Toward Predicting Flood Event Peak Discharge in Ungauged Basins by Learning Universal Hydrological Behaviors with Machine Learning

AKHIL SANJAY POTDAR,[a] PIERRE-EMMANUEL KIRSTETTER,[a,b,c,d,e] DEVON WOODS,[c] AND MANABENDRA SAHARIA[f]

[a] *Data Science and Analytics Institute, University of Oklahoma, Norman, Oklahoma*
[b] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[c] *School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, Oklahoma*
[d] *Advanced Radar Research Center, University of Oklahoma, Norman, Oklahoma*
[e] *NOAA/National Severe Storms Laboratory, Norman, Oklahoma*
[f] *Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India*

(Manuscript received 15 December 2020, in final form 8 June 2021)

ABSTRACT: In the hydrological sciences, the outstanding challenge of regional modeling requires to capture common and event-specific hydrologic behaviors driven by rainfall spatial variability and catchment physiography during floods. The overall objective of this study is to develop robust understanding and predictive capability of how rainfall spatial variability influences flood peak discharge relative to basin physiography. A machine-learning approach is used on a high-resolution dataset of rainfall and flooding events spanning 10 years, with rainfall events and basins of widely varying characteristics selected across the continental United States. It overcomes major limitations in prior studies that were based on limited observations or hydrological model simulations. This study explores first-order dependencies in the relationships between peak discharge, rainfall variability, and basin physiography, and it sheds light on these complex interactions using a multidimensional statistical modeling approach. Among different machine-learning techniques, XGBoost is used to determine the significant physiographical and rainfall characteristics that influence peak discharge through variable importance analysis. A parsimonious model with low bias and variance is created that can be deployed in the future for flash flood forecasting. The results confirm that, although the spatial organization of rainfall within a basin has a major influence on basin response, basin physiography is the primary driver of peak discharge. These findings have unprecedented spatial and temporal representativeness in terms of flood characterization across basins. An improved understanding of subbasin scale rainfall spatial variability will aid in robust flash flood characterization as well as with identifying basins that could most benefit from distributed hydrologic modeling.

SIGNIFICANCE STATEMENT: To improve understanding of the effect of precipitation on floods, a machine-learning workflow is designed to scrutinize hydrological processes and is applied on a database of flood events over the United States. The model accurately reproduces observed maximum streamflow. It reflects physical hydrologic behavior that is consistent across basins, thereby addressing the challenge of regional modeling and improving upon traditional hydrological models. Rainfall spatial variability has a major influence on flood peak discharge, although basin geomorphology is the primary driver. This grants an improved understanding of the conditions under which floods are generated, allowing for better predictions and warning.

KEYWORDS: Atmosphere; Streamflow; Precipitation; Hydrology; Hydrometeorology; Radars/radar observations; Hydrologic models; Flood events; Data science; Machine learning

1. Introduction

Flood hazards are ranked as the third most frequent type of natural disasters behind severe storms and tropical cyclones. They have caused estimated economic losses of $146.5 billion in the United States in the last 40 years (about $450 per person in the United States), with that total steadily increasing (Smith 2020). Most fatalities associated with flooding are attributed to flash floods (Ashley and Ashley 2008) Higher flood risk is expected, along with more intense precipitation events globally, under climate change (Sillmann et al. 2013). Defined as rapid rises of water along an existing waterway, flash floods typically occur within 6 h and often within 3 h of a rainfall event (NWS 2010). Discharge at the outlet increases suddenly under the integrated influences of specific hydrological processes, all of which show variable effects under different basin geomorphological, climatological, and spatiotemporal conditions (Saharia et al. 2017). While emergency managers require timely and accurate information on impacted areas, capturing the dynamics of these multifactorial processes to predict flash floods is both critical and difficult.

Corresponding author: Pierre-Emmanuel Kirstetter, pierre.kirstetter@noaa.gov

DOI: 10.1175/JHM-D-20-0302.1

In the hydrological sciences, the fundamental and long-standing challenge of regional modeling hampers the characterization and prediction of flash floods over large areas, which is increasingly critical to preventing further loss of life as hydrologic regimes change. Specifically, the challenge is to extrapolate modeled hydrological information from gauged to ungauged basins or maintain modeling skills when calibrating a hydrologic model from a single basin to multiple basins together (e.g., Blöschl and Sivapalan 1995; Blöschl et al. 2013; Hrachowitz et al. 2013). The observed hydrologic response (discharge at the basin outlet) results from unobserved, nonlinear, and integrated contributions of surface and subsurface processes. This endemic lack of observational hydrologic constraints leaves a considerable range of options to adjust hydrologic model parameters to reproduce the observed behavior of a given watershed (problem of equifinality; Beven 2001), while regional modeling involves capturing different hydrologic behaviors that vary with watershed characteristics such as geomorphology, climatology, geology, pedology, etc. Plus, at the flash flood scale process dynamics are driven by the spatial and temporal distribution of rainfall. Existing operational methods of regional flash flood monitoring focus on specific hydrologic processes. For example, the Flash Flood Guidance (FFG) system used worldwide estimates runoff generation (Sweeney 1992), However, FFG only addresses parts of the flood's characteristics and does not focus on water propagation overland or along streams. It misses any occurrence of flooding downstream of the rainfall, especially the delay, magnitude, and duration of the flood. Because a flood forecasting system needs to describe these characteristics ahead of time, modern instances of forecasting systems such as the Ensemble Framework for Flash Flood Forecasting (EF5; Gourley et al. 2017), the National Water Model (Salas et al. 2018), and the Global Flood Awareness System (GLOFAS; Alfieri et al. 2013; https://www.globalfloods.eu) use hydrologic models to interpret and predict flood characteristics through simplified representations of the processes such as water routing that take place in the watershed. Such models are calibrated at local points where streamflow observations are available, leaving large areas potentially without accurate hydrological simulations.

