Journal of Hydrology 605 (2022) 127297

Contents lists available at ScienceDirect

Journal of Hydrology

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

Improving streamflow prediction in the WRF-Hydro model with LSTM networks

Kyeungwoo Cho, Yeonjoo Kim *

Department of Civil and Environmental Engineering, Yonsei University, Seoul 03722, South Korea

ABSTRACT

Researchers have attempted to use machine learning algorithms to replace physically based models for streamflow prediction. Although existing studies have contributed to improving machine learning methods, they still have weaknesses, such as large dataset requirements and overfitting. Therefore, we propose an approach that combines the Weather Research and Forecasting hydrological modeling system (WRF-Hydro) and the Long Short-Term Memory (LSTM) network, i.e., WRF-Hydro-LSTM, to improve streamflow simulations. In this approach, LSTM was employed to predict the residual errors of WRF-Hydro; in contrast, the conventional approach with LSTM predicts streamflow directly. Here, we performed numerical experiments to predict the inflow of Soyangho Lake in South Korea using WRF-Hydro-LSTM, WRF-Hydro-only, and LSTM-only. WRF-Hydro-LSTM and LSTM-only showed better results (NSE = 0.95 and R greater than 0.96) compared to WRF-Hydro-only (NSE = 0.72 and R = 0.88); however, in terms of the percent bias, WRF-Hydro-LSTM had a better value (1.75) than LSTM-only (17.36). While the LSTM-only follows objective functions and not physical principles, WRF-Hydro-LSTM simulates residual errors and efficiently decreases uncertainties that are inherent with conventional methods. Furthermore, a sensitivity test on the training dataset indicated that the correlation coefficient and NSE value were not overly sensitive, but the PBIAS value differed substantially depending on the training set. This study demonstrates that WRF-Hydro-LSTM is particularly useful for representing real-world physical constraints and thus can potentially improve streamflow prediction compared to using either of the two approaches exclusively.

1. Introduction

Streamflow describes the volume of available freshwater and the flood discharge. Therefore, accurate streamflow prediction is important for water resource management. Predicted basin-scale streamflow can be used for effectively managing water resources by controlling the water storage of the reservoir, which plays an important role in water power generation and flood alleviation as well as in irrigation schemes (Wedgbrow et al., 2002). To achieve reliable predictions, numerous hydrological models have been developed and utilized to simulate streamflow (Chiew et al., 2003; Asefa et al., 2006; Naabil et al., 2017), which can be classified as empirical, conceptual, and physically based. While lumped models, including empirical and conceptual models, use a limited number of parameters and basin average input data, physically-based distributed hydrological models are configured with various physically relevant and spatially distributed parameters to represent the spatial distribution of metrological, environmental, and physical characteristics, as well as multiple principles of hydrological systems to represent simplified real-world conditions (Schumann, 1993; Moradkhani & Sorooshian, 2008; El Hassan et al., 2013). Thus, distributed models are expected to simulate a hydrological cycle that obeys the laws of nature better than the conceptual rainfall–runoff models (Donnelly-Makowecki & Moore, 1999; El Hassan et al., 2013) and have strengths for hydrological forecasting (Carpenter & Georgakakos, 2006). Furthermore, the use of physically interpretable parameters in physically based distributed models enables us to understand various hydrological processes in relatively high temporal and spatial resolutions, including not only streamflow processes, but also evapotranspiration processes and soil moisture dynamics (Devia et al., 2015).

Although considerable efforts have been made to improve the performance of hydrological models (Boyle et al., 2000; Chuck et al., 2004; Vrugt et al., 2005; Crochemore et al., 2016), they are still limited by uncertainties in their datasets, heterogeneity in their parameters and variables, scale effects, and nonlinearities in their process dynamics, the structure of which is mostly based on unknown or complex boundary conditions and initial states (Liu & Gupta, 2007; Paniconi & Putti, 2015; Lane et al., 2019). Brigode et al. (2013) has shown that the major source of uncertainty in hydrological modeling is the lack of model robustness, which is the result of unsuitable model structures and inappropriate calibration procedures. Additionally, the annual and interannual variability in streamflow and effects of climate change also increases these uncertainties, which further complicates water resource management (Chiew et al., 2003).

The Weather Research and Forecasting hydrological modeling

* Corresponding author.
E-mail address: yeonjoo.kim@yonsei.ac.kr (Y. Kim).

https://doi.org/10.1016/j.jhydrol.2021.127297
Received 7 April 2021; Received in revised form 24 November 2021; Accepted 27 November 2021
Available online 13 December 2021

system (WRF-Hydro) couples a fully distributed land surface model with routing components, which is used to simulate the streamflow of watersheds. It includes multiple conceptual and physically based approaches, high-resolution terrestrial hydrological processes, and energy fluxes. Accordingly, the model has been used for water cycle analysis, replication of observations, and in forecast systems in numerous fields (Koren et al., 1999; Yucel et al., 2015; Senatore et al., 2015; Arnault et al., 2016; Givati et al., 2016; Xiang et al., 2017; Maidment, 2017; Kerandi et al., 2018).

Similar to other physically-based models, the WRF-Hydro model requires parameter calibration for optimal performance. Earlier studies mainly applied manual calibration through trial-and-error methods. For instance, Yucel et al. (2015) simulated the streamflow for specific flood events in the western Black Sea region of Turkey using a manual and stepwise calibration method. By selecting six parameters affecting the shape of the hydrograph, they calibrated the parameters and simulated a ~ 10 d period of flood events. Their results showed considerable improvements in the fit of the hydrograph, as demonstrated by a 22.2% reduction in the root mean squared error. This substantiated the capability of the WRF-Hydro model to simulate major features of flood events with parameter calibration. Naabil et al. (2017) also employed a manual calibration method using two variables—the infiltration partition parameter and Manning's roughness parameter—to simulate the stream discharge of the Tono Basin in West Africa for a year. The calibrated model was able to capture the time of the maximum peak discharge in the simulated year. However, their results also indicated that a yearlong simulation requires additional improvements to capture other major flood event features, such as the volume and peak flow rate. Recently, enhanced calibration methods that utilize Parameter Estimation Tool software (Senatore et al., 2015) or implement additional channel infiltration function (Lahmers et al., 2019) have also been employed to improve the performance of the WRF-Hydro model. These parameter calibration methods, however, still have performance inadequacies in terms of input forcing datasets and residual error reduction, which arise not only from parameter uncertainties but also from other uncertainties inherent in the model's structural limitations.

Following recent developments in sophisticated machine learning techniques, many studies in the field of hydrology have focused on adopting these techniques to interpolate meteorological data (Appelhans et al., 2015), modeling groundwater (Xu et al., 2014; Xu & Valocchi, 2015; Sahoo et al., 2017), and forecasting streamflow (Rasouli et al., 2012; Shortridge et al., 2016). Among many efforts, long short-term memory (LSTM), which is a recurrent neural network (RNN), has been developed and widely used to better process long-term sequential datasets (Hochreiter & Schmidhuber, 1997; Greff et al., 2015). Kratzert et al. (2018) employed LSTM networks to simulate runoff using catchment attributes and meteorology for large-sample datasets of 241 catchments. They compared this model to the Sacramento Soil Moisture Accounting Model (SAC-SMA) of New England, South Atlantic–Gulf, Arkansas–White–Red, and Pacific Northwest regions of the United States. The results revealed that the LSTM model could predict runoff from climate observations with an accuracy comparable to that of the SAC-SMA model. However, they also reported that LSTM relied strongly on large datasets, and as a data-driven model, it had no explicit representation of real-world hydrological processes. Later, to optimize the parameters, the LSTM with the ant lion optimizer (ALO) model (LSTM-ALO) was employed by Yuan et al. (2018) to predict monthly runoff in the Astor River Basin in Pakistan. The study demonstrated that the LSTM-ALO model improved the model performance by demonstrating the sensitivity of LSTM to datasets and parameter values. Likewise, multiple studies have applied and improved machine learning algorithms to replace physically based models, showing that the LSTM model exhibits a comparable or better performance in simulating runoff (Yuan et al., 2018; Kratzert et al., 2018; Xiang et al., 2020).

Recent studies have shown the potential of machine learning techniques in overcoming the limitations or errors in hydrologic models (Xu et al., 2014; Konapala et al., 2020; Yang et al., 2020). To improve the prediction of physically based regional groundwater flow models in the Republican River basin of the United States, Xu et al. (2014) introduced a combination of two machine learning methods: instance-based weighting and support vector regression. The results showed that the data-driven models based on machine learning algorithms could predict errors in outputs from the groundwater models, thus correcting the simulated outputs by removing the errors. Konapala et al. (2020) demonstrated the potential of a hybrid model to simulate streamflow using the Sacramento soil moisture accounting model (SAC) and LSTM to build two different hybrid models. They compared different combinations of the SAC and LSTM models in 531 catchments across the Conterminous United States. They showed considerable improvements in the Nash–Sutcliffe efficiency (NSE) value with the hybrid models. Furthermore, Yang et al. (2020) employed ANNs coupled with a categorization approach (CA) to simulate streamflow in data-limited areas. To generate training data they used a geomorphology-based hydrological model (GBHM) and computer vision to extract spatial features to use as inputs. The prediction results of their combined model showed improved NSE and percent bias (PBIAS) values when compared to the GBHM standalone simulation. This study also shows the potential of various applications, including machine learning techniques, for overcoming the limitations in previous hydrological models.

