Regional Streamflow Drought Forecasting in the Colorado River Basin using Deep Neural Network Models

Scott Hamshaw, Machine Learning Specialist, U.S. Geological Survey, Integrated Modeling and Prediction Division, Bristol, VT, shamshaw@usgs.gov

Phillip Goodling, Hydrologist, U.S. Geological Survey, MD-DE-DC Water Science Center, Cantonsville, MD, pgoodling@usgs.gov

Konrad Hafen, Hydrologist, U.S. Geological Survey, Idaho Water Science Center, Boise, ID, khafen@usgs.gov

John Hammond, Research Hydrologist, U.S. Geological Survey, MD-DE-DC Water Science Center, Cantonsville, MD, jhammond@usgs.gov

Ryan McShane, Hydrologist, U.S. Geological Survey, WY-MT Water Science Center, Bozeman, MT, rmcshane@usgs.gov

Roy Sando, Physical Scientist, WY-MT Water Science Center, Bozeman, MT, tsando@usgs.gov

Apoorva Shastry, Contractor to the U.S. Geological Survey, Earth Science Division, Moffett Field, CA, ashastry@contractor.usgs.gov

Caelan Simeone, Hydrologist, U.S. Geological Survey, Oregon Water Science Center, Portland, OR, csimeone@usgs.gov

David Watkins, Machine Learning Engineer, U.S. Geological Survey, Integrated Information Dissemination Division, Davis, CA, wwatkins@usgs.gov

Ellie White, Data Scientist, U.S. Geological Survey, Integrated Information Dissemination Division, Owensboro, KY, ewhite@usgs.gov

Michael Wieczorek, Geographer, U.S. Geological Survey, MD-DE-DC Water Science Center, Baltimore, MD, mewieczo@usgs.gov

Abstract

Process-based, large-scale (e.g., conterminous United States [CONUS]) hydrologic models have struggled to achieve reliable streamflow drought performance in arid regions and for low-flow periods. Deep learning has recently seen broad implementation in streamflow prediction and forecasting research projects throughout the world with performance often equaling or exceeding that of process-based models. Deep learning models are a possible approach to increase the accuracy of streamflow drought predictions and to expand the spatial coverage of river locations with available streamflow drought forecasts.

As part of a multi-component Data-Driven Drought Prediction project, the U.S. Geological Survey is developing and testing deep learning models for streamflow drought forecasting. In this work, we present preliminary results of a deep learning model capable of predicting streamflow drought occurrence at ungaged locations for the Colorado River Basin (CRB). A long short-term memory (LSTM) neural network model was trained using 40 years (1980-2020) of daily streamflow data from 425 streamgages within and surrounding the CRB using static watershed attributes as well as meteorological and remotely sensed dynamic forcing inputs. Model tests were performed to evaluate model accuracy for now-casting streamflow drought conditions at ungaged locations and for forecasting drought conditions at lead times ranging from 0 to 14 days. Nearly all model configurations showed behavioral performance for predicting daily streamflow percentiles. Comparisons of LSTM model performance for predicting drought using fixed drought thresholds (calculated over all days and years) and variable drought thresholds (unique threshold calculated for each day of the year) identify differences in model skill between locations with implications for model design.

Introduction

Hydrological drought is a significant and recurring problem facing water resource managers in the western United States and beyond. Streamflow droughts, one component of hydrological drought, have increased in duration and deficit volume in the western United States in recent decades (Hammond and others, 2022). Streamflow drought materializes "as a lack of water in the hydrological system, manifesting itself as abnormally low streamflow in rivers and streams" (Van Loon, 2015). The different (*i*) time periods, (*ii*) time units (ex. daily vs monthly), and (*iii*) approach (standardized index vs threshold method) one can use to define streamflow as being abnormally low have led to multiple definitions of streamflow drought depending on location and water use sector. The threshold method compares a daily streamflow value to either a fixed threshold or a variable threshold, such as one that changes over the course of a year to correspond to streamflow seasonality (Van Loon, 2015). The fixed and variable threshold methods have been utilized by the U.S. Geological Survey (USGS) to characterize streamflow drought throughout the CONUS (Hammond and others, 2022), with the National Drought Monitor (droughtmonitor.unl.edu/) using streamflow percentiles based on the variable threshold method as one component of assessing drought across the CONUS. The choice of fixed or variable threshold approach to characterize drought leads to differences in timing and intensity of defined drought events that can serve diverse stakeholder needs. Choice of drought definition has also been suggested to have implications in design of drought early warning systems (Sutanto and Van Lanen, 2021).

The Colorado River Basin (CRB) has been the recent focus of significant scientific and public attention related to sustained drought conditions. Persistent streamflow drought in the CRB has resulted in large reservoirs reaching record low levels, necessitating new water management strategies (Wheeler and others, 2022). To assist water resource managers in the CRB, multiple existing modeling programs including the Bureau of Reclamation Colorado River 24-Month Study Projections (www.usbr.gov/lc/region/g4000/riverops/24ms-projections.html), the National Weather Service Colorado River Basin Forecast Center's water supply estimates (www.cbrfc.noaa.gov/wsup/graph/west/map/esp_map.html), and the Natural Resource Conservation Service water supply estimates (www.nrcs.usda.gov/wps/portal/wcc/home/waterSupply/waterSupplyForecasts) all forecast water availability at different time scales and locations. However, there remain additional river and stream locations in the CRB not served by existing forecast products and a need for models specifically focused on streamflow drought early warning. In the Data-Driven Drought Prediction project, the USGS is working to build complementary modeling and forecasting capacity for hydrological drought in the CRB and is specifically employing machine learning (ML) models due to their promising capabilities in hydrological prediction (Shen and others, 2021).