Regional modeling can be addressed through two types of approaches: model-dependent methods and data-driven methods (Razavi and Coulibaly 2013). Models can be classified into three categories based on how these hydrological processes are described: empirical, physical, and conceptual (Solomatine and Wagener 2011). Traditional empirical (statistical) models are built from the joint analysis of precipitation (input) and discharge (response) time series data to derive statistical equations that represent the input-response behavior of a catchment. Their predictive power is limited at the basin outlet (Devia et al. 2015). Conceptual models synthesize hydrologic processes into an a priori parametric structure. Parameters are imperfect representations of physical processes that are calibrated using catchment observations. They can be challenging to interpret, although spatially distributed parameters allow for the representation of spatial variations of hydrologic properties and processes (distributed models). An example of this is the conceptual representation of the water balance used in the EF5 modeling framework (Flamig et al. 2020; Gourley et al. 2017), a state-of-the-art system developed by the University of Oklahoma and the NOAA National Severe Storms Laboratory for flash flood prediction at the U.S. National Weather Service. Physically based models are mechanistic and designed to represent the physical processes of the system. The rationale expects a degree of physical realism to the extent that the laws of conservation of mass, momentum, and energy are maintained. The model's structure and parameters are designed a priori based on the understanding of the basin physics. As such the selected parameters are not calibrated, making diagnosis difficult. One example is the routing component within EF5 (Vergara et al. 2016). As an alternative, emphasis on the advent of new data driven models through the integration of machine learning is growing (Solomatine and Wagener 2011). Recent data-driven approaches involving machine learning [e.g., long short-term memory (LSTM)] show high potential for joint learning from diverse hydrological and meteorological training data to streamflow simulations but are often limited to predictions of time series at particular sites (see review from Mosavi et al. 2018) and seldom address spatiotemporal predictions (Kratzert et al. 2019; Hu et al. 2019).

The novel approach of this study owes itself to several factors. For the first time to best our knowledge, a ''physics informed statistical'' machine-learning modeling approach is designed and evaluated for regional modeling at the hydrological event scale. The event scale is relevant to identify fundamental hydrologic processes that allow for space and time transferability across watersheds, especially for flash floods. Runoff and the transport of water through the channel network and the magnitude of the peak discharge is controlled by the physiography of the basin and the variability of rainfall at the subbasin scale. One goal of this study is to understand the relative impact of rainfall variability and catchment features on flooding. The contribution of hydrologic processes driven by watershed attributes are accounted for and represent their integrated response at the basin outlet. Datasets are gathered that represent the geomorphological, climatological, and spatiotemporal precipitation attributes, thus incorporating the physics of the hydrologic system. Specifically, precipitation variability described as moments is shown to provide a refined description and a deeper understanding of rainfall forcing (Zoccatelli et al. 2011; Smith et al. 2005). A large-sample dataset allows for regression modeling using gradient boosted trees to identify the multivariate relationships that exist between the dependent (peak discharge) and independent watershed attributes. The use of physical constraints along with proper methods to mitigate overfitting eases the generalization of the model for diverse catchments. This original data-driven framework is proposed for scrutinizing hydrological processes over a large-scale dataset to examine interbasin hydrologic consistency and improve upon the limited representativeness of traditional empirical models. The machine-learning modeling performed in this study diverts from time series analysis to create a universal model owing to the dataset that combines multitudes hydrologic events and basins. It undertakes no prior assumptions in the design of the model structure nor with the calibration of parameters that often hinder the transferability of conceptual and physically based models across basins. The framework allows for the performance of prescriptive and predictive analytics to learn universal hydrological signatures

and their dependencies from the data, to estimate the cause factors given the prescribed results, and to predict hydrological responses. Flood studies rarely combine prescriptive and predictive analytics. While biases arise from a priori structure designs and parameter choices in hydrological models, impacting the applicability over ungauged basins, this approach inherently calibrates the parameters to the data. Once calibrated by data training, the machine-learning model requires no further tuning. New insights into catchment's response are provided by the objective selection of the most relevant predictors by the machine-learning framework, and the subsequent analysis of their impact. Such a model can be used as a diagnostic tool to identify and interpret key hydrological processes, a feat that other conceptual models cannot claim. Through the study of feature importance, simple and parsimonious models can be designed to represent the basin physics similarly to mechanistic models. This novel approach seeks to provide high predictive power, interpretation, and versatility over diverse basins to address the important hydrologic sciences challenge of characterization of floods in ungauged basins.

