# A Hybrid Physics–AI Model to Improve Hydrological Forecasts

YANAN DUAN,[a] SATHISH AKULA,[b] SANJIV KUMAR,[a] WONJUN LEE,[c] AND SEPIDEH KHAJEHEI[d]

[a] *Earth System Science Program, College of Forestry, Wildlife, and Environment, Auburn University, Auburn, Alabama*
[b] *Department of Computer Science and Software Engineering, Auburn University, Alabama*
[c] *Department of Computer Science, California State University, Northridge, California*
[d] *Descartes Labs, Inc., Santa Fe, New Mexico*

(Manuscript received 26 March 2022, in final form 19 September 2022)

ABSTRACT: The National Oceanic and Atmospheric Administration has developed a very high-resolution streamflow forecast using National Water Model (NWM) for 2.7 million stream locations in the United States. However, considerable challenges exist for quantifying uncertainty at ungauged locations and forecast reliability. A data science approach is presented to address the challenge. The long-range daily streamflow forecasts are analyzed from December 2018 to August 2021 for Alabama and Georgia. The forecast is evaluated at 389 observed USGS stream gauging locations using standard deterministic metrics. Next, the forecast errors are grouped using watersheds' biophysical characteristics, including drainage area, land use, soil type, and topographic index. The NWM forecasts are more skillful for larger and forested watersheds than smaller and urban watersheds. The NWM forecast considerably overestimates the streamflow in the urban watersheds. The classification and regression tree analysis confirm the dependency of the forecast errors on the biophysical characteristics. A densely connected neural network model consisting of six layers [deep learning (DL)] is developed using biophysical characteristics, NWM forecast as inputs, and the forecast errors as outputs. The DL model successfully learns location invariant transferrable knowledge from the domain trained in the gauged locations and applies the learned model to estimate forecast errors at the ungauged locations. A temporal and spatial split of the gauged data shows that the probability of capturing the observations in the forecast range improved significantly in the hybrid NWM-DL model (82% ± 3%) than in the NWM-only forecast (21% ± 1%). A trade-off between overly constrained NWM forecast and increased forecast uncertainty range in the DL model is noted.

SIGNIFICANCE STATEMENT: A hybrid biophysical–artificial intelligence (physics–AI) model is developed from the first principle to estimate streamflow forecast errors at ungauged locations, improving the forecast's reliability. The first principle refers to identifying the need for the hybrid physics–AI model, determining physically interpretable and machine identifiable model inputs, followed by the deep learning (DL) model development and its evaluations, and finally, a biophysical interpretation of the hybrid model. A very high-resolution National Water Model (NWM) forecast, developed by the National Oceanic and Atmospheric Administration, serves as the biophysical component of the hybrid model. Out of 2.7 million daily forecasts, less than 1% of the forecasts can be verified using the traditional hydrological method of comparing the forecast with the observations, motivating the need for the AI technique to improve forecast reliability at millions of ungauged locations. An exploratory analysis followed by the classification and regression tree analysis successfully determines the dependency of the forecast errors on the biophysical attributes, which along with the NWM forecast, are used for the DL model development. The hybrid model is evaluated in a subtropical humid climate of Alabama and Georgia in the United States. Long-term streamflow forecasts from zero-day lead to 30-day lead forecasts are archived and analyzed for 979 days (December 2018–August 2021) and 389 USGS gauging stations. The forecast reliability is assessed as the probability of capturing the observations in its ensemble range. As a result, the forecast reliability increased from 21% (±1%) in the NWM only forecasts to 82% (±3%) in the hybrid physics–AI model.

KEYWORDS: Seasonal forecasting; Hydrologic models; Data science

## 1. Introduction

### a. Big picture

This study addresses a big-data challenge in near-term hydrological forecasting at its "tipping point" for technological innovation. The National Water Model (NWM), developed by the National Oceanic and Atmospheric Administration (NOAA), provides 0–30 days streamflow forecast data for 2.7 million streams across the conterminous United States (Gochis et al. 2018; Hooper et al. 2017). However, there are only 10 330 gauging stations where observations are available (Eberts et al. 2019). That means the traditional hydrological

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-22-0023.s1.

Yanan Duan and Sathish Akula contributed equally as first authors to the manuscript.

*Corresponding author*: Sanjiv Kumar, szk0139@auburn.edu

DOI: 10.1175/AIES-D-22-0023.1 e220023

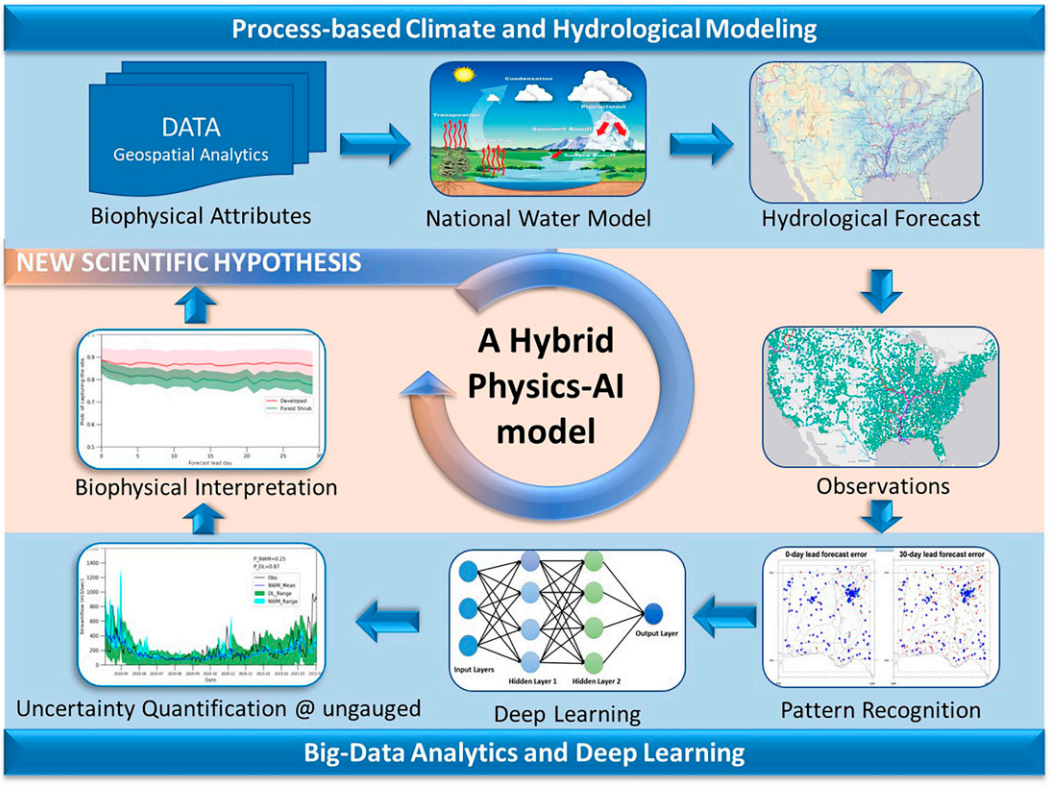


FIG. 1. A hybrid physics–AI model for improving hydrological forecast. Data science workflow and connections among different components are shown. A two-way workflow is shown here: Process-based climate and hydrological modeling provide machine identifiable and physically interpretable quantity to the deep learning model, which in turn provide an improved forecast, and its biophysical interpretation can develop a new scientific hypothesis. Each of these components is described in detail in the subsequent figures and main text.

method of comparing the forecast with the observation for uncertainty quantification is not applicable for 99.6% of the forecasted sites. Further, it is humanly impossible to understand the error structure and calibrate the model performance across many sites using manual routines. Finally, the forecast data are only available in real time (for two days only) and then removed from the system because of high data volume. Hence, a real-time forecast evaluation and adaptive model calibration routine capability are required.

A long-term goal of this study is to develop an efficient computational framework (Fig. 1) that provides a set of statistical and real-time deep learning models for uncertainty quantification in large-scale hydrological forecast data and analyses. The deep learning method can provide scientific breakthroughs in hydrological forecasting (Shen et al. 2018). Especially when there is not enough training dataset with labels (observations) due to the lack of expensive infrastructure to collect data, the deep learning method can still provide high performance with transfer learning. It has been widely used in the small-data setting and can transfer knowledge of feature representation (e.g., knowledge of data) or knowledge of parameters (e.g., knowledge of model; Li et al. 2021). In this work, the deep learning model learns location invariant transferrable knowledge from the domain trained in the gauged sites and applies the learned model to improve hydrological forecasts at the ungauged sites. We hypothesize that the domain invariant properties can be related to biophysical attributes, for example, topography, land use, soil types, and initial condition and climate forcing uncertainty in the forecast (Fig. 2). Hence a systematic evaluation of deep learning methods for hydrological forecasting application can bring theoretical advances and feedback to the hydrological model improvement (Fig. 1).

Processing of high-resolution NWM forecasts demands highly intensive resources in terms of storage and computation. In addition to restricted storage, big-data computation for analysis takes a long time due to slow input/output (I/O) operations. Furthermore, many I/O operations are required to train or validate datasets for uncertainty quantification using deep learning. Therefore, if the assessment and uncertainty quantification is provided as a service proposed here, the conventional hydrological applications are not affordable for an unknown number of service requests since it is very hard to *scale out* (Dragoni et al. 2017). Furthermore, the conventional server-side applications are a single executable artifact (i.e., monolith) whose modules cannot be executed independently. This makes monolithic applications challenging to use in a container-based cloud computing system (Dragoni et al. 2017). Here, a new analytics platform supported by container-based software-defined storage (CB-SDS)

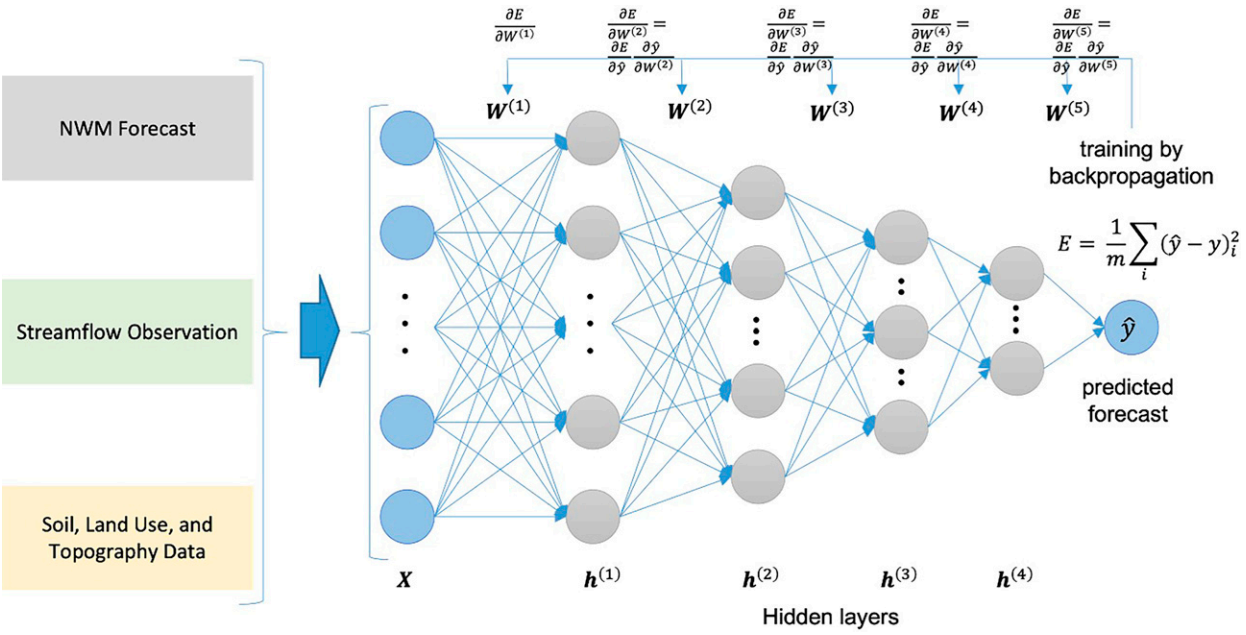


FIG. 2. A general setup of deep neural network for learning error models at gauged watersheds.

can be helpful. The CB-SDS technology provides data storage space to end users and end-point applications dynamically and flexibly using containers while hiding the complexity of storage resource management (Lee and Kumar 2016). The new analytics platform will be described in part two of this work.

