - JJA; and fall: September to November- SON) for both the CRPS and MAE, respectively. These figures include the results of all forecast issue days within their respective periods (i.e., all forecasts generated during the summer days – 122 initial dates of forecast - are represented in the summer boxplots in Figs. 5 and 6). Lower CRPS and MAE values indicate more accurate forecasts.

The annual results indicate that LSTM performs better than both CEQUEAU-DA and CEQUEAU-OL for both CRPS and MAE values up to the 8-day lead-time (Fig. 5). On average, the median LSTM performance is 22 % (42 %) better for CRPS (MAE) compared to CEQUEAU-DA, and 37 % (37 %) better for CRPS (MAE) compared to CEQUEAU-OL across all lead times.

The LSTM model outperforms CEQUEAU during the first three days of the forecast, as seen in the third quartile of the LSTM CRPS which is inferior (better) to the median CRPS of the CEQUEAU forecasts for the

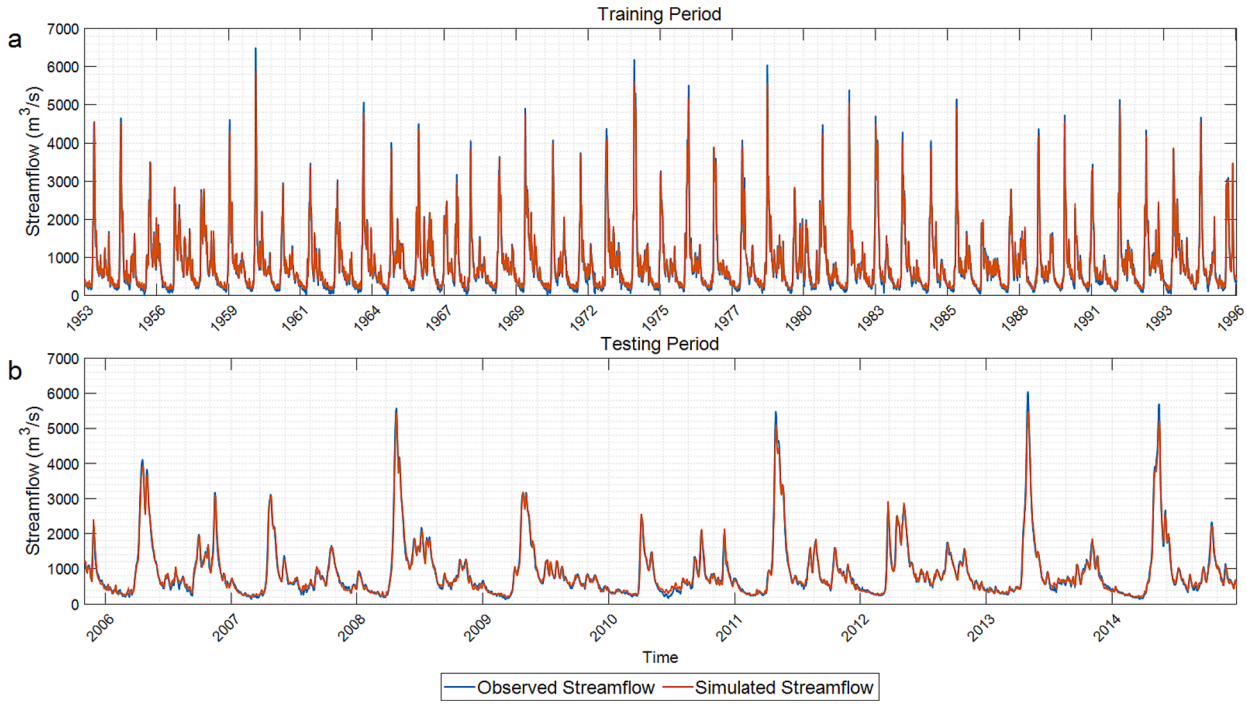


Fig. 3. Comparison between observed and simulated streamflow from the LSTM models over the 1953–1996 training period (a) and the 2005–2015 testing period (b), using a 1-day lead-time.

first two days, and a similar trend can be observed for MAE. This is supported by a non-parametric Wilcoxon test for median equality, which shows the results on an annual basis for both skill scores, as depicted in Table 2a.

For days 4–9, the LSTM model still performs better than CEQUEAU, but by a smaller margin. The MAE results, on an annual scale, show that LSTM has significantly lower MAE values than CEQUEAU (using both DA and OL methods) for the nine lead-times (Table 2b). The CRPS results also support these findings, as shown in Fig. 4. However, after the 3-day lead-time, there is no significant difference between LSTM and CEQUEAU-DA for the CRPS.

Additionally, CEQUEAU-DA generally provides better CRPS and MAE values than the OL scenario when looking at forecast quality over the entire year, which is expected. It should also be noted that the LSTM displays progressively wider spread (i.e. wider interquartile range) in CRPS and MAE as lead-time increases (for example, interquartile ranges of CRPS for days 1, 3, 5 are 39, 82, and 119 m^3/s, respectively). This is likely attributed to the fact that the LSTM model has the exact same starting conditions for every member of the ensemble during a forecast, and only the forecasted weather can contribute to the variability, as will be discussed further.

The ensemble forecast skill is evaluated and investigated on a seasonal basis (Figs. 6 and 7). The results reveal that the LSTM model performs well with lower CRPS and MAE values than both CEQUEAU-DA and CEQUEAU-OL up to a lead time of five days for CRPS values (see Table 2a for significance test results) and eight days for MAE values (Table 2b) for all seasons. However, from day six, the overall LSTM performance decreases and is now inferior to the two CEQUEAU model versions. The only exception is the summer season, for which the LSTM provides ensemble forecasts with better accuracy than CEQUEAU-DA and CEQUEAU-OL until the 8-day lead time. The Wilcoxon test results show decreasing significance with longer lead-times in general for the CEQUEAU-DA and CEQUEAU-OL comparisons (Table 2).

4.3. Forecast performance evaluation

Ensemble streamflow forecasts generated with the LSTM model are compared to those produced by CEQUEAU-DA and CEQUEAU-OL models for up to 9-day lead-times. The performance is evaluated for each season by selecting a typical forecast event and using the forecast issue date with the median streamflow observation of that season to ensure representativeness (Fig. 8). Ensemble forecasts using the three methods (LSTM, CEQUEAU-DA and CEQUEAU-OL) are then evaluated for these selected events.

The results in Fig. 8 depict the added value of the LSTM model to the overall quality of ensemble forecasts in different seasons for various lead-times. In winter, LSTM provides an almost unbiased forecast up to approximately the 6-day lead-time, compared to the CEQUEAU forecasts. The CRPS for the LSTM forecast is also lower than that of the CEQUEAU forecasts over the entire period due to the ensemble having a lower spread. The initial states are well simulated by all three forecasting models and are not a factor in the bias related to initial conditions.