The paper is organized as follows. Section 2 describes the study region and the flood, physiographic, and rainfall datasets. Section 3 proposes the method, and the results are discussed in section 4 in terms of model selection and evaluation. The summary and conclusions are provided in section 5.

2. Data

a. Predictand: Peak discharge

Flash floods are characterized by the observed *peak discharge* during a hydrological event at the basin outlet. Discharge time series information from the U.S. Geological Survey automated stream gauges are curated in the Unified Flash Flood Database (National Severe Storms Laboratory; https://blog.nssl.noaa.gov/flash/database/) to describe individual flooding events and identify flooding peak discharge values at more than 10 000 locations across the United States. The time series information is collected at intervals ranging from 5 to 60 min, and it is suitable for analyzing the impact of rainfall spatial variability on floods. A subset of 3490 stream gauge locations is used, with stages jointly defined by the USGS, the National Weather Service, and local stakeholders and corresponding to action, minor, moderate, and major flooding. Action stage is defined as the stage at which NWS forecasters take "mitigation action for possible significant hydrologic activity" and it often corresponds to bankfull conditions (https://www.nws.noaa.gov/directives/sym/pd01009050curr.pdf). This dataset covers diverse climatologic, hydrologic, and weather conditions, which makes it a representative flash flood database over the United States (Gourley et al. 2013).

Gauges that are affected by regulation or diversion are further screened out using the regulation codes supplied by the USGS. In this database, a flood event is defined as the period when streamflow is above the defined action stage for that gauge. If there is a 24-h period with discharge values below action stage, then the events are considered as separate. For each flood event, the database includes the start and end time (UTC) when the flow first exceeded and dropped below the action stage threshold, respectively, along with the time (UTC) and magnitude of peak flow ($m^3 s^{-1}$) that is used in this study. The maximum basin area in this study is approximately 45 000 km^2 with a median area of 890 km^2, suitable for analyzing the impact of rainfall spatial variability on floods.

A dataset of 21 143 flooding events was finalized for the study, that capture the influence of rainfall spatial variability on catchment response, over 1113 basins across diverse climatological and physiological conditions in the contiguous United States (CONUS).

b. Predictors: Geomorphology and climatology

A natural flood generally occurs with intense precipitation or snowmelt, but the transport of water through the channel network and the magnitude of the peak discharge are regulated by the physiography of the basin and the variability of rainfall at the subbasin scale. The physical depiction of the catchment behavior is enriched through the integration of geomorphologic and climatological characteristics that represent the integrated contribution of hydrologic processes at the basin outlet. The database comprises attributes representing various properties such as geomorphology, topography, climatology, vegetation, and soil with 50 variables from the 902 basins.

As potential explanatory variables of flash flooding, geomorphological parameters were derived from the National Elevation Dataset (NED; http://ned.usgs.gov/) digital elevation model (DEM) across the CONUS. To ensure compatibility between DEM-based flow accumulations and the actual river network, flow accumulation and direction was extracted by delineating basins with USGS stations, and the National Hydrography Dataset (NHD; http://nhd.usgs.gov/) was used to resample the 30-m DEM to a 1-km grid. The geomorphologic parameters for delineated catchments were extracted from these grids. Variables representing soil properties such as mean depth-to-bedrock and K factor (erodibility) were derived from the State Soil Geographic (STATSGO) database (Miller and White 1998). The National Land Cover Dataset (Fry et al. 2011) was used to estimate land cover and land use data such as the runoff curve number. Last, hydroclimatic variables (e.g., mean annual and seasonality of precipitation and temperature) were extracted from the WorldClim database (https://www.worldclim.org/data/bioclim.html). The spatially distributed basin attributes included in this study are provided in Table 1.

To reduce input uncertainties in modeling results, catchments were selected based on climatological snow percent of total precipitation. Only basins that get less than 20% of their annual precipitation from snowpack were included. For basins that get greater than 20% of its annual precipitation from snowpack, only events in summer months (May–October) were included. The dataset is devoid of missing values and is numerical in all its attributes. The representativeness of the dataset has been previously demonstrated by Saharia et al. (2017) by mapping basin flashiness over the United States to predict flash flooding severity in ungauged regions.

c. Predictors: Precipitation spatial variability

The spatial distribution of a rainfall event over a basin is described through precipitation moments computed from the Multi-Radar Multi-Sensor (MRMS) precipitation reanalysis (Zhang et al. 2016; Zhang and Gourley 2018). The MRMS

TABLE 1. Important predictors for the study.