*b. This study*

Part one of this work describes a scientific basis and initial results for a testbed region in the southeastern United States (Fig. 3). We have downloaded and archived the NWM forecast data for 1000 days from December 2018 to August 2021 (a total of 979 days). The forecast skills are assessed in the humid subtropical climate at 389 gauged locations in Alabama and Georgia. The three research objectives are 1) to evaluate the NWM streamflow and soil moisture forecast; 2) to investigate the relationship between watershed biophysical attributes and forecast errors; 3) to develop a deep learning model for uncertainty quantification at ungauged basins.

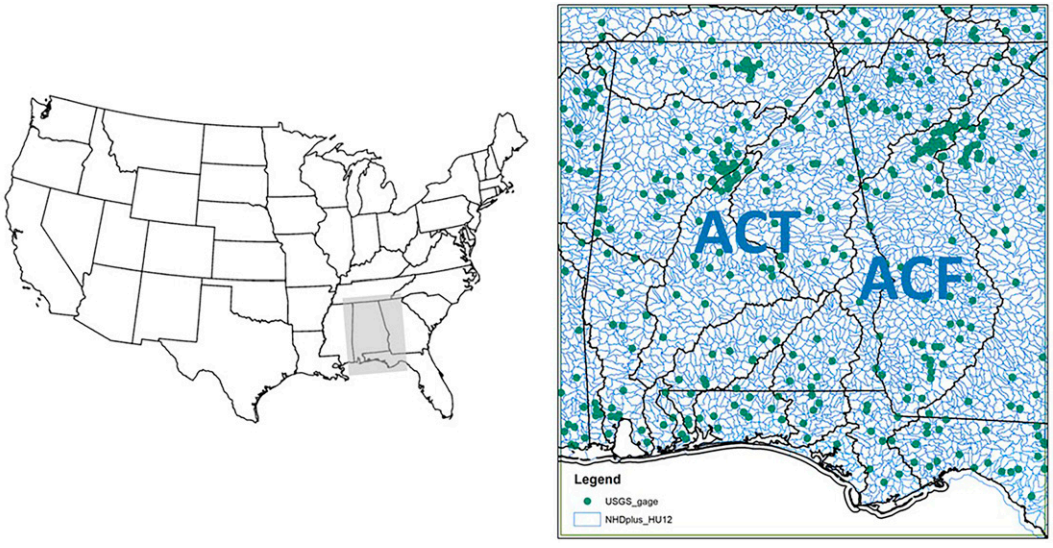


FIG. 3. Study area—Alabama and western Georgia in the United States, USGS gauges, and local watershed boundaries. The figure shows the study area overlaid with HUC12 (local watersheds), HUC6 watershed boundaries (regional watershed), and USGS stream gauging stations (gray dots). Two major river basins, the ACT and ACF, are also shown.

TABLE 1. List of the datasets.

| Dataset | Source | Remarks |
|---|---|---|
| Streamflow and soil moisture forecast | NOAA NWM Forecast | 979 days (December 2018–August 2021)[a] |
| Streamflow observations | USGS | 454 USGS gauges[b] |
| Soil moisture observations | SMAP-L2 | 9-km resolution[c] |
| Biophysical characteristics | | |
| Drainage area | USGS StreamStats | 389 watersheds[d] |
| Land use | NWM WRF processing system | The NWM domain setup file was downloaded from NWM shared FTP point. |
| Soil type | | |
| Topographic index | Elevation data | |
| Distance from coastline | Computed in ArcGIS | It is the shortest distance from gauges to the gulf coastline. |

[a] The channel routing file "nwm.tHHz.long_range.channel_rt_M.fLLL.conus.nc" provides streamflow outputs; and land model output file "nwm.tHHz.long_range.land_M.fLLL.conus.nc" provides soil moisture output. HH refers to the forecast initialization hour: 00, 06, 12, and 18; M refers to the ensemble member (1–4), and LLL refers to the forecast lead hours (0–720). We used the "COMID" attribute to join the NWM reach ID with the corresponding USGS gauging station and using National Hydrographic Dataset flowlines.
[b] The 454 USGS gauges are shown in Fig. 3.
[c] The SMAP soil moisture is L2 half-orbit enhanced 5-cm soil moisture with 9-km resolution, available from Descartes Lab.
[d] Quality-controlled watershed delineation where the difference between the StreamStats and USGS drainage area is less than 5%.

Prediction at ungauged basins has been studied in hydrological science (Hrachowitz et al. 2013). The fundamental concept is to transfer model parameters from a hydrologically similar basin to the ungauged location, also known as the regionalization approach (Wagener and Wheater 2006). The homogeneity between gauged and ungauged basins is essential in the regionalization approach. They include homogeneous climate forcing, geology, landform, land use, and soil texture (Sivapalan et al. 2003). Singh et al. (2014) used classification and regression tree (CART) analysis to determine the relationship between watershed characteristics and the performance of the transfer model. They found that similarity in elevation, climate, and streamflow characteristics (base flow/runoff) was the dominant control for the successful parameter transfer across 83 watersheds in the United States.

The application of machine learning (ML) to improve hydrological and climate forecasts have increased. Frame et al. (2021) have developed a postprocessing tool for NWM forecast using a long-term short-term memory model, an ML technique. Kratzert et al. (2019) found improvement in ML model performance with the catchment attributes as the additional input parameters (in addition to meteorological inputs). Barnes et al. (2019) have employed an artificial neural network (ANN) to extract climate change signals from the model uncertainty and internal climate variability. Mayer and Barnes (2021) have used the ANN to identify teleconnection patterns that can potentially improve subseasonal forecasts in the midlatitude regions. This study uses big-data techniques to 1) understand the forecast error structure and 2) build a deep learning model to quantify uncertainty at ungauged watersheds. Thus, we utilize the strengths of both process-based modeling (NWM) and big-data technology to improve hydrological forecasting.

## 2. Data and method

The study involves archiving and preprocessing the NWM forecasts, preparing explanatory variables representing the biophysical characteristics, analyzing spatiotemporal characteristics of the forecast errors at gauged locations, relating the forecast errors to the biophysical attributes, and finally developing the deep learning model for ungauged watersheds.

### *a. Study area*

We developed a prototype test bed in Alabama and Georgia (Fig. 3), representing the humid subtropical climate in the southeastern United States. Alabama–Georgia has rainfall-dominated hydrology with wet winter and dry fall seasons with an annual average rainfall of 1407 mm (PRISM climate data). Major land-cover types: forest and woodland (47.4%), agricultural vegetation (7.2%), shrub and hay (19.2%), developed land (8.6%), and open water (4.3%) [source: National Land Cover Database (NLCD) 2016]. We have overlaid the study area with a 12-digit Hydrologic Unit Code (HUC12), that is, local subwatershed boundaries. The test bed consists of 441 active U.S. Geological Survey (USGS) gauging stations and 3592 HUC12 watersheds, that is, 12.2% of all HUC12 are instrumented.

### *b. Datasets*

Table 1 lists the datasets employed in this study.

#### 1) NWM FORECAST

The NWM provides a high-resolution (1 km) streamflow forecast at subcontinental scales (Gochis et al. 2013). Components of NWM include the Noah land surface model with multiparameterization (Noah-MP), subsurface and terrain routing module, and gridded diffusive wave channel and reservoir routing modules. Our earlier work (Duan and Kumar 2020) describes the model components (also see Gochis et al. 2013). Here, we describe the streamflow forecast attributes.

We have evaluated the long-range streamflow forecast (0–30 days) from December 2018 to August 2021. Each day, a

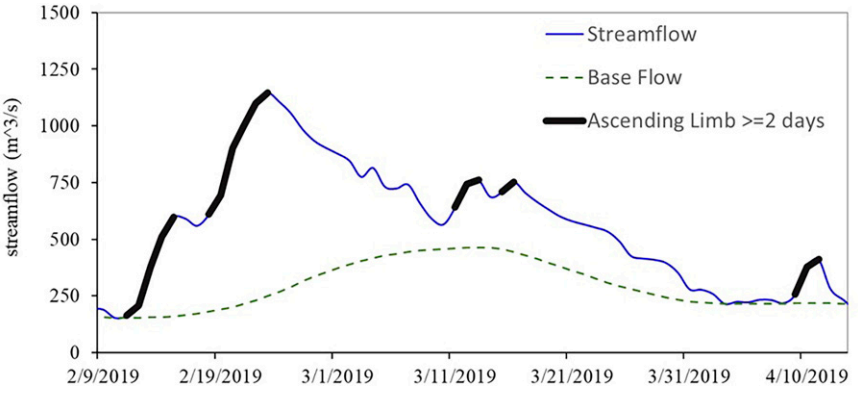


FIG. 4. The hydrograph characteristics: time-to-peak flow and the baseflow component of the streamflow. Observed streamflow at USGS Gauge ID: 02397000, Coosa River (Mayo's Bar) near Rome is shown. Ascending limb ≥ two days were included in the time-to-peak calculation, and base flow is the contribution of the subsurface flow to the streamflow (see text).

16-member ensemble forecast is initialized with the observationally constrained initial condition of streamflow and soil moisture states and forecasted for the next 30 forecast days using meteorological forcing from Climate Forecast System, version 2, data (Saha et al. 2014). The NWM daily forecast is available at NOAA NCEP central operations (https://www.nco.ncep.noaa.gov/pmb/products/nwm/). However, because of high data volume, the forecast data are only available in real time (for two days only) and then removed from the NOAA FTP folder. So, we download the NWM soil moisture and streamflow forecast daily and archive data for the research area (~10 TB data).