This study proposes a hybrid model combining a machine learning algorithm (LSTM) and the WRF-hydro model (hereafter referred to as WRF-Hydro-LSTM) to improve streamflow prediction and demonstrates improved prediction skills with the suggested approach and the model sensitivity of different datasets and hyperparameters over Soyangho Lake, South Korea. In this study, LSTM was used to predict the residual errors between observations and WRF-hydro simulations. It is distinguished from existing studies of correcting biases of streamflow prediction using statistical approaches (e.g., Hashino et al., 2007; Farmer et al., 2018) which are limited in simulating unprecedented or extreme streamflow because the models consider those conditions to be outliers. In particular, we present how the model's performance can vary with different datasets and hyperparameters, which have not been addressed in previous related studies (e.g., Konapala et al., 2020). The simulations using WRF-Hydro and LSTM individually (hereafter referred to as WRF-Hydro-only and LSTM-only, respectively) were also compared with the simulations of the proposed model, WRF-Hydro-LSTM.

2. Methods

2.1. WRF-Hydro model

The WRF-Hydro model was first introduced by the National Center for Atmospheric Research (NCAR) under the Research Applications Laboratory as a hydrological extension package that could be coupled with the WRF model (Gochis & Chen, 2003). It was designed to improve simulations representing terrestrial hydrological processes in the WRF model. Currently, the WRF-Hydro model itself has the capabilities of a distributed and physically based hydrological modeling system. It not only couples regional atmospheric models (i.e., the WRF model) but also acts as an uncoupled, stand-alone, land-surface hydrological model. The uncoupled mode, which is the model configuration used in this study, utilizes one-way process using a meteorological forcing dataset, while the coupled mode has a two-way interaction process between the climate model and the components of the WRF-Hydro model.

There are two options for the land surface model in the WRF-Hydro model: Noah and Noah-Multiparameterization (Noah-MP) land surface models. The WRF-Hydro model can simulate the three-dimensional distribution of discharge, including under the land surface, whereas previous models had only been able to simulate one-dimensional discharges (Gochis et al., 2018). Because the WRF-Hydro model uses Noah land surface models and can simulate three-dimensional distribution, it is also called the Noah distribution or NCAR distributed hydrological modeling system (Gochis & Chen, 2003). The available physics options include land surface parameterization, surface overland flow, saturated subsurface flow, conceptual or empirical base flow, channel routing, and reservoir routing (Gochis et al., 2018). Using these options, the WRF-Hydro model can simulate terrestrial hydrological processes and energy fluxes with multi-spatiotemporal resolution. In this study, we activated surface overland, saturated subsurface, and channel routing for the routing module. The steepest descent for the surface overland flow routing, diffusive wave for the channel routing, and exponential bucket for the base flow model were used.

2.2. LSTM networks

The RNN model is widely used to manipulate sequential data in machine learning (Lipton, 2015). However, RNN has limitations in terms of learning long-term dependencies due to gradient vanishing during multiple backpropagation processes. The LSTM network has been devised to mitigate the weaknesses of RNN models for long-term sequential datasets (Hochreiter & Schmidhuber, 1997; Greff et al., 2015). LSTM is a memory-strengthened version of RNN that can process long-term sequential data with a low level of gradient vanishing, compared to other algorithms. With long-term memory, the LSTM algorithm can predict multivariate time-series data with high accuracy (Hochreiter & Schmidhuber, 1997). The LSTM block structure can model time-series prediction, including long-term dependencies. As such, this study employed LSTM networks, which have an outstanding ability to process time-series data (Fang et al., 2017; Kratzert et al., 2018; Yuan et al., 2018). In the WRF-Hydro-LSTM approach, LSTM provides residual error prediction, which is difficult to predict because it is a consequence of multiple processes.

The LSTM network has a modified RNN structure that has an additional cell-state structure in the hidden layer, which is known as the LSTM block, to enable the learning of long-term time dependencies. The main principle of LSTM can be explained using memory cells and nonlinear gates, which maintain its state and regulate the flow of information in the cell-state structure (Greff et al., 2015) with a constant error carousel (CEC, *C*), an input gate (*I*), an output gate (*O*), and a forget gate (*F*) (Fig. 1. The CEC is a neuron that keeps the error constant and prevents gradient vanishing due to backpropagation. The input gate and output gate serve as signal filters for distinguishing the time-dependent and time-independent signals to prevent input and output weight conflicts. By installing the input gate in front of the CEC and the output gate at the back, past information can be used selectively when required by the signal. To control the value of the CEC, the forget gate receives an error from the CEC and causes the CEC to forget the error when it is considered unnecessary.

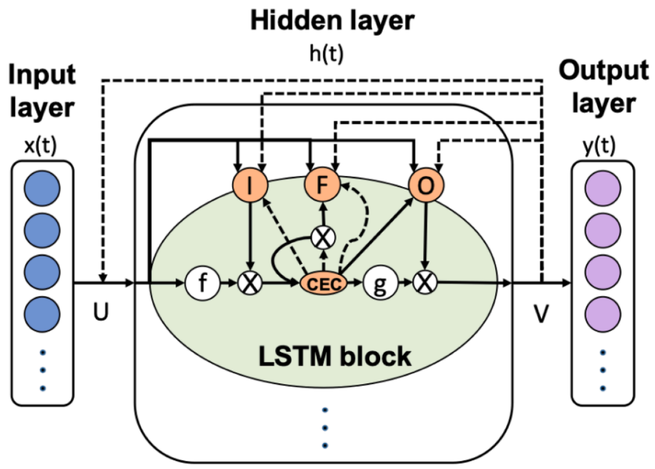

Fig. 1. Schematic of the Long Short-Term Memory network (LSTM) structure and LSTM block. Solid lines indicate the propagation flow; dotted lines indicate the backpropagation flow (Olah, 2015; Graves, 2013).

For demonstrating effective and reasonable modeling performance in the LSTM model, we need to fine-tune the hyperparameters and select proper optimizers, loss functions, and activation functions. The hyperparameters are parameter sets that are used to control the quality of learning, which depend on the structure of the machine learning algorithm. As the machine learning process in the LSTM model is mostly based on stochastic gradient descent optimization, the principal parameters for efficient and accurate learning performance are considered as the number of hidden layers, number of epochs and nodes (neurons) in the hidden layer, learning rate, and batch size (Kratzert et al., 2019; Xiang et al., 2020). An epoch indicates a complete training process using the entire training dataset, batch size is the total amount of training data allotted in a single batch, learning rate is the step size in the training process, and a node is a computational unit that organizes a network in a hidden layer. For these variables, there is not an exact reference number or calculation method; therefore, their values should be set by user preference depending on the type of dataset or via a trial-and-error method.

2.3. WRF-Hydro-LSTM approach

We developed the WRF-Hydro-LSTM approach (Fig. 2 to improve streamflow prediction using the traditional WRF-Hydro model while preserving the physical constraints in the model. Therefore, to minimize the impact on physical constraints and maximize the strength of the LSTM, we used the LSTM for predicting the difference between the observed and WRF-Hydro-simulated inflows (residual errors) from the meteorological inputs. As previously noted, residual errors are consequences of multiple sources and are difficult to quantify, which makes it difficult to predict and remove them. However, this approach was developed based on the assumption that the residual errors in the model

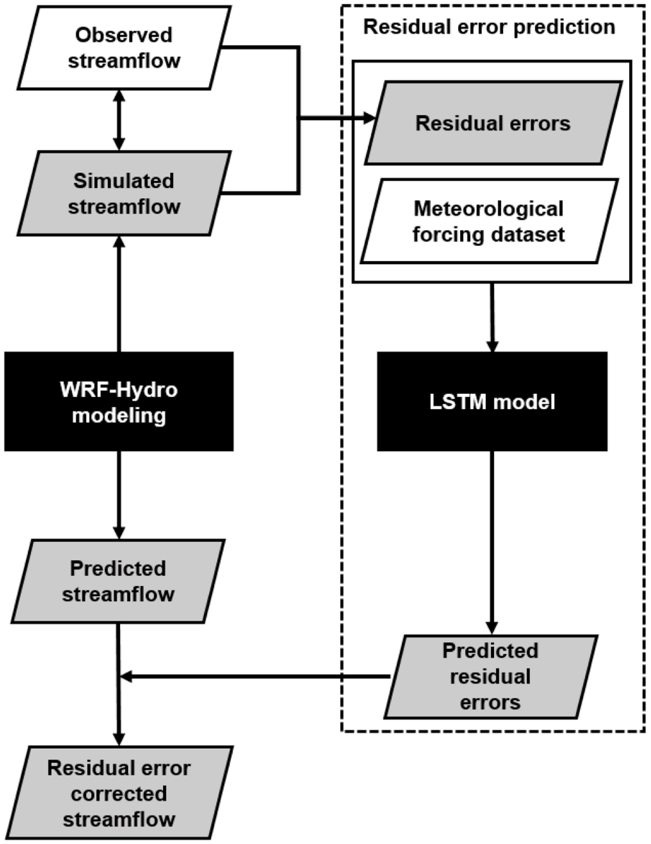
Observed streamflow
Simulated streamflow
WRF-Hydro modeling
Predicted streamflow
Residual error corrected streamflow
Residual error prediction
Residual errors
Meteorological forcing dataset
LSTM model
Predicted residual errors

Fig. 2. Schematic of the WRF-Hydro-LSTM process.

result would have a pattern. In our approach, LSTM modeling succeeds the WRF-Hydro simulation step and processes inputs independently, as shown in Fig. 2.