Type	Variable	Meaning
Geomorphological	Area	Total upstream area that contributes runoff (estimated from adigital elevation model and flow grids)
	G1	Catchment first-order moment of flow distance (Zoccatelli et al. 2011)
	G2	Catchment second-order moment of flow distance (Zoccatelli et al. 2011)
	River length	Measured along a line centered from the basin outlet to the intersection of the extended main channel and the basin boundary
	Relief ratio	Relief is the difference in elevation between the outlet and the highest point in the basin; relief ratio is relief divided by the basin length; it is a measure of the basinwide river slope; the higher the relief ratio is, the higher is the runoff
	Ruggedness	Drainage density multiplied by relief (Strahler 1964)
	Slope to outlet	Local slope computed at a distance of 1 km over the basin outlet
	Rock volume	Volume of rock at the outlet
Precipitation moments	Activated basin	Portion of the basin where rainfall occurs
	Rainfall volume	Product of activated basin area and average rainfall
	Product mean	Mean product of accumulated rainfall and flow distance of the activated basin [Eq. (3)]
	Flow distance (mean)	Mean flow distance of the activated basin
Climatological	bio_10	Mean temperature of warmest quarter
	bio_15	Precipitation seasonality: coefficient of variation of monthly mean precipitation
	Snow percentage	Percentage of snow in the gauge
	Temp (mean)	Annual mean 2-m temperature

reanalysis derives precipitation data at fine spatial and temporal scale (0.01° and 5 min) for a period from 2001 to 2011 over the CONUS from the NEXRAD data archive available from Amazon Web Services (https://aws.amazon.com/public-datasets/nexrad/). The radar coverage is not uniform across the country, especially in the western CONUS, and the accuracy of surface precipitation estimation decreases with radar beam height (Kirstetter et al. 2010, 2013). To further reduce input uncertainties in modeling results and to ensure that only high-quality MRMS rainfall data are included, all events that fall in basins with mean radar beam height greater than 2 km above the ground level were discarded.

Formulations for precipitation spatial variability as moments are used at the event scale to represent rainfall forcing and characterize its impact on each flooding event reported in the Unified Flash Flood Database between 2002 and 2011. These metrics map precipitation organization onto the flow distance, that is, distance measured from any point in the basin to the basin outlet along the flow path (Zoccatelli et al. 2011). The rain-activated basin is the fraction of the watershed that experiences a rain event. The accumulated precipitation, the flow distance, and the product of accumulated precipitation and flow distance are characterized by their moments (mean, standard deviation, skewness, and kurtosis) conditioned on the activated basin. The precipitation moments are computed on the rainfall accumulated before the peak time of the hydrograph over a duration corresponding to 1.5 times the lag time. The lag time is defined as the duration between the time of the centroid of effective rainfall over the basin and the peak time of the hydrograph. Such duration has been selected following a sensitivity analysis on precipitation moments and duration performed on a simulated hydrologic database by Emmanuel et al. (2015). Additional metrics were included such as the *rainfall volume*, computed as the product between the activated basin area and the event averaged precipitation, to capture the volume of water collected from precipitation by the basin that is expected to contribute to the peak discharge.

To summarize, the precipitation variables and the large number of geomorphological and climatological variables included in this study allows us to explore how rainfall spatial variability, the watershed physiography and climatology impact hydrologic events maximum streamflow over a wide variety of situations. These physically based variables are used as predictors in the machine-learning framework to train a representative model that captures various yet common local and regional hydrological behaviors.

3. Method

a. *Machine-learning approach*

The curated large-sample dataset allows for regression modeling using gradient boosted trees to identify the multivariate relationships that exist between the predictand (peak discharge) and predictors representing the geomorphological, climatological, and spatiotemporal attributes of precipitation. The most relevant physically based predictors are objectively selected through the machine-learning training procedure. Statistical analysis helps understand and interpret the predictor–predictand relationships and make predictions on unseen data in order to test the robustness the model in the context of ungauged basins and regional modeling. Previous hydrological prediction studies have applied deep learning algorithms such as LSTMs on streamflow time series (e.g., Kratzert et al. 2018; Mosavi et al. 2018). Extreme Gradient Boosting (XGBoost) has been used to forecast streamflow on single basins (Ni et al. 2020; Yu et al. 2020; Gauch et al. 2019). In the present study, the unique event-based dataset requires an algorithm adapted to regression benchmark studies

(Orzechowski et al. 2018). XGBoost is a machine-learning technique for predictor selection and multivariate regression analysis that produces a prediction model in the form of an ensemble of weak prediction decision trees. It is a supervised learning algorithm designed for fast computational time, especially on large datasets. XGBoost is a form of gradient-boosted decision trees that can generate new models based on the prediction of the residual's errors of prior models. The term ''gradient boosting'' refers to the utilization of a gradient descent to minimize the loss when adding additional models (Brownlee 2016). This algorithm sequentially builds and generalizes models by optimizing of a differentiable loss function (root-mean-square error). It combines the benefits of the tree-based and gradient boosted models to overcome multicollinearity. XGBoost is well known for superior performance in terms of prediction accuracy, lower overfitting, and robustness toward correlated predictors with respect to other techniques such as random forests, elastic net, or least absolute shrinkage and selection operator (LASSO; Fernández-Delgado et al. 2019). Empirical models have been observed to overfit during training (Devia et al. 2015), and the use of proper methods to mitigate overfitting eases the generalization of the model for diverse catchments. XGBoost was selected after testing against other algorithms such as random forest, the results of which are not shown for the sake of brevity. It is important that this technique is used through an unbiased method that trains a representative model learning universal hydrologic behaviors across scales. For this purpose, a training strategy is designed that applies multifold cross validation across multiple independent subsets of training, testing, and validation datasets. Reproducibility across different basins is monitored through various regression evaluation metrics that quantify the model performances. Note that the predictors are not transformed because tree-based regression algorithms are invariant to monotonic transformations of the independent variables (Segal 1988).