In spring, forecasts vary significantly depending on the forecast model. The LSTM has much less variability than the CEQUEAU-OL forecast, which is in turn less variable than the CEQUEAU-DA implementation. The latter has more variability primarily due to the assimilation of initial states, making it more sensitive to input meteorological data variations. Additionally, the LSTM forecasts for the first days display less bias than the hydrological model counterparts, owing to the integration of recent streamflow as inputs. It can also be seen that up to a lead-time of approximately five days, the LSTM forecast shows smaller ensemble error than the other models. However, in the following days, the small spread and increasing bias of the LSTM ensemble members heavily penalized the CRPS score, making it worse overall than the CEQUEAU implementations. This is also seen in Fig. 6b, where the LSTM generally produces better forecasts for shorter lead-times only. In summer, LSTM decreases the spread of the ensemble forecasts over the 9-day lead-time, allowing it to get closer to the observed streamflow compared to both the CEQUEAU-DA and CEQUEAU-OL ensembles except during

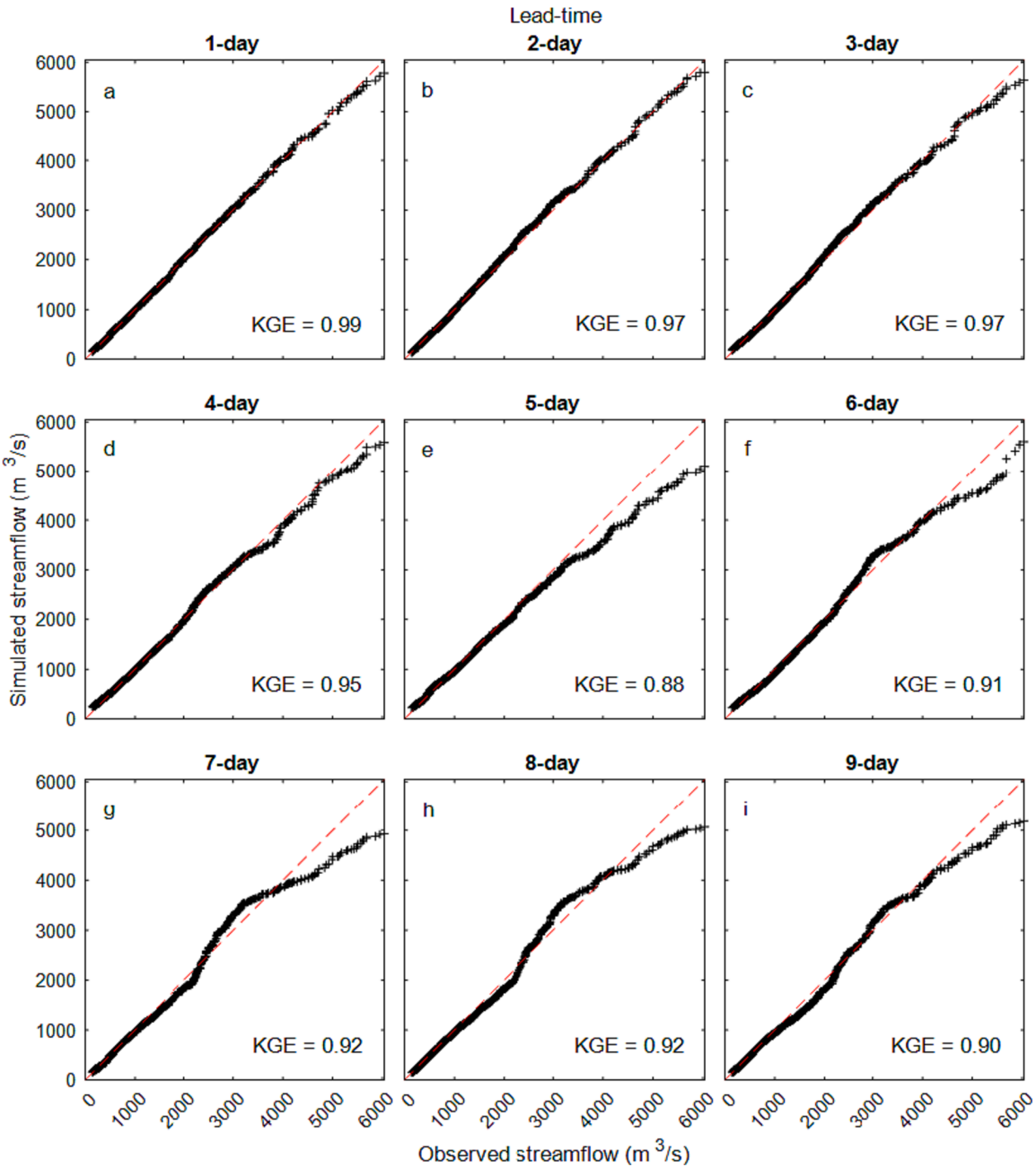


Fig. 4. Quantile-quantile plot showing observed streamflow against simulated values from the nine LSTM models over the testing period, from 1 to 9 days in lead-time from a) to i). The 1:1 slope (dashed red line) is added for comparison purposes, representing a perfect match between observed and simulated streamflow. Each panel contains 3341 points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the last forecast days, contributing to the improvements in CRPS (see Fig. 6c). In fall, the comparison of LSTM-ensemble forecasts with the CEQUEAU-DA and CEQUEAU-OL ensembles shows that LSTM generates reliable forecasts that encompass the observation with very little spread for the entire forecast duration. The CEQUEAU-DA, however, has initial conditions that are more representative of the current state and are more saturated (i.e., more reactive) than CEQUEAU-OL for that given period.

Overall, the main difference between the LSTM and CEQUEAU-based forecasts is that the LSTM is more confident in its forecasting, generating forecasts with less spread. When the forecast is unbiased, the forecast skill is better than that of the hydrological models. However, when the LSTM model generates a biased forecast, the low spread makes all 50 members biased, thus increasing the CRPS and MAE error metric values.