As presented in Fig. 2, we initially generated the streamflow simulation data using the WRF-Hydro model (i.e., WRF-Hydro-only approach), followed by the calculation of residual errors between the simulated and observed data for the LSTM model input. With residual errors and the meteorological forcing dataset, we optimized the hyperparameters and trained the LSTM model. Finally, we predicted the residual errors of the future time step using a trained LSTM model and applied the predicted values to the WRF-Hydro simulation for the future time step. Thus, the WRF-Hydro-LSTM approach improved the conventional WRF-Hydro model simulation by applying predictive but difficult-to-quantify errors from multiple uncertainties.

2.4. Case study from the Soyangho Lake basin

2.4.1. Study area

We applied the WRF-Hydro-LSTM approach for the Soyangho Lake basin located in the northeastern region of South Korea, which extends over the province of Gangwon-do (Fig. 3a). The Soyangho Lake basin is the main source of water and hydroelectricity in the Seoul metropolitan region, which is the most densely populated region in South Korea. The total area of the basin is 2703 km^2, with a mean annual precipitation of 1154 mm (1981–2010) (National Water Resources Management Information System, 2003). Soyangho Lake was created by the Soyang Dam, which has a storage capacity of 2900 million tons of water. It is the largest rockfill dam in Asia and serves multiple purposes, including hydroelectricity generation, water supply, and flood control. The mean annual inflow of Soyangho Lake was 2.148×10^9m^3, and its streamflow was the target of the simulations performed in this study. The Korea Water Resources Corporation routinely measures streamflow using gauges and provides inflow and outflow data for the dam. Owing to the absence of any other hydraulic structure in the upstream area of Soyangho Lake, we used the hourly inflow rate for the calibration and evaluation to exclude anthropogenic impacts, such as outflow control in the upstream region.

2.4.2. Meteorological forcings and terrain datasets

In this study, the Noah-MP model within the WRF-Hydro model was configured with a grid size of 1500 m, including the entire watershed area of Soyangho Lake. For the routing module, nested routing grid cells within the land surface model with a grid size of 100 m were used, and the number of grid cells was 660 × 660.

For meteorological data (2013–2018), we used the Local Data Assimilation and Prediction System (LDAPS) datasets, which was the product of a numerical weather prediction model that employed a three-dimensional data assimilation technique with a 1.5 km horizontal spatial resolution and 70 vertical layers. LDAPS provided the 3 h data of precipitation, temperature, surface pressure, u and v components of wind, specific humidity, and longwave and shortwave radiations, which are the meteorological forcings required to drive the WRF-Hydro model. Among these variables, precipitation is the most important variable for streamflow simulation. The data assimilation product (LDAPS data) for precipitation was replaced with observational data gridded using the Parameter-elevation Regressions on Independent Slopes Model (PRISM) method (Daly et al., 2007; Daly et al., 2008; Kim et al., 2013). PRISM considers distance, elevation, cluster, vertical layer, topographic facet, coastal proximity, topographic position, and effective terrain of the station by weights. The data consistency among the meteorological variables might be compromised to some extent, but it is necessary to improve the precipitation data because it is critical for accurately predicting streamflow. The observed gauge-based precipitation data from the Automated Synoptic Observing System (ASOS) stations in the upstream area, Injae, Sokcho, Daegwallyeong, and Gangneung-1, 2 (Fig. 3b), were used because these stations directly affected the streamflow of the basin and were interpolated in the grid data at a 1.5 km resolution.

Terrain datasets were required for the WRF-Hydro model. Each grid of the domain for the land surface model contained static terrestrial data, such as those for land use, soil type, topographic height, and vegetation fraction. Gridded data was interpolated horizontally using the WRF Preprocessing System (WPS) geographical static dataset with a 30″ resolution, which consisted of the United States Geological Survey datasets and the Moderate Resolution Imaging Spectroradiometer (MODIS). For the land-use data, we used 20 MODIS categories. A reconnaissance soil map provided by the Rural Development Administration of Korea was used to classify the soil types and their characteristics (Fig. 3c). Soil types were classified and their material properties were updated accordingly. Reconnaissance soil maps with scales of 1:50,000 contain soil characteristic data for soil type, slope, soil fraction, drainage, effective soil depth, and land-use type. These maps were used to extract two-dimensional soil texture type information and aggregated for a resolution of 1.5 km.

Furthermore, we used spatial topographic data and lake polygons from SHuttle Elevation Derivatives (HydroSHEDs) at multiple scales for the routing module. In this study, hydrologically conditioned elevation (CON) data from HydroSHEDs with a resolution of 3″ were used for the terrain elevation data. CON data is a result of DEM postprocesses such as the deepening of open water surfaces, weeding of the coastal zone, stream burning, filtering, molding of valley courses, sink filling, and carving through barriers (Lehner et al., 2008). Using these terrain

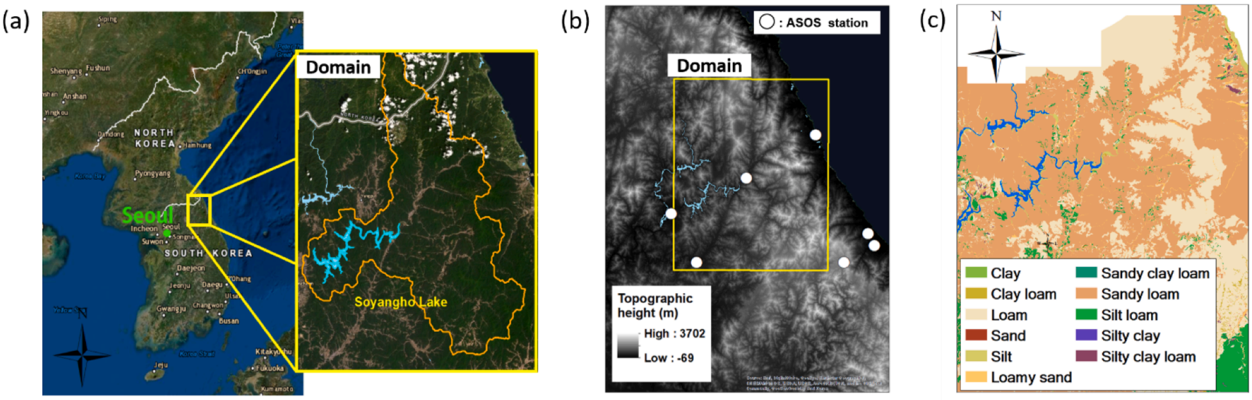

Fig. 3. a) Location of Soyangho Lake, South Korea; b) locations of the Automated Synoptic Observing Systems (ASOS) referenced in this study, topographic elevation data, and lake polygon file (from HydroSHEDs and HydroLAKES); and c) the soil texture map.

datasets, we generated flow directions, watersheds, stream and reservoir networks, one-dimensional parameter files for lake and stream networks, nest grid cells for routing, and the lake's geographical information. Moreover, HydroLAKES data, which represent a global digital map of lakes larger than 10 ha, were employed to improve the ability of the model to reflect the role of the lake (Messager et al., 2016).

2.4.3. Experimental designs

We performed three different types of model simulations: WRF-Hydro-only, WRF-Hydro-LSTM, and LSTM-only for the Soyangho Lake basin. Initially, we performed a WRF-Hydro-only simulation from meteorological data from 2013 to 2018 using eight meteorological forcing data variables (Table 1).

Based on the calculated residual errors of the WRF-Hydro-only model with respect to the observations and meteorological variables, the LSTM model was trained with data from 2013 to 2016 and validated for 2017 (Table 1). Notably, the average of the spatially variant meteorological datasets was used for training the LSTM model. The trained LSTM model with optimized hyperparameters predicted the residual errors and then applied them to the WRF-Hydro-only streamflow prediction for 2018. Please refer to Section 2.4.4 for details on the optimized hyperparameters of the LSTMs.

Finally, we performed an LSTM-only simulation to predict streamflow. Similar to the WRF-Hydro-LSTM approach, the LSTM model was trained, but this was undertaken with the LDAPS meteorological dataset and observed streamflow in the WRF-Hydro-LSTM model (Table 1). In addition, the hyperparameter values were optimized for the LSTM-only case. Therefore, all the settings were the same as in the WRF-Hydro-LSTM approach, except for the target dataset and values for the hyperparameters. We then predicted the streamflow for 2018 using the trained LSTM model.

The streamflow prediction of the WRF-Hydro-LSTM approach was evaluate compared to that of WRF-Hydro-only and LSTM-only. We conducted a comparative analysis using three commonly used metrics for hydrological model evaluation: PBIAS, NSE, and the correlation coefficient (r). The metrics are computed as follows:

$$PBIAS = \frac{\sum_{t=1}^{N}(Q_{sim}^{t} - Q_{obs}^{t})}{\sum_{t=1}^{N}Q_{obs}^{t}} \times 100\% \tag{1}$$

$$NSE = 1 - \frac{\sum_{t=1}^{N}(Q_{sim}^{t} - Q_{obs}^{t})}{\sum_{t=1}^{N}(Q_{sim}^{t} - \overline{Q}_{obs})} \tag{2}$$

$$r = \frac{\sum_{t=1}^{N}(Q_{obs}^{t} - \overline{Q}_{obs})(Q_{sim}^{t} - \overline{Q}_{sim})}{\sqrt{\sum_{t=1}^{N}(Q_{obs}^{t} - \overline{Q}_{obs})^{2}}\sqrt{\sum_{t=1}^{N}(Q_{sim}^{t} - \overline{Q}_{sim})^{2}}} \tag{3}$$

where $\overline{Q}_{sim}$ is the mean of the simulated streamflow, $\overline{Q}_{obs}$ is the mean of the observed streamflow, Q_{sim}^{t} is the simulated streamflow at time t, and Q_{obs}^{t} is the observed streamflow at time t. PBIAS is widely used to evaluate hydrological models by indicating the magnitude of over- or under-estimations compared to the observations. The NSE and r values were used to evaluate the predictive power of the model and the linear correlation between the observations and predictions, respectively. For the WRF-Hydro-LSTM and LSTM-only models, we used the 50th percentile result while repeating the same experiment 30 times.