b. *Data partitioning and model training*

To objectively assess the model performance and ability to predict in ungauged basins, and reduce overfitting, the main dataset is partitioned into training, validation, and test datasets. Exploratory data analysis is performed using the training dataset, while hyperparameter tuning (e.g., learning rate and depth of the tree) and model comparisons are done with the validation dataset, and the final model is objectively assessed with the testing dataset. The best model should explain as much variance in the data as possible (i.e., maximize the coefficient of determination R^2) and minimize the overall bias [mean relative error (MRE)], and minimize overfitting. Performance metrics target systematic discrepancies and random errors in the model predictions with respect to observations. MRE is used to quantify systematic error, and the root-mean-square error is used to describe the random error. To quantify overfitting, an ''accuracy loss'' is introduced as the difference in R^2 values obtained when comparing the model predictions with the training and test data. Deeper insight on the model performance is provided by the conditional bias computed along each predictor. It allows us to evaluate whether the model is unbiased multidimensionally.

Data partitioning was performed using stratified random sampling and by splitting data into training, validation, and testing sets using a 70:15:15 ratio. This sampling technique ensures that the subsamples are representative of each other.

During the training step, an XGBoost model derives trees defined by varying depth and number of nodes according to the user's specifications. The ensemble of trees that are generated (or learned) while training becomes the set of parameters for the predictive model. These trees are defined by hyperparameters such as subsample ratios of predictors, learning rate, and max depth. Hyperparameters are tuned to achieve the best model performance (i.e., reducing bias and variance). Cross validation is applied to identify the best model and to improve the model representativeness by using all observations for both training and validation (cf. Fig. 1). Among the possible designs of k-fold cross validation, $k = 4$ was selected as a trade-off between model refinement and computational time. The training dataset (sample) is randomly partitioned into four equal-sized subsamples. Cross validation is performed four times, each time using a different subsample as the inner-fold validation set for checking the model performance, and the remaining three subsamples as inner-fold training sets. The four results are averaged to produce a single estimation for a single hyperparameter combination.

Searching for the best parameter combination in the hyperparameter space occurs by random selection. Parameter combinations are selected 50 times, and each undergoes a fourfold cross validation (cf. Fig. 1). Essentially, 200 (50×4) models are tested to find the best hyperparameters. Once the best hyperparameter combination is identified these settings are used to train the predictive model on the training data. The model performance is then checked on the validation dataset.

An ensemble of models gives more insights into the predictor importance and reduces bias induced (Beven and Binley 1992). An uncertainty analysis implemented as a Monte Carlo experiment is applied to avoid any bias associated with dataset partitioning, to objectively extract a set of empirical models and to identify the most important predictors. The entire method described in Fig. 1 (gray area) is performed 40 times with different subsets of training and validation splits to obtain 40 different models. Note that this approach is different from the generalized likelihood uncertainty estimator (GLUE), which applies on conceptual models to quantify the prediction uncertainty that results from their design and structure. A ''repeated random subsampling validation'' is used to generate multiple random splits of the dataset into training and validation data. For each split, a model fits the training data by identifying the best hyperparameter combination and by assessing the predictive accuracy on the validation data. With 40 such unique random splits, a total of 8000 models are created. Monte Carlo cross validation allows us to keep the proportion of the training/validation split independent from the number of iterations (i.e., the number of partitions) to ensure proper representation of the training and the validation data (Xu and Liang 2001).

c. *Predictor ranking, selection, and importance*

Predictor selection is performed recursively after initial modeling of data. Gradient boosted trees quantify each predictor importance by measuring the mean decrease in impurity

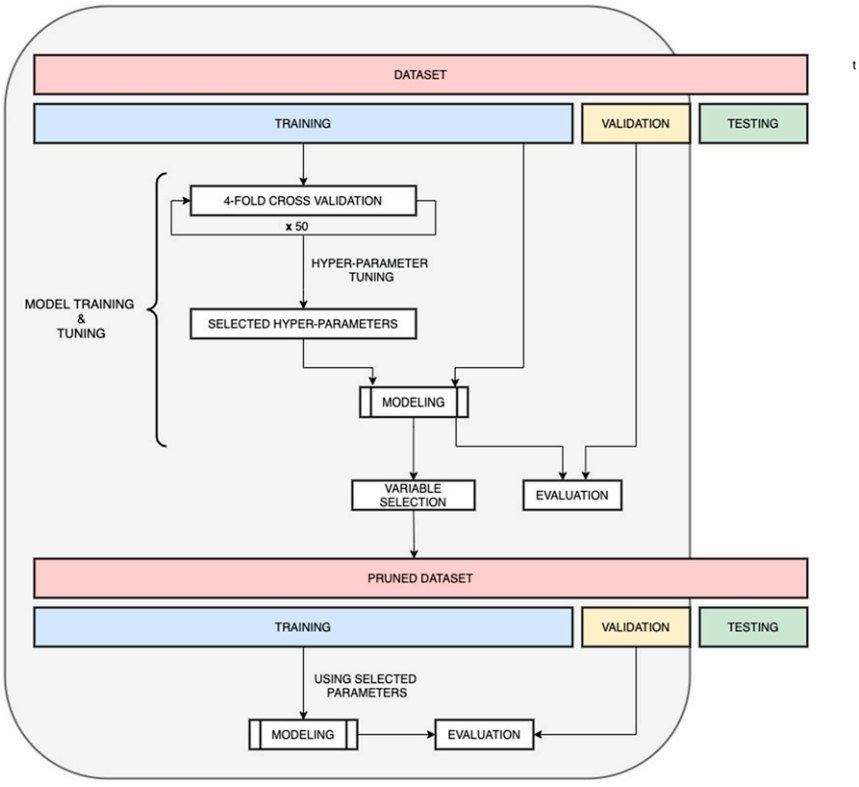

FIG. 1. Modeling method.

