WaterBench-Iowa: A Large-scale Benchmark Dataset for Data-Driven Streamflow Forecasting

Ibrahim Demir[1,2], Zhongrun Xiang[1], Bekir Demiray[3], Muhammed Sit[3]

[1] Department of Civil and Environmental Engineering, University of Iowa, Iowa City, 52246 USA
[2] Department of Electrical and Computer Engineering, University of Iowa, Iowa City, 52246 USA
[3] Interdisciplinary Graduate Program in Informatics, University of Iowa, Iowa City, 52246 USA

Correspondence to: Zhongrun Xiang (zhongrun-xiang@uiowa.edu)

Abstract. This study proposes a comprehensive benchmark dataset for streamflow forecasting, WaterBench-Iowa, that follows FAIR data principles that is prepared with a focus on convenience for utilizing in data-driven and machine learning studies, and provides benchmark performance for state-of-art deep learning architectures on the dataset for comparative analysis. By aggregating the datasets of streamflow, precipitation, watershed area, slope, soil types, and evapotranspiration from federal agencies and state organizations (i.e., NASA, NOAA, USGS, and Iowa Flood Center), we provided the WaterBench-Iowa for hourly streamflow forecast studies. This dataset has a high temporal and spatial resolution with rich metadata and relational information, which can be used for a variety of deep learning and machine learning research. We defined a sample benchmark task of predicting the hourly streamflow for the next five days for future comparative studies, and provided benchmark results on this task with sample linear regression and deep learning models, including Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and S2S (Sequence-to-sequence). Our benchmark model results show a median NSE of 0.74 and a median KGE of 0.79 among 125 watersheds for the 120-hr ahead streamflow prediction task. WaterBench-Iowa makes up for the lack of unified benchmarks in earth science research and can be accessed at Zenodo https://doi.org/10.5281/zenodo.7087806.

1 Introduction

Deep learning, a set of artificial neural networks (ANN) based algorithms for supervised and unsupervised modeling, has been widely used and recognized as a powerful approach within many scientific disciplines for technological and predictive progress (Goodfellow et al., 2016). As conventional machine learning techniques were deemed limited in learning the representations of high-dimensional datasets from their raw form, by providing universal approximator models (Cybenko, 1989; Hornik et al., 1989; Leshno et al., 1993), deep neural networks increased scientists' ability to model both linear and non-linear problems without time-intensive data engineering processes by domain experts (LeCun et al., 2015). Deep learning's predictive modeling capabilities have led to improvements in various fields, including image recognition and synthesis (Demiray et al., 2021), speech recognition, language modeling, and time-series prediction.

Flooding is a significant concern for many areas in the world as it is on an upward trend due to climate change. The 1998 Bangladesh flood, the Iowa flood of 2008, and the 2013 North India floods show how catastrophic and both economically and

psychologically devastating floods can be for populations in respective regions. In order to maximize the preparedness for floods and minimize their effects after the disaster (Yildirim and Demir, 2021), weather and flood forecasting stands as a perennial research interest for hydrologists and data scientists. Streamflow prediction and runoff modeling are research efforts where the water from the land or channel over time is being modeled and forecasted using previous data points for a location or nearby locations with similar characteristics. Although this effort is conventionally carried out with physically based models that require extensive computational (Agliamzanov et al., 2020) and data resources, it is critical for flood mitigation and decision support (Xu et al., 2020).

Being a time-series prediction task, in essence, flood forecasting takes advantage of the practicality and efficacy that deep learning brings to predictive modeling. Both time-series adaptations of deep learning models intended for natural language processing, and time-series focused deep neural network implementations make this possible by proposing methodologies that put the sequential nature of time-series datasets into good use. Recurrent neural network (RNN) architectures such as Long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and Gated recurrent unit (GRU) networks (Chung et al., 2014), and Attention based sequence-to-sequence networks (Vaswani et al., 2017) are pronounced starting point for deep neural network architectures for most time-series forecasting tasks.

Supervised learning, whether it be deep or not, is the most common form of machine learning (LeCun et al., 2015), and supervised learning tasks, such as flood forecasting, need a dataset of previously recorded or labeled entries for the task. That dataset typically consists of X and y values where X values are the input that the model expects, and y values are the output values the model returns. A supervised learning model is trained using a loss function that measures the similarity or difference of the y values from the dataset (actual ys) and the outputs of the model (predicted ys). During a typical training process, predicted ys get closer to the actual ys in time, hence the name training. As a quintessential part of any supervised learning task, training neural network models on established datasets is common among deep learning practitioners and researchers (Goodfellow et al., 2016). For most tasks that deep learning researchers tackle today, there are vast amounts of benchmark datasets available freely for research. While computer vision datasets such as Imagenet (Deng et al., 2009), Ms-celeb-1m (Guo et al., 2016), Adobe-240fps (Su et al., 2017), and Vimeo-90K (Xue et al., 2019) and similarly time-series datasets namely, automobile parts demand dataset Parts (Seeger et al., 2016), electricity and traffic (Yu et al., 2016) have been widely used to test proposed neural network architectures, to the best of our knowledge. There are not many specific datasets that are published for geoscience studies (Ebert-Uphoff et al., 2017) and specifically for flood and streamflow forecasting.

The number of studies in hydrology and water resources, and particularly in flood forecasting that employ deep learning, has been gaining interest in the last several years (Sit et al., 2020). Flood forecasting studies in the literature, due to the aforementioned sequential nature, have vastly employed RNNs and LSTMs. Kratzert et al. (2018) utilize LSTM networks for daily runoff prediction using meteorological datasets. Furthermore, Kratzert et al. (2019) applied a similar approach for ungauged US locations. Bai et al. (2019) incorporate a stack autoencoder with LSTM for daily streamflow measurements from data for a week. Xiang et al. (2020) predict the next 24-hours of hourly streamflow rate by utilizing an encoder-decoder sequence-to-sequence neural network that also uses rainfall products. Xiang and Demir (2020), moreover, extend their study

and develop a model that forecasts the hourly streamflow rate for the next five days using three days of historic data. They also incorporate upstream sensors into their proposed network. Using the same dataset, Xiang et al. (2021), explore the generalization of sequence-to-sequence encoder-decoder networks in flood forecasting. Sit and Demir (2019) predict hourly sensor measurements for 24 hours using data from the upstream sensor network and historic stage height measurements. And finally Sit et al. (2021a), utilize graph neural networks for streamflow forecasting for a small watershed in Iowa. To sum up, deep learning models such as LSTM have been used in meteorology and hydrology studies of soil moisture modeling (Fang et al., 2017), water table depth prediction (Zhang et al., 2018), rainfall-runoff modeling (Hu et al., 2018; Kratzert et al., 2018), streamflow forecasting (Xiang et al., 2020), etc. As is presented by perspective studies (Reichstein et al. 2019), deep learning models such as LSTM can extract spatial-temporal features automatically to gain further process understanding of Earth system science problems. Therefore, we pay great attention to the application of LSTM and its variant models in this research.