Machine learning, and in particular deep learning models, have been rapidly adopted in hydrologic modeling such as daily streamflow prediction (Shen, 2018). As demonstrated by

Kratzert and others (2019a) and additional studies (Feng and others, 2020; Frame and others, 2021; Nevo and others, 2022), deep learning models have equaled and exceeded performance of regional process-based hydrologic models such as WRF-Hydro (i.e., National Water Model) for daily streamflow prediction. Regional scale long short-term memory (LSTM) models have been a widely adopted deep learning model because they are particularly adept at predicting daily streamflow from a combination of dynamic forcing data and static watershed attributes (Klotz and others, 2022). However, LSTM models trained to predict daily streamflow have also shown inconsistent performance in more arid regions (Feng and others, 2020), regulated watersheds (Ouyang and others, 2021), and during low-flow or drought periods (Kratzert and others, 2019b) – locations and periods considered generally to be more challenging to model. These conditions are also difficult for process-based models to accurately simulate. This research aims to address the need for further ML hydrological model development for streamflow drought situations. We additionally note that there is to the best of our knowledge no comparable studies on ML models being used to directly predict a streamflow percentile and this USGS Data-Driven Drought Prediction project seeks to contribute additional research in this area.

In this work, we highlight one component of the larger effort – testing a baseline LSTM model for predicting streamflow drought occurrence in ungaged locations within and adjacent to the CRB. We compare a model that directly predicts a daily streamflow percentile (with periods below a threshold corresponding to drought) to a model trained to predict streamflow, which is then used to estimate the percentile and drought conditions. We also compare the performance at different forecast time horizons and in a prediction in ungaged basins (PUB) approach.

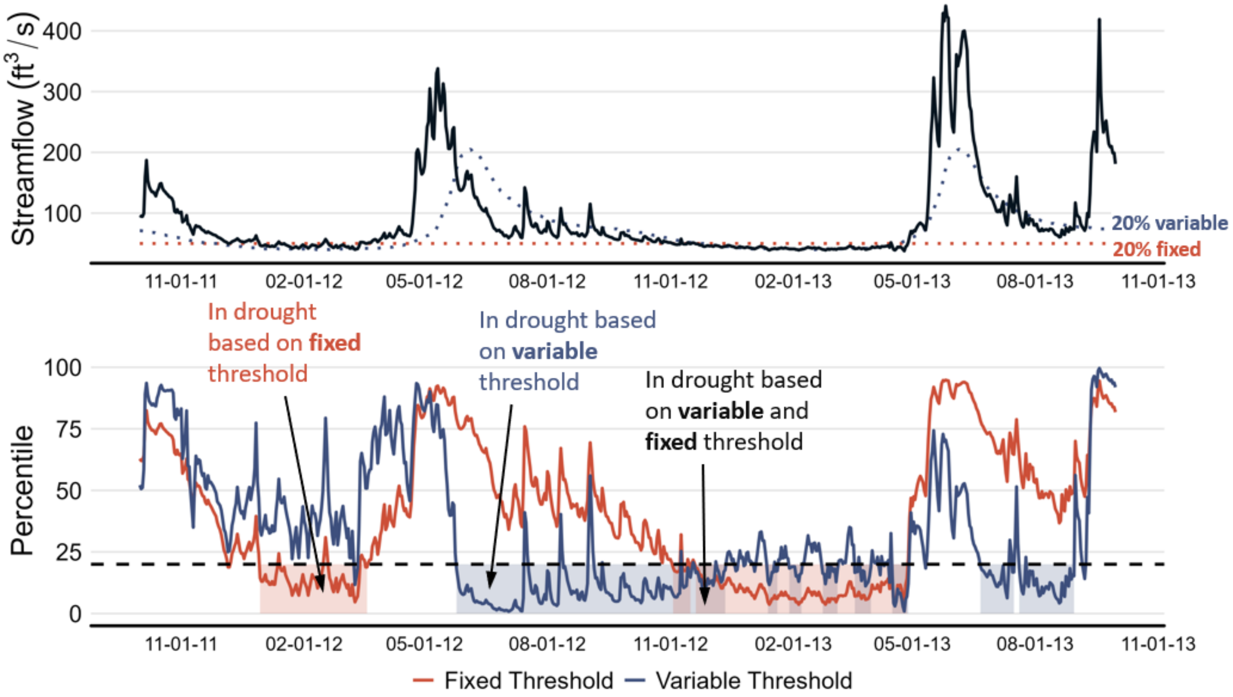

Figure 1. Example streamflow (USGS, 2022) and derived streamflow percentiles for USGS Streamgage 0929500, Yellowstone River near Altonah, UT. The 20% fixed and variable thresholds (dotted lines upper panel) are shown in relation to streamflow from the 2012 and 2013 water years. The corresponding streamflow percentiles can be used to identify drought events at the 20% severity level for fixed and variable threshold, respectively (shaded areas in lower panel).

Data & Study Location

The modelling target data for this work were daily streamflow and daily streamflow percentiles from 425 streamgages located within the CRB and surrounding area. The 425 streamgages in the dataset encompass all streamgages that met two criteria for streamflow records: (a) include at least 95% of days in each year and (b) have at least 8 of 10 complete years for all decades (e.g., 1990–1999) in the period from 1980-2020. The streamflow percentiles and associated streamflow drought metrics are available from Simeone (2022). The daily percentiles selected for this work were the fixed (long-term) and variable (moving 30-day window) threshold percentiles (Figure 1). Drought periods are defined using thresholds of severity including 2%, 5%, 10%, 20%, and 30%. We selected the 20% threshold for drought to use in evaluating model performance in this study to highlight the model performance corresponding to moderate drought events. Additionally, daily streamflow, converted to runoff in mm/d, was used as a modeling target. Input data variables (features) included gridded meteorological data that were aggregated to basin averages (Table 1) but did not include at-site streamflow as this component of the project focused on prediction at ungaged time and locations. Additionally, 27 static watershed attributes available for the National Hydrography (NHD) NHDPlus Version 2.1 catchments (Wieczorek and others, 2018) were used as model inputs (Table 2). Model inputs were re-scaled prior to being used in modeling by z-score normalization for input features and min-max normalization for the target variable. Model input data for the streamgages used in this study are available from Wieczorek and others (2023).