(variance) in a forest of trees. It is used to rank predictors as the more the reduction obtained with a given predictor, the higher the importance of the predictor.

Each model provides its own unique predictor importance ranking (Fig. 3). The 40 rankings are aggregated to identify the most important predictors in terms of frequency of occurrence (i.e., how often a predictor is selected) and accumulated importance across the 40 models. A given predictor selected at least once in the 40 iterations is deemed as important and the accumulated importance is used to rank the predictors overall. With this definition, 32 of the 50 predictors are identified as important.

The predictors whose importance is greater than or equal to the mean of all the predictor importance values are selected to create a pruned training/validation/testing dataset, upon which a more parsimonious model is trained. Performance metrics between the parsimonious model and the initial model are systematically compared and found to be negligible. Table 2 presents descriptive statistics for the 16 predictors in the final set.

4. Results

a. Model selection

A model should be general enough that it explains the variance in the entire dataset to a satisfactory level, while minimizing overfitting. Among the 40 models generated, we choose the model with the highest R^2 value on the validation dataset, that keeps the training-to-validation performance loss to less than 15%.

Another substantiation of selection is made with regard to the physical realism of the model. It should reflect physical consistency in terms of predictor importance model partial terms. Among the categories of processes that impact the hydrologic response of a basin, geomorphology is expected to have the greatest impact, followed by the spatial distribution of precipitation (precipitation moments), and finally the climatology. A predictor importance plot provides the overall ranking of predictors across the 40 models and allows us to check their physical consistency, along with our final chosen model. By combining the predictors by their category of processes, a representative value of importance for each category is extracted through the median value.

b. Model performance

Through the selection method, a model is chosen whose performance on the validation dataset is $R^2 = 0.78$. Predictions from this model are compared with the observed peak values from the testing dataset that was untouched in all the selection and ranking procedures. It provides an assessment of the model's representativeness and its predictive power in ungauged basins. Figure 2a shows a scatterplot comparing the predicted and observed peak discharge values that display data pairs grouped along the 1:1 line. A corresponding density-colored scatterplot of the simulated versus observed unit peak

TABLE 2. Descriptive statistics for the predictors.

Predictors	Mean	Std	Min	25%	50%	75%	Max
Area	2130	4020	22	269	888	1980	45 600
River length	84 000	75 400	10 100	35 600	60 300	102 000	659 000
Relief ratio	0.006	0.011	0.000 431	0.001 61	0.002 64	0.006 14	0.168
Slope to outlet	0.0107	0.0107	0.000 577	0.005 34	0.007 81	0.0113	0.192
Temp	12.7	3.77	−0.90	10.1	12.3	15.5	22.7
Ruggedness	0.185	0.245	0.004 21	0.0557	0.107	0.182	2.40
Rock volume	6.84	9.23	0	1.00	3.00	8.00	64.0
Activated basin	1270	2710	2.00	155	483	1260	53 800
Flow distance (mean)	40 900	41 400	1690	16 300	28 600	48 900	491 000
Product mean	321	646	0.0935	8.95	94.2	376	28 900
G1	41 100	38 000	3020	17 100	29 600	49 700	276 000
G2	3 840 000 000	8 820 000 000	11 000 000	356 000 000	1 070 000 000	3 050 000 000	95 900 000 000
Snow percentage	11.6	8.80	0	3.90	11.7	17.0	72.7
bio_10	23.0	2.60	8.34	21.4	23.0	24.8	29.8
bio_15	24.7	13.7	6.02	14.9	20.1	32.8	91.1
Rainfall volume	13.5	49.6	0.000 066 0	0.114	1.74	10.1	3410

discharges (Fig. 2b) can be qualitatively compared with Fig. 2a in Gourley et al. (2017) and suggests superior performance in predicting peak flows as compared with the CREST water balance module with kinematic wave routing. Good scores are reported with a R^2 value of 0.77, mean relative error of 0.02%, root-mean-square error of 157.2 $m^3 s^{-1}$, and limited overfitting as indicated by a train-versus-test R^2 loss of 15%. With such a low mean relative error the model is almost unbiased overall.