A sensitivity test was conducted to examine the impact of the training dataset duration and determine the years for which the data needed to be used to evaluate model performance. Consequently, we used four years of training data (2013–2016), which was the longest sequential period possible with available data, and we labeled it according to the starting and ending training periods (TP1316). The test sets were configured with a total of 10 cases with different training periods (TP13, 14, 15, 16, 1314, 1415, 1516, 1315, 1416, and 1316). For the validation, we utilized data from 2017 because it is the last year before the prediction year (2018). It is impossible to have information about the prediction period (i.e., whether it will be a wet, dry, or normal year) beforehand; thus, we sought to improve the performance using the closest year's data, which is expected to have a climate trend similar to that of the prediction period under climate change (Trenberth, 2011). Furthermore, the validation year (2017) and the prediction year (2018) were wet years. We then performed an additional validation year for a case with a dry year, 2014. We examined how the characteristics of the validation year influence the model performance with the cases using 2013, 2015, 2016, and 2017 for training and 2014 for validation (TP13 + 1517).

Table 1
LSTM input variables.

Variable	Unit	Case
Specific humidity	kg/kg	WRF-Hydro-LSTM & LSTM-only
Incoming shortwave radiation	W/m^2	
Incoming longwave radiation	W/m^2	
Near-surface wind in the *u*-component	m/s	
Near-surface wind in the *v*-component	m/s	
Air temperature	K	
Surface pressure	Pa	
Precipitation rate	mm/s or $kg/m^2/s$	
Residual error (Observed – simulated (WRF-Hydro) inflow)	m^3/s	WRF-Hydro-LSTM
Observed inflow	m^3/s	LSTM-only

2.4.4. Hyperparameter tuning

For the LSTM model, the number of nodes per hidden layer, batch size, learning rate, and number of epochs must be determined, as introduced in Section 2.2. In this study, the random search method (Bergstra & Bengio, 2012) was used to identify a set of optimal values for hyperparameters because they are nonlinearly correlated and have no fixed range. We generated 200 random hyperparameter values for the candidate set within a certain range (Table 2) and evaluated the values for both the WRF-Hydro-LSTM and LSTM-only methods. The range was configured based on experience and results from previous studies (Stathakis, 2009; Masters & Luschi, 2018). As the LSTM model shuffles the split data for each epoch, the starting point for training changes, and the training results may vary within a certain range. Therefore, we permitted 10 iterations to evaluate each candidate set and selected the optimal set, which had the lowest mean squared error in the validation process.

In this study, two LSTM hidden layers and an adaptive moment estimation optimizer (Adam) (Kingma & Ba, 2014) were used for stable and efficient model optimization. Adam, the latest version of the optimization algorithm for stochastic gradient descent, can rapidly and accurately search for optimal machine learning values. We set the Adam parameters in the Keras API using default values (0.001 for the learning rate, 0.9 for the momentum decay rate, and 0.999 for the adaptive term decay rate). To train the model with updated weights, we employed the

Table 2
Random search ranges for hyperparameter tuning.

Hyperparameter	Range
Number of nodes (1st LSTM layer)	8, 16, 24, ⋯, 256 (Step size: 8)
Number of nodes (2nd LSTM layer)	8, 16, 24, ⋯, 256 (Step size: 8)
Batch size	8, 16, 24, ⋯, 200 (Step size: 8)
Learning rate	0.001, 0.0005, 0.0001
Number of epochs	10, 11, 12, ⋯, 100 (Step size: 1)

mean squared error for the loss function and mean absolute error to optimize the objective function.

3. Results and discussion

3.1. Analysis of the WRF-Hydro model

We first evaluated the performance of the WRF-hydro model for peak flow in 2018 (Fig. 4). The hydrograph indicates that the simulated values at the peak point were slightly lower and decreased more slowly than the observed values. Moreover, in comparison with the observations, the NSE of the WRF-Hydro model was 0.93, r was 0.96, and PBIAS was −6.12, suggesting that the model captured a total volume of inflow that is quite close to the observations, albeit with a slight overestimation. These results show that the WRF-hydro model is capable of simulating streamflow with effective predictive skills for a specific short period of a few weeks.

We then evaluated the performance of a one-year streamflow simulation using the same model (Fig. 5). The evaluation results for a year (PBIAS = 11.78, NSE = 0.72, and r = 0.88) indicated a slight decrease in performance compared to the short-term case (Fig. 5a). The results were still acceptable in hydrological modeling according to Moriasi et al. (2007) because the NSE was higher than 0.5, PBIAS was less than 25, r and NSE decreased slightly, and the PBIAS increased from −6.12 to 11.78, suggesting considerable underestimation of streamflow for an extended period. We identified multiple peaks, indicating a large difference between the observations and model simulation results (e.g., between May and June as well as at the beginning of July), and the sum of the differences (overestimated and underestimated values) contributed to the relatively large PBIAS.

3.2. Evaluation of streamflow prediction skills

We evaluated the performance of WRF-Hydro-LSTM in comparison to WRF-Hydro-only and LSTM-only for simulating the 2018 streamflow in the Soyangho Lake basin (Fig. 5). The streamflow simulated with WRF-Hydro-LSTM and LSTM-only had a hydrograph shape markedly similar to that of the observations unlike the streamflow simulated with WRF-Hydro-only, although some of the peak flows were still captured with limitations in both WRF-Hydro-LSTM and LSTM-only. The metrics also indicated the superiority of both WRF-Hydro-LSTM and LSTM-only over WRF-Hydro-only, with an NSE of 0.95 and r above 0.97. This result was predicted as the loss function for optimization in the LSTM model and directly associated with the NSE and r, which was set with the mean squared error. WRF-Hydro-LSTM showed a decrease in PBIAS, whereas LSTM-only showed an increase in PBIAS in comparison to WRF-Hydro-only. The PBIAS values indicate that the underestimation bias of WRF-Hydro-only was improved in WRF-Hydro-LSTM; otherwise, LSTM-only was considerably more biased toward underestimation than WRF-Hydro-only. Consequently, the volume of streamflow accumulated for 2018 (Fig. 6), which is closely related to PBIAS, showed that out of the three models, only WRF-Hydro-LSTM captured the observed trends reasonably. Similarly, we noted that the performance of LSTM-only is poor in low flows, thus exerting a relatively smaller effect on the loss function than high flows.

Because LSTM uses a loss function for regression in the training process without any physical constraints, such as conservation principles, it can violate the laws of physics while minimizing losses. This could further worsen the results. In this study, we found that these characteristics of LSTM, evaluated by the range of values and standard deviation (SD), could limit the results owing to the variability of datasets. While LSTM-only, with a wide range of fluctuations in values and high variability between dry and wet seasons, directly predicted the streamflow (range: 0–2325, SD: 174.8, 2018), the LSTM model in the WRF-Hydro-LSTM approach could only predict residual error, whose values fluctuate less seasonally (range: −713–1067, SD: 109.8 in 2018). Among the different roles of LSTM, the WRF-Hydro-LSTM approach simulated the streamflow in a manner markedly closer to the observations, while systematically limiting the impacts of possible errors from the LSTM model and conserving physical constraints as opposed to LSTM-only.

3.3. Training data sensitivity of WRF-Hydro-LSTM and LSTM-only

We also examined how the duration and timing of the training data influenced prediction skills. Fig. 7 demonstrates that climate conditions, such as precipitation, show strong interannual variability, and hence, the training data from different years (i.e., dry or wet years) could influence the streamflow prediction skills of LSTM models. Notably, in this study, the LSTM models (i.e., WRF-Hydro-LSTM and LSTM-only) were trained with data from the years 2013 to 2016 and validated with data from 2017, and the trained models were then evaluated for the year 2018 in the default case (TP1316). We conducted sensitivity tests with various training datasets, utilizing 30 iterations per dataset, over numerous time periods (1, 2, 3, and 4 years between 2013 and 2016).

The NSE and r values in all cases of WRF-Hydro-LSTM and LSTM-only were found to be close to 1 for all evaluations with the three metrics (NSE, r, and PBIAS) (Fig. 8), without considerable differences between the cases. Such model performance was foreseeable because we used the mean squared error for the loss function to train the LSTM models. Furthermore, the ranges of NSE and r values in the 30 iterations of LSTM-only were much wider than those in the WRF-Hydro-LSTM

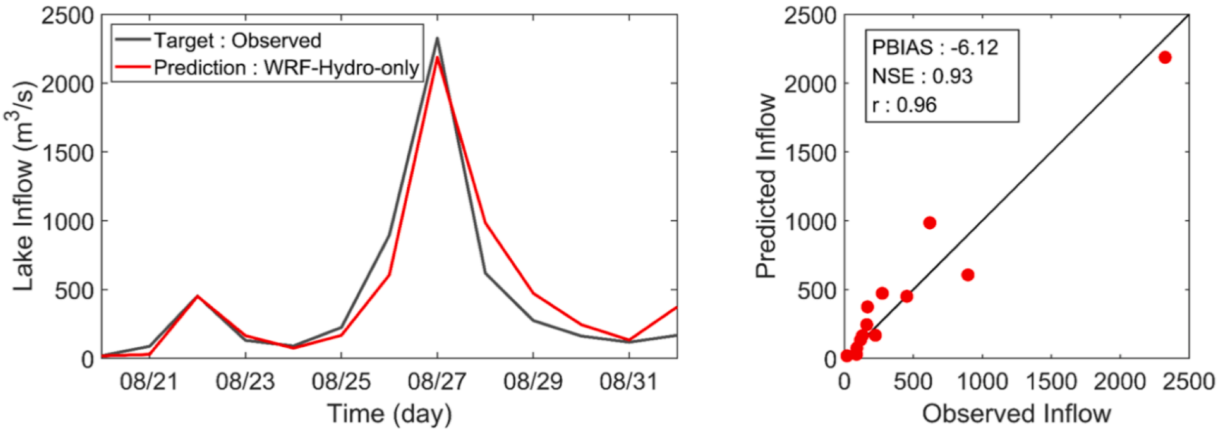

Fig. 4. Comparison between the observations and WRF-Hydro-simulated streamflow time series at Soyangho Lake for the period from August 19, 2018 to September 1, 2018.