#### 2) STREAMFLOW AND SOIL MOISTURE OBSERVATIONS

The observed streamflow is obtained from the USGS. Soil moisture data are Soil Moisture Active Passive (SMAP) L2 half-orbit enhanced 5-cm soil moisture with 9-km resolution (access provided by the Descartes Lab). It is derived from the enhanced SMAP L1C_TB_E product (Chan et al. 2018). The daily soil moisture data are averaged across all available images in one day.

#### 3) BIOPHYSICAL CHARACTERISTICS

To compute basin-average biophysical attributes, we obtained the watershed boundary shapefile from USGS StreamStats (https://streamstats.usgs.gov/ss/). The watershed boundary having a drainage area that matches with the USGS drainage area (bias < 5%) is included in the analysis; 389 watershed boundaries meet the criteria (out of 454 gauges). Most of the watersheds along the coastal region were not delineated well. The gridded land use (NLCD 2016; Homer et al. 2020) and soil texture (CONUS-SOIL; Miller and White 1998) data are from the Weather Research and Forecasting processing system and clipped to the watershed boundaries to obtain the biophysical characteristics.

### c. *Hydrograph characteristics*

We used "base flow to total flow ratio" and "time to peak" to evaluate hydrograph characteristics (Fig. 4). The baseflow component is supplied by subsurface/groundwater drainage, which is relatively more stable than the stormflow/surface runoff component driven by rainfall events. We computed base flow using the soil and water assessment tool baseflow filter program that uses a recursive digital filter to separate the base flow from the daily streamflow record (Arnold et al. 1995; Arnold and Allen 1999). The baseflow filter program comes as a stand-alone program with daily streamflow observation or forecast as inputs and base flow as outputs; we used the base flow from the third-pass filter (see Fig. 4) (more details in supplemental text T1).

The "time to peak" is a critical parameter for flood forecasting, and it measures the time (in days) for an $e$-fold increase in the streamflow after a rainfall event. We used a method similar to the baseflow recession constant method (Vogel and Kroll 1996) applied for the ascending limb of the hydrograph, that is, $(Q_t - Q_{t-1}) > 0$ with the following assumptions $dQ/dt = Q_t - Q_{t-1}$, $Q = (Q_t + Q_{t+1})/2$, error is distributed normally in the log space. We used a continuous ascending threshold of 2 days or greater (Fig. 4):

$$\mathrm{t2p} = \exp\left[-\frac{1}{m}\sum_{t=1}^{m}\ln\frac{Q_t - Q_{t-1}}{(Q_t + Q_{t+1})/2}\right], \tag{1}$$

where $Q_t$ is observed or simulated streamflow on the $t$th day.

### d. *Evaluation metrics*

The metrics to evaluate model performance are anomaly correlation coefficient (ACC) [Eq. (2)], normalized root-mean-square error (nRMSE) [Eq. (3)], and biases percentage (PBIAS) [Eq. (4)];

$$\mathrm{ACC} = \frac{\overline{(f - c_f)(o - c_o)}}{\sqrt{\overline{(f - c_f)^2} \times \overline{(f - c_o)^2}}}, \tag{2}$$

where $f$ is forecast, $c_f$ is forecast climatology, $o$ is observation, $c_o$ is observation climatology, and overbar denotes the average quantity:

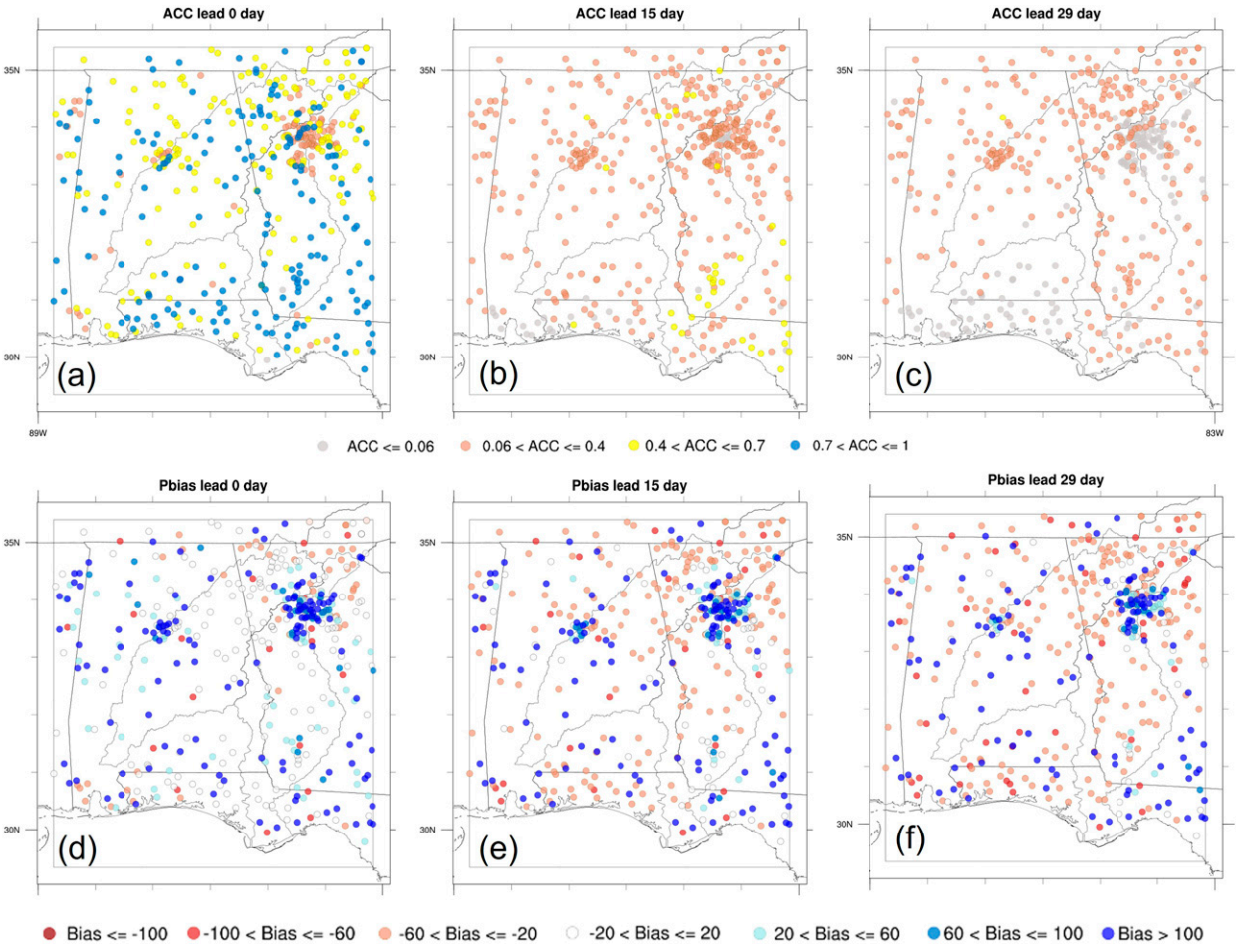


FIG. 5. The NWM evaluation at 389 USGS gauges in the Southeast. The first row shows the ACC between the NWM forecast and observations at (a) 0-, (b) 15-, and (c) 29-day leads. (d)–(f) As in (a)–(c), but for the PBIAS between NWM and observation for the same lead time.

$$\text{nRMSE} = \frac{\sqrt{1/n \times \sum_{i=1}^{n} (f_i - o_i)^2}}{c_o}. \tag{3}$$

PBIAS [Eq. (4)] is a relative level of the streamflow over/underestimating. The positive value indicates overestimating.

$$\text{PBIAS} = 100 \times \frac{\sum_{i=1}^{N} (f_i - o_i)}{\sum_{i=1}^{N} o_i}. \tag{4}$$

*e. AI*

This is a broader effort that automates intellectual tasks normally performed by humans. A subset of artificial intelligence (AI) is machine learning, which uses traditional methods such as mathematical statistics and rules to improve model performance from data (Sarker 2021). In this work, we used a CART to investigate the dependency of the forecast errors on biophysical attributes. CART is a traditional machine learning method that recursively partitions the data space into two branches at each node using the Gini impurity criterion, and the root node is recursively divided to have the highest impurity (Loh 2011). We used the CART R package "rpart" (version 4.1–15), with four independent variables (land use, soil texture, drainage area, topographic index), and nRMSE or ACC as the dependent variables. The CART parameter is the package default.

A subset of machine learning is deep learning (DL), which emphasizes learning successive layers of increasingly meaningful representations (Sarker 2021). We developed a deep learning model as a function of biophysical attributes and NWM forecast (inputs) to predict forecast errors (outputs) at ungauged basins where observations are unavailable. However, the biophysical attributes and NWM forecast are available [Eq. (5) and Fig. 2]:

$$\text{abs}(f_i - O_i) = \begin{bmatrix} W_1 & \cdots & W_m & W_f \end{bmatrix} \begin{bmatrix} \text{BP}_1 \\ \vdots \\ \text{BP}_m \\ f_i \end{bmatrix}. \tag{5}$$