The forecasted hydrographs of each model are compared in Fig. 9. The results illustrate that the LSTM is more accurate at shorter lead-times (1-day to 3-day) than the CEQUEAU-DA and CEQUEAU-OL forecasts throughout the year. Note that since forecasts are generated every 3-days, it was required to generate and stich series of 1- to 3-day forecasts together to produce a complete hydrograph. Similar results are found for the 4- to 6-day and 7- to 9-day forecasts presented in the

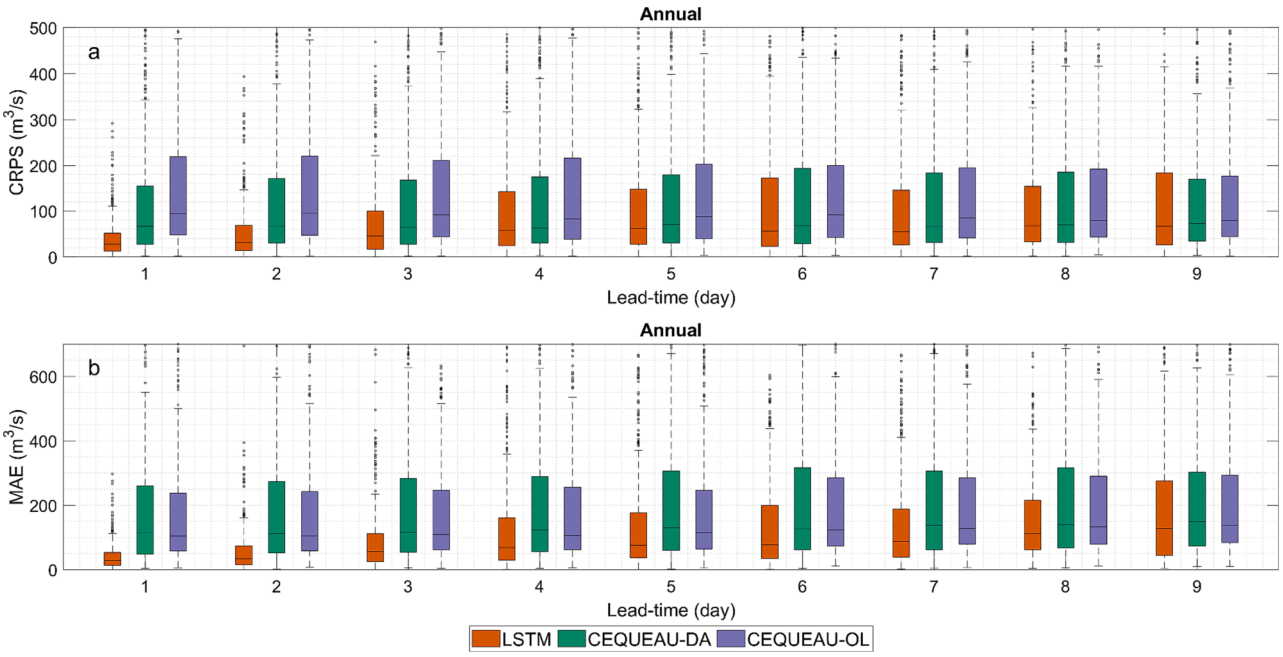


Fig. 5. CRPS (a) and MAE (b) of the annual streamflow ensemble forecasts for the LSTM model (orange), CEQUEAU-DA (DA; green), and CEQUEAU-OL (OL; purple) over the 2015–2019 forecasting period. Each boxplot contains 536 forecasts, corresponding to one forecast every three days over the study period. The center horizontal line in each boxplot represents the median, box edges represent the 25th and 75th percentiles, and whiskers represent the extreme values not considered as outliers. Dots outside of the whiskers are outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

supplementary materials Fig. S3A and S3B. The longer lead-times demonstrate that the LSTM errors are primarily caused by temporal shifts, while the CEQUEAU errors are mainly attributed to amplitude errors, but with better timing.

5. Discussion

5.1. Comparison between CEQUEAU and LSTM

The CEQUEAU and LSTM models share some similarities but differ in key aspects when it comes to forecasting streamflow. CEQUEAU-DA initiates forecasts with an improved estimate of initial states, resulting in a more sensitive model from the first lead-time. On the other hand, the LSTM model is initialized deterministically using past weather and streamflow observations, resulting in a lack of uncertainty in the 1-day lead-time (Fig. 8). Both models are trained on historical data, but CEQUEAU is calibrated over a historical time period and used to generate forecasts based on incoming weather forecasts, while the LSTM model is trained on a set of two inputs (precipitation and temperature) and one output (streamflow).

CEQUEAU-DA produces a wider spread of forecast results compared to the LSTM model. Both LSTM and CEQUEAU (both DA and OL) display CRPS and MAE errors worsening with increasing lead-times. This is likely due to the decreasing reliability of weather forecasts as lead-times increase, resulting in less skillful streamflow forecasts from both LSTM and CEQUEAU. Overall, the results indicate that LSTM provided better forecast results compared to the CEQUEAU model, while also requiring simpler implementation and no data assimilation. However, the LSTM model has the limitation of needing to be trained multiple times for different lead-times. Furthermore, while the LSTM model may perform well on a single catchment as seen with the LSJ, it might also benefit from training over several basins to be robust and efficient at capturing extreme events, such as in Kratzert et al. (2018, 2019).

Finally, it is important to note that the LSTM model was able to successfully improve spring forecasts for up to 3 days compared to CEQUEAU-DA and CEQUEAU-OL. This is useful as spring melt is one of the most important processes to capture correctly for hydropower management in this system. However, it can also be seen that the CEQUEAU model outperforms the LSTM model in spring for longer lead-times. This could be due to the LSTM model not having sufficient data to successfully represent the streamflow patterns on longer timescales, and to the CEQUEAU model's mass conservation constraints that ensure that snow that has accumulated will melt eventually, leading to smaller biases in overall streamflow on longer time scales.

5.2. On the necessity of performing data assimilation

Although LSTMs do not typically implement DA directly, they do have a recurrent state that allows for the ingestion of near real-time observations (Nearing et al., 2022). In this study, DA was not performed directly on the LSTM models. Instead, rather than implementing DA directly on the LSTM model states, an autoregressive component is integrated into the LSTM model by providing observed streamflow on the timesteps immediately preceding the forecast. This means that the result of the previous simulation has no impact on the current forecast and the LSTM breaks continuity in the forecasting stage. This method, along with a proper DA implementation, was done in Nearing et al. (2022), in which it is shown that the autoregressive approach is more accurate than the DA approach. However, this method is also sensitive to missing data, as any missing streamflow observation would preclude the model from being used in forecasting mode for that day. Some methods have been proposed to tackle this problem with success in Nearing et al. (2022). Hydrological models, on the other hand, ensure the forecast is done in one single run, which is an advantage over LSTMs using observed streamflow as inputs. The LSTM used in this study therefore trades continuity for not having to resort to the complex step of DA given the number of memory states in the LSTM. By providing observed streamflow in the training and forecasting steps, the LSTM can learn to minimize initial state errors, similar to the DA step in a hydrological model using observed streamflow to adjust its initial states.