Most of the studies mentioned here acquire several raw data products, whether in terms of rainfall measurements, physical features of the studied area, or stage height/discharge measurements, from authorities and build their own dataset benefiting from their expertise in the area. There are several datasets and benchmarks in other earth science studies, i.e., air quality forecast dataset, 3D cloud detection dataset, and LANL earthquake prediction dataset. One of the early user-friendly datasets in earth science is the Beijing PM2.5 Data. It was published in 2017, and it includes the hourly air quality PM2.5 data from the U.S. Embassy in Beijing and meteorological data from Beijing Capital International Airport. After the dataset have released, researchers developed different novel machine learning and deep learning models, including support vector machines (Zhu et al., 2018, Liu et al., 2019), recurrent neural networks (Athira et al., 2018), attention-based LSTM (Li et al., 2019), interpretable deep learning (Guo et al., 2018), hybrid deep learning (Du et al., 2018), convolutional networks (Tao et al., 2019), and stacked LSTM (Sagheer and Kotb, 2019) on this specific dataset. While knowledge of the application domain is essential to find scientifically robust ways to prepare the input data and to interpret the results of machine learning models, such knowledge is not always accessible to deep learning experts. If there are well-defined benchmark datasets with a clear description of the machine learning task to solve and have well-defined and domain-science informed evaluation metrics, then it becomes possible for non-domain experts to solve such challenges and introduce novel machine learning methods to the field. Furthermore, these papers used the same dataset, and therefore, the results are comparable. Thus, scientists could focus more on modeling and improving on the basis of existing papers rather than collecting their own datasets. A benchmark in hydrology will no doubt enhance the application and development speed of deep learning studies in the water resources field. For improved generic deep learning-based flood forecasting models, scientists must expand on previous work, and this can be done with the same testing setup and evaluation mechanism. There are some studies in the literature of hydrology in limited numbers that construct their neural network architecture around the CAMELS dataset (Newman et al., 2014). CAMELS is a vast dataset that includes meteorological and observed streamflow data points for the United States, albeit not in an easy-to-use and ideal format for deep learning research. It contains 671 catchments in the contiguous US that are minimally impacted by human activities. It includes features such as topography, climate, streamflow, land cover, soil, and geology on a watershed scale, and the hydrometeorological time-series data ranges from 1980 to 2014 on a daily basis. The data is generated from

different sources, including Daymet, NLDAS, and Maurer. CAMELS aggregated these datasets at the watershed level. The researchers also did the model simulation using physically-based models such as the NWS model, and SNOW-17/SAC-SMA; however these modeling results are not shared as a benchmark. Even though there is a dataset that could be used for predictive deep learning rainfall-runoff modeling, there is still a lack of accessible datasets for benchmarking purposes (Masley et al., 2020). There remains a need for a dataset that is more convenient to use in deep learning research given that most of the deep learning researchers are not domain experts. The limited usage of CAMELS in the literature also predicates the challenges the CAMELS dataset presents for deep learning research.

Another dataset for flood forecasting is FlowDB (Godfried et al., 2020). Unlike CAMELS, there are not many studies that report their performance over FlowDB yet as the dataset is recently published. FlowDB is an hourly precipitation and river flow dataset that also includes a subset dataset for flash floods. The subset dataset includes injury costs and damage estimations for flash flood events. FlowDB gathers river flow data from the USGS and precipitation data from many agencies, including the USGS, NOAA, and ASOS. Additionally, the data FlowDB provides regarding flash floods uses NSSL Flash by NOAA.

This study proposes a flood forecasting dataset that is prepared with a focus on convenience for utilizing in data-driven and machine learning studies and provides benchmark performance for state-of-art deep learning architectures on the dataset for comparative analysis. Our dataset follows FAIR data principles (Wilkinson et al., 2016), which means it is findable and accessible through DOI, and the data is richly described with references. WaterBench provides data from 125 catchments in the state of Iowa. The precipitation time-series data ranges from October 2011 to September 2018 along with catchment-based features such as topography, soil type, and slopes. Even though the dataset was designed in a way to eliminate most of the preprocessing and data engineering tasks out of the way for machine learning applications and research, it could be used in other studies with similar goals, such as physically based modeling. Similarly, the dataset could be used by combining it with other benchmark datasets such as IowaRain (Sit et al., 2021b) utilizing cloud-based rainfall products (Seo et al., 2019). WaterBench is different from CAMELS with a higher temporal resolution. In addition, it focuses on the state of Iowa, and many large catchments in WaterBench contain multiple USGS gauges, which helps to better represent the river structure, and upstream-downstream relations in deep learning algorithms. The rest of this paper is structured as follows; the dataset preparation phase and methodology employed in that phase are discussed in section 2. Section 3 gives a list of tasks that could be tackled using this dataset and presents the performance of several neural network implementations in flood forecasting tasks. In the last section, conclusions are discussed.

2 Methodology and Dataset

2.1. Study Area

The State of Iowa is located in the Midwest of the United States. It has abundant and diversified water resources with 71,655 miles of rivers and streams from border to border (Iowa DNR, 2004). In 2008, the Eastern Iowa was devastated with flooding which caused over $6 billion in property losses. Streamflow monitoring and forecasting are consequently critical for Iowa for

better water resources and disaster management. In addition, agricultural-based activities in Iowa have a low pavement rate with limited human influence, which makes it a suitable area for rainfall-runoff studies.

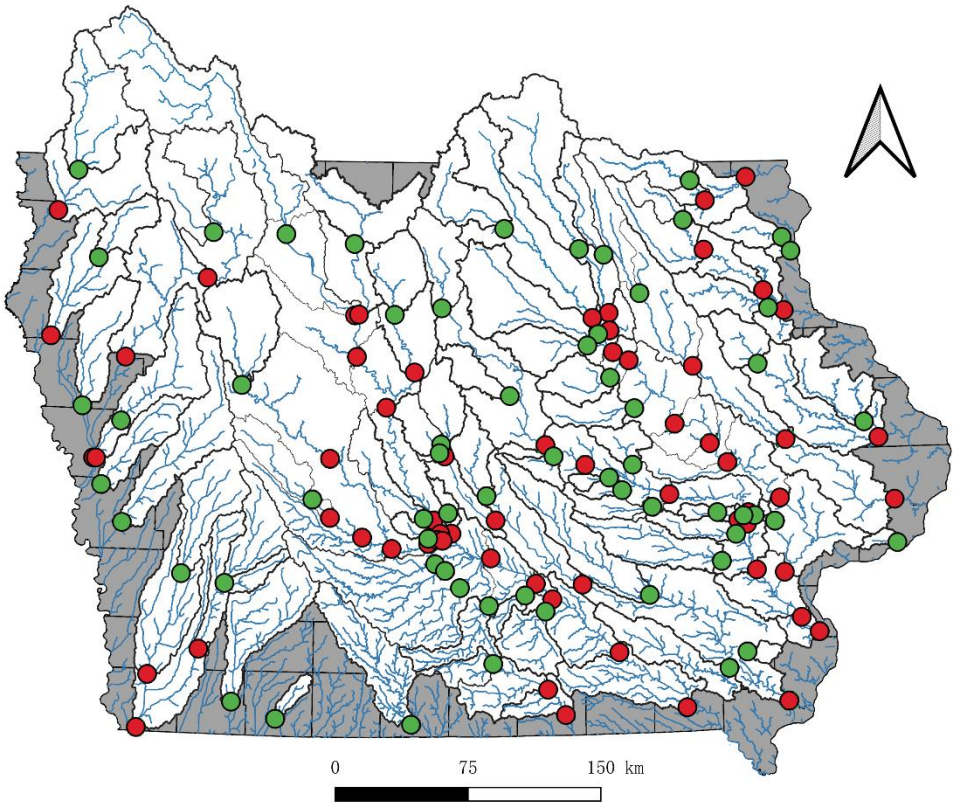

Figure 1. The location of 125 USGS gauges in the State of Iowa for upstream sub-basins (green dot) and large downstream basins (red dot).