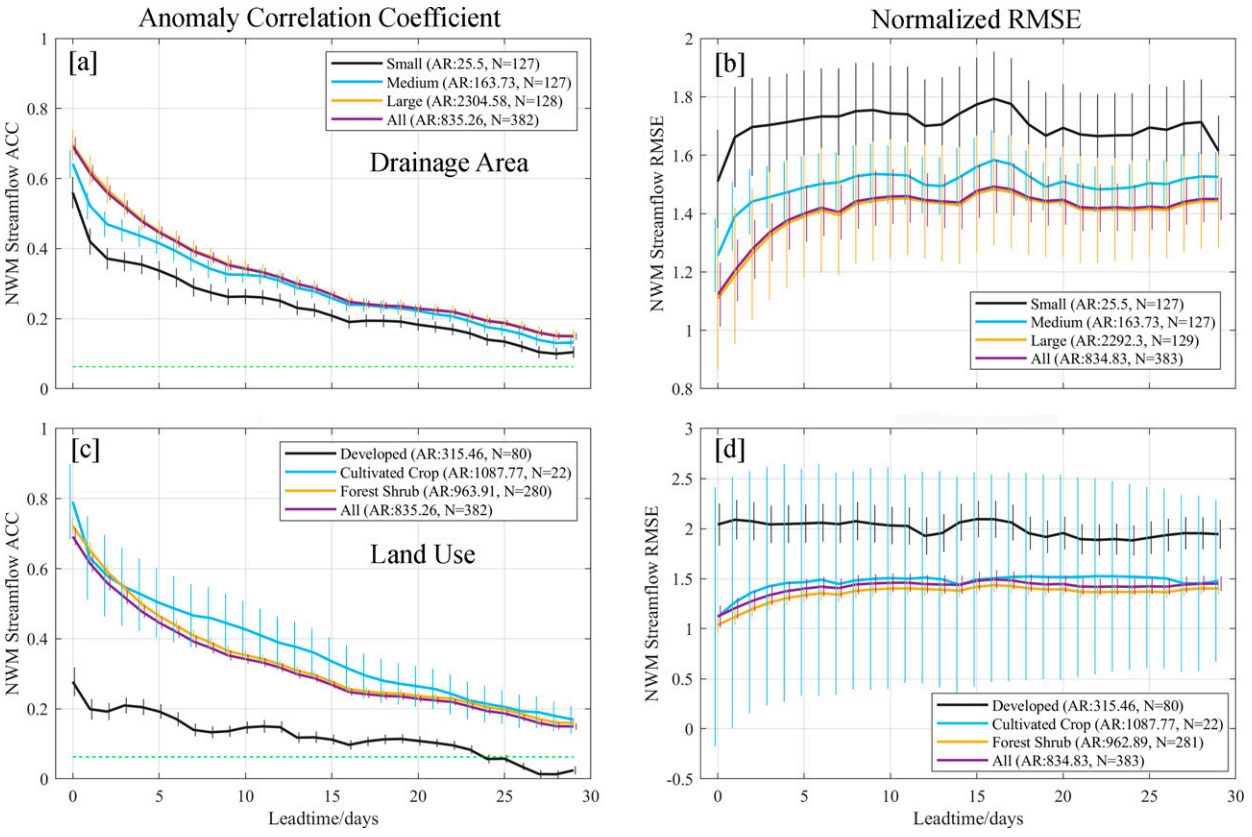


FIG. 6. Effects of biophysical attributes (drainage area and land use) on streamflow forecast skill using ACC and nRMSE metric. (a),(b) Watershed grouped under small, medium, and large categories; their respective average AR and sample size ($N$) are shown in the legend. The area unit is square miles. (c),(d) Watershed grouped under major land-use categories: forested, agriculture/cultivated crop, and developed/urban watersheds. The bar is 95% uncertainty range with a normal distribution assumption.

Here, $W_1, \ldots, W_m$, and $W_f$ are the biophysical and the flow coefficients obtained from the DL model and $BP_1, \ldots, BP_m$, and $f_i$ are the corresponding biophysical attributes, NWM forecast at a given lead time, and the watershed, and $m = 13$ corresponding to land use, soil texture classes for 10 different depths, topographic index, and drainage area.

The computation process of the deep learning model goes through several layers, as shown in Fig. 2. We used a densely connected neural network consisting of six layers: one input layer, four intermediate hidden layers, and one output layer (Fig. 2). The first layer receives the NWM streamflow forecast and biophysical attributes as input features and then extracts more meaningful representations out of input. This extracting process of representation progresses at the following four intermediatory layers. Finally, the last layer computes the output value: the absolute difference between NWM and observed flow. The first and intermediate layers use rectified linear unit (ReLU) that zeroes out negative values as their activation function. The ReLU activation function provides the model with a much richer hypothesis space that would benefit from deep representations since it introduces nonlinearity into the network (Nair and Hinton 2010). The difference between the network's predicted output and the actual output is continuously reduced by training the densely connected neural network with the observation dataset consisting of approximately 150 000 data samples.

We used the RMSprop optimizer, a gradient-based optimization technique, and a stochastic technique for minibatch learning (Tieleman and Hinton 2012). RMSprop deals with vanishing gradients of very complex neural networks by using the average of squared gradients to normalize the gradients; this means that the learning rate changes over time. Finally, the network is compiled with mean square error as the loss function and the mean absolute error (MAE) as the model evaluation metric.

### 1) DL HYPERPARAMETERS

Our DL model sets the batch size as 1024 and the number of epochs as 13 to achieve optimum model performance (Fig. S1 in the online supplemental material) and prevent overfitting. We determined these hyperparameters after several trials to achieve similar mean absolute error performance in the training and validation data. The validation MAE drops initially, then if the validation MAE starts increasing and the difference between validation MAE and training MAE is greater than the training MAE, the overfitting starts (Chollet 2018). Further, we selected 1024 as the batch size because our training data samples are large (~150 000; each day and each site forecast were treated as individual data sample). Therefore, we determined

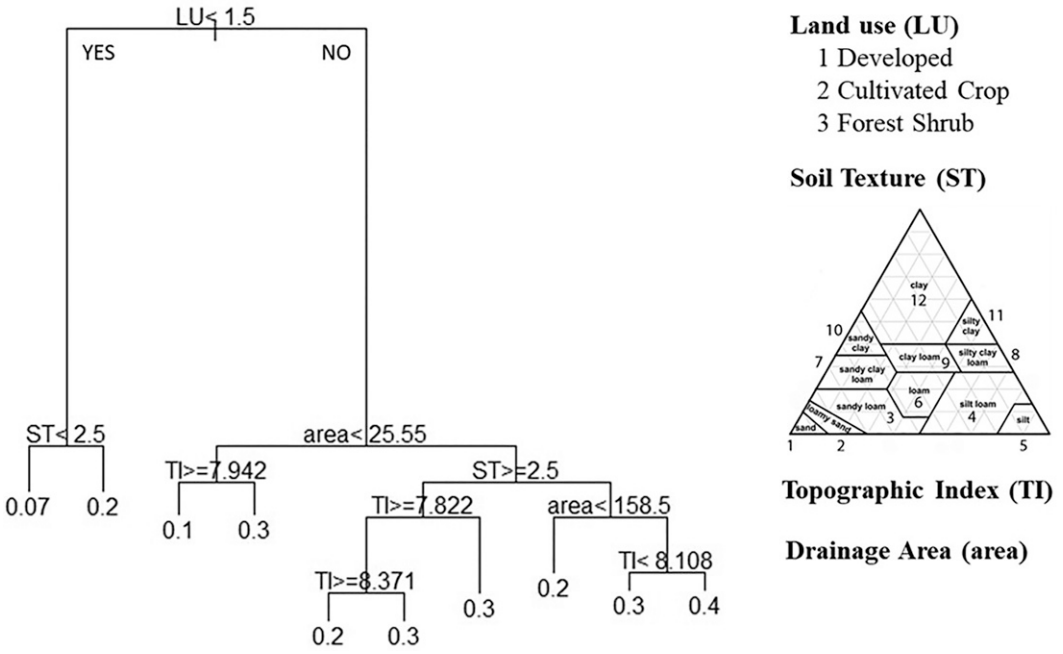


FIG. 7. The NWM performance dependency on biophysical attributes classified using CART. The CART results for the streamflow ACC of the 14-days lead forecast. The left branch is yes, and the right branch is no in each bifurcation.

the optimal batch size from empirical experiments without any memory issues and big degradation of generalization. A full description of the model in the form of a Jupyter notebook is appended in the supplemental materials.

#### 2) DL MODEL EVALUATION

We evaluated the DL model performance using a two-part analysis where part 1 divides the observations temporarily, and part 2 divides it spatially. For temporal split analysis, the first half of streamflow observations (11 December 2018–2 April 2020) is treated as the gauged (labeled data), and the second half (3 April 2020–15 August 2021) as the ungauged (predicted data); and therefore, assessing the DL model performance for ungauged data. Additionally, we developed a separate DL model for each lead day (0–29 lead days) that incorporates lead-time dependency, that is, smaller forecast errors at the shorter lead time versus a larger forecast error at a longer lead time.

For spatial split analysis—we split the data into 90% gauged and 10% ungauged data for the entire period (11 December 2018–15 August 2021) and repeated the process 10 times to sample all stations in model evaluations. We also investigated the sensitivity of the DL model performance for various split levels, for example, 65/35, 70/30, …, 95/5 (gauged/ungauged), and found that the DL model is not sensitive to the spatial split level, generally (not shown), and selected 90/10 as the optimal level.

## 3. Results

### *a. The NWM streamflow forecast skill and its dependency on biophysical attributes*

The NWM can provide potentially skillful forecasts showing spatial and temporal dependency (Figs. 5 and 6). Generally, shorter lead-time forecasts are more skillful than longer lead times. Figure 5 compares the NWM forecast with USGS observations at 389 gauges and 0, 14, and 29 lead days using ACC and percentage biases metric. The points are the corresponding watershed outlets, and their relative upstream/downstream position does not necessarily represent the increasing drainage area, that is, the two near points may have much different basin areas. As expected, the forecast skill decreases with increasing lead time. However, the ACC remains statistically significant for most watersheds (300 out of 389) at a 30-day lead.

Spatial clustering in the forecast biases (Figs. 5d–f) suggests the forecast skill's dependency on biophysical attributes. There are two large biases (>100%) clusters: 1) in the Apalachicola–Chattahoochee–Flint (ACF) northeast basin boundary with a small drainage area, and 2) in the Alabama–Coosa–Tallapoosa (ACT) basin central-west boundary. The clusters represent the most prominent cities: Birmingham (Alabama) and Atlanta (Georgia). There are also other large biases forecasts scattered in the study area. Overall, 24% of watersheds (93 out of 389 watersheds) show large biases, which develop at 0 days and remain large throughout the forecast period (Figs. 5d–f).

An exploratory analysis of the forecast skill confirms its dependency on biophysical attributes. Figure 6 shows the ACC and nRMSE for the watershed grouped as per their drainage area (AR) and land-use types. A higher ACC and smaller nRMSE show better skill in the forecast. The NWM forecasts are less skillful for small watersheds (AR: 25.5 $mi^2$) than medium (AR: 164 $mi^2$) and large (AR: 2305 $mi^2$) watersheds. The NWM forecast performs poorly for the urban/developed watersheds compared with the agriculture and forested watersheds (Figs. 6c,d). This is a