Both the hydrological model with DA and the LSTM model provided skilled streamflow simulations and forecasts. This study found that for

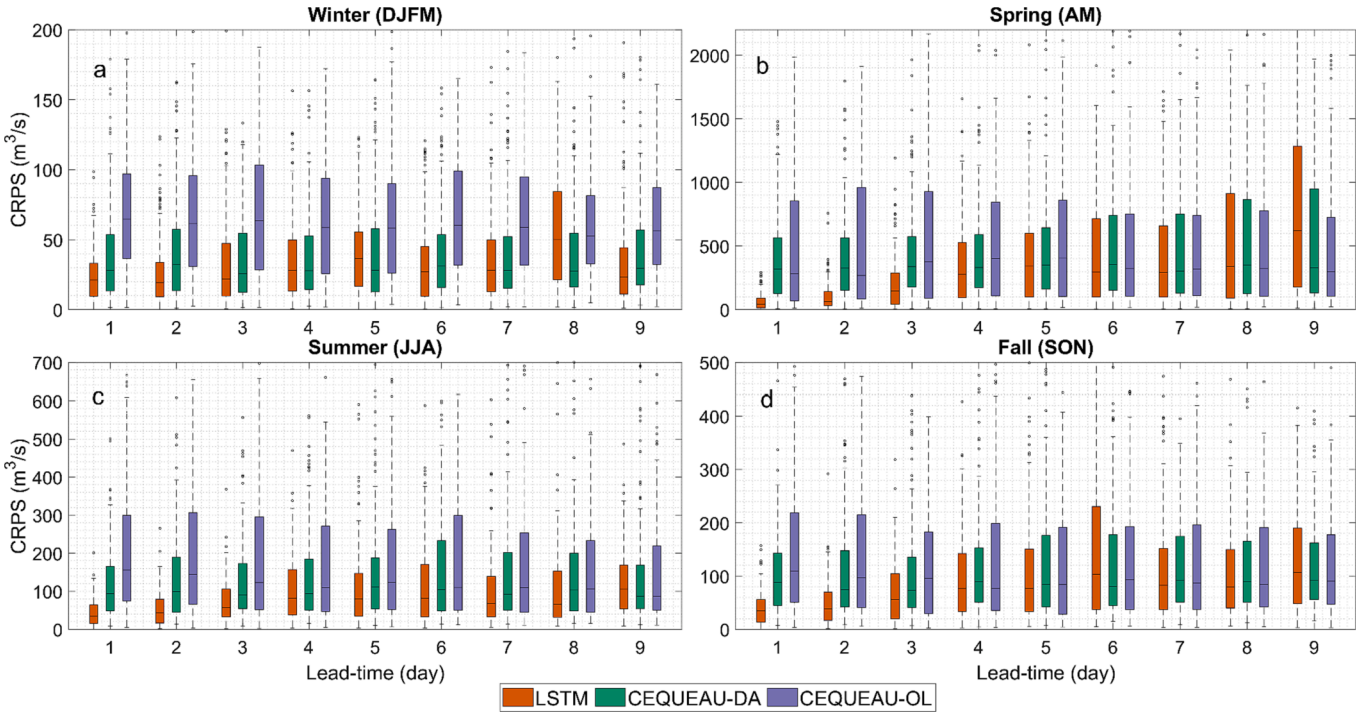


Fig. 6. CRPS of the seasonal streamflow ensemble forecasts for the LSTM model (orange), CEQUEAU-DA (green), and CEQUEAU-OL (orange) over each season of the 2015–2019 forecasting period: winter: (December to March; DJFM - a), spring (April and May; AM - b) summer (June to August; JJA - c), and fall (September to November; SON - d). The number of points in each boxplot represents the number of issued forecasts for that season, equal to 191, 101, 122, and 122 for Winter, Spring, Summer, and Fall, respectively. Note that the y-axis ranges are different in all panels due to the large differences between seasons and some outliers are not shown for clarity's sake. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the catchment studied, the LSTM outperforms the hydrological model for short-term forecasts (up to 5–9 days lead-time depending on the season) when using the CRPS and MAE metrics. Fig. 4a and 4b showed that the LSTM performed better than the CEQUEAU-DA and CEQUEAU-OL models for all lead-times with a significant improvement in forecast skill.

5.3. *Effects of excluding streamflow observations from the LSTM forecasting model*

In this study, the LSTM used observed streamflow up to the forecast issue date to provide information on the current hydrological state. The use of observed streamflow as a predictor is not a new concept in hydrology; in fact, it has been a useful proxy for short-term forecasts over the last decade (Cloke and Pappenberger, 2009). Different studies have applied various methods such as using historical streamflow observations directly to estimate future outcomes (e.g., Rajagopalan et al., 2010), applying regression models (e.g., Seo et al., 2006; Hopson and Webster, 2010; Bogner et al., 2016), using autoregressive models (e.g., ARMA, ARMAX, GARCH,; Zhang et al., 2015; Amiri, 2015) and using artificial neural networks (e.g., Coulibaly et al., 2000; Machado et al., 2011).

Recent studies have focused on using LSTM for its ability to simulate hydrological processes and provide high-quality streamflow simulations and forecasts from historical weather data. In this study, streamflow was added as an input to the LSTM as a form of data integration. By training the model to combine the information from observed streamflow and weather up to the forecast issue date, as well as forecasted weather, the LSTM is able to make accurate streamflow forecasts for shorter periods. However, as the lead-time increases, the impact of observed streamflow diminishes because the streamflow observations are not available past the forecast issue date. This is reflected in the progressive widening of the CRPS and MAE values as lead-times increase (see Figs. 4 to 6). To improve performance for longer lead-times, the LSTM must be trained independently for each lead-time, with the observed streamflow lagged by the appropriate number of days. This makes the training process more labor-intensive and the model is less able to rely on the observed streamflow for predicting the forecasted streamflow, which results in progressively worse CRPS and MAE scores, similar to the forecasts issued using the hydrological model. Another possible explanation for this behavior is that of persistence in the streamflow observations. Indeed, using a simple persistence model, we find that we obtain Nash-Sutcliffe metric values of 0.9 and 0.87 for 1- and 2- day ahead deterministic forecasts. However, by day 7, the NSE drops to 0.58. This means that observed streamflow for a given day has a strong predictive power for the first few following days before losing skill. Of course, this test is only applicable for deterministic forecasts, but it informs on the predictive power of integrating streamflow observations as inputs to the LSTM model.