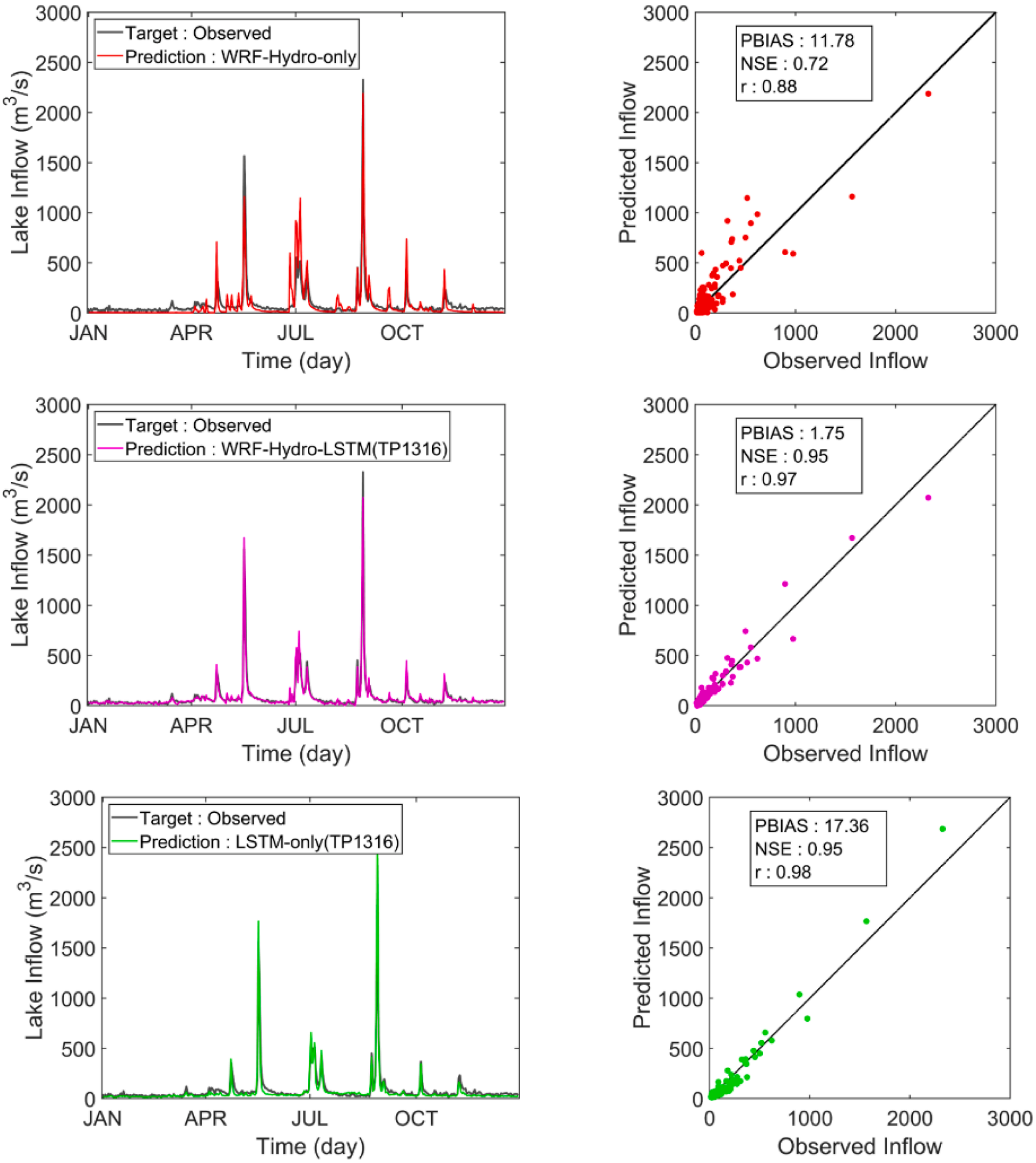

Fig. 5. Comparison of the target and predicted time series in the LSTM network-based machine learning process. a) Results of the WRF-Hydro-only simulation; b) corrected results of the WRF-Hydro simulation with predicted residuals (WRF-Hydro-LSTM); and c) results of the lake inflow prediction using LSTM-only.

model, indicating that WRF-Hydro-LSTM provides a more stable prediction than LSTM-only. These aspects can be explained by the difference in target value and the influence of the LSTM model prediction between WRF-Hydro-LSTM and LSTM-only. The WRF-Hydro-LSTM model uses residual errors as target values for simulation and predicts streamflow as the sum of the two components. In contrast, LSTM-only uses streamflow data as the target value and predicts the streamflow directly. This may lead to more systematic errors owing to the wide range of fluctuations in the streamflow values.

The PBIAS values in Fig. 8 vary markedly according to the training dataset used for the LSTM models. In particular, the prediction results based on PBIAS were substantially influenced by the configuration of the training dataset. The best configuration of the training dataset can be represented by the consistency and similarity of the training dataset with the target dataset. In WRF-Hydro-LSTM, the cases ending in 2015 showed wide variation and large absolute PBIAS. The cases using the years 2013 or 2016 as the training data performed relatively better than the cases ending in 2015. Table 3 shows the smallest absolute PBIAS in cases using 2016 as the training set (TP16, TP1516, and TP1416). This was followed by TP13 and TP1314, which included 2013 in the training set. These findings can be attributed to the fact that the prediction year (i.e., 2018), with a relatively large residual error between the observation inflow and WRF-Hydro-simulated inflow, is analogous to the patterns in the years 2013 and 2016 (Fig. 7).

In the case of LSTM-only, the cases with training datasets that either begin or end in 2015 (dry year) showed unstable PBIAS, and the smallest absolute PBIAS in cases mainly included the year 2013 (wet year) in the training set (i.e., TP13 and TP1315, followed by TP1415, TP15, and TP1314) (Fig. 8 and Table 3). This tendency was similar but less prominent when compared to the WRF-Hydro-LSTM model. These

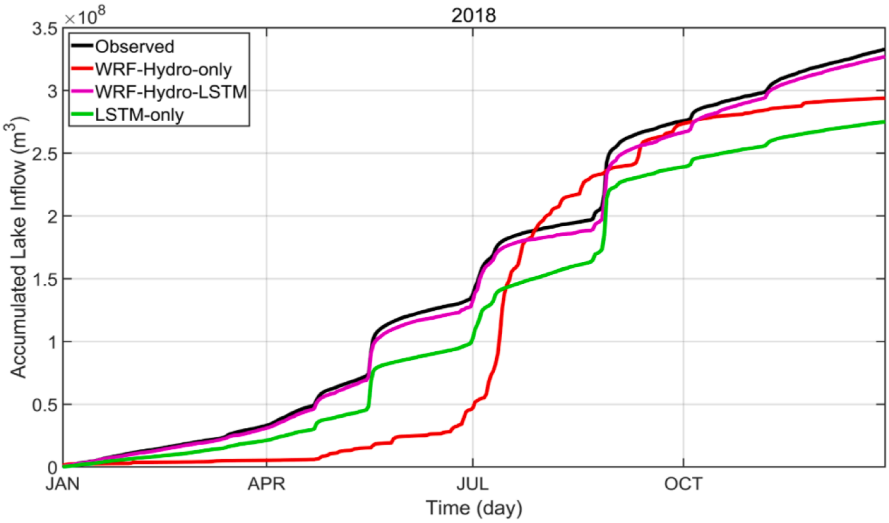

Fig. 6. Comparison of the cumulative lake inflow between the simulations and observations.