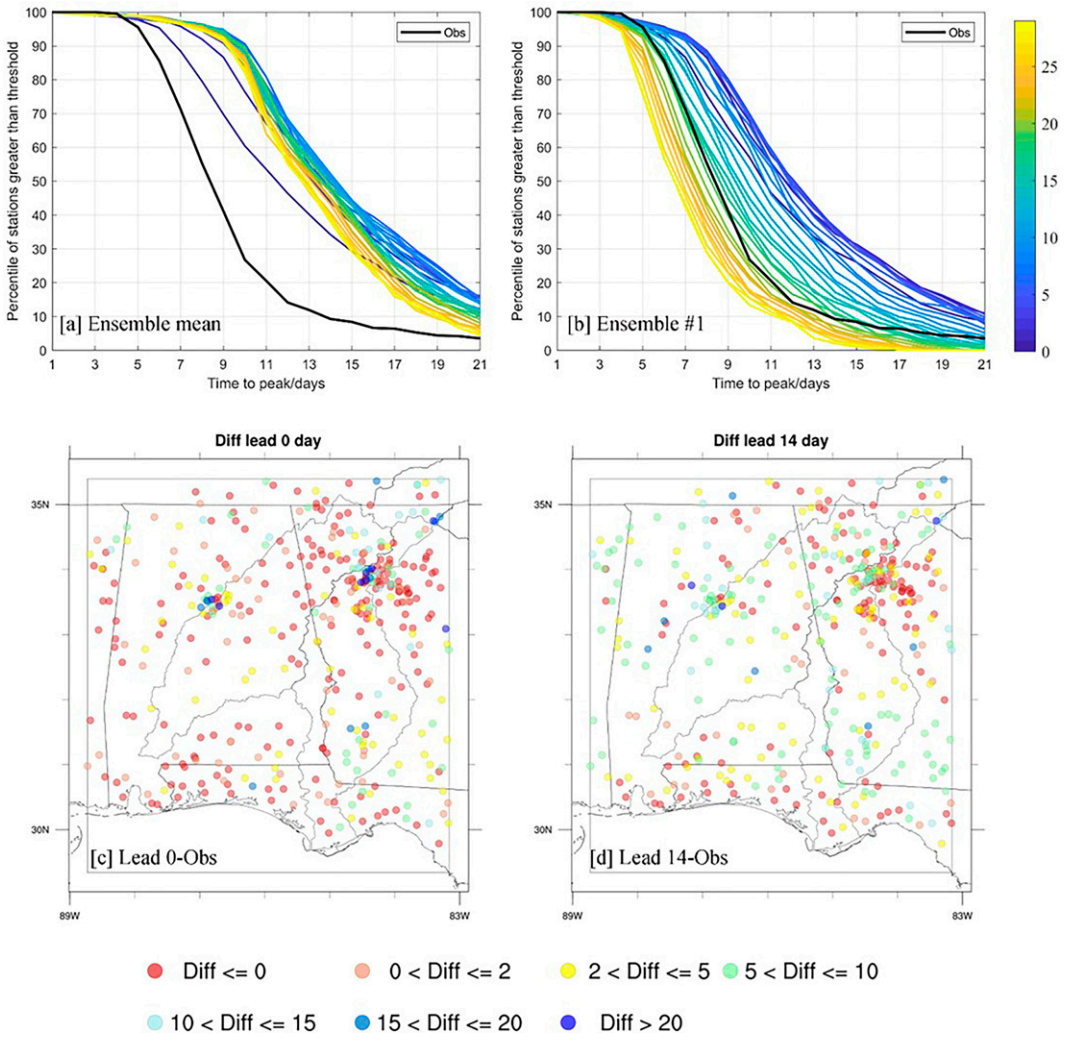


FIG. 8. Time to peak in the NWM forecast. (a) Time to peak of NWM 16 ensemble mean and observation. (b) Time to peak of NWM first ensemble member and observation. The color bar is forecast lead day from 0 to 29. (c) The forecast time to peak minus observation for the lead zero-day forecast. (d) The forecast time to peak minus observation for the lead 14-day forecast.

discouraging result for NWM forecast application in urban flooding. The urban/developed watersheds show a significantly smaller ACC (less than 0.2) and larger nRMSE (>2.0) than forested or agricultural watersheds. However, most watersheds (282 out of 389) are forested (major land-cover types); hence the overall skill looks similar to the forested watershed skill.

A generally higher nRMSE (>1) suggests considerable biases in the NWM forecast (Figs. 6b,d), and most of these biases develop at early lead (e.g., 0-day lead). For large and forested watersheds, the nRMSE increases from ~1.1 at 0-day lead to ~1.5 at 10-day lead, then it remains relatively stable. The higher forecast biases can be related to the biases in seasonal climate forecast data (CFSv2) that provide climate forcing for the NWM model forecast. Duan and Kumar (2020) used the observed meteorological forcing to drive the NWM and found smaller biases. Figures 6b and 6d emphasize the need for developing a biases correction methodology for the NWM forecast (outside our scope of the study).

### b. *CART of forecast errors*

The CART algorithm identifies land use (LU), soil texture (ST) for the first layer (0–5 cm), drainage area (area), and topographic index (TI) as major determinants for forecast skill classification, that is, forecast skills are significantly different between the partitioned groups (Fig. 7). Despite the large forecast errors (e.g., Figs. 6b,d), the CART successfully identifies the forecast skill's biophysical dependency. This is an encouraging result and provides a scientific basis for AI/ML applications. As expected from the exploratory analysis (Fig. 6), the land-use type is the primary determinant, and forested and

agricultural watersheds show better performance than developed watersheds.

The CART results are generally not sensitive to the lead-time dependency and evaluation metrics (nRMSE and ACC). For example, land-use type remains a primary determinant for classifying the forecast errors across both nRMSE and ACC metrics and at all lead times (supplemental Figs. S2, S3). ACC is sensitive to temporal variability but insensitive to the magnitude of forecast error (Smith et al. 2019). The squared ACC is interpreted as the potential forecast skill rather than the real skill, especially compared with mean square error (Murphy and Epstein 1989). Factor sequence determining the tree classification using the nRMSE metric is land use, topographic index, and soil texture, whereas using the ACC metric is land use, soil texture, area, and topographic index (supplemental Figs. S2, S3).

### c. *Hydrograph characteristics*

The water moves slower in the NWM ensemble mean forecast (Fig. 8). For example, the median time to peak is 8.5 days for the observations but 11.5–14.5 days for the NWM ensemble mean forecast (Fig. 8a). The ensemble mean (16-member ensemble average) forecast moves slower than the individual ensemble forecast, which shows comparable performance for longer-lead forecasts (Fig. 8b), suggesting biases in time to peak can be related to initial condition effects that have a more substantial influence on the shorter lead time (Duan and Kumar 2020). The spatial clustering of time-to-peak biases is also noted.

The NWM performs better for the base flow to total flow ratio, as most stations (forested) show biases within ±10% (Fig. 9). The base flow to total flow ratio bias does not show lead-time dependency. The NWM overestimates the base flow to total flow ratio in the urban watershed by 20%–40% (Fig. 9b), and it underestimates the ratio in the agricultural watersheds by a smaller magnitude (~10%). Base flows are also underestimated in the coastal watersheds (Fig. 9a and supplemental Fig. S4).

### d. *Soil moisture forecast evaluation*

A gridded evaluation of soil moisture forecast supports biophysical and lead-time dependency of the hydrological forecast (Figs. 10 and 11). We regridded NWM soil moisture forecast (1 km) to the SMAP resolution (9 km) using the area-average method (see supplemental text T2 for details). Statistically significant ACC between NWM forecast and SMAP data are found up to 30 days lead forecast; however, it decreases with the increasing lead time. There is a considerable drop in the ACC from 0th-day lead (0.65 for all) to 1st-day lead (0.45) soil moisture forecast; then it drops smoothly to 0.28 at the 29th-day lead forecast (Fig. 11a). The urban area has a smaller skill (see Atlanta and Birmingham's area marked with circles) than forested and agricultural areas.

Despite a gradual drop in the ACC, the nRMSE did not increase with the increasing lead time for the soil moisture

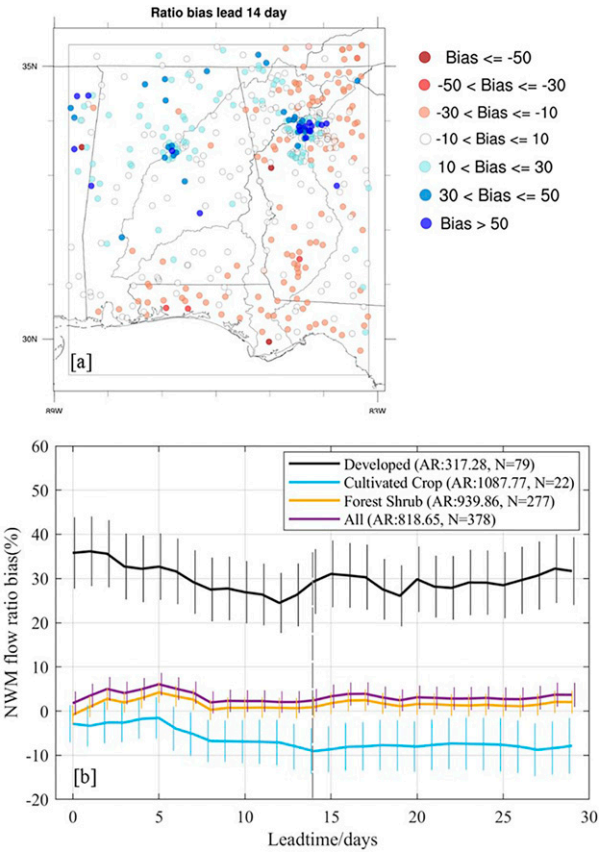


FIG. 9. Base flow to streamflow (flow ratio) ratio in the NWM. Evaluation of base flow to total flow ratio in the NWM forecast. (a) Spatial distribution of the ratio biases for 14-days lead NWM forecast and (b) the ratio biases grouped by land-use categories. Their respective AR and sample size (*N*) is shown in the legend. The area unit is a square mile.

forecast (Fig. 11b). In fact, there is a gradual decrease in nRMSE with increasing lead time. These results can be related to the uncoupled model configuration, that is, the NWM soil moisture does not feedback to the CFSv2 climate forecast model. Because of the decreasing influence of the initial soil moisture anomalies and the same climate forcing (CFSv2), the nRMSE may decrease with increasing lead time, that is, at the shorter lead time, there are two sources of uncertainty (initial condition and climate forcing) that reduces to only one source (climate forcing) at the longer lead time (Duan and Kumar 2020).

The CART analysis for ACC shows the TI as the primary determinant, followed by the soil texture type (Fig. 11c). The grid cells having TI less than 5.5 show poorer ACC than grid cells with higher TI, that is, the model performance is better in the valley area than the ridge area where TI is smaller. ACC decreases from clay to sand soil types. Land-use type is the only determinant using nRMSE as the evaluation criterion, with forest land-use type showing the smallest nRMSE and developed area showing the largest nRMSE (Fig. 11d).