The United States Geological Survey (USGS) has over a hundred streamflow gauges in the state of Iowa for monitoring the streamflow rate in different streams. The measurements from the USGS are typically recorded at 15- to 60- minute intervals in Iowa. Due to the site maintenance or shutdowns, the coverage of the USGS streamflow gauges changes over the years. In this dataset, we selected all USGS gauges in the State of Iowa with available data from October 1st, 2011 (the water year 2012) to September 30th, 2018 (the water year 2018).

As shown in Figure 1, red dots are located at the outlets of larger basins with multiple USGS gauges, which are divided into several smaller upstream sub-basins. The green dots are located at the outlets of the most upstream sub-basins. Thus, considering the connectivity of the streams, the relationship of these gauges in one watershed can be represented as a tree structure.

2.2. Dataset Features

WaterBench includes detailed metadata and time-series features for each catchment. These datasets are available in .csv format for each catchment. The details of the datasets with data source, type, resolution and units are shown in Table 1. The statistics of the data, including the watershed size, concentration-time (the longest streamflow path in the catchment), slope, and four soil types, are shown in Table 2 and Figure 2. For each catchment, we provide static data (area, slope, travel time, etc.) as well as time-series for streamflow, precipitation, and ET.

Table 1. The details of datasets with data source, type, resolution and units

Datasets	**Data Type**	**Sources**	**Resolution**		**Unit**
			Spatial	**Temporal**	
Area	GIS shapefile	IFC (Krajewski et al., 2017)	Station based	constant	km^2
Slope	Hillslope data		Hillslope based	constant	%
Travel time	Reach shapefile		Station based	constant	hour
ET	Estimation from historical data		State based	monthly	mm / month
Soil types	Soil data	NASA (Post et al., 2000)	0.5-degree grid	constant	%
Streamflow Rate	USGS gage measurement	USGS	Station based	15-60 mins	ft^3/s
Precipitation	Stage IV multi-sensor measurement	NOAA (Lin, 2011)	4km grid	hourly	mm/hr

Table 2. The minimum, maximum, median, and standard deviation (SD) of the watershed area, concentration time, average slope, and percentage of soil types including loam, silt, sandy clay loam, and silty clay loam among 125 USGS gauges in the State of Iowa.

	Area (km^2)	**Concentration Time (hr)**	**Slope**	**Loam**	**Silt**	**Sandy clay loam**	**Silty clay loam**
Min	6	2	0.38%	0%	0%	0%	0%
Max	36,453	315	4.32%	98%	100%	84%	93%
Mean	5,405	77	1.97%	33%	31%	18%	18%
Median	1,918	53	1.80%	33%	21%	4%	7%
SD	8,320	68	0.80%	28%	30%	24%	23%

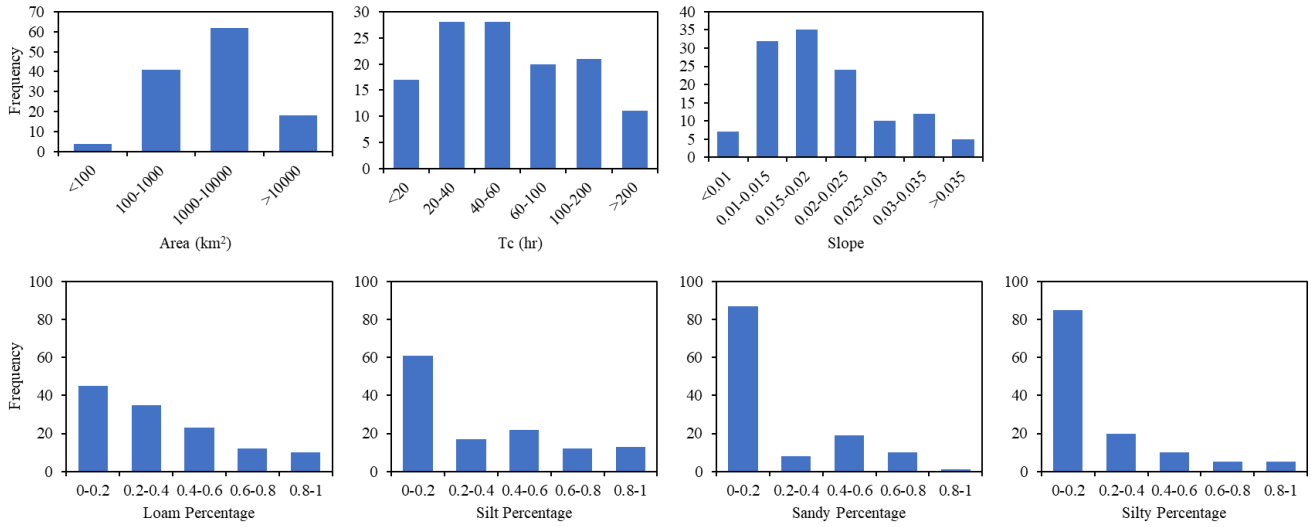

Figure 2. Histograms of the catchment area (a), concentration time (b), average slope (c), and percentage of soil types including loam (d), silt (e), sandy clay loam (f), and silty clay loam (g) for 125 USGS gauges in the State of Iowa.

Table 3. Summary statistics for precipitation and streamflow among 125 catchments from water year 2012 to 2018. Missing rate as a limitation.

	Annual Total Precipitation (mm)	Max. Hourly Precipitation (mm)	Annual Mean Streamflow (m3/s)	Missing Rate of Precipitation (Raw Data)	Missing Rate of Streamflow (Raw Data)
Min	794	9.1	3	0.02%	0.69%
Max	1,056	60.0	12,963	0.04%	33.14%
Mean	952	24.8	1,926	0.02%	15.16%
Median	961	22.2	608	0.02%	16.14%
SD	57	10.3	2,864	0.01%	6.4%

As it is shown in Table 3, all 125 catchments share similar precipitation ranges from 794 to 1056, with a small standard deviation of 57. Geologically, all the catchments are located in two HUC watersheds, the Upper Mississippi and Missouri, and the study results may not be applicable to other regions in the U.S. However, the modeling algorithms and the neural network architectures normally apply to a broad spectrum of problems, and they would be useful in other regions. WaterBench-Iowa is also subject to a relatively high missing data rate for streamflow since the reliable hourly dataset is limited by the USGS for some of the watersheds in Iowa. In the following sections, we will discuss the details of specific datasets and features.

2.2.1 Area

In the water cycle, precipitation is the main driving force of the streamflow. Based on the 90m digital elevation model (DEM), only the precipitation in a certain area will contribute to a stream. Each measuring station has its corresponding area, which can be calculated from the watershed boundary shapefiles. Since the total precipitation amount is the product of precipitation intensity and area, in the same watersheds, upstream sub-basins typically have lower streamflow rates than the larger basins. In WaterBench, the boundary shapefiles of each watershed are obtained from the Iowa Flood Information System (IFIS), a system operated by the Iowa Flood Center (IFC). Moreover, the area is calculated from the shapefiles in the unit of square kilometers. Thus, the area contains one value per station, and it is available in the column of "area" in the "{station_id}_data.csv" files.