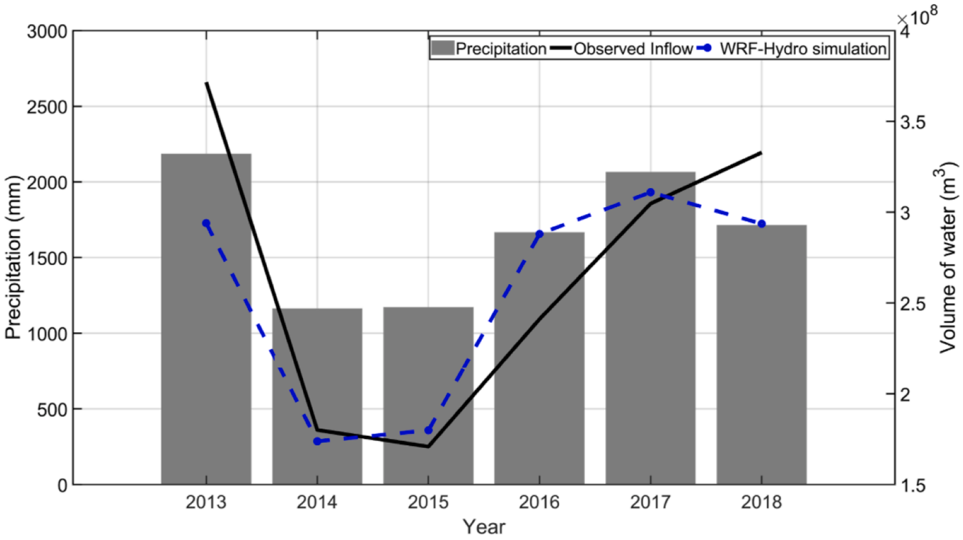

Fig. 7. Trends in total annual precipitation, observed inflow, and simulated inflow using the WRF-Hydro model.

findings suggest that the training data should have a pattern similar to that of the target data for the best performance of LSTM (the year 2018, wet year). Furthermore, we identified the importance of data consistency from the results of TP1415 and TP1314. According to the similarity between the target year 2018 (wet year) and 2013 (wet year), TP1314 should have produced a better result than TP1415; however, it showed an even higher value for the PBIAS (Table 3). This result can be explained by the consistency of data responsible for stabilizing the LSTM training process, as clearly shown in the results of TP1415 and TP1315. Two consecutive dry years (i.e., the years 2014 and 2015) yielded a better result, whereas the cases with nonconsecutive composition produced a higher value of PBIAS, even though they included the year 2013 (i.e., TP1316 and TP1314). We also found that the number of years included in the training dataset that had a similar pattern to the target year was more important than the length of the training period. For example, TP1316 (default case) in the WRF-Hydro-LSTM model had the longest training period and showed the best performance among all the cases. TP16 and TP1516 yielded equivalent results, although they used a shorter period (Table 3).

Finally, we employed another case using 2014 (dry year) for validation and 2013, 2015, 2016, and 2017 for the training period (TP13 + 1517) (Table 3). Regardless of the characteristics of the validation year, two distinct cases, the wet year (2017) and the dry year (2014), showed similarly successful performance with their PBIAS, NSE, and r values (TP1316 vs. TP13 + 1517 in Table 3). This result shows that our LSTM approaches had stable performance regardless of the characteristics (i.e., wet or dry) of the validation year.

3.4. Hyperparameter tuning for WRF-Hydro-LSTM and LSTM-only

Hyperparameter tuning for each model is required to utilize the LSTM for the LSTM-only and WRF-Hydro-LSTM models. We tuned hyperparameters separately for all different cases with LSTM-only and WRF-Hydro-LSTM, and all cases showed different optimal values for each hyperparameter (Table 4). Focusing on the default case (TP1316, evaluated in Section 3.2) in Table 4 revealed that the WRF-Hydro-LSTM model used more nodes for the first hidden layer and smaller values for the batch size and learning rate than LSTM-only. A large number of nodes in the hidden layer indicate that the data requires a greater capacity for wide training networks. The smaller batch size and learning rate indicate that the training data are difficult to generalize, thus requiring smaller gradient steps for updating the weights (Goodfellow

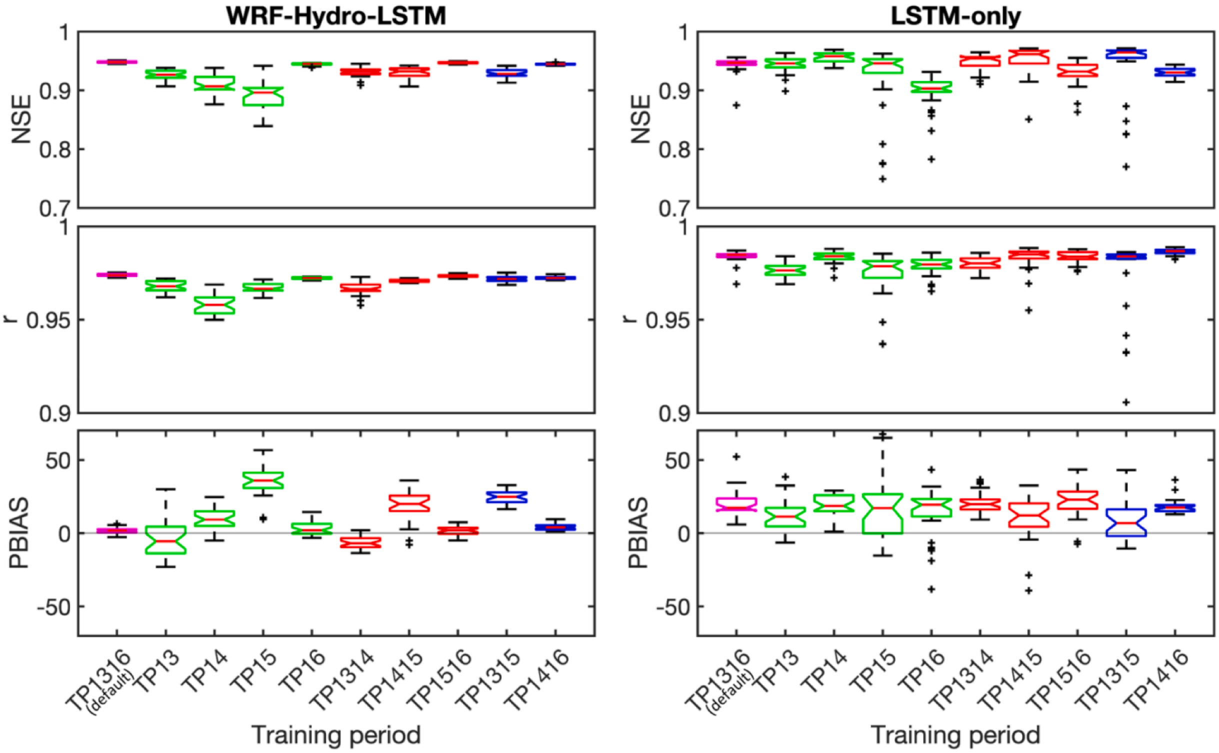

Fig. 8. Comparisons among the different training periods (TP) cases (1, 2, 3, and 4 years between 2013 and 2016) with their optimized combination of hyperparameters (30 iterations per case).

Table 3
Comparison of WRF-Hydro-LSTM and LSTM-only inflow prediction for Soyangho Lake in 2018 with different training periods (50th percentile value among 30 iterations; * indicates the case with the lowest PBIAS among cases with the same length of training period).

			WRF-Hydro-LSTM			LSTM-only		
Case	Training Period	Validation Period	PBIAS	NSE	r	PBIAS	NSE	r
TP1316 (default)	2013–2016	2017	1.78	0.95	0.97	17.33	0.95	0.98
TP13	2013		−5.57	0.93	0.97	11.30*	0.95*	0.98*
TP14	2014		9.29	0.91	0.96	18.60	0.96	0.98
TP15	2015		35.86	0.90	0.97	17.17	0.95	0.98
TP16	2016		1.98*	0.94*	0.97*	19.40	0.90	0.98
TP1314	2013–2014		−6.94	0.93	0.97	19.77	0.95	0.98
TP1415	2014–2015		19.96	0.93	0.97	12.14*	0.96*	0.98*
TP1516	2015–2016		1.91*	0.95*	0.97*	22.87	0.93	0.98
TP1315	2013–2015		24.84	0.93	0.97	6.83*	0.96*	0.98*
TP1416	2014–2016		4.02*	0.94*	0.97*	17.44	0.93	0.99
TP13 + 1517	2013, 2015–2017	2014	1.18	0.95	0.98	5.03	0.96	0.99

Table 4
Optimized values for hyperparameters of different models and cases with the 2017 validation period using the random search method (N1, N2: Number of nodes in the first and second hidden layers, respectively; BS: Batch size; LR: Learning rate; EP: Epochs).

		WRF-Hydro-LSTM					LSTM-only				
Case	Training period	N1	N2	BS	LR	EP	N1	N2	BS	LR	EP
TP1316 (default)	2013–2016	176	88	16	0.0001	56	64	96	120	0.0005	70
TP13	2013	32	120	24	0.001	64	184	56	168	0.0005	99
TP14	2014	16	16	8	0.0005	74	248	48	168	0.001	91
TP15	2015	112	120	24	0.001	76	80	192	136	0.001	87
TP16	2016	240	104	16	0.0001	90	104	80	176	0.001	95
TP1314	2013–2014	192	88	16	0.001	86	120	216	16	0.0001	83
TP1415	2014–2015	104	136	24	0.001	81	144	176	176	0.001	82
TP1516	2015–2016	224	200	8	0.0001	98	144	104	168	0.0005	76
TP1315	2013–2015	160	200	24	0.001	99	192	104	168	0.0005	93
TP1416	2014–2016	96	88	16	0.0001	75	184	64	192	0.0005	88

et al., 2016). This result suggests that the prediction of residual error (WRF-Hydro-LSTM) requires a more substantial workload than that of streamflow (LSTM-only).

Moreover, Table 4 also shows that the optimal values for each case varied regardless of the duration and specific year used for the training (the details of data sensitivity test in Section 3.3), and thus, there is no noticeable tendency among the cases. Because all cases showed different combinations of numbers without clear characteristics, we found that employing the random search method was efficient and appropriate for finding optimal values. Nevertheless, the results indicate that the case with data restricted to training the LSTM model for optimized prediction skill leads to a large number of nodes, small batch sizes, and low learning rates. The WRF-Hydro-LSTM cases tended to have smaller batch sizes compared to the LSTM-only cases, implying that the residual errors require much more iterations for training the LSTM model than for streamflow. While streamflow has many variable values, limiting the performance of LSTM, it has factors that can be generalized, such as streamflow seasonality and the correlation between precipitation data (r: 0.8841 in 2018). The generalized pattern in the training dataset makes training the LSTM model easier. In contrast, residual errors have an ambiguous correlation with other forcing variables, including precipitation ($r = –0.48$ in 2018); therefore, training with residual errors requires more effort than training with streamflow.