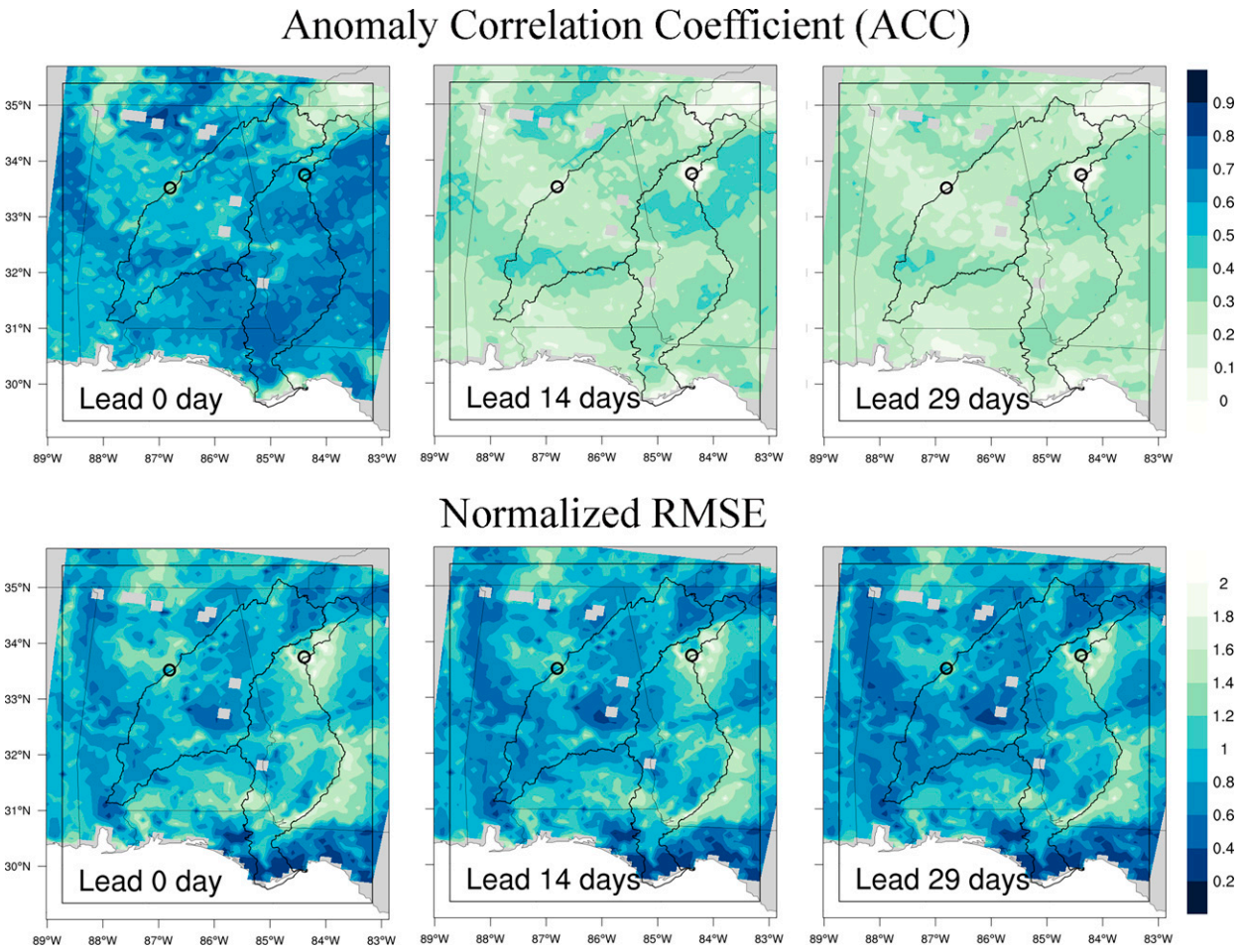


FIG. 10. Evaluation of the soil moisture forecast in the NWM and its comparison with SMAP soil moisture observation. (top) ACC; (bottom) nRMSE.

The hydrological forecast evaluations (streamflow and soil moisture; sections 3a–3d) demonstrated the dependency of the forecast errors on the biophysical attributes that are machine identifiable too (e.g., CART results), therefore providing a scientific basis for the AI/ML model development.

### e. *DL model performance*

The DL model improves the forecast reliability by combining the AI with the physically based NWM forecast. The forecast reliability is measured as the probability of capturing the observation in the forecast ensemble range [Eq. (6)]:

$$\text{P\_Model} = \frac{\sum_i x_i}{N}, \tag{6}$$

where $i = 1, 2, \ldots, N$ is the number of observations in the evaluation period, and $x_i = 1$ if the observed streamflow is within the forecast ensemble range, otherwise $x_i = 0$. The DL forecast range is obtained by adding and subtracting one absolute forecasted error [Eq. (5)] from the NWM ensemble mean forecast. The NWM forecast range is obtained from its 16-member ensemble forecast from maximum to minimum.

#### 1) TEMPORAL SPLIT ANALYSIS

The DL-based forecast significantly improves the model's ability to capture observations. Figure 12a shows a 10-day lead forecast during the first year of the model evaluation period for the Coosa River (Mayo's Bar) near Rome, Georgia. The probability of capturing the observation improves from 0.25 in the NWM forecast to 0.83 in the DL forecast. Please note that the model evaluation period (3 April 2020–15 August 2021) is independent of the DL model calibration period (11 April 2018–2 April 2020); hence the evaluation period can be treated as ungauged data.

Similarly, we assessed reliability for all 384 watersheds for the entire evaluation period (3 April 2020–15 August 2021) and 0–29 days lead forecasts. Figure 12b shows the forecast reliability averaged across all watersheds as a function of forecast lead time. The DL-based forecast reliability ranges

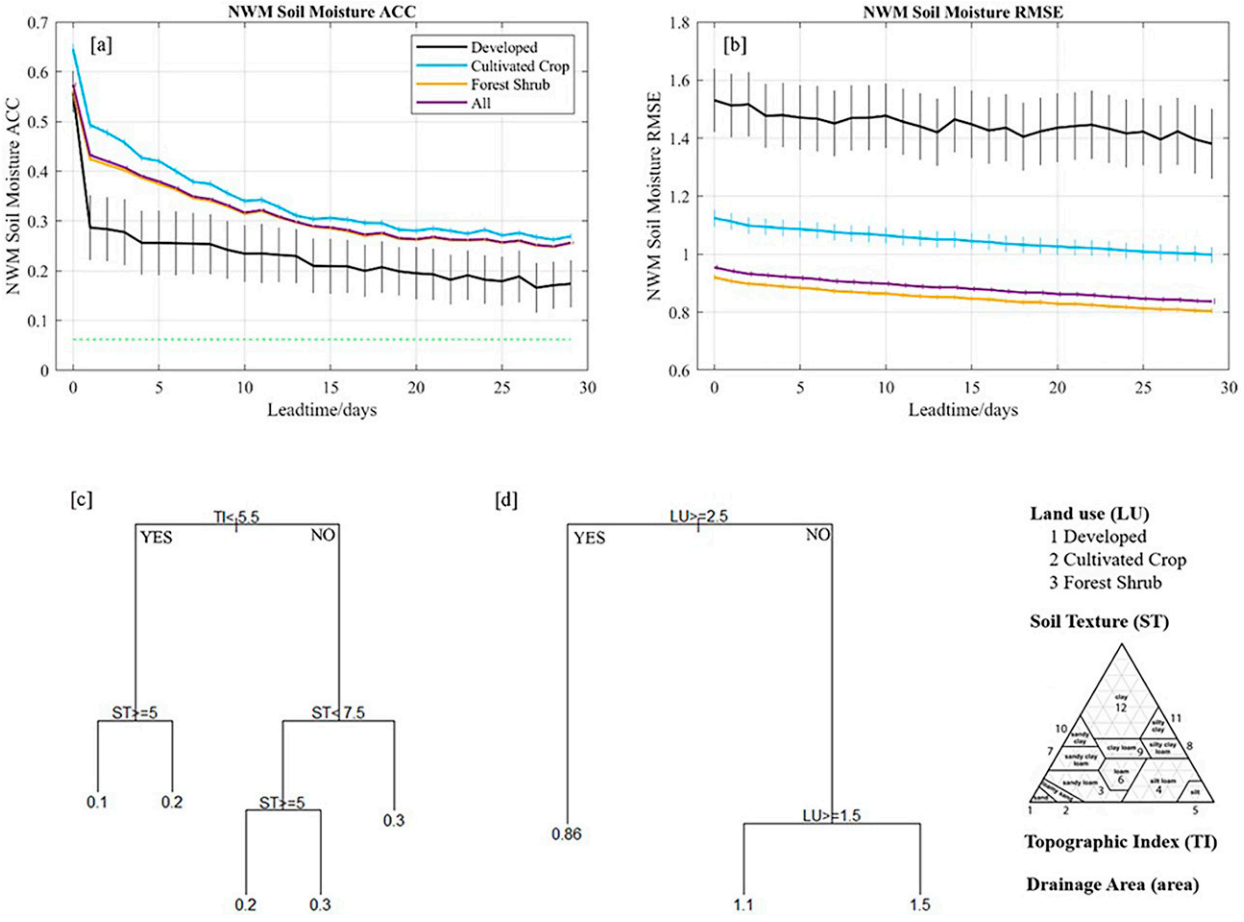


FIG. 11. The soil moisture forecast skill dependency on biophysical attributes. (a) The NWM soil moisture ACC is classified by land-use categories. (b) The NWM soil moisture nRMSE is classified by land-use categories. (c) The CART results for the soil moisture ACC of 14-day lead forecast. (d) The CART results for the soil moisture nRMSE of 14-day lead forecast. The left branch is yes, and the right branch is no in each bifurcation.

from 0.86 ± 0.03 for the 0-day lead forecast to 0.79 ± 0.03 for the 29-day lead forecast. The corresponding reliability in the NWM forecast is 0.10 ± 0.01 for the 0-day lead forecast and 0.31 ± 0.01 for the 29-day lead forecast. Thus, the DL model increases the reliability of the forecast. A slight increase in the NWM forecast's reliability at a longer lead time can be due to a larger ensemble spread than the shorter lead time with a constrained forecast due to initial condition effects. The DL model performance is similar between calibration and evaluation periods (not shown).

A comparison with the long-term monthly mean climatology forecast (gray shading and dashed lines in Fig. 12a) shows that the DL model shows an improved performance in capturing the observations. The long-term monthly mean is obtained from the USGS, and the corresponding MAE is calculated from the daily observed values (December 2018–August 2021) as its average absolute departure from the long-term monthly mean. As a result, the probability of capturing the observation in the climatology forecast is 0.74, which is less than the DL forecast (0.83). Similarly, averaged across the 384 sites, the probability of capturing the observations is (0.70 ± 0.01) in the climatological forecast, which is also less than the DL forecast (0.82 ± 0.03, Fig. 12b).

However, the forecast uncertainty range also increased in the DL forecast, probably expected from its design [Eq. (5)], that is, the DL model was designed to predict the forecast errors. For example, the DL forecast range is comparable to or slightly larger (smaller) than the climatology forecast ranges for the low-flow (high flow) observations in Fig. 12a. The average ratio of the DL forecast range to the climatology forecast range is 1.45 in Fig. 12a. The main advantage of the DL model is its ability to predict forecast errors at ungauged locations.