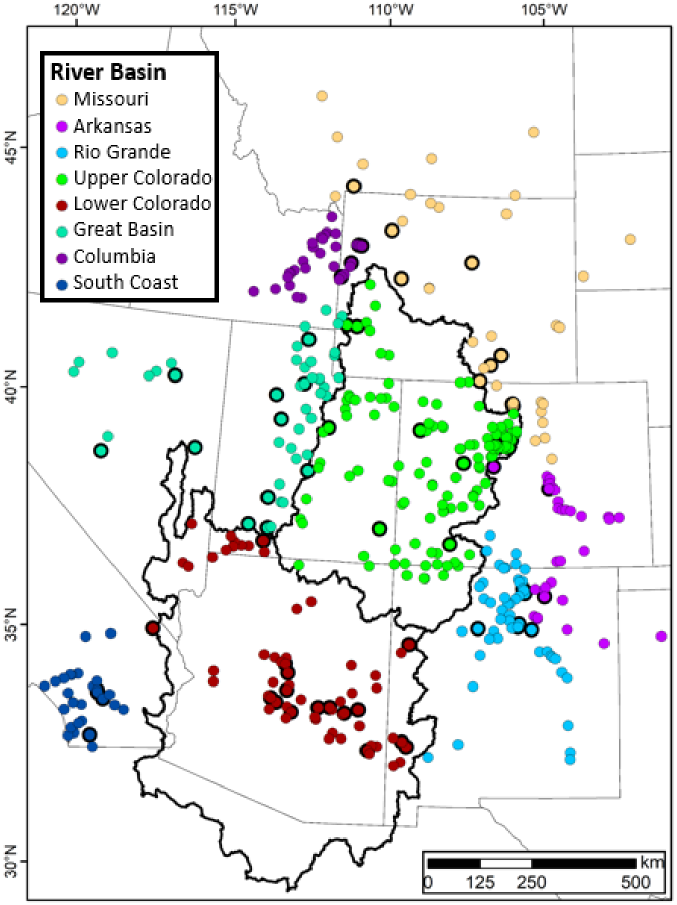

Figure 2. Map of streamgage locations with sufficient observations records to have streamflow percentiles available for use in the streamflow drought model. Dots with black border indicate reference gages in the USGS Hydro-Climatic Data Network (HCDN).

The study area includes the upper and lower CRB as well as surrounding areas within a 100-mile buffer. The 425 streamgages (Figure 2) include 205 within the CRB and 225 located in watersheds that drain any land within the buffer. The purpose of using data from additional streamgages nearby the CRB was to increase the number of streamgages available for training and evaluating models. From the 425 streamgages, 26 were withheld as an unbiased test dataset to be used in evaluation of future model development. Additionally, 17 streamgages were withheld for lacking input data or being located on the mainstem of the Colorado River below Lake Powell.

Table 1. Gridded meteorology and climatology datasets used as model inputs. Data were aggregated to a single basin average value for each day for each streamgage

Variable	Units	Source	Reference
Minimum Temperature	°C	gridMET	(Abatzoglou, 2013)
Maximum Temperature	°C		
Precipitation	mm		
Evapotranspiration (Reference - grass)	mm		
Standardized Precipitation Evapotranspiration Index (SPEI)	unitless		
Snow Water Equivalent (SWE)	mm	NASA NSIDC	(Broxton and others, 2019)
Soil Moisture (0-10 mm depth)	kg/m^2	NASA NLDAS	(Mitchell, 2004)
Soil Moisture (10-40 mm depth)	kg/m^2		
Soil Moisture (40 – 100 mm depth)	kg/m^2		

Table 2. Static watershed attributes used as inputs to the regional deep learning models. (Wieczorek and others, 2018)

Attribute	Description	Attribute	Description
SqKm	Drainage Area in square kilometers	MIRAD_2012	% of watershed in irrigated agriculture (2012)
ELEV_MEAN	Mean elevation	FRESHWATER_WD	Total freshwater withdrawals
ELEV_MAX	Maximum elevation	SANDAVE	Average % of sand in soil
MINWD6190	Average of minimum monthly number of days of measurable precipitation (1961-1990)	CLAYAVE	Average % of clay in soil
MAXWD6190	Average of maximum monthly number of days of measurable precipitation (1961-1990)	SILTAVE	Average % of silt in soil
RF7100	Mean annual average for the Rainfall and Runoff factor (1971-2000)	HGA	Percentage of Hydrologic Group A soil
ARTIFICIAL	Percentage of all flowline reach that is an artificial reach	HGB	Percentage of Hydrologic Group B soil
AET	Mean annual evapotranspiration	HGC	Percentage of Hydrologic Group C soil
RH	Average relative humidity	HGD	Percentage of Hydrologic Group D soil
WB5100_ANN	Average annual runoff (1951-2000)	ROCKDEP	Average range in total soil thickness
MAXP6190	Maximum average annual precipitation (1961-1990)	CONTACT	Subsurface flow contact time index
CWD	Average number of consecutive days with measurable precipitation	STREAM_SLOPE	Average flowline slope
RECHG	Mean annual natural ground-water recharge	TOTAL_ROAD_DENS	Density of all road types
BFI	Base flow index	NLCD19_FOREST	2019 watershed % of land use in forest
TWI	Topographic wetness index	NLCD19_WETLAND	2019 watershed % of land use in wetlands
EWT	Average depth to water table	DI_EROM	reservoir storage intensity in units of days
SATOF	Percentage of Dunne overland flow as a percent of total flow	DI_PMC	degree of regulation

Methods

Deep Learning Model

Separate LSTM deep learning models were trained and tested for multiple different modeling target variables, forecast horizons, and an ungaged application. The LSTM model is a type of recurrent neural network that is designed to learn from sequential data (e.g., natural language, time series). The Python open source package N*euralHydrology* (Kratzert and others, 2022) was used for all model training. *NeuralHydrology* is a deep learning package based on the PyTorch machine learning framework and features an LSTM model implementation that can be trained on GPU computing instances. Details on the LSTM architecture and implementation can be found in Kratzert and others (2018). We used the USGS Cloud Hosting Solutions (CHS) implementation of Amazon Web Services Sagemaker computing platform for model training.