2.2.2 Time of Concentration

The time of concentration provides the dimension of stream length for a watershed. In WaterBench, the time of concentration is defined as the longest length over the velocity, which is the time the water concentrates from the most distant point from the watershed outlet. The velocity used in this study is a constant value of 0.75 m/s, which was found appropriate for Iowa catchments (Mandapaka et al., 2009; Mantilla et al., 2011), and has been successfully used in many hydrologic models (Fonley et al., 2016; Sloan et al., 2017). Thus, for a long and narrow watershed, it may have a small watershed area but a large time of concentration. In WaterBench, the time of concentration is obtained from the IFIS with the unit of hours. Thus, the time of concentration contains one value per station, and it is available in the column of "travel_time" in the "{station_id}_data.csv" files.

2.2.3 Slope

The slope is one of the topographic features that represents the slope gradient in percentage. A steep slope may cause a higher velocity and lower infiltration rate, which normally causes a larger streamflow rate during a precipitation event. The original file, hillslope map, is calculated by IFC (Sit et al., 2019), which split the land of Iowa into over 600,000 hydrologic units using the algorithm developed by Mantilla and Gupta (2005). In WaterBench, the average slope is calculated from the mean value of the hillslopes in each catchment (Gericke and Du, 2012). Thus, the slope is a constant value per watershed, and it is available in the column of "slope" in the "{station_id}_data.csv" files.

2.2.4 Soil Type

Soil type is one of the topographic features that represents the proportions of 12 different soil types on the land. Normally, the sandy soil has the largest infiltration rate, and the clay has the least infiltration rate. The original file, global soil types, is available from NASA (Post et al., 2000). It is a 2-D map with a spatial resolution of 0.5 degrees. The soil type proportion is then calculated using the weighted average for each watershed. It needs attention that four dominant soil types, including the

loam, silt, sandy clay loam, and silty clay loam, contribute to 99.91% of the area in Iowa. Thus, only these four soil types are considered in the dataset. The percentage of each soil type is constant in the time series dataset for each station in the columns of "loam", "silt", "sandy_clay_loam", and "silty_clay_loam" in the "{station_id}_data.csv" files.

2.2.5 Streamflow Rate

The streamflow rate is a variable measured by the USGS in the unit of cubic feet per second. The data was acquired from the USGS National Water Information System. There are nearly 200 real-time streamflow measuring stations in Iowa. After removing the stations established after 2011 or permanently closed before 2018, a total of 125 stations are selected, as shown in Figure 1. For each station, streamflow data was aggregated to hourly values. The original data contains a few missing values due to station system failures or internet outages. For the stations located in the northern part of Iowa, the river may freeze and have no flow rate measurement over the winter, and all missing values were reported as -9999 by the USGS. In the dataset, each watershed has two columns, with the first column representing the timestamp from 2011/10/01 00:00 to 2018/9/30 23:00, and the second column representing the the streamflow values. Thus, the streamflow rate contains 61,368 values per station, and they are available in the column of "discharge" in the "{station_id}_data.csv" files.

2.2.6 Precipitation Volume

Many station-based and satellite datasets have been measuring precipitation over the years. After comparisons, it is found that NOAA's Stage IV multi-sensor measurement is the most accurate (Seo et al., 2018) in the state of Iowa. The Stage IV multi-sensor provides the hourly precipitation amount with a 4km-grid spatial resolution. The catchment level average precipitation is then calculated at each hour. Since there is no rainfall or snowfall most of the time, most precipitation values in the dataset are 0. In the dataset, we provide the hourly catchment-averaged precipitation data for each station from 2011/10/01 00:00 to 2018/9/30 23:00. Thus, the precipitation data contains 61,368 values per station, and they are available in the column of "precipitation" in the "{station_id}_data.csv" files.

2.2.7 Evapotranspiration (ET)

ET represents the evaporation and plant transpiration from the land in the water cycle. It is one of the major losses of precipitated water. Since there is no high-resolution real-time ET dataset available, we used the monthly estimation from the historical measurement data in the past decades (Krajewski et al., 2017) as an empirical dataset. This is a monthly-based dataset for the entire state of Iowa, and successfully captures the seasonal effects in the state of Iowa. In the dataset, we applied the ET value for each timestamp from 2011/10/01 00:00 to 2018/9/30 23:00. Thus, the ET data contains 61,368 values for all stations, and they are available in the column of "et" in the "{station_id}_data.csv" files.

2.2.8 Watershed Relationship

Since many USGS measurement gauges are in the same watershed, many catchments in WaterBench-Iowa are not independent, and a relationship tree is given in the "catchment_relationship.csv". The csv file represents a disconnected directed graph with each row representing an edge. 63 out of 125 catchments have one or more upstream, as shown in the relationship, which are relatively large catchments. The remaining 62 catchments are specified as the very upstream catchments which have only one stream gage. Since these catchments have no overlapping area, the catchments in our dataset form a disconnected graph. Since the catchments have overlapping areas, the watershed ID 646 has the largest connected subgraph with 27 upstream catchments. With upstream-downstream relationships, WaterBench-Iowa supports the cutting-edge studies such as graph neural networks.

3. Benchmark Tasks and Metrics

In this section, we define a sample benchmark task of predicting the hourly streamflow for the next five days for future comparative studies. This task forecasts the future hourly floods at each hour as National Water Model does. At each hour t, we predict the streamflow for the next 5 days from hour t+1 to t+120 using all the data we can obtain at time t. In this task, we ignore the errors in the rainfall forecast, and use all the data including the topology data, past three days' precipitation and streamflow data, and the future five days' precipitation data as input, to predict the streamflow for the next 120 hours at the watershed outlet. Thus, we made 5-day predictions at each hour in the training and test datasets, and evaluated the results on different lead times from hour 1 to hour 120. This task is a typical regression modeling of time series data. Therefore, we suggest the traditional Ridge regression model and three deep learning models for modeling in this benchmark. Please refer to the recent studies for the detailed model structures such as LSTM (Kratzert et al., 2018), GRU (Gao et al., 2020) and S2S (Xiang & Demir, 2020).

We take two separate approaches to tackle this problem. The first approach involves a separate deep learning model for each of the available watersheds, while the second one is to build a single large regional model that carries out the same task for all available watersheds. For this specific task, we selected the last water year as the test set, and the rest as the training set. We further formatted the original dataset into a ready-to-use structure for each watershed with four files named as train_x, train_y, test_x, test_y. Thus, a total of 500 files for 125 watersheds are provided for this specific task. Since general statistics such as mean squared error (MSE) and root mean squared error (RMSE) are not dimensionless, the metrics for this study are Nash-Sutcliffe efficiency (NSE) and Kling-Gupta efficiency (KGE). They are both dimensionless statistics that are widely used in hydrological studies, and can be used to compare between watersheds. Both NSE and KGE range from negative infinity to 1, and the closer to 1 the better. The equations 1 and 2 for NSE and KGE are shown below:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \mu_Y)^2} \qquad \text{Eq. 1}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\frac{\sigma_{\hat{Y}}}{\sigma_Y} - 1)^2 + (\frac{\mu_{\hat{Y}}}{\mu_Y} - 1)^2} \qquad \text{Eq. 2}$$

where: Y_i is the observation at the time i; $\hat{Y}_i$ is the model result at the time i; n is the total number of observations; r is the Pearson correlation coefficient; σ is the standard deviation; μ is the mean; σ_Y is the standard deviation of all the observations; $\sigma_{\hat{Y}}$ is the standard deviation of model forecasts; μ_Y is the mean of all the observations; and $\mu_{\hat{Y}}$ is the mean of all model forecasts. Both NSE and KGE are dimensionless and in the range of (-∞,1]. For both metrics, the closer to 1, better the model performs. We calculate the NSE and KGE based on the test year for each prediction hour. Since we predict the streamflow for the next 120 hours at each hour, there will be 120 different NSE and KGE values for different hours at each watershed for the lead time from 1 to 120 hours. It should be noted that since the watersheds here are not filtered, it is possible for some watersheds to be greatly affected by human activities, including mitigation, construction, irrigation, urban drainage, etc. activities in watersheds. Thus, a median value of all 125 watersheds is meaningful to report as a widely employed practice within other hydrology studies (Kratzert et al., 2018, Xiang et al., 2020). In addition, since the prediction accuracy typically decreases when the lead time increases, the median NSE and KGE of 125 stations at the 120-hr ahead predictions are the lowest. Thus, the 120-hr ahead prediction scores are the most important metric that can represent the overall model performance on this task.