Additional tests were conducted to overcome the need for training one LSTM per lead-time in order to improve forecasting performance. One test involved training the LSTM using only historical weather data as the input, rather than incorporating observed streamflow data. This approach was found to provide good results for only two days of lead-time. However, for longer lead-times, the LSTM performed worse than the hydrological models (CEQUEAU-DA and CEQUEAU-OL; see Fig. S4). This could be due to the fact that the forecast data from the ECMWF could have different statistical properties than the observations used to train the LSTM, leading to a bias in the forecast. Incorporating observed streamflow data may help to minimize this effect by reducing the weight of the weather component in the forecasting chain.

5.4. *Limitations*

This study has a few limitations that should be considered. First, the study was performed on a single catchment used for hydropower generation. It would be interesting to apply the methodology to a larger sample of catchments to investigate the generalizability of the approach

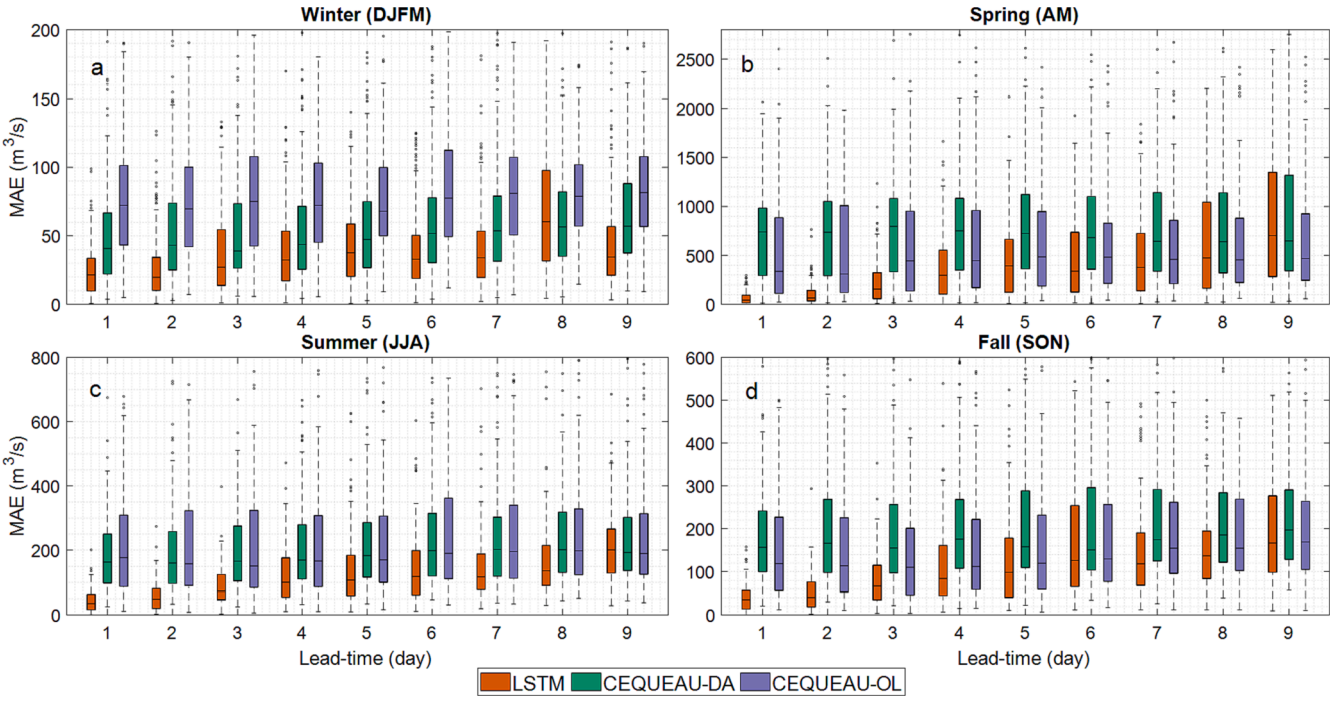


Fig. 7. MAE of the seasonal streamflow ensemble forecasts for the LSTM model (orange), CEQUEAU-DA (green), and CEQUEAU-OL (purple) over each season of the 2015–2019 forecasting period: winter: (December to March; DJFM - a), spring (April and May; AM - b) summer (June to August; JJA - c), and fall (September to November; SON - d). The number of points in each boxplot represents the number of issued forecasts for that season, equal to 191, 101, 122, and 122 for Winter, Spring, Summer, and Fall, respectively. Note that the y-axis ranges are different in all panels due to the large differences between seasons and some outliers are not shown, for clarity's sake. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Results of a Wilcoxon rank sum statistical test with a significance level of 1 % for CRPS (a) and MAE (b) results between the forecasts generated by the three models used in this study (LSTM, CEQUEAU-DA and CEQUEAU-OL). A value of H = 0 indicates equal medians between the forecasts of the groups defined in each column, whereas a value of H = 1 indicates that the null hypothesis is rejected, indicating different medians for the two groups. Performance is evaluated per lead-time.

(a) CRPS

	H - Result of the hypothesis test									
	Winter		Spring		Summer		Fall		Annual	
Lead-time	LSTM vs. DA	LSTM vs. OL	LSTM vs. DA	LSTM vs. OL	LSTM vs. DA	LSTM vs. OL	LSTM vs. DA	LSTM vs. OL	LSTM vs. DA	LSTM vs. OL
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	0	1	1	1	1	1	1	1	1	1
4	0	1	0	0	0	1	0	0	0	1
5	0	1	0	0	1	1	0	0	0	1
6	0	1	0	0	0	1	0	0	0	1
7	0	1	0	0	1	1	0	0	0	1
8	1	0	0	0	1	1	0	0	0	1
9	1	1	0	1	0	0	0	0	0	1

(b) MAE

	H - Result of the hypothesis test									
	Winter		Spring		Summer		Fall		Annual	
Lead-time	LSTM vs. DA	LSTM vs. OL	LSTM vs. DA	LSTM vs. OL	LSTM vs. DA	LSTM vs. OL	LSTM vs. DA	LSTM vs. OL	LSTM vs. DA	LSTM vs. OL
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	0	1	1
5	1	1	1	0	1	1	1	0	1	1
6	1	1	1	0	1	1	1	0	1	1
7	1	1	1	0	1	1	1	1	1	1
8	0	1	0	0	1	1	1	0	1	1
9	1	1	0	0	0	0	0	0	1	1