LSTM models, like other deep learning models, have multiple settings (commonly referred to as hyperparameters) that can be tuned for optimal performance. In this preliminary work, we have not performed an exhaustive hyperparameter search, but instead adopted hyperparameter settings based on previous studies of daily streamflow prediction (e.g., Kratzert and others, 2019a). For all models, a sequence length of 270 days was used with a single output prediction (i.e., the model uses the previous 270 days of inputs to make a single prediction at time *t*, *t*+7, or *t*+14). Additional LSTM hyperparameters were set as follows: 192 hidden units; single LSTM layer; 0.4 dropout before output layer; 0.1 standard deviation of normalized target data added noise (to reduce overfitting); and learning rate schedule of 0: 1e-3, 1: 5e-4, 5: 1e-4, 10-: 5e-5; and batch size of 512. Models were trained for 30 epochs, which was found to be sufficient for model convergence. The training (loss) function used differed according to the target variable with average Nash-Sutcliffe efficiency (NSE) used for daily streamflow (Kratzert and others, 2019b), and symmetric mean absolute percentage error (SMAPE) (Smyl, 2017) used for fixed and variable percentiles.

Experiment Design

The project design consists of training individual LSTM models for each combination of model target variable, forecast horizon, and training/validation data configuration (Table 3). For all models, training data consisted of the period from 01-Oct-1981 to 31-Mar-2005 and validation period of 01-Apr-2005 to 31-Mar-2014. It is typical in machine learning modeling to have a three-way train/validation/test split of data (Subramanian, 2018). As described earlier, a set of 26 streamgages were set aside for future use as a test data partition, along with the time period from 01-Apr-2014 to 31-Mar-2020. Therefore, in this preliminary work, the train and validation data partitions can be seen as equivalent to a simple training and testing split, however, for consistency within the overall project, we will continue to refer to them as training and validation here.

Models trained using all streamgages were trained using data from all 384 streamgages and then validated over the same 384 gages and represent a model scenario of making predictions at locations where past observations are available (i.e., ungaged in time). For the prediction in ungaged basins (PUB) models, spatial cross-validation was used where streamgages were assigned into geospatial groups based on if their drainage areas overlapped by 50% or more. Geospatial groups were then randomly assigned into k = 10 cross-validation folds. The PUB LSTM models are trained on all data from k-1 folds. This is repeated k times resulting in out-of-sample (ungaged) predictions for all 384 streamgages.

Table 3. Experiment setup for LSTM model runs

Model Run	Modeling Target Variable	Forecast Horizon	Streamgages used in model validation
Streamflow-0d	Daily Streamflow (mm/d)	0 days	All streamgages
Streamflow-7d	Daily Streamflow (mm/d)	7 days	All streamgages
Streamflow-14d	Daily Streamflow (mm/d)	14 days	All streamgages
PUB-Streamflow-0d	Daily Streamflow (mm/d)	0 days	Streamgages withheld in training
Fixed-0d	Fixed Percentile	0 days	All streamgages
Fixed-7d	Fixed Percentile	7 days	All streamgages
Fixed-14d	Fixed Percentile	14 days	All streamgages
Variable-0d	Variable Percentile	0 days	All streamgages
Variable-7d	Variable Percentile	7 days	All streamgages
Variable-14d	Variable Percentile	14 days	All streamgages
PUB-FIX-0d	Fixed Percentile	0 days	Streamgages withheld in training
PUB-VAR-0d	Variable Percentile	0 days	Streamgages withheld in training

To assess the effectiveness of using a model trained on daily streamflow for predicting drought, we converted the daily streamflow to an estimated streamflow percentile using linear interpolation. The transformed model predictions are then able to be evaluated using the same metrics as models directly trained on the streamflow percentile data. The transformed streamflow (Q) models are identified as Q-to-Fixed-0d and Q-to-Variable-0d.

Evaluation

Model predictions were evaluated using three performance metrics that capture the ability of the model to simulate the overall time series of streamflow percentiles as well as the robustness of correctly predicting drought occurrence. Regression performance metrics of overall percentile prediction utilized were NSE and Kling Gupta Efficiency (KGE) (see Knoben and others, 2019). For evaluating accuracy of predictions of streamflow drought conditions, we classified drought and non-drought periods for both modeled and observed time series using the 20th percentile threshold (i.e., moderate to exceptional drought levels combined), then used Cohen's kappa to quantify how well the modeled predictions of drought periods match observations. Cohen's kappa is a measure of inter-rater reliability (Landis and Koch, 1977), in which the agreement between two raters, here observed and modeled drought occurrence, can be determined. Cohen's kappa values range between -1 and 1 with values below 0 indicating no agreement and 1 being perfect agreement. Fixed model predictions were evaluated on fixed threshold drought periods, and variable model predictions were evaluated on variable threshold drought periods. Each metric was calculated on a per streamgage basis and overall minimum, median, and maximum values for each model run were tabulated.

Results

LSTM Model Performance for Predicting Daily Streamflow

The LSTM models trained on daily streamflow showed acceptable predictive ability at all lead times (0, 7, and 14 days) with median NSE scores ranging from 0.47 to 0.69 and KGE scores ranging from 0.60 to 0.69 (Figure 3 and Table 4). For daily streamflow prediction, KGE scores above 0.3 and NSE scores above 0.5 can be considered "behavioral" and show meaningful modeling capability (Knoben and others, 2019). There are several sites where the LSTM model performance is worse than assuming the mean flow at that site as evidenced by a non-trivial number of KGE values below -0.41 (or NSE below 0). Given that many streamgages used in model evaluation have moderate to significant flow regulation, model predictions at some streamgage locations are inaccurate.

Table 4. Summary statistics for model performance during validation period for models trained using daily streamflow (mm/day) as the target variable.