4. Conclusions

This study aims to improve the streamflow prediction of the WRF-Hydro model using a machine learning algorithm (i.e., LSTM), particularly for improving long-term streamflow predictions. Our analysis in Soyangho Lake, South Korea demonstrates that the strength of machine learning to be used in improving physically based model, which is not addressed much before. The proposed WRF-Hydro-LSTM predicts differences between WRF-Hydro predictions and observations through embedded LSTM model and considerably improve WRF-Hydro predictions.

WRF-Hydro-LSTM and LSTM-only show improved results in terms of the NSE and correlation coefficient compared to WRF-Hydro-only; however, in terms of the PBIAS, WRF-Hydro-LSTM had a better value than LSTM-only. These dissimilarities in PBIAS values were mainly attributable to the several limitations of LSTM, which tends to generate more systematic errors with highly variable datasets and optimizes the weights based on the given loss function without any physical constraints. Although LSTM makes nonlinear modeling easier and allows a relatively short computational time, LSTM-only is limited in extreme event predictions, and thus, WRF-Hydro-LSTM is considered the most suitable for streamflow prediction.

Furthermore, we performed a sensitivity test on the training dataset in terms of its length and elements (1, 2, 3, or 4 years of data with different component years). The results indicated that the correlation coefficient and NSE value were not overly sensitive to different training sets because we used the mean squared error for the loss function, but the PBIAS value differed substantially depending on the training set. Both WRF-Hydro-LSTM and LSTM-only displayed improved performance when the training set contained specific years (i.e., the years 2013 and 2015 for LSTM-only and 2016 for WRF-Hydro-LSTM) that had input characteristics (i.e., precipitation, streamflow, and residual errors) similar to those of the target year. More years similar to the target year in trend and composition in the training dataset was also crucial for the prediction result. In addition, a test case with different validation years (2014) supported that the model performance may vary with different validation and target years quantitatively, not qualitatively. As such, it is usually impossible to recognize the trend of the target data for prediction; thus, we recommend that using large data would be advantageous for the ideal performance of the LSTM model.

Our approach has several limitations that need to be improved for clarity and performance. We evaluated the model performance through the representative basin, Soyangho Lake, South Korea, with accumulated observational data (2013–2018). Notwithstanding the limited period of data and number of basins used, we argue that the evaluation results demonstrate the strength of the combined structure of the WRF-Hydro-LSTM and the potential of LSTM application. Although there was a short period for training and calibration, meteorological conditions such as precipitation in South Korea show significant yearly seasonal variations, and most years have dry periods in winter and spring with heavy rain periods in summer. The data period may limit the performance to some extent, but the 6-year data is acceptable given the data availability from the KLDAPS dataset. However, the LSTM model, as a data-driven model, has high data dependency, which requires an accurate and sufficient training dataset. Limited available historical meteorological data for training the model should be complemented by techniques such as cross-validation and an increased number of training iterations.

In recent years, hydrologists have attempted to apply machine learning algorithms in lieu of physically based models because of their efficiency and favorable performance. Given this trend, we showed that machine learning algorithms can be used to improve streamflow prediction through physically based modeling. Although the machine learning approach has limitations, it has more potential in this method than when used directly to make streamflow predictions. Although the approach continues to be developed with new types of algorithms, such as physics-guided neural networks (PGNN), they still include a limited number of options. Read et al. (2019) attempted to develop a PGNN to improve the prediction of lake water temperatures, but they also had to simplify the physics of the system owing to this limitation. Therefore, our proposed approach improves machine learning limitations and utilizes its strengths for improvement during application. Furthermore, because our approach requires only residual errors and meteorological forcing data to drive the physically based model, it can be applied to other models and easily used to efficiently improve simulations of any basin.

We also note that WRF-Hydro facilitates applying the trained LSTM to the upstream and downstream streamflow, i.e., to the stream without a gauge. With transfer learning for deep learning neural networks (Weiss et al., 2016), a pretrained model for one domain with a set of hyperparameters can be fine-tuned on any relevant domain or problem. Although in this study we focused on understanding model sensitivity by training the model separately for all different cases, a well-developed WRF-Hydro-LSTM model for one basin can be applied to different locations, in the same or another basin, using a transfer learning technique. The applicability of the trained WRF-Hydro-LSTM should be further investigated in a future study.

Data availability

The datasets for this study are accessible online. The HydroSHEDS-conditioned DEM and Lake polygons are available at http://www.hydrosheds.org, the inflow observation data for the Soyangho Lake is available at http://www.water.or.kr, LDAPS and ASOS meteorological forcing datasets are available at http://data.kma.go.kr, and the soil map is available at http://soil.rda.go.kr.

CRediT authorship contribution statement

Kyeungwoo Cho: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – original draft. **Yeonjoo Kim:** Conceptualization, Investigation, Methodology, Writing – review & editing.

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.

Acknowledgements

This work was based on a Master's thesis by Kyeungwoo Cho at Yonsei University (February 2020), advised by Yeonjoo Kim. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (2018R1A1A3A04079419 and 2020R1A2C2007670), and by the Technology Advancement Research Program through the Korea Agency for Infrastructure Technology Advancement (KAIA) funded by the Ministry of Land, Infrastructure, and Transport (21CTAP-C163541-01).

References

Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. *Spatial Statistics* 14, 91–113. https://doi.org/10.1016/j.spasta.2015.05.008.

Arnault, J., Wagner, S., Rummler, T., Fersch, B., Bliefernicht, J., Andresen, S., Kunstmann, H., 2016. Role of runoff-infiltration partitioning and resolved overland flow on land-atmosphere feedbacks: a case study with the WRF-Hydro coupled modeling system for West Africa. J. Hydrometeorol. 17 (5), 1489–1516. https://doi.org/10.1175/JHM-D-15-0089.1.

Asefa, T., Kemblowski, M., McKee, M., Khalil, A., 2006. Multi-time scale streamflow predictions: the support vector machines approach. J. Hydrol. 318 (1–4), 7–16. https://doi.org/10.1016/j.jhydrol.2005.06.001.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13 (10), 281–305.

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resour. Res. 36 (12), 3663–3674. https://doi.org/10.1029/2000WR900207.

Brigode, P., Oudin, L., Perrin, C., 2013. Hydrological model parameter instability: a source of additional uncertainty in estimating the hydrological impacts of climate change? J. Hydrol. 476, 410–425. https://doi.org/10.1016/j.jhydrol.2012.11.012.

Carpenter, T.M., Georgakakos, K.P., 2006. Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. J. Hydrol. 329 (1), 174–185. https://doi.org/10.1016/j.jhydrol.2006.02.013.

Chiew, F., Zhou, S., & Mcmahon, T. (2003). Use of seasonal streamflow forecasts in water resources management. *J. Hydrol. 270*((1-2)), 135–144. https://doi.org/10.1016/S0022-1694(02)00292-5.

Chuck, K., Joana, L., Brad, A., M., V. R. (2004). Developing a Watershed Characteristics Database to Improve Low Streamflow Prediction. *J. Hydrol. Eng. 9*(2), 116–125. 10.1061/(ASCE)1084-0699(2004)9:2(116).

Crochemore, L., Ramos, M.-H., Pappenberger, F., 2016. Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrol. Earth Syst. Sci. 20 (9), 3601–3618. https://doi.org/10.5194/hess-20-3601-2016.

Daly, C., Smith, J. W., Smith, J. I., & McKane, R. B. (2007). High-resolution spatial modeling of daily weather elements for a catchment in the Oregon cascade mountains, United States, J. Appl. Meteorol. Climatol. 46(10), 1565-1586. Retrieved Jul 6, 2021, from https://journals.ametsoc.org/view/journals/apme/46/10/jam2548.1.xml.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Pasteris, P. P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol. 28 (15), 2031–2064. https://doi.org/10.1002/joc.1688.

Devia, G.K., Ganasri, B.P., Dwarakish, G.S., 2015. A review on hydrological models. Aquat. Procedia 4, 1001–1007. https://doi.org/10.1016/j.aqpro.2015.02.126.

Donnelly-Makowecki, L.M., Moore, R.D., 1999. Hierarchical testing of three rainfall–runoff models in small forested catchments. J. Hydrol. 219 (3), 136–152. https://doi.org/10.1016/S0022-1694(99)00056-6.

El Hassan, A.A., Sharif, H.O., Jackson, T., Chintalapudi, S., 2013. Performance of a conceptual and physically based model in simulating the response of a semi-urbanized watershed in San Antonio. Texas. Hydrological Processes 27 (24), 3394–3408. https://doi.org/10.1002/hyp.9443.

Fang, K., Shen, C., Kifer, D., Yang, X. (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. Geophys. Res. Lett. 44(21), 11,11-30,39. 10.1002/2017GL075619.

Farmer, W.H., Over, T.M., Kiang, 2018. Bias correction of simulated historical daily streamflow at ungauged locations by using independently estimated flow duration curves. Hydrol. Earth Syst. Sci. 22, 5741–5758. https://doi.org/10.5194/hess-22-5741-2018.