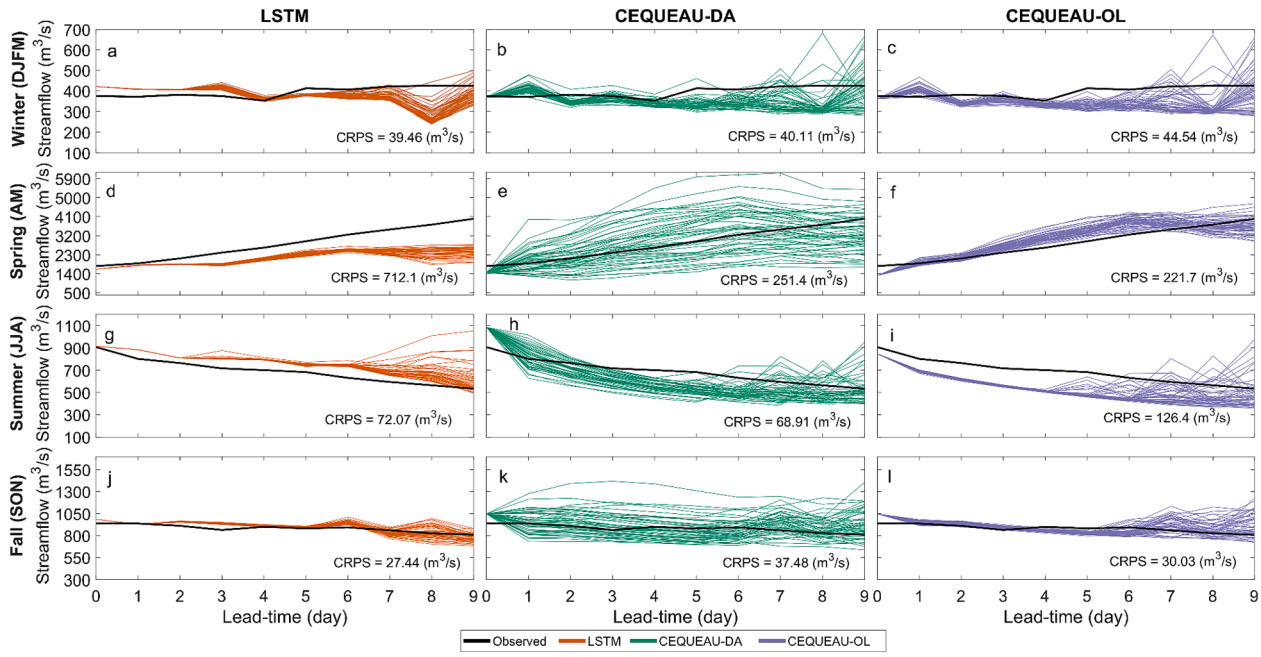


Fig. 8. Forecasted streamflow ensembles for each season (winter: a, b, c; spring: d, e, f; summer: g, h, i; fall: j, k, l) as generated by the LSTM model (left column; a, d, g, j), CEQUEAU-DA (middle column; b, e, h, k), and CEQUEAU-OL (right column; c, f, i, l) over a 9-day lead-time. For each row (i.e., season), the forecast date chosen for display is that which corresponds to the day where the observed flow is the median value of all observations for that season. The exact dates are Jan-26–2016 (Winter), Apr-26–2017 (Spring); Jul-30–2016 (Summer) and Nov-21—2015 (Fall).

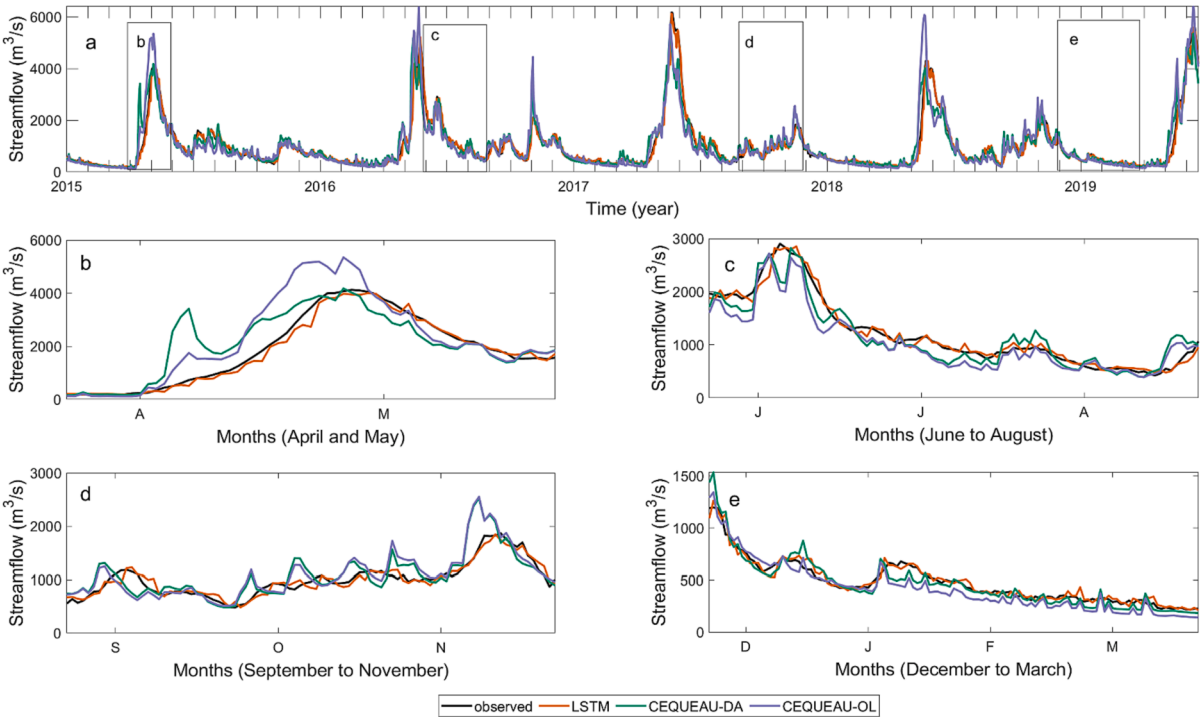


Fig. 9. Hydrographs generated from successive 1- to 3-day lead-time streamflow forecasts averaged over the 50 members for the LSTM (orange), CEQUEAU-DA (green), and CEQUEAU-OL (purple) models and observations (black) over the period between January 2015 and May 2019 (a). The focus is placed on the individual seasons of spring (b), summer (c), fall (d) and winter (e). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to other catchments. This would require access to forecast data and long observational time-series, which should be considered in future research efforts. Second, the choice of LSTM model hyperparameters was based largely on expert knowledge and trial-and-error, and the resulting models were of high quality and able to outperform the traditional hydrological model implementations, which was sufficient for the scope of this paper. However, there are now better approaches to optimize the choice of hyperparameters. This could provide a path forward for better models in the future. It is also important to note that the LSTM trained in this study used a wide array of combinations of ERA5 meteorological

time series. However, the most robust model tested only made use of precipitation and temperature. Adding more variables made the training more difficult and either converge to worse values or not converge at all. It could be possible to improve on these results by adding more types of data and finding LSTM model structures and hyperparameters that can make use of the extra input data; however, this was not successfully implemented in this study.