	KGE			NSE		
Model	**Min**	**Median**	**Max**	**Min**	**Median**	**Max**
Streamflow-0d	-10.88	0.69	0.96	-35.83	0.69	0.96
Streamflow-7d	-30.93	0.61	0.94	-9.00	0.60	0.90
Streamflow-14d	-29.20	0.60	0.92	-2.84	0.53	0.89
PUB-Streamflow-0d	-51.21	0.42	0.97	-217.67	0.47	0.94

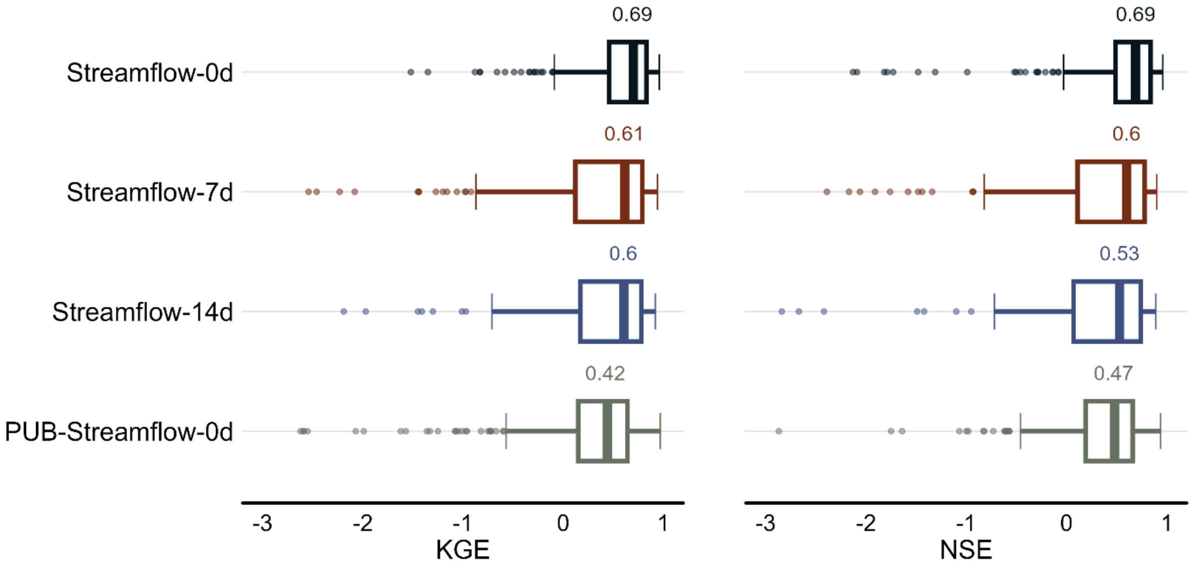

Figure 3. Model performance during validation period for models trained using daily streamflow (mm/day) as the target variable. The PUB (prediction in ungaged basins) model indicates model spatial cross-validation is done to provide predictions at locations not used at all in model training. Note, outlier points below -3.0 values are clipped to visualize quartiles more clearly.

LSTM Model Performance for Predicting Streamflow Percentiles and Drought Occurrence

LSTM models trained directly on streamflow percentiles showed behavioral model performance across lead times and threshold type (Table 5 and Figures 4 and 5). Fixed threshold LSTM models had accuracy (median KGE 0.67 to 0.72) very similar to daily streamflow models (median KGE 0.60 to 0.69) and higher than variable threshold models (median KGE 0.37 to 0.51). Given that fixed threshold percentiles are a re-scaled daily streamflow, performance would be expected to be on par with daily streamflow prediction. We also observed greater loss in predictive accuracy at longer lead times for variable percentiles than for fixed percentiles (Table 5, Figures 4 to 6)

Table 5. Summary statistics for model performance during validation period for predicting daily streamflow percentiles as the target variable.

	NSE			KGE		
Model	**Min**	**Median**	**Max**	**Min**	**Median**	**Max**
Fixed-0d	-0.92	0.62	0.96	-1.64	0.72	0.97
Fixed-7d	-0.90	0.53	0.92	-1.54	0.67	0.95
Fixed-14d	-0.89	0.52	0.92	-1.58	0.67	0.95
Variable-0d	-1.17	0.24	0.77	-1.69	0.51	0.85
Variable-7d	-1.05	0.11	0.60	-1.68	0.43	0.77
Variable-14d	-0.97	0.01	0.63	-1.86	0.37	0.78
Q to Fixed-0d	-2.23	0.48	0.93	-2.36	0.52	0.96
Q to Variable-0d	-1.33	0.04	0.72	-2.40	0.29	0.92
PUB-Fixed-7d	-1.65	0.40	0.92	-1.62	0.58	0.95
PUB-Variable-7d	-1.09	-0.17	0.69	-1.23	0.46	0.81

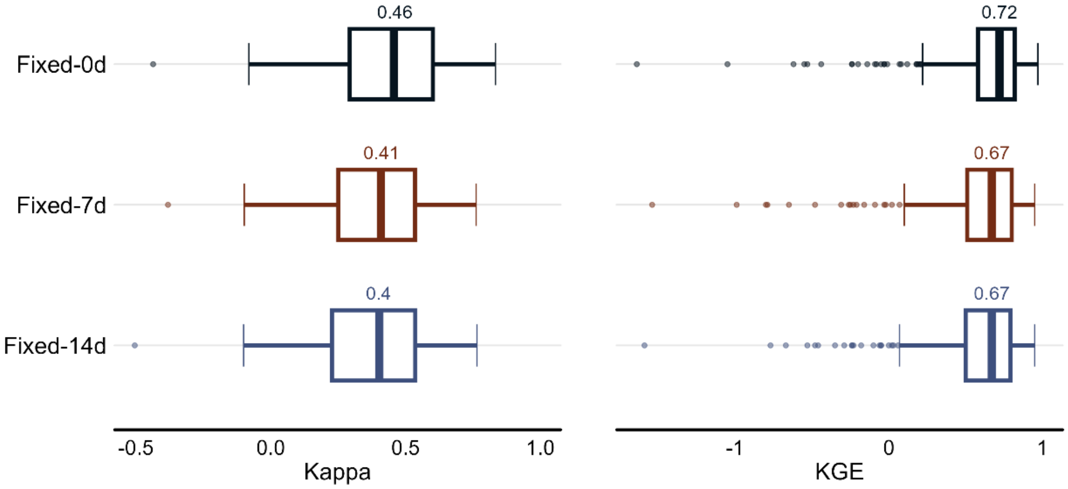

Figure 4. Model performance during validation period for models trained using fixed threshold daily streamflow percentiles as the target variable.