4. Benchmark Results and Discussion

To provide baseline results over the sample benchmark task and two approaches defined in the previous section, we employed a linear regression model using Ridge regression, and three deep learning models using LSTM, GRU, and sequence-to-sequence (S2S) network architectures. For the first approach, we considered each watershed independent and trained one model for each watershed. Thus, the relationship between the watersheds is not used in this benchmark. The median NSE and KGE scores among 125 watersheds at each hour are shown in Figure 3 and Table 4. As shown in the figure and the table, the Ridge regression has a high accuracy in the first 24 hours since the streamflow rates normally do not change too much in one day, and they are relatively easy to predict. The metrics for the medium-range show that the model using GRU has the best performance. The NSE and KGE histograms of GRU show that for most of the watersheds the GRU model performs well and only in a limited number of watersheds the GRU model give negative scores. The standard deviations show a relatively stable results in all prediction hours using deep learning models. However, the Ridge model shows higher standard deviations and lower model performance than deep learning models over 48 hours.

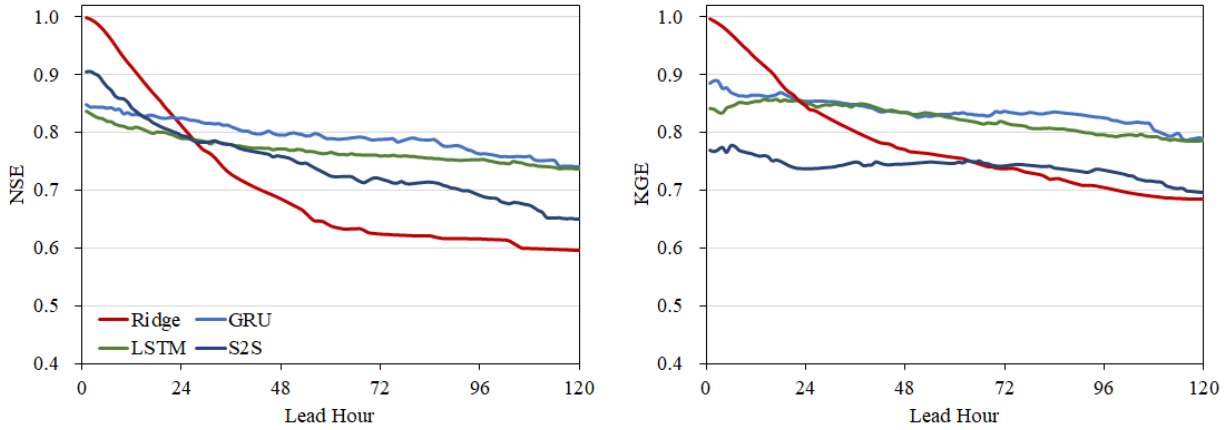

Figure 3. The median NSE and KGE among 125 watersheds in 125 different models at the prediction of the next 1 to 120 hours.

Table 4. The median (standard deviation) NSE and KGE among 125 watersheds at the prediction hours 1, 6, 12, 24, 48, 72, 96, and 120 in 125 different models.

			NSE			KGE		
Hour	**Ridge**	**GRU**	**LSTM**	**S2S**	**Ridge**	**GRU**	**LSTM**	**S2S**
1	**1(0.05)**	0.85(0.49)	0.84(0.72)	0.91(0.2)	**1(0.04)**	0.88(0.34)	0.84(0.34)	0.77(0.21)
6	**0.97(0.56)**	0.84(0.47)	0.82(0.7)	0.88(0.41)	**0.97(0.23)**	0.87(0.32)	0.84(0.33)	0.78(0.25)
12	**0.91(1.35)**	0.83(0.48)	0.81(0.66)	0.84(0.57)	**0.93(0.4)**	0.86(0.31)	0.85(0.32)	0.76(0.29)
24	0.81(2.44)	**0.83(0.47)**	0.79(0.61)	0.79(0.72)	0.85(0.58)	**0.85(0.3)**	0.85(0.3)	0.74(0.34)
48	0.69(2.91)	**0.8(0.46)**	0.77(0.59)	0.76(0.9)	0.77(0.66)	**0.83(0.28)**	0.83(0.29)	0.75(0.38)
72	0.62(2.89)	**0.79(0.45)**	0.76(0.65)	0.72(0.91)	0.74(0.66)	**0.84(0.28)**	0.82(0.3)	0.74(0.41)
96	0.62(2.7)	**0.76(0.43)**	0.75(0.56)	0.69(0.95)	0.71(0.63)	**0.82(0.28)**	0.8(0.29)	0.74(0.42)
120	0.6(2.6)	**0.74(0.43)**	0.74(0.51)	0.65(0.93)	0.69(0.62)	**0.79(0.28)**	0.79(0.3)	0.7(0.41)

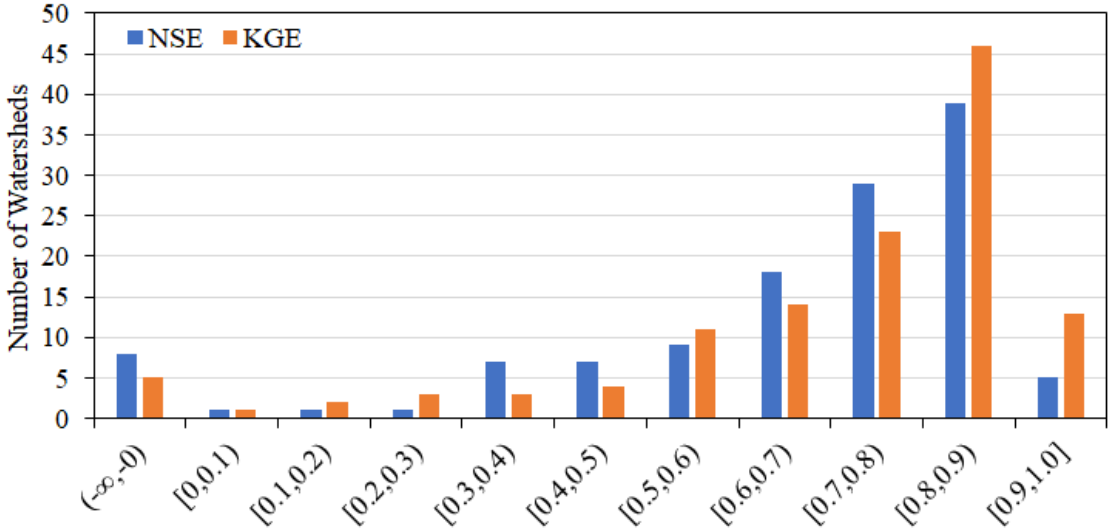

Figure 4. Histogram of the GRU model performance.