Givati, A., Gochis, D., Rummler, T., Kunstmann, H., 2016. Comparing one-way and two-way coupled hydrometeorological forecasting systems for flood forecasting in the mediterranean region. Hydrology 3 (2), 19. https://doi.org/10.3390/hydrology3020019.

Gochis, D.J., M. Barlage, A. Dugger, K. FitzGerald, L. Karsten, M. McAllister, J. McCreight, J. Mills, A. RefieeiNasab, L. Read, K. Sampson, D. Yates, & W. Yu. (2018). The WRF-Hydro Modeling System Technical Description, (Version 5.0). NCAR Technical Note. 107 pages. Available online at https://ral.ucar.edu/sites/default/files/public/WRFHydroV5TechnicalDescription.pdf.

Gochis, D.J., Chen, F. (2003). Hydrological Enhancements to the Community Noah Land Surface Model (No. NCAR/TN-454+STR). University Corporation for Atmospheric Research. 10.5065/D60P0X00S.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. Retrieved from. http://www.deeplearningbook.org.

Graves, A. (2013) Generating Sequences With Recurrent Neural Networks. arXiv preprint arXiv:1308.0850. 1–43. Retrieved from https://arxiv.org/abs/1308.0850.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J. (2015). LSTM: Search Space Odyssey. *CoRR*, *abs/1503.0*(10), 2222–2232. Retrieved from http://arxiv.org/abs/1503.04069.

Hashino, T., Bradley, A.A., Schwartz, S.S., 2007. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. Hydrology and Earth System Sciences 11 (2), 939–950. https://doi.org/10.5194/hess-11-939-2007.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Konapala, G., Kao, S.C., Painter, S.L., Lu, D., 2020. Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environ. Res. Lett. 15 (10) https://doi.org/10.1088/1748-9326/aba927.

Kerandi, N., Arnault, J., Laux, P., Wagner, S., Kitheka, J., Kunstmann, H., 2018. Joint atmospheric-terrestrial water balances for East Africa: a WRF-Hydro case study for the upper Tana River basin. Theor. Appl. Climatol. 131 (3), 1337–1355. https://doi.org/10.1007/s00704-017-2050-8.

Kim, M.K., Lee, D.H., Kim, J., 2013. Production and validation of daily grid data with 1 km resolution in South Korea. J. Clim. Res. 8, 13–25 (In Korean with English abstract).

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. *CoRR*, *abs/1412.6*. Retrieved from https://arxiv.org/abs/1412.6980.

Koren, V., Schaake, J., Mitchell, K., Duan, Q.Y., Chen, F., Baker, J.M., 1999. A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. J. Geophys. Res. Atmos. 104 (D16), 19569–19585. https://doi.org/10.1029/1999JD900232.

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. Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for large-scale hydrological modeling. Hydrol. Earth Syst. Sci. Discuss. 1–32 https://doi.org/10.5194/hess-2019-368.

Lahmers, T.M., Gupta, H., Castro, C.L., Gochis, D.J., Yates, D., Dugger, A., Hazenberg, P., 2019. Enhancing the structure of the WRF-hydro hydrologic model for semiarid environments. J. Hydrometeorol. 20 (4), 691–714. https://doi.org/10.1175/JHM-D-18-0064.1.

Lane, R.A., Coxon, G., Freer, J.E., Wagener, T., Johnes, P.J., Bloomfield, J.P., Reaney, S. M., 2019. Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain. Hydrol. Earth Syst. Sci. 23 (10), 4011–4032. https://doi.org/10.5194/hess-23-4011-2019.

Lehner, B., Verdin, K.L., Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data. Eos, Transa. Am. Geophys. Union 89 (10), 93–94. https://doi.org/10.1029/2008EO100001.

Lipton, Z.C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning. *CoRR*, *abs/1506.00019*. Retrieved from http://arxiv.org/abs/1506.00019.

Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour. Res. 43 (7) https://doi.org/10.1029/2006WR005756.

Maidment, D.R., 2017. Conceptual framework for the national flood interoperability experiment. JAWRA J. Am. Water Resour. Assoc. 53 (2), 245–257. https://doi.org/10.1111/1752-1688.12474.

Messager, M.L., Lehner, B., Grill, G., Nedeva, I., Schmitt, O., 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nat. Commun. 7 (1), 13603. https://doi.org/10.1038/ncomms13603.

Masters, D., Luschi, C. (2018). Revisiting Small Batch Training for Deep Neural Networks. *CoRR*, *abs/1804.07612*. Retrieved from http://arxiv.org/abs/1804.07612.

Moradkhani, H., Sorooshian, S. (2008). General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis. In S. Sorooshian, K.-L. Hsu, E. Coppola, B. Tomassetti, M. Verdecchia, & G. Visconti (Eds.), *Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models* (pp. 1–24). 10.1007/978-3-540-77843-1_1.

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. *Trans. ASABE*, *50*(3), 885–900. 10.13031/2013.23153.

Naabil, E., Lamptey, B.L., Arnault, J., Olufayo, A., Kunstmann, H., 2017. Water resources management using the WRF-Hydro modelling system: Case-study of the Tono dam in West Africa. J. Hydrol.: Reg. Stud. 12, 196–209. https://doi.org/10.1016/j.ejrh.2017.05.010.

National Water Resources Management Information System 2003, Han River Flood Control Office, South Korea, accessed 1 October 2020, <wamis.go.kr>.

Olah, C., 2015. Understanding LSTM Networks. Retrieved from. http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Paniconi, C., Putti, M., 2015. Physically based modeling in catchment hydrology at 50: survey and outlook. Water Resour. Res. 51 (9), 7090–7129. https://doi.org/10.1002/2015WR017780.

Rasouli, K., Hsieh, W., Cannon, A., 2012. Daily streamflow forecasting by machine learning methods with weather and climate inputs. J. Hydrol. 414–415, 284–293. https://doi.org/10.1016/j.jhydrol.2011.10.039.

Read, J.S., Jia, X., Willard, J., Appling, A.P., Zwart, J.A., Oliver, S.K., Kumar, V., 2019. Process-guided deep learning predictions of lake water temperature. Water Resour. Res. 55 (11), 9173–9190. https://doi.org/10.1029/2019WR024922.

Sahoo, S., Russo, T.A., Elliott, J., Foster, I., 2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. Water Resour. Res. 53 (5), 3878–3895. https://doi.org/10.1002/2016WR019933.

Schumann, A.H., 1993. Development of conceptual semi-distributed hydrological models and estimation of their parameters with the aid of GIS. Hydrol. Sci. J. 38 (6), 519–528. https://doi.org/10.1080/02626669309492702.

Senatore, A., Mendicino, G., Gochis, D.J., Yu, W., Yates, D.N., Kunstmann, H., 2015. Fully coupled atmosphere-hydrology simulations for the central Mediterranean: Impact of enhanced hydrological parameterization for short and long time scales. J. Adv. Model. Earth Syst. 7 (4), 1693–1715. https://doi.org/10.1002/2015MS000510.

Shortridge, J.E., Guikema, S.D., Zaitchik, B.F., 2016. Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. Hydrol. Earth Syst. Sci. 20 (7), 2611–2628. https://doi.org/10.5194/hess-20-2611-2016.

Stathakis, D., 2009. How many hidden layers and nodes? Int. J. Remote Sens. 30 (8), 2133–2147. https://doi.org/10.1080/01431160802549278.

Vrugt, J.A., Diks, C.G.H., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling. Combin. Strengths Glob. Optimiz. Data Assimilation 41 (1), 1–17. https://doi.org/10.1029/2004WR003059.

Wedgbrow, C.S., Wilby, R.L., Fox, H.R., O'Hare, G., 2002. Prospects for seasonal forecasting of summer drought and low river flow anomalies in England and Wales. Int. J. Climatol. 22 (2), 219–236. https://doi.org/10.1002/joc.735.

Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. J. Big Data 3 (1), 9. https://doi.org/10.1186/s40537-016-0043-6.

Xiang, T., Vivoni, E.R., Gochis, D.J., Mascaro, G., 2017. On the diurnal cycle of surface energy fluxes in the North American monsoon region using the WRF-Hydro model. J. Geophys. Res.: Atmos. 122 (17), 9024–9049. https://doi.org/10.1002/2017JD026472.

Xiang, Z., Yan, J., Demir, I. (2020). A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resour. Res. 56*(1), e2019WR025326. 10.1029/2019WR025326.

Xu, T., Valocchi, A.J., Choi, J., Amir, E., 2014. Use of machine learning methods to reduce predictive error of groundwater models. Groundwater 52 (3), 448–460. https://doi.org/10.1111/gwat.12061.

Xu, T., Valocchi, A.J., 2015. Data-driven methods to improve baseflow prediction of a regional groundwater model. Comput. Geosci. 85, 124–136. https://doi.org/10.1016/j.cageo.2015.05.016.

Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., Zhao, B., 2020. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. J. Hydrol. 590 (March), 125206 https://doi.org/10.1016/j.jhydrol.2020.125206.

Yuan, X., Chen, C., Lei, X., Yuan, Y., Muhammad Adnan, R., 2018. Monthly runoff forecasting based on LSTM–ALO model. Stoch. Env. Res. Risk Assess. 32 (8), 2199–2212. https://doi.org/10.1007/s00477-018-1560-y.

Yucel, I., Onen, A., Yilmaz, K., Gochis, D., 2015. Calibration and evaluation of a flood forecasting system: utility of numerical weather prediction model, data assimilation and satellite-based rainfall. J. Hydrol. 523 https://doi.org/10.1016/j.jhydrol.2015.01.042.