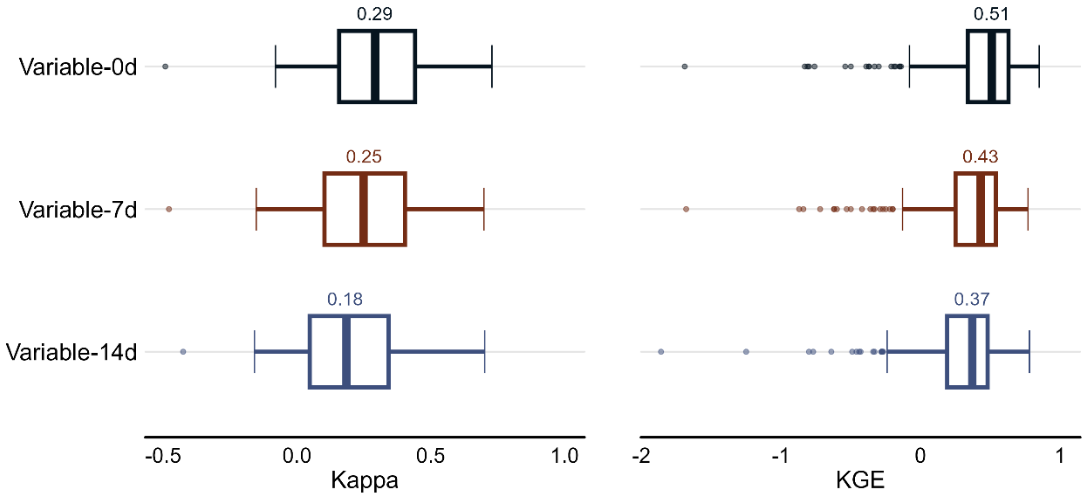

Figure 5. Model performance during validation period for models trained using variable threshold daily streamflow percentiles as the target variable.

Predicted streamflow percentiles from the LSTM models trained directly on percentiles (fixed and variable threshold) were also compared to percentiles estimated from the daily streamflow predictions (models in Table 5 and Figure 6) at the 0-day lead time. Estimated percentiles from the daily streamflow models had lower accuracy than models trained directly on percentiles with median KGE of 0.52 and 0.29, for variable and fixed thresholds respectively. Results show that training the models on the target of interest (streamflow percentiles) yields better performing models than training on a variable that then needs post processing.

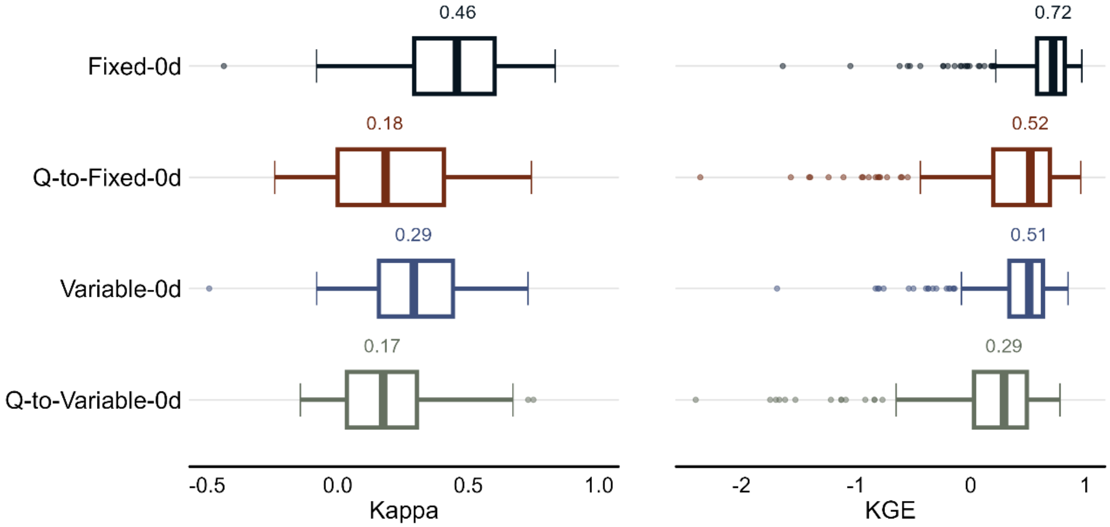

Figure 6. Model performance at the 0-day lead time during validation period for models trained directly using streamflow percentiles compared to predictions calculated from daily streamflow predictions.

Percentile model performance, for fixed and variable drought thresholds (i.e., a predicted and observed streamflow percentile are both less than 20%), was fair with accuracy measured by median values of Cohen's kappa between 0.18 and 0.46 (Table 6 and Figures 4 and 5). Like the prediction of percentiles generally, the drought prediction for fixed percentile models was more accurate than for variable percentile models. Additionally, for both fixed and variable thresholds, the modeling of percentiles directly resulted in more accurate predictions compared to estimates derived from daily streamflow predictions (Figure 6 and Table 6).

Table 6. Summary statistics for model drought occurrence prediction accuracy at the 20th percentile threshold.