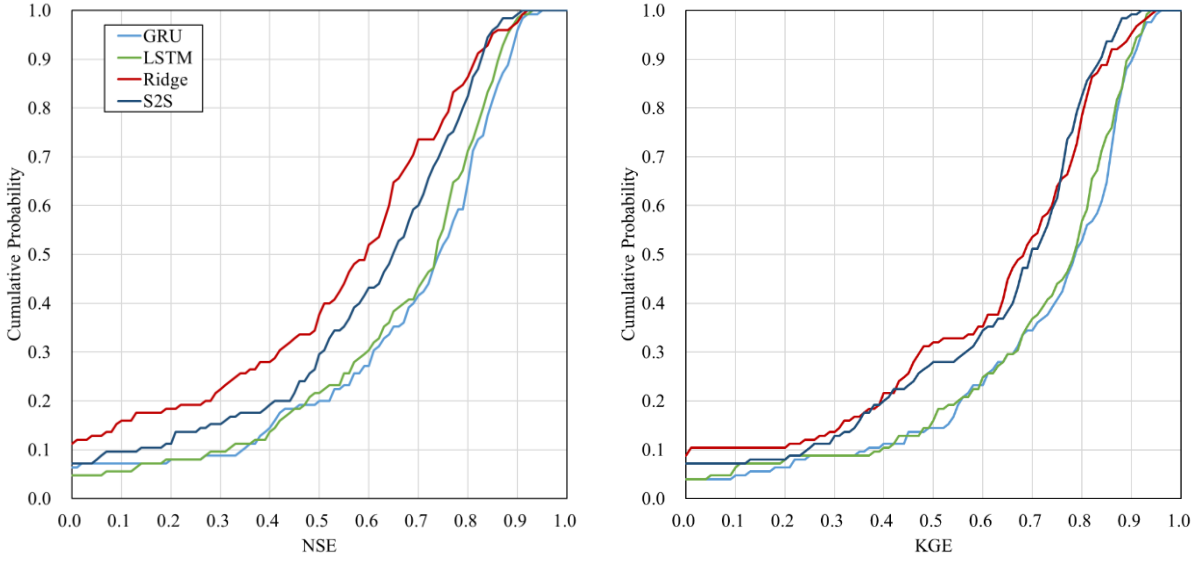

Figure 5. Cumulative probability curve of the NSE and KGE at the 120hr ahead predictions.

Figure 5 shows the cumulative distribution of the NSE and KGE among the 125 catchments at the lead time of 120 hours in addition to the median value for all 125 catchments. The results suggest that there is a large standard deviation between catchments, and that negative NSE and KGE values occur in 10% of the catchments. These catchments with negative NSE or KGE values are small (Figure 7), so it is very challenging to predict the streamflow over five days.

As for the second approach, we attempted to develop single regional models for all 125 watersheds since they share similar physical attributes. As shown in Figure 6, a single model on all 125 watersheds is possible with the physical features

including area, slope, travel time, and soil types using the customized NSE loss function (Xiang et al., 2021). Among four models, similar to the first approach, the performance of Ridge regression is hard to beat at first. Nevertheless, the deep learning model S2S starts to show a better performance starting the second day. Table 5 shows the detailed results of the regional model. Regional modeling using deep learning is more difficult as seen by the decline in model performance and greater standard deviations compared to the basin modeling results in Table 4.

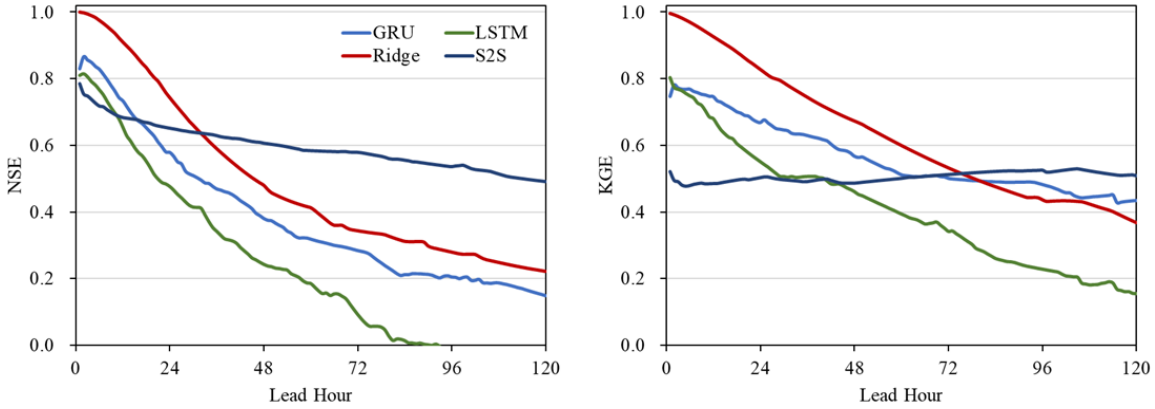

Figure 6. The median NSE and KGE among 125 watersheds using one regional model at the prediction of the next 1 to 120 hours.

Table 5. The median (standard deviation) NSE and KGE among 125 watersheds at the prediction hour 1, 6, 12, 24, 48, 72, 96, and 120 using one regional model.

	NSE				KGE			
Hour	**Ridge**	**GRU**	**LSTM**	**S2S**	**Ridge**	**GRU**	**LSTM**	**S2S**
1	**1(0.1)**	0.83(0.4)	0.81(0.48)	0.79(216.13)	**1(0.1)**	0.75(0.25)	0.8(0.39)	0.52(27.69)
6	**0.97(3.99)**	0.83(0.45)	0.77(0.6)	0.72(155.71)	**0.97(0.8)**	0.77(0.24)	0.75(0.4)	0.48(23.37)
12	**0.91(16.86)**	0.73(0.45)	0.65(0.79)	0.68(160.17)	**0.93(1.84)**	0.75(0.23)	0.68(0.42)	0.49(23.58)
24	**0.74(62.36)**	0.58(0.47)	0.48(1.37)	0.65(165.67)	**0.83(4.01)**	0.67(0.25)	0.55(0.46)	0.5(23.82)
48	0.48(189.98)	0.38(0.71)	0.24(2.91)	**0.61(167.32)**	**0.67(7.84)**	0.57(0.32)	0.46(0.6)	0.49(23.25)
72	0.34(320.38)	0.29(0.88)	0.09(4.83)	**0.58(171.55)**	**0.53(11.16)**	0.5(0.36)	0.34(0.76)	0.51(23.11)
96	0.28(448.45)	0.21(1.01)	-0.04(6.83)	**0.54(173.72)**	0.44(14.2)	0.48(0.39)	0.23(0.91)	**0.53(23.17)**
120	0.22(568)	0.15(1.23)	-0.24(8.91)	**0.49(178.42)**	0.37(16.81)	0.43(0.45)	0.16(1.05)	**0.51(23.22)**