The DL model performance also shows the biophysical dependency (Fig. 13). The DL model performance is generally the reverse of the NWM performance because the DL model is trained to predict the error between the NWM forecast and the observations [Eq. (5)]. For example, the DL model reliability is 7% higher for the developed watersheds (P_DL = 0.87 ± 0.06, averaged across 0–29 lead days) compared to the predominantly forested and shrub watersheds (P_DL = 0.80 ± 0.03)

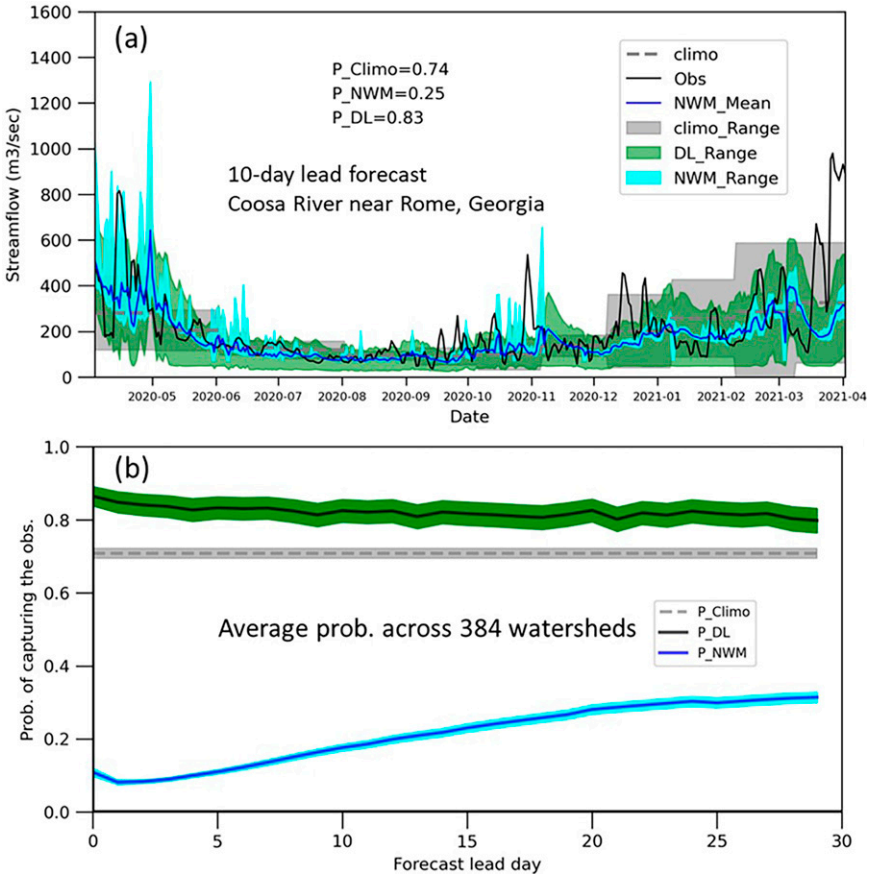


FIG. 12. DL model temporal split evaluation. (a) The model performance for the 10-day lead forecast for the Coosa River (Mayo's Bar) near Rome (USGS Gauge ID: 02397000) and the period of 3 Apr 2020–2 Apr 2021; see text. (b) The probability of capturing the observed streamflow model performing over 30 lead days for all the 384 stations. Shading shows a 95% uncertainty range.

(Fig. 13a). Similarly, the DL performs 6% better for the predominantly sandy watershed (ST $<$ 2.5, P_DL = 0.87 $\pm$ 0.06), than clay-loam watersheds (ST $>$ 2.5, P_DL = 0.81 $\pm$ 0.04) (Fig. 13b).

#### 2) SPATIAL SPLIT ANALYSIS

The spatial split analysis shows comparatively similar DL model performance as found in temporal split analysis (cf. Figs. 12–14). Taking Coosa River near Rome, Georgia, as an ungauged site, that is, this site was not included in the DL model development as one of the 10% withheld sites. Then, the DL model was developed using the remaining 90% sites, and the model was used to predict forecast error for the Coosa River. For the 10-day lead forecast from 3 April 2020 to 15 August 2021 (for comparison), the DL model captured the observed streamflow 87% times (Fig. 14a), comparable to or even slightly better than the temporal split analysis (83%). Averaged across all 384 sites, the spatial split analysis showed similar performance (82% $\pm$ 1%, Fig. 14b) to the temporal split analysis (82% $\pm$ 3%, Fig. 12b).

## 4. Conclusions and discussion

This study demonstrated the potential for improving hydrological forecast by combining a physically based model with artificial intelligence (AI) techniques. As a result, the forecast reliability increased from 21% ($\pm$1%) in the NWM only forecasts to 82% ($\pm$3%) DL model forecast (Fig. 12b). It is important to note that the DL model is not independent, but the DL model is built using the NWM forecast and observations, that is, the proposed technique combines the strength of both the physically based model and AI techniques. A more reliable NWM-DL forecast can potentially inform decision-making.

The DL forecast range is larger than the NWM forecast range (Fig. 12a). Our attempt to constrain the DL model's forecast range was unsuccessful. It is likely that most of the forecast uncertainty is due to seasonal climate forecast data (CFSv2), and the DL model was not designed, at least in this study, to improve the seasonal climate forecast. For example, Frame et al. (2021) improved the forecast skill using the DL model and the observed climate forcing.

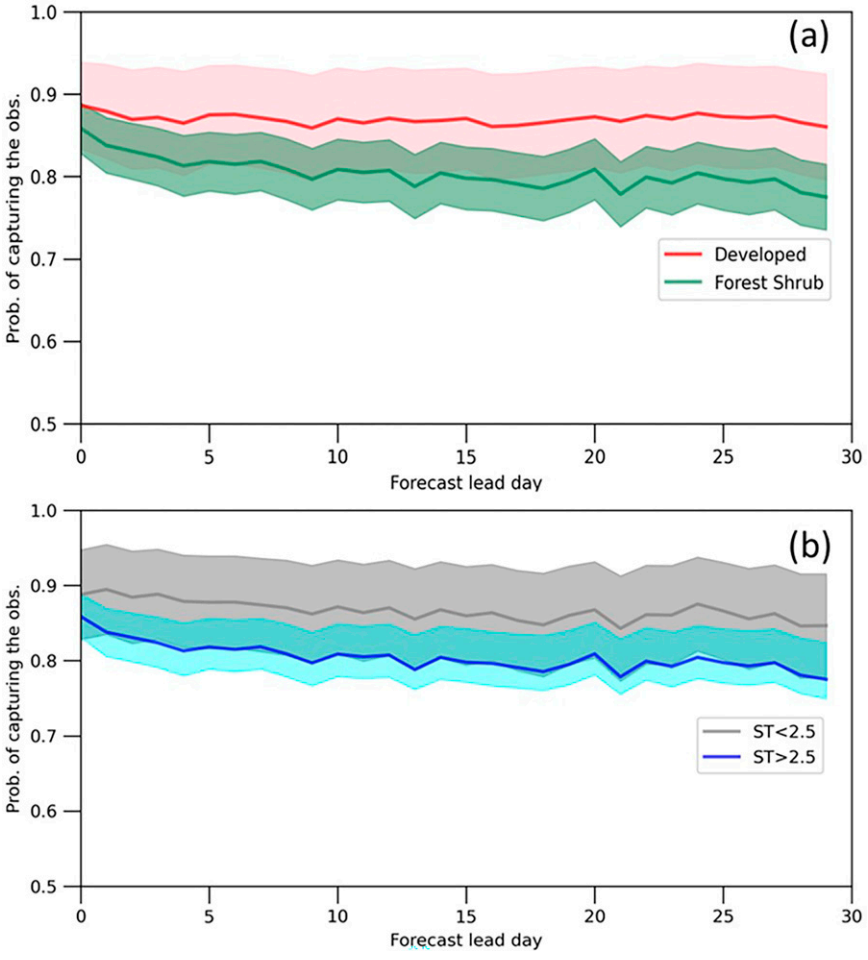


FIG. 13. DL model captures the biophysical dependency of the forecast uncertainty. (a) The probability of DL model prediction across two land-use categories: Developed watersheds ($N = 80$) and forested and shrub watersheds ($N = 280$). (b) As in (a), but for soil texture categories: ST < 2.5 ($N = 77$), and ST > 2.5 ($N = 303$). Shading shows a 95% uncertainty range.

We used a simplistic measure of the forecast reliability, that is, observations are contained in the forecast ensemble spread [Eq. (6)]. The DL model provided a dynamic forecast error that mostly contained observations (~82%). The main advantage of the DL model is that it can be applied to ungauged sites, as demonstrated in Fig. 14. Other probabilistic forecast skill metrics not included here are the Brier score, continuous ranked probability score, relative operating characteristics score, and forecast convergence score (Brum and Schwanenberg 2022). Future studies may include one or more of these measures in the DL model performance (e.g., Weyn et al. 2021).

We used the NWM forecast as one of the inputs to the DL model (Fig. 2) instead of the individual climate variable, for example, precipitation and temperature data. Kumar et al. (2013) found that the crop suitability index that combines climate, soil, land use, and topographic characteristics into a single index is a better predictor of the cropland spatial distribution in the United States than the individual driver variables. Similarly, the NWM combines the various climate inputs in a biophysically constrained way to provide the flow, an input to the DL model.

The big-data analytics identified three critical areas for potential improvement in the NWM model: 1) effects of urbanization are not well captured in the NWM forecast; 2) underestimation of the base flow in the agricultural watersheds that can be related to the irrigation effect; 3) tidal processes may affect base flow in the coastal watersheds (supplemental Fig. S4).

The conceptual framing of this study, for example, Fig. 1, emphasizes a complementary contribution of the process-based climate-hydrological model and AI techniques. The proposed and demonstrated complementary contributions can apply to a broad range of climate science problems. For example, the AI/ML technique can heuristically search the ensemble members from a very large sample, and the selected ensemble from the process-based model can show the skillful decadal climate forecast (e.g., Smith et al. 2020). Future improvements in the DL methodology (Fig. 2) may include spatially distributed

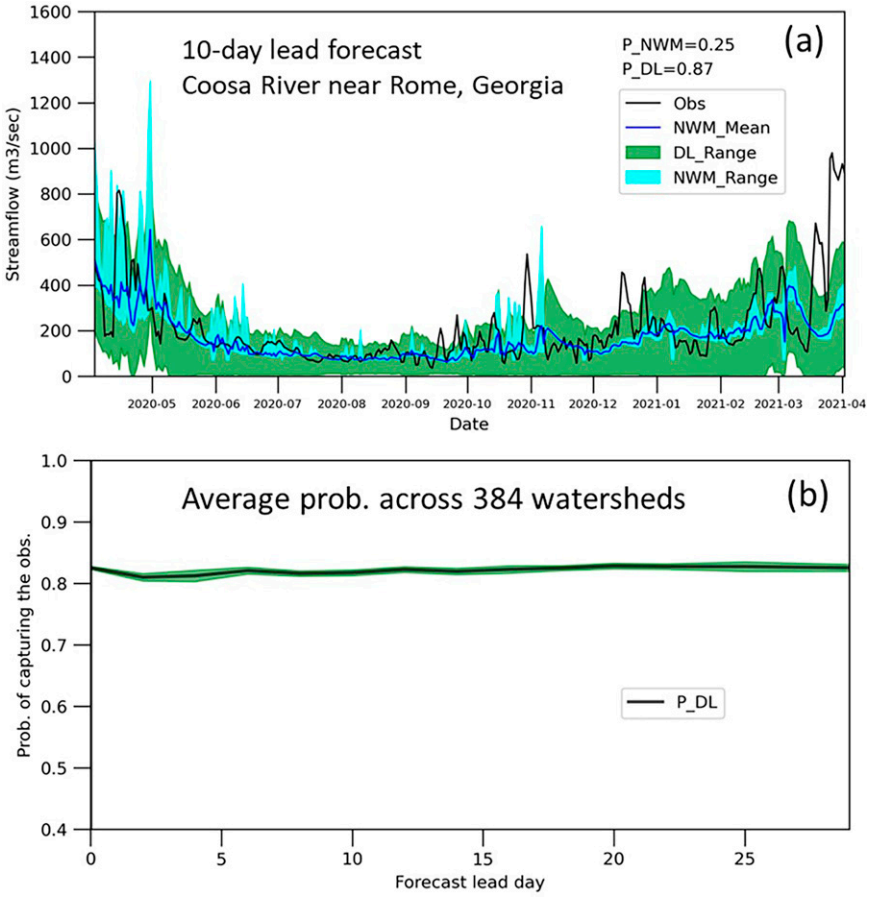


FIG. 14. DL model spatial split evaluation. (a) The model performance for the 10-day lead forecast for the Coosa River (Mayo's Bar) near Rome (USGS Gauge ID: 02397000) and the period of 3 Apr 2020–2 Apr 2021. (b) The probability of capturing the observed streamflow model performing over 30 lead days for all the 384 stations. Shading shows a 95% uncertainty range.

biophysical attributes and an analysis correction-based additive inflation method (Crawford et al. 2020).