	Cohen's kappa		
Model	**Min**	**Median**	**Max**
Fixed-0d	-0.44	0.46	0.83
Fixed-7d	-0.38	0.41	0.76
Fixed-14d	-0.50	0.40	0.76
Variable-0d	-0.49	0.29	0.73
Variable-7d	-0.47	0.25	0.70
Variable-14d	-0.43	0.18	0.70
Q to Fixed-0d	-0.24	0.18	0.74
Q to Variable-0d	-0.14	0.17	0.75
PUB-Fixed-0d	-0.48	0.30	0.79
PUB-Variable-0d	-0.33	0.24	0.78

Evaluation of Model Performance for Prediction in Ungaged Basins (PUB) Task

Evaluating the predictive capacity of the LSTM models in a true ungaged scenario, where no data from the streamgages considered in validation were used in training, the model performance expectedly decreased for both variable and fixed thresholds (Figure 7 and Table 6) but was still behavioral. A greater decrease in accuracy was observed for the fixed threshold percentile models compared to variable threshold.

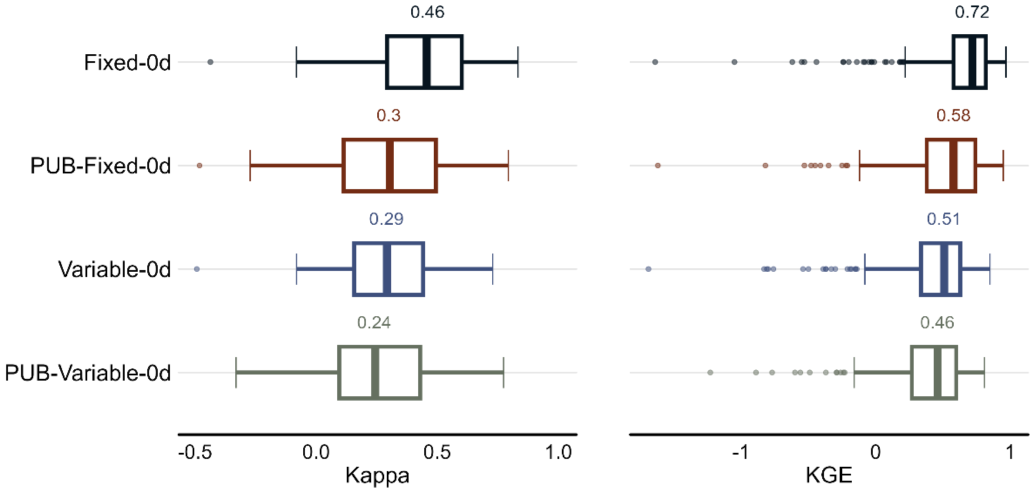

Figure 7. Model performance at the 0-day lead time during validation period for models trained directly on streamflow percentiles. The PUB (prediction in ungaged basins) model indicates model spatial cross-validation is done to provide predictions at locations and is not used at all in model training.

Discussion

LSTM Modeling of Daily Streamflow in the Colorado River Basin Region

Error metrics for daily streamflow prediction in this work are comparable to existing national-scale studies on daily streamflow prediction using LSTM models. Several LSTM daily streamflow studies have been undertaken using the CAMELS basin dataset (Addor and others, 2017) of reference hydrological basins such as Xie and others (2021) and Frame and others (2021) that reported a median KGE of 0.75 and 0.74, respectively (equivalent to the Streamflow-0d model test in this work) across 531 basins in the CONUS. Given the preliminary results presented here are for a model that has not gone through a hyperparameter model tuning experiment yet, these results show promise for the compiled dataset to predict daily streamflow magnitudes across the CRB region. Additionally, the comparability of performance given that the 384 streamgages used in this study include basins with moderate to substantial flow regulation and more arid climate, whereas the CAMELS basins are minimally altered and include humid climate basins, indicates that there is potential for application of LSTM models to streamflow drought prediction in the CRB. However, the results for the PUB model evaluations highlight the challenge in building a reliable ungaged streamflow prediction model in the climatologically diverse and anthropogenically altered CRB region. Kratzert and others (2019a) reported a PUB LSTM median NSE of 0.69 across 531 CAMELS basins, which is higher than the results achieved here (median NSE of 0.47). Our work is more comparable to Ouyang and others (2021) who performed several experiments testing PUB prediction using different subsets of streamgages (with varying degrees of flow regulation) and reported highly variable median NSE values (depending on streamgage subsets used in training and testing). This highlights the effect that flow regulation has on streamflow predictability; complimentary work could explore how flow regulation and watershed attributes correlate to LSTM model performance.

LSTM Modeling of Streamflow Percentiles

Streamflow drought is dependent on the definition selected (e.g., a fixed or variable threshold) and that definition has a noticeable effect on LSTM models predictive capacity. Fixed thresholds were considerably easier to predict than variable thresholds using gridded meteorology and watershed attributes. The difficulty in predicting variable threshold percentiles could be partially attributed to the inputs (i.e. observation values) and the target data being transformed to a deviation from a seasonally adjusted baseline. Investing the transformation of input variables and how it affects model performance could be an area for further research.. Streamflow drought is directly measured based on streamflow magnitudes, and one benefit of training a model on streamflow is that a single LSTM model can then be used to estimate a variety of streamflow drought calculations by post-processing the modeled streamflow. However, in all cases, our results showed that LSTM models trained directly on streamflow percentiles improved predictive ability. It is likely that the processing of streamflow into percentiles, which necessarily re-scales streamflow, results in improved performance of estimating lower percentile values. One drawback to this approach is that streamgages used for training must have sufficient records available to estimate percentiles reliably. In our work, streamgages must have been operational for at least 40 years which resulted in data from numerous gaging stations being excluded from use.

Conclusions

The LSTM models developed in this work show promise for ML-based prediction of streamflow drought using gridded meteorology and remote sensing data as inputs. We found when using a standard LSTM model with typical loss functions (e.g., NSE) a model trained directly on calculated streamflow percentiles shows greater accuracy in predicting percentiles and drought occurrence relative to models trained on streamflow that is then post-processed. Fixed percentile drought periods were more accurately predicted than variable thresholds, likely due to more predictable seasonal patterns. Preliminary model performance for drought prediction was generally found to be fair to moderate based on Cohen's Kappa metric, which highlights that streamflow drought prediction is a generally challenging prediction problem, especially in the CRB region. However, model (hyperparameter) tuning, loss function optimization for low-flow/percentile conditions, and use of model ensembles are likely to result in improvements that support the potential for this modeling approach. Additionally, the use of more remotely sensed data inputs, forecasted weather data, and/or using nearby streamflow and reservoir information are also likely to improve predictive performance. A USGS data-driven drought prediction project is implementing these modeling improvements to further build early warning capacity for streamflow drought within the Colorado River Basin region.