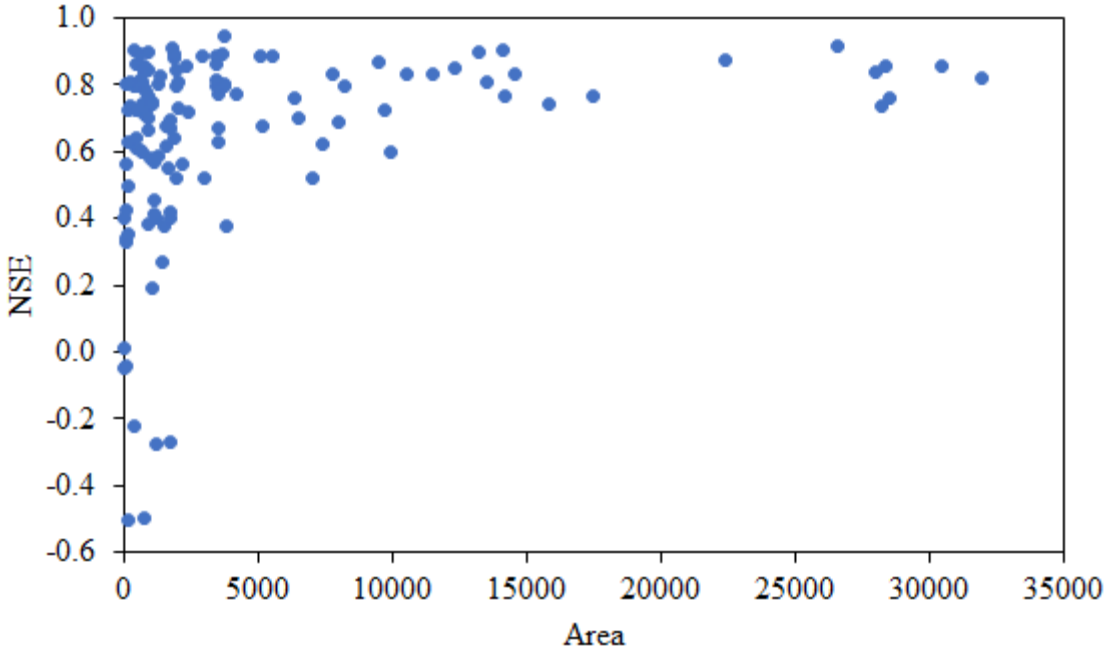

Figure 7. The distribution of the 120-hr ahead prediction using the best model in our benchmark (GRU for the single station).

As shown in the results, there are two major limitations. First, the model efficiency is low on the first day. It is shown in Figure 3 and Table 4 that the deep learning models do not show a higher accuracy in the first several hours compared to the Ridge model. Some hydrological studies have also shown that the basic persistence model (Streamflow t+n = Streamflow t) is hard-to-beat for short-range predictions when n is smaller than 12 hours (Krajewski et al., 2020). Thus, it is hard to make both short-range and medium-range predictions accurate in one model. The second limitation is the scale effect, where the large basins have better model performance on the streamflow forecast and the small basins are hard to predict. The results show that as watersheds get larger, the predictions become easier and better. This means the small watersheds, typically representing the middle and upper reaches, are harder to predict. Figure 7 shows the drainage area and 120-hr ahead prediction performance in NSE for 125 watersheds. The scale effect observed in our benchmark indicates the prediction in small watersheds is still a challenge.

Although a lot of metadata is provided in our dataset, as a benchmark, our study does not consider complex pretreatment nor models with domain knowledge in hydrology. Some recent studies have shown that the moving average for smoothing, the consideration of time lag, the consideration of watershed upstream-downstream connections, and other deep learning model architectures may be effective for a better prediction. However, these studies are based on their own datasets, and the results cannot be directly compared. We encourage researchers to conduct comparisons based on the WaterBench-Iowa.

5. Conclusion

In this study, by aggregating the datasets of watershed area, slope, soil types, streamflow, precipitation, and ET from NASA, NOAA, USGS, and IFC, we presented a dataset, namely WaterBench-Iowa, that is prepared for an hourly streamflow forecast task. This dataset has a high temporal resolution with abundant geographic and relational information, which can be used for a variety of deep learning and machine learning application research. We defined a sample streamflow forecasting task for the next 120 hours and provided example benchmark results on this task with a traditional linear and three custom deep learning models.

WaterBench-Iowa is not filtered and thus represents an actual streamflow forecast problem as much as possible. Although the data is limited to the Midwest, we believe that any studies on this dataset could provide insights for other streamflow forecasting and rainfall-runoff modeling studies in other watersheds. With the open-source release of WaterBench-Iowa (https://github.com/uihilab/WaterBench), this work provides a comparable benchmark, which to some extent makes up for the lack of a unified benchmark in hydrological and water resources research. We highly encourage other researchers to use the WaterBench-Iowa in their hydrological modeling research studies.

6. Data and Code Availability

The data and codes that support this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.7087806 (Demir et al., 2022a). The dataset covers the 125 catchments in Iowa, U.S. with seven different features, including precipitation, streamflow rate and ET with available data from October 1st, 2011 (the water year 2012) to September 30th, 2018 (the water year 2018). The original files of the dataset, metadata, and sample codes can be downloaded from archive files. Four different models, including Ridge, LSTM, GRU, S2S, used in our paper are provided with ready-to-run Python Jupyter Notebooks as well (Demir et al., 2022). The most recent code version can be found at https://github.com/uihilab/WaterBench (Demir et al., 2022b). It is welcome to send us feedback by filing an issue on the repository.

7. Acknowledgments

The work reported in this study was made possible by the support of members of the Iowa Flood Center and the Department of Civil and Environmental Engineering at the University of Iowa. This research received no external funding. We sincerely appreciate all the valuable comments and suggestions from the editors and reviewers, which helped us improve the quality of the manuscript.

Reference

Agliamzanov, R., Sit, M., & Demir, I. (2020). Hydrology@ Home: a distributed volunteer computing framework for hydrological research and applications. Journal of Hydroinformatics, 22(2), 235-248.

Athira, V., Geetha, P., Vinayakumar, R., & Soman, K. P. (2018). Deepairnet: Applying recurrent networks for air quality prediction. *Procedia computer science*, *132*, 1394-1403.

Bai, Y., Bezak, N., Sapač, K., Klun, M., & Zhang, J. (2019). Short-term streamflow forecasting using the feature-enhanced regression model. *Water Resources Management*, *33*(14), 4783-4797.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303-314.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

Du, S., Li, T., Yang, Y., & Horng, S. J. (2018). Deep air quality forecasting using hybrid deep learning framework. arXiv preprint arXiv:1812.04783.

Ebert-Uphoff, I., Thompson, D.R., Demir, I., Gel, Y.R., Karpatne, A., Guereque, M., Kumar, V., Cabral-Cano, E. and Smyth, P., (2017), September. A vision for the development of benchmarks to bridge geoscience and data science. In 17th International Workshop on Climate Informatics.

Fonley, M., Mantilla, R., Small, S. J., & Curtu, R. (2016). On the propagation of diel signals in river networks using analytic solutions of flow equations. *Hydrology & Earth System Sciences*, *20*(7).

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., & Lin, Q. (2020). Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. *Journal of Hydrology*, *589*, 125188.

Gericke, O. J., & Du Plessis, J. A. (2012). Catchment parameter analysis in flood hydrology using GIS applications. Journal of the South African Institution of Civil Engineering, 54(2), 15-26.

Godfried, I., Mahajan, K., Wang, M., Li, K., & Tiwari, P. (2020). FlowDB a large scale precipitation, river, and flash flood dataset. *arXiv preprint arXiv:2012.11154*.

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

Guo, T., Lin, T., & Lu, Y. (2018). An interpretable LSTM neural network for autoregressive exogenous model. arXiv preprint arXiv:1804.05251.