The corresponding model that was trained on the pruned dataset that contains only selected variables (see sections 3c and 4c) shows robust performance on the test dataset, with R^2 equal to 0.76 with a train-versus-test R^2 loss of 10%, mean relative error of 0.02%, and root-mean-square error of 160 $m^3 s^{-1}$. This parsimonious model runs faster because it relies on fewer predictors and can be considered as a candidate for real-time prediction of peak discharge.

c. *Feature selection*

The predictor importance plot in Fig. 3 highlights the most important predictors in the final model and allows us to assess the physical consistency of the machine-learning approach. Among these important predictors are nine geomorphological variables, three variables associated with precipitation moments, and four climatological variables. The six most important predictors are geomorphologic attributes, followed by the precipitation moment. It indicates that, as expected, the primary driver of peak discharge is basin physiography, and the spatial organization of rainfall within the basin also has a major influence on the basin response. The rainfall volume is expected to be the most important precipitation moment as it relates to the peak discharge through the volume of water processed by the basin during an event. Precipitation seasonality and mean temperature during the warmest

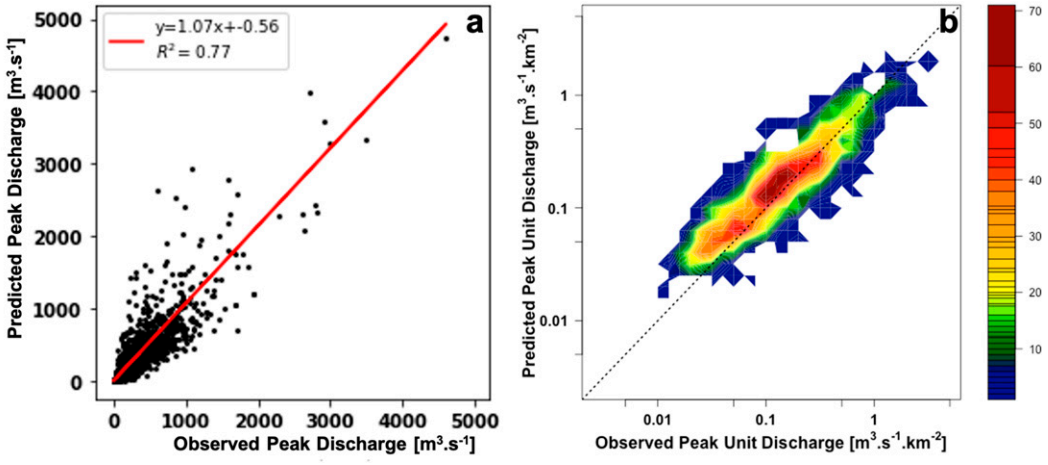

FIG. 2. (a) Predicted vs observed peak discharge values (test dataset) and (b) density-colored scatterplot of the simulated vs observed unit peak discharges (test dataset).

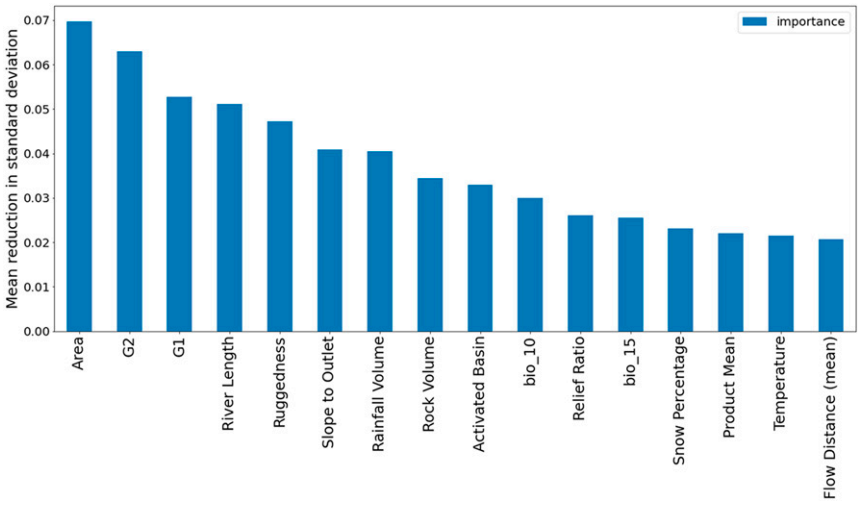

FIG. 3. Predictor importance plot.

quarter highlight the climatological background driving the atmospheric processes associated with flooding.

d. Conditional bias

A conditional bias plot shows the model prediction bias conditioned on each predictor to diagnose and understand the relative dependence of the model accuracy on the predictor values. The global bias (see section 4b) may result from a balance between conditional overestimations and underestimations under various predictors' values. By assessing the extent to which the model is conditionally unbiased with respect to each predictor, conditional bias allows us to assess the model physical realism and whether it captures the multivariate relationships that exist between peak discharge and independent precipitation, climatology, and watershed attributes. Figures 4, 5, and 6 show the conditional biases relative to the observed peak discharge values in the testing dataset for the geomorphological, precipitation variability, and climatological predictors, respectively. To allow a comparison between variables, conditional bias values are computed at percentile bins of each variable. Conditional biases are within 10% for almost all predictors' values, indicating that the model is mostly multidimensionally unbiased and captures the relationships between peak discharge precipitation variability, climatology, and watershed geomorphology.

Figure 4 shows that overestimation is more frequent for low geomorphological values (e.g., small watersheds and low ruggedness) until the 20th percentile. Otherwise, conditional biases tend to be within 5%.