Finally, it is important to recall that this study relies on ERA5 data that is not available in real-time but has a few days lag between real-time and the reanalysis emission. This means that while the results are good in this study, applicability in real-time forecasting would require replacing the most recent observations with another source than ERA5 data, such as station observations.

6. Conclusion

This study evaluates the potential of the LSTM in simulating and forecasting streamflow of the Lac-Saint-Jean catchment in Canada. It compares the LSTM to the operational CEQUEAU model used by Rio Tinto for this catchment, which is set up in open-loop mode (CEQUEAU-OL) and with a data assimilation scheme (CEQUEAU-DA), which is used as a benchmark. The main findings of this study are as follows:

(1) The LSTM achieves good performance in the training and testing periods for lead times up to 9 days with a KGE higher than 0.88.
(2) The LSTM provides more skillful ensemble forecasts compared to CEQUEAU-OL and CEQUEAU-DA, as CRPS and MAE results show lower values for the LSTM, all percentiles considered.
(3) The LSTM forecasts display tighter spreads than the CEQUEAU-based forecasts, likely due to the strong influence of the observed streamflow from the previous days used as a predictor, as opposed to the DA implementation that contains uncertainty.
(4) The LSTM eliminates the need for integrating a DA process, typically required by traditional hydrological models, while still providing high-quality forecasts.

The findings of this study have highlighted the advantages, limitations, and specific evaluation of the LSTM performance in streamflow forecasting for all seasons. Overall, this study shows that LSTM is a promising model for forecasting short-term streamflow, and confirms previous findings in other regions and catchments. It is recommended that this LSTM forecasting model be implemented for all seasons and lead-times up to 9 days, except for spring, for which the model should not be used for forecasts beyond 6–7 days of lead-time. It is likely that more advanced deep learning networks and data integration strategies will lead to even more significant improvements, such as training LSTM models on a large set of catchments to increase the size of the dataset (as in Kratzert et al., 2019). However, this study demonstrates that, in this snow-dominated North American catchment, LSTM models can provide short-term streamflow forecasts with better accuracy than those generated by more complex distributed hydrological models.

CRediT authorship contribution statement

Behmard Sabzipour: Conceptualization, Methodology, Software, Validation, Investigation, Formal analysis, Writing – original draft. **Richard Arsenault:** Software, Validation, Supervision, Funding acquisition. **Magali Troin:** Writing – review & editing, Visualization, Data curation, Software. **Jean-Luc Martel:** Validation, Writing – review & editing, Visualization. **François Brissette:** Supervision, Writing – review & editing, Funding acquisition. **Frédéric Brunet:** Methodology, Software, Writing – review & editing. **Juliane Mai:** Writing – review & editing, Visualization.

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.

Data availability

Data will be made available on request.

Acknowledgements

This study was partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) under the Collaborative Research and Development grant CRDPJ-522126-17. The authors would like to thank the anonymous reviewers who contributed to improving this manuscript through their insightful and constructive comments. The authors would also like to thank Rio Tinto for sharing their hydrometeorological data on the Lac-Saint-Jean catchment and the European Center for Medium-Range Weather Forecasts (ECMWF) for providing access to the historical forecast data from their MARS computing and archiving facilities. In this study, the ERA5 reanalysis dataset produced by Hersbach et al., (2018) was used. It has been downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview. The base map in Fig. 1 was created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license. Copyright © Esri. All rights reserved. For more information about Esri® software, please visit www.esri.com.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2023.130380.

References

Amiri, E., 2015. Forecasting daily river flows using nonlinear time series models. J. Hydrol. 527, 1054–1072.
Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., Lettenmaier, D.P., 2016. Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. Water Resour. Res. 52 (6), 4209–4225.
Arsenault, R., Côté, P., 2019. Analysis of the effects of biases in ensemble streamflow prediction (ESP) forecasts on electricity production in hydropower reservoir management. Hydrol. Earth Syst. Sci. 23 (6), 2735–2750. https://doi.org/10.5194/hess-23-2735-2019.
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.
Bergeron, J., Leconte, R., Trudel, M., Farhoodi, S., 2021. On the choice of metric to calibrate time-invariant ensemble kalman filter hyper-parameters for discharge data assimilation and its impact on discharge forecast modelling. Hydrology 8 (1), 36.
Bogner, K., Liechti, K., Zappa, M., 2016. Post-processing of stream flows in Switzerland with an emphasis on low flows and floods. Water 8 (4), 115.
Boucher, M.-A., Tremblay, D., Delorme, L., Perreault, L., Anctil, F., 2012. Hydro-economic assessment of hydrological forecasting systems. J. Hydrol. 416, 133–144.
Brajard, J., Carrassi, A., Bocquet, M., Bertino, L., 2020. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. J. Comput. Sci. 44, 101171.
Cassagnole, M., Ramos, M.-H., Zalachori, I., Thirel, G., Garçon, R., Gailhard, J., Ouillon, T., 2021. Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs – a conceptual approach. Hydrol. Earth Syst. Sci. 25 (2), 1033–1052. https://doi.org/10.5194/hess-25-1033-2021.
Cho, K., Kim, Y., 2022. Improving streamflow prediction in the WRF-Hydro model with LSTM networks. J. Hydrol. 605, 127297.
Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. J. Hydrol. 375 (3–4), 613–626.
Coulibaly, P., Anctil, F., Bobée, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J. Hydrol. 230 (3–4), 244–257.
Damavandi, H.G., Shah, R., Stampoulis, D., Wei, Y., Boscovic, D., Sabo, J., 2019. Accurate prediction of streamflow using long short-term memory network: a case