Guo, Y., Zhang, L., Hu, Y., He, X. and Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In European conference on computer vision (pp. 87-102). Springer, Cham.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), pp.1735-1780.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), pp.359-366.

Demir, Ibrahim, Xiang, Zhongrun, Demiray, Bekir Z, & Sit, Muhammed. (2022a). WaterBench-Iowa: A Large-scale Benchmark Dataset for Data-Driven Streamflow Forecasting [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7087806

Demir I., Xiang, Z., Demiray, B. Z., and Sit, M. (2022b). WaterBench [dataset]. available at: www.github.com/uihilab/WaterBench, last access: Feb 10, June 2022.

Iowa Department of Natural Resources. (2004). Chapter 1 Iowa's Water Resources. http://www.iowadnr.gov/portals/idnr/uploads/water/watershed/files/nonpoint%20plan/nps04.pdf

Kabir, S., Patidar, S., & Pender, G. (2020, April). Investigating capabilities of machine learning techniques in forecasting stream flow. In *Proceedings of the Institution of Civil Engineers-Water Management* (Vol. 173, No. 2, pp. 69-86). Thomas Telford Ltd.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005-6022.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344-11354.

Krajewski, W. F., Ceynar, D., Demir, I., Goska, R., et al. (2017). Real-time flood forecasting and information system for the state of Iowa. *Bulletin of the American Meteorological Society*, *98*(3), 539–554. https://doi.org/10.1175/BAMS-D-15-00243.1

Krajewski, W. F., Ghimire, G. R., & Quintero, F. (2020). Streamflow Forecasting without Models. *Journal of Hydrometeorology*, *21*(8), 1689-1704.

LeCun, Y. (1989). Generalization and network design strategies. *Connectionism in perspective*, *19*, 143-155.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, *6*(6), 861-867.

Li, Y., Zhu, Z., Kong, D., Han, H., & Zhao, Y. (2019). EA-LSTM: Evolutionary attention-based LSTM for time series prediction. Knowledge-Based Systems, 181, 104785.

Liu, W., Guo, G., Chen, F., & Chen, Y. (2019). Meteorological pattern analysis assisted daily PM2. 5 grades prediction using SVM optimized by PSO algorithm. Atmospheric Pollution Research, 10(5), 1482-1491.

Maskey, M., H. Alemohammad, K. J. Murphy, and R. Ramachandran (2020), Advancing AI for Earth science: A data systems perspective, Eos, 101, https://doi.org/10.1029/2020EO151245. Published on 06 November 2020.

Mandapaka, P. V., Krajewski, W. F., Mantilla, R., & Gupta, V. K. (2009). Dissecting the effect of rainfall variability on the statistical structure of peak flows. Advances in Water Resources, 32(10), 1508-1525.

Mantilla, R., & Gupta, V. K. (2005). A GIS numerical framework to study the process basis of scaling statistics in river networks. *IEEE Geoscience and Remote sensing letters*, *2*(4), 404-408.

Mantilla, R., Gupta, V. K., & Troutman, B. M. (2011). Scaling of peak flows with constant flow velocity in random self-similar networks. *Nonlinear Processes in Geophysics*, *18*(4).

Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R. and Blodgett, D. (2014). A large sample watershed-scale hydrometeorological dataset for the contiguous USA. UCAR/NCAR, Boulder, CO.

Ni, L., Wang, D., Singh, V. P., Wu, J., Wang, Y., Tao, Y., & Zhang, J. (2020). Streamflow and rainfall forecasting by two long short-term memory-based models. *Journal of Hydrology*, *583*, 124296.

Post, W.M., and L. Zobler. (2000). Global Soil Types, 0.5-Degree Grid (Modified Zobler). ORNL DAAC. Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/540

Sagheer, A., & Kotb, M. (2019). Unsupervised pre-training of a Deep LStM-based Stacked Autoencoder for Multivariate time Series forecasting problems. Scientific Reports, 9(1), 1-16.

Seeger, M., Salinas, D., & Flunkert, V. (2016, December). Bayesian intermittent demand forecasting for large inventories. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 4653-4661).

Seo, B. C., Krajewski, W. F., Quintero, F., ElSaadani, M., Goska, R., Cunha, L. K., ... & Petersen, W. A. (2018). Comprehensive evaluation of the IFloodS radar rainfall products for hydrologic applications. *Journal of Hydrometeorology*, *19*(11), 1793-1813.

Seo, B. C., Keem, M., Hammond, R., Demir, I., & Krajewski, W. F. (2019). A pilot infrastructure for searching rainfall metadata and generating rainfall product using the big data of NEXRAD. Environmental modelling & software, 117, 69-75.

Sit, M., & Demir, I. (2019). Decentralized flood forecasting using deep neural networks. *arXiv preprint arXiv:1902.02308*.

Sit, M., Sermet, Y., & Demir, I. (2019). Optimized watershed delineation library for server-side and client-side web applications. Open Geospatial Data, Software and Standards, 4(1), 1-10.

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology*, *82*(12), 2635-2670.

Sit, M., Demiray, B., & Demir, I. (2021a). Short-term hourly streamflow prediction with graph convolutional gru networks. arXiv preprint arXiv:2107.07039.

Sit, M., Seo, B. C., & Demir, I. (2021b). Iowarain: A statewide rain event dataset based on weather radars and quantitative precipitation estimation. arXiv preprint arXiv:2107.03432.

Sloan, B. P., Mantilla, R., Fonley, M., & Basu, N. B. (2017). Hydrologic impacts of subsurface drainage from the field to watershed scale. *Hydrological Processes*, *31*(17), 3017-3028.

Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., & Wang, O. (2017). Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1279-1288).

Tao, Q., Liu, F., Li, Y., & Sidorov, D. (2019). Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. IEEE Access, 7, 76690-76698.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, J. H., Lin, G. F., Chang, M. J., Huang, I. H., & Chen, Y. R. (2019). Real-time water-level forecasting using dilated causal convolutional neural networks. *Water Resources Management*, *33*(11), 3759-3780.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

Xiang, Z., Demir, I., Mantilla, R., & Krajewski, W. F. (2021). A Regional Semi-Distributed Streamflow Model Using Deep Learning. EarthArXiv. https://doi.org/10.31223/X5GW3V

Xiang, Z., & Demir, I. (2020). Distributed long-term hourly streamflow predictions using deep learning–A case study for State of Iowa. *Environmental Modelling & Software*, *131*, 104761.

Xiang, Z., Yan, J., & Demir, I. (2020). A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water resources research*, *56*(1), e2019WR025326.

Xu, H., Windsor, M., Muste, M., & Demir, I. (2020). A web-based decision support system for collaborative mitigation of multiple water-related hazards using serious gaming. Journal of environmental management, 255, 109887.

Xue, T., Chen, B., Wu, J., Wei, D., & Freeman, W. T. (2019). Video enhancement with task-oriented flow. *International Journal of Computer Vision*, *127*(8), 1106-1125.

Yildirim, E., & Demir, I. (2021). An Integrated Flood Risk Assessment and Mitigation Framework: A Case Study for Middle Cedar River Basin, Iowa, US. International Journal of Disaster Risk Reduction, 56, 102113.

Yu, H. F., Rao, N., & Dhillon, I. S. (2016, December). Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction. In *NIPS* (pp. 847-855).

Zhu, S., Lian, X., Wei, L., Che, J., Shen, X., Yang, L., ... & Li, J. (2018). PM2. 5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. Atmospheric Environment, 183, 20-32.