*Acknowledgments.* S. K. and W. L. acknowledge support from the USDA NIFA Grant 2020-67021-32476. S. K., Y. D., and S. A. acknowledge the support of the USDA Hatch Grant ALA031-1-18023. We also acknowledge the support of the Descartes Lab. This work was partially completed with computing resources provided by the Auburn University Easley Cluster. Finally, we would like to acknowledge high-performance computing support from Cheyenne (https://doi.org/10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. S. K. conceived the idea and analysis methods. S. K. and W. L. supervised the hybrid physics–AL model development. Y. D. downloaded, archived, and conducted the hydrological analysis and developed the cleaned data for the ML application. S. A. developed the D. L. model and evaluated its performance. SKH provided access to the Descartes Lab computational platform and SMAP data access. Y. D. and S. K. wrote the draft manuscript. All authors contributed to manuscript development and writing.

*Data availability statement.* The repository is available at https://github.com/cwsauburn/NWM_DL. This repository contains the following underlying data:

- raw_data folder: It contains all the raw data required to train and validate the DL models.
- temporal_DLModels Folder: It contains trained DL models for 0 to 30 forecast day lead data; these models were generated using the temporal split of the data.
- spatial_0–15 Folder: It contains trained DL models for 0 to 15 forecast day lead data; these models were generated using the spatial split of the data.
- spatial_15–30 Folder: It contains trained DL models for 15 to 30 forecast day lead data; these models were generated using the spatial split of the data.
- data_setup_scripts: It contains the scripts to set up the input data to train the DL model for 0 to 30 days of lead streamflow data.

- train&predict_scripts: This folder contains different machine learning scripts used to train the model and predict the streamflow of ungauged sites for all the 30-day lead data.
- temporal_graph_data_scripts: It contains the script for the temporal graphs and the raw data used to generate these graphs (Figs. 12 and 13).
- spatial_graph_data_scripts: It contains the script for the spatial graphs and the spatial data used to generate these graphs (Fig. 14).

Due to the large size of the raw NWM forecast data (10 TB), it can be obtained on request by contacting the paper's authors (primary contact: szk0139@auburn.edu).

## REFERENCES


Arnold, J. G., and P. M. Allen, 1999: Automated methods for estimating baseflow and ground water recharge from streamflow records 1. *J. Amer. Water Res. Assoc.*, **35**, 411–424, https://doi.org/10.1111/j.1752-1688.1999.tb03599.x.

——, ——, R. Muttiah, and G. Bernhardt, 1995: Automated base flow separation and recession analysis techniques. *Groundwater*, **33**, 1010–1018, https://doi.org/10.1111/j.1745-6584.1995.tb00046.x.

Barnes, E. A., J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2019: Viewing forced climate patterns through an AI lens. *Geophys. Res. Lett.*, **46**, 13 389–13 398, https://doi.org/10.1029/2019GL084944.

Brum, M., and D. Schwanenberg, 2022: Long-term evaluation of the Sub-seasonal to Seasonal (S2S) dataset and derived hydrological forecasts at the catchment scale. *EGUsphere*, https://doi.org/10.5194/egusphere-2022-419.

Chan, S. K., and Coauthors, 2018: Development and assessment of the SMAP enhanced passive soil moisture product. *Remote Sens. Environ.*, **204**, 931–941, https://doi.org/10.1016/j.rse.2017.08.025.

Chollet, F., 2018: *Deep Learning with Python*. Manning Publications, 384 pp.

Crawford, W., S. Frolov, J. McLay, C. A. Reynolds, N. Barton, B. Ruston, and C. H. Bishop, 2020: Using analysis corrections to address model error in atmospheric forecasts. *Mon. Wea. Rev.*, **148**, 3729–3745, https://doi.org/10.1175/MWR-D-20-0008.1.

Dragoni, N., S. Giallorenzo, A. L. Lafuente, M. Mazzar, F. Montesi, R. Mustafin, and L. Safina, 2017: Microservices: Yesterday, today, and tomorrow. *Present and Ulterior Software Engineering*, M. Mazzara and B. Meyer, Eds., Springer, 195–216.

Duan, Y., and S. Kumar, 2020: Predictability of seasonal streamflow and soil moisture in national water model and a humid Alabama–Coosa–Tallapoosa River basin. *J. Hydrometeor.*, **21**, 1447–1467, https://doi.org/10.1175/JHM-D-19-0206.1.

Eberts, S. M., M. D. Woodside, M. N. Landers, and C. R. Wagner, 2019: Monitoring the pulse of our nation's rivers and streams—The U.S. Geological Survey Streamgaging Network. USGS Fact Sheet 2018–3081, 2 pp., https://doi.org/10.3133/fs20183081.

Frame, J. M., F. Kratzert, A. Raney II, M. Rahman, F. R. Salas, and G. S. Nearing, 2021: Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *J. Amer. Water Res. Assoc.*, **57**, 885–905, https://doi.org/10.1111/1752-1688.12964.

Gochis, D., W. Yu, and D. Yates, 2013: The NCAR WRF-Hydro Technical Description and User's Guide, version 1.0. NCAR Tech. Doc., 120 pp., https://ral.ucar.edu/sites/default/files/public/projects/wrf_hydro/WRF_Hydro_Technical_Description_and%20User_Guide_v1.0.pdf.

——, and Coauthors, 2018: The NCAR WRF-Hydro Modeling System Technical Description (version 5.0). NCAR Tech. Note, 107 pp., https://ral.ucar.edu/sites/default/files/public/WRF-HydroV5TechnicalDescription.pdf.

Homer, C., and Coauthors, 2020: Conterminous United States land cover change patterns 2001–2016 from the 2016 National Land Cover Database. *ISPRS J. Photogramm. Remote Sens.*, **162**, 184–199, https://doi.org/10.1016/j.isprsjprs.2020.02.019.

Hooper, R. P., G. S. Nearing, and L. S. Condon, 2017: Using the national water model as a hypothesis-testing tool. *Open Water J.*, **4**, 3.

Hrachowitz, M., and Coauthors, 2013: A decade of predictions in ungauged basins (PUB)—A review. *Hydrol. Sci. J.*, **58**, 1198–1255, https://doi.org/10.1080/02626667.2013.803183.

Kratzert, F., D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing, 2019: Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resour. Res.*, **55**, 11 344–11 354, https://doi.org/10.1029/2019WR026065.

Kumar, S., V. Merwade, P. S. C. Rao, and B. C. Pijanowski, 2013: Characterizing long-term land use/cover change in the United States from 1850 to 2000 using a nonlinear bi-analytical model. *Ambio*, **42**, 285–297, https://doi.org/10.1007/s13280-012-0354-6.

Lee, W., and S. Kumar, 2016: Software-defined storage-based data infrastructure supportive of hydroclimatology simulation containers: A survey. *Data Sci. Eng.*, **1**, 65–72, https://doi.org/10.1007/s41019-016-0008-y.

Li, Y., T. Liu, D. Jiang, and T. Meng, 2021: Transfer-learning-based network traffic automatic generation framework. *2021 Sixth Int. Conf. on Intelligent Computing and Signal Processing (ICSP)*, Xi'an, China, IEEE, 851–854, https://ieeexplore.ieee.org/document/9408767.

Loh, W. Y., 2011: Classification and regression trees. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery*, **1**, 14–23, https://doi.org/10.1002/widm.8.

Mayer, K. J., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*, **48**, e2020GL092092, https://doi.org/10.1029/2020GL092092.

Miller, D. A., and R. A. White, 1998: A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. *Earth Interact.*, **2**, https://doi.org/10.1175/1087-3562(1998)002<0001:ACUSMS>2.3.CO;2.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117**, 572–582, https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Nair, V., and G. E. Hinton, 2010: Rectified linear units improve restricted Boltzmann machines. *Proc. 27th Int. Conf. on Machine Learning*, Haifa, Israel, ICML, 807–814, https://icml.cc/Conferences/2010/papers/432.pdf.

Saha, S., and Coauthors, 2014: The NCEP climate forecast system version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Sarker, I. H., 2021: Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions.

*SN Comput. Sci.*, **2**, 420, https://doi.org/10.1007/s42979-021-00815-1.

Shen, C., and Coauthors, 2018: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrol. Earth Syst. Sci.*, **22**, 5639–565, https://doi.org/10.5194/hess-22-5639-2018.

Singh, R., S. A. Archfield, and T. Wagener, 2014: Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments—A comparative hydrology approach. *J. Hydrol.*, **517**, 985–996, https://doi.org/10.1016/j.jhydrol.2014.06.030.

Sivapalan, M., and Coauthors, 2003: IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrol. Sci. J.*, **48**, 857–880, https://doi.org/10.1623/hysj.48.6.857.51421.

Smith, D. M., and Coauthors, 2019: Robust skill of decadal climate predictions. *npj Climate Atmos. Sci.*, **2**, 13, https://doi.org/10.1038/s41612-019-0071-y.

——, and Coauthors, 2020: North Atlantic climate far more predictable than models imply. *Nature*, **583**, 796–800, https://doi.org/10.1038/s41586-020-2525-0.

Tieleman, T., and G. Hinton, 2012: Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude. COURSERA: Neural Networks for Machine Learning, 4, 26–31.

Vogel, R. M., and C. N. Kroll, 1996: Estimation of baseflow recession constants. *Water Resour. Manage.*, **10**, 303–320, https://doi.org/10.1007/BF00508898.

Wagener, T., and H. S. Wheater, 2006: Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. *J. Hydrol.*, **320**, 132–154, https://doi.org/10.1016/j.jhydrol.2005.07.015.

Weyn, J. A., D. R. Durran, R. Caruana, and N. Cresswell-Clay, 2021: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002502, https://doi.org/10.1029/2021MS002502.