Acknowledgements

This research was funded by the U.S. Geological Survey (USGS) Water Availability and Use Science Program as part of the Water Resources Mission Area Data-Driven Drought Prediction Project. Computing resources were provided by USGS Cloud Hosting Solutions. Thanks to reviewers for their suggestions and edits that improved this article. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Open data

The data and model code that support the findings of this study are available at doi.org/10.5066/P97NIH7Y

References

Abatzoglou, J.T., 2013, Development of gridded surface meteorological data for ecological applications and modelling: International Journal of Climatology, v. 33, no. 1, p. 121–131.

Addor, N., Newman, A.J., Mizukami, N., and Clark, M.P., 2017, The CAMELS data set: catchment attributes and meteorology for large-sample studies: Hydrology and Earth System Sciences, v. 21, no. 10, p. 5293–5313.

Broxton, P., Zeng, X., and Dawson, N., 2019, Daily 4 km Gridded SWE and Snow Depth from Assimilated In-Situ and Modeled Data over the Conterminous US, Version 1:

Feng, D., Fang, K., and Shen, C., 2020, Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales: Water Resources Research, v. 56, no. 9.

Frame, J.M., Kratzert, F., Raney, A., Rahman, M., Salas, F.R., and Nearing, G.S., 2021, Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics: JAWRA Journal of the American Water Resources Association, p. 1752– 1688.12964.

Hammond, J.C., Simeone, C., Hecht, J.S., Hodgkins, G.A., Lombard, M., McCabe, G., Wolock, D., Wieczorek, M., Olson, C., Caldwell, T., Dudley, R., and Price, A.N., 2022, Going Beyond Low Flows: Streamflow Drought Deficit and Duration Illuminate Distinct Spatiotemporal Drought Patterns and Trends in the U.S. During the Last Century: Water Resources Research, v. 58, no. 9.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G., 2022, Uncertainty Estimation with Deep Learning for Rainfall–Runoff Modelling: Catchment hydrology/Modelling approaches, accessed March 13, 2022, at https://doi.org/10.5194/hess-26-1673-2022.

Knoben, W.J.M., Freer, J.E., and Woods, R.A., 2019, Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores: Hydrology and Earth System Sciences, v. 23, no. 10, p. 4323–4331.

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D., 2022, NeuralHydrology — A Python library for Deep Learningresearch in hydrology: Journal of Open Source Software, v. 7, no. 71, p. 4050.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M., 2018, Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks: Hydrology and Earth System Sciences, v. 22, no. 11, p. 6005–6022.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., and Nearing, G.S., 2019a, Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning: Water Resources Research, v. 55, no. 12, p. 11344–11354.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G., 2019b, Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets: Hydrology and Earth System Sciences, v. 23, no. 12, p. 5089–5110.

Landis, J.R., and Koch, G.G., 1977, The Measurement of Observer Agreement for Categorical Data: Biometrics, v. 33, no. 1, p. 159.

Mitchell, K.E., 2004, The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system: Journal of Geophysical Research, v. 109, no. D7, p. D07S90.

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., and others, 2022, Flood forecasting with machine learning models in an operational framework: Hydrology and Earth System Sciences, v. 26, no. 15, p. 4013–4032.

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., and Shen, C., 2021, Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy: Journal of Hydrology, v. 599, p. 126455.

Shen, C., 2018, A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists: Water Resources Research, v. 54, no. 11, p. 8558–8593.

Shen, C., Chen, X., and Laloy, E., 2021, Editorial: Broadening the Use of Machine Learning in Hydrology: Frontiers in Water, v. 3, p. 681023.

Simeone, C.E., 2022, Streamflow Drought Metrics for Select United States Geological Survey Streamgages for Three Different Time Periods from 1921 - 2020: U.S. Geological Survey data release, https://doi.org/10.5066/P92FAASD.

Smyl, S., 2017, Ensemble of specialized neural networks for time series forecasting. In 37th international symposium on forecasting.

Subramanian, V., 2018, Deep learning with PyTorch: a practical approach to building neural network models using PyTorch: Packt Publishing, Birmingham, UK.

Sutanto, S.J., and Van Lanen, H.A.J., 2021, Streamflow drought: implication of drought definitions and its application for drought forecasting: Hydrology and Earth System Sciences, v. 25, no. 7, p. 3991–4023.

Van Loon, A.F., 2015, Hydrological drought explained: WIREs Water, v. 2, no. 4, p. 359–392.

U.S. Geological Survey., USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed February 22, 2022, at http://dx.doi.org/10.5066/F7P55KJN

Wheeler, K.G., Udall, B., Wang, J., Kuhn, E., Salehabadi, H., and Schmidt, J.C., 2022, What will it take to stabilize the Colorado River? Science, v. 377, no. 6604, p. 373–375.

Wieczorek, M.E., Hafen, K.C., and Staub, L.E., 2023, Data-Driven Drought Prediction Project Model Inputs for Upper Colorado Portion of the National Hydrologic Geo-Spatial Fabric version 1.1 and Select U.S. Geological Survey Streamgage Basins:

Wieczorek, M.E., Schwarz, G.E., and Jackson, S.E., 2018, Select Attributes for NHDPlus Version 2.1 Reach Catchments and Modified Network Routed Upstream Watersheds for the Conterminous United States: U.S. Geological Survey data release, https://doi.org/10.5066/F7765D7V.

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., and Shen, C., 2021, Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships: Journal of Hydrology, v. 603, p. 127043.