study in the Brazos River Basin in Texas. Internat. J. Environ. Sci. Dev. 10 (10), 294–300.
Day, G.N., 1985. Extended streamflow forecasting using NWSRFS. J. Water Resour. Plann. Manage. 111 (2), 157–170.
Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367.
Fang, K., Shen, C., 2020. Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. J. Hydrometeorol. 21 (3), 399–413.
Feng, D., Fang, K., Shen, C., 2020. Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resour. Res. 56 (9), e2019WR026793.
Girihagama, L., Naveed Khaliq, M., Lamontagne, P., Perdikaris, J., Roy, R., Sushama, L., Elshorbagy, A., 2022. Streamflow modelling and forecasting for Canadian watersheds using LSTM networks with attention mechanism. Neural Comput. Appl. 34 (22), 19995–20015.
Granata, F., Di Nunno, F., 2023. Neuroforecasting of daily streamflows in the UK for short-and medium-term horizons: A novel insight. J. Hydrol. 624, 129888.
Granata, F., Di Nunno, F., de Marinis, G., 2022. Stacked machine learning algorithms and bidirectional long short-term memory networks for multi-step ahead streamflow forecasting: A comparative study. J. Hydrol. 613, 128431.
Grawinkel, M., Nagel, L., Mäsker, M., Padua, F., Brinkmann, A., Sorth, L., 2015. Analysis of the ECMWF Storage Landscape. In: 13th USENIX Conference on File and Storage Technologies (FAST 15) (pp. 15–27).
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1–2), 80–91.
Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast. 15 (5), 559–570.
Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., 2020. The ERA5 global reanalysis. Q. J. R. Meteorolog. Soc. 146 (730), 1999–2049.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hopson, T.M., Webster, P.J., 2010. A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. J. Hydrometeorol. 11 (3), 618–641.
Hu, C., Wu, Q., Li, H., Jian, S., Li, N., Lou, Z., 2018. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. Water 10 (11), 1543. https://doi.org/10.3390/w10111543.
Hunt, K.M., Matthews, G.R., Pappenberger, F., Prudhomme, C., 2022. Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. Hydrol. Earth Syst. Sci. Discuss. 1–30.
Kao, I.-F., Zhou, Y., Chang, L.-C., Chang, F.-J., 2020. Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting. J. Hydrol. 583, 124631.
Khoshkalam, Y., Rousseau, A.N., Rahmani, F., Shen, C., Abbasnezhadi, K., 2023. Applying transfer learning techniques to enhance the accuracy of streamflow prediction produced by long Short-term memory networks with data integration. J. Hydrol. 622, 129682.
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. https://doi.org/10.5194/hess-22-6005-2018.
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, 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.
Le, X.H., Nguyen, D.H., Jung, S., Yeon, M., Lee, G., 2021. Comparison of deep learning techniques for river streamflow forecasting. IEEE Access 9, 71805–71820.
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. https://doi.org/10.5194/hess-25-5517-2021.
Loiselle, G., Martel, J.L., Poulin, A., Lachance-Cloutier, S., Turcotte, R., Fournier, J., Mai, J., Arsenault, R., 2021. A semi-empirical wind set-up forecasting model for Lake Champlain. Hydrol. Processes 35 (6), e14240.
Machado, F., Mine, M., Kaviski, E., Fill, H., 2011. Monthly rainfall–runoff modelling using artificial neural networks. Hydrol. Sci. J.-J. Des. Sci. Hydrol. 56 (3), 349–361.
Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environ. Modell. Software 25 (8), 891–909.
Mai, J., Arsenault, R., Tolson, B.A., Latraverse, M., Demeester, K., 2020. Application of parameter screening to derive optimal initial state adjustments for streamflow forecasting. Water Resour. Res. 56 (9) p. e2020WR027960.
Mather, A.L., Johnson, R.L., 2016. Forecasting Turbidity during Streamflow Events for Two Mid-Atlantic U.S. Streams. Water Resour. Manage. 30 (13), 4899–4912. https://doi.org/10.1007/s11269-016-1460-1.
Morin, G., Paquet, P., 2007. Modèle hydrologique CEQUEAU, rapport de recherche no R000926. Université du Québec, INRS-Eau, Terre et Environnement, Québec.
Nearing, G.S., Klotz, D., Frame, J.M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A.K., Shalev, G., Nevo, S., 2022. Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. Hydrol. Earth Syst. Sci. 26 (21), 5493–5513.
Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., 2022. Flood forecasting with machine learning models in an operational framework. Hydrol. Earth Syst. Sci. 26 (15), 4013–4032.
Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. J. Hydrol. 303 (1–4), 290–306.
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.
Rajagopalan, B., Salas, J.D., Lall, U., 2010. Stochastic methods for modeling precipitation and streamflow. In: Sivakumar, B., Berndtsson, R. (Eds.), Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting. World Scientific, pp. 17–52.
Sabzipour, B., Arsenault, R., Troin, M., Martel, J.L., Brissette, F., 2023. Sensitivity analysis of the hyperparameters of an ensemble Kalman filter application on a semi-distributed hydrological model for streamflow forecasting. J. Hydrol. 626, 130251.
Sahoo, B.B., Jha, R., Singh, A., Kumar, D., 2019. Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophys. 67 (5), 1471–1481.
Seo, D.-J., Herr, H., Schaake, J., 2006. A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrol. Earth Syst. Sci. Discuss. 3 (4), 1987–2035.
Shen, C., Lawson, K., 2021. Applications of deep learning in hydrology. Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences, pp. 283–297.
Sorooshian, S., Duan, Q., Gupta, V.K., 1993. Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model. Water Resour. Res. 29 (4), 1185–1194.
Tang, S., Sun, F., Liu, W., Wang, H., Feng, Y., Li, Z., 2023. Optimal postprocessing strategies with LSTM for global streamflow prediction in ungauged basins. Water Resour. Res. 59 (7).
Tounsi, A., Abdelkader, M., Temimi, M., 2023. Assessing the simulation of streamflow with the LSTM model across the continental United States using the MOPEX dataset. Neural Comput. & Applic. 35 (30), 22469–22486.
Troin, M., Arsenault, R., Wood, A.W., Brissette, F., Martel, J.L., 2021. Generating Ensemble Streamflow Forecasts: A Review of Methods and Approaches Over the Past 40 Years. Water Resour. Res. 57 (7) https://doi.org/10.1029/2020wr028392.
Twedt, T. M., Schaake Jr, J. C., & Peck, E. L. (1977). National Weather Service extended streamflow prediction [USA]. Proceedings Western Snow Conference.
Valéry, A. (2010). Modélisation précipitations débit sous influence nivale: Elaboration d'un module neige et évaluation sur 380 bassins versants Doctorat Hydrobiologie, Institut des Sciences et Industries du Vivant et de l'Environnement AgroParisTech). du Vivant et de …].
Xu, W., Jiang, Y., Zhang, X., Li, Y., Zhang, R., Fu, G., 2020. Using long short-term memory networks for river flow prediction. Hydrol. Res. 51 (6), 1358–1376.
Zhang, X., Peng, Y., Zhang, C., Wang, B., 2015. Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences. J. Hydrol. 530, 137–152.
Zhang, J., Zhu, Y., Zhang, X., Ye, M., Yang, J., 2018. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. J. Hydrol. 561, 918–929.