For precipitation variability (Fig. 5), the bias conditioned by rainfall volume shows a decreasing trend starting around +10% below the 10th percentile of rainfall volume and ending around −2% above the 80th percentile. Above the 50th percentile, conditional biases tend to be within 5%.

Conditional biases associated with climatological predictors do not show a particular trend (Fig. 6) but display higher variability around the 0% bias line than their geomorphology and precipitation counterparts. Overall, these results indicate that the model simultaneously and equally captures differences between hydrologic behaviors in different

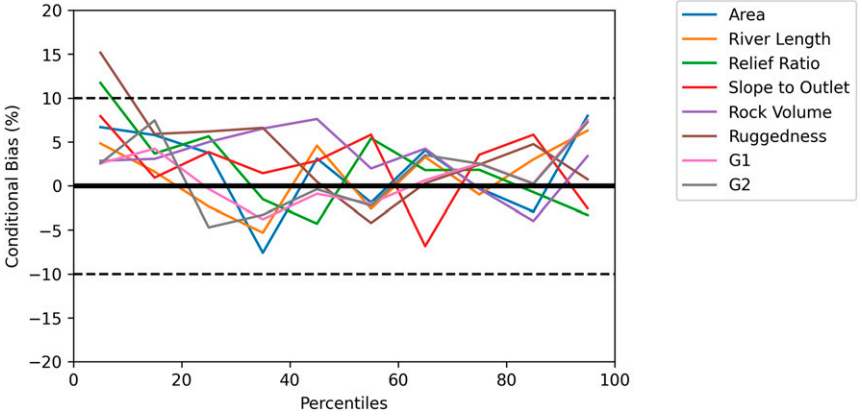

FIG. 4. Conditional bias for geomorphological predictors.

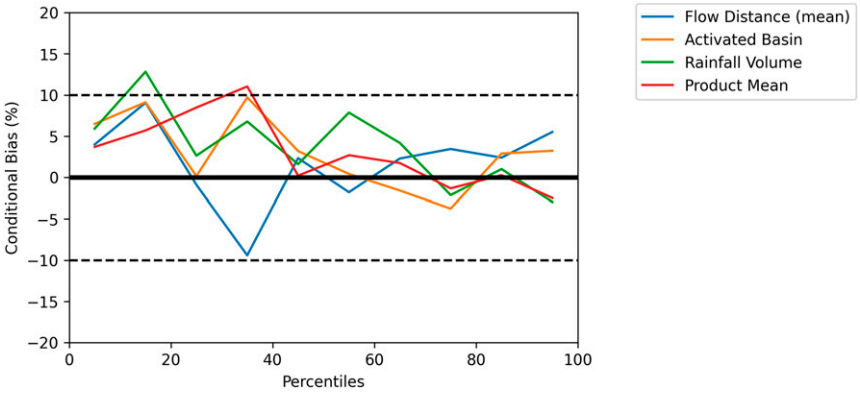

FIG. 5. Conditional bias for precipitation spatiotemporal variability predictors.

catchments, climatologies, and precipitation forcing conditions to a satisfactory extend.

e. *ALE plots*

Accumulated local effects (ALE) plots provide additional insight on the influence of factors on the hydrological response by quantifying the model response to the predictor values. It allows us to compare the relative impact of different factors (geomorphology, spatial distribution of precipitation, and climatology) with respect to flood generation mechanisms. Also, in addition to the importance plots that indicate a global importance (Fig. 3), ALE plots provide the conditional impacts for each predictor.

An ALE plot highlights the average impact of a given predictor on the model predictions (Molnar 2019). ALE plots are unbiased and valid even when predictors are correlated. They help reduce complex prediction functions to a newer function that solely depends on the predictor of interest. To understand the influence of a given predictor in multivariable functions, differences in prediction are calculated for the predictor, which are averaged to define the partial derivative of that predictor and filter out the interaction with correlated predictors, before it is centered to improve the interpretation. Partial derivatives are computed by holding all the other predictors constant and are conditional to the predictor's values. The predictor is binned into intervals to compute prediction differences. Bins are based on percentile values taken by the predictors to ensure uniformity across bins and predictors. The differences in the prediction relay the predictor's effect in terms of partial derivative for each individual instance in a bin. The partial derivatives are conditionally averaged over each bin to estimate the local effects. The local effects are accumulated across all bins to derive ALE values, that are finally centered.

The ALE plots for geomorphological attributes, spatial distribution of precipitation, and climatology are provided in Figs. 7, 8, and 9, respectively. To interpret the ALE values, one should consider the value on the y axis as the conditional effect of the given predictor, when compared with the overall mean prediction for that bin. For instance, in Fig. 6 the difference in peak discharge is $-65\,\mathrm{cm^3\,s^{-1}}$ for the 10th percentile of the predictor area, that is, the prediction is lower by $50\,\mathrm{cm^3\,s^{-1}}$ in comparison with the mean prediction involving all predictors. To allow a comparison between variables, ALE values are computed at percentile bins of each variable. Consistent with the importance plot (Fig. 3), in general the geomorphological ALE plots display larger ranges of variations than the precipitation and the climatological ALE plots, indicating that the geomorphological predictors have a higher impact on the model output (i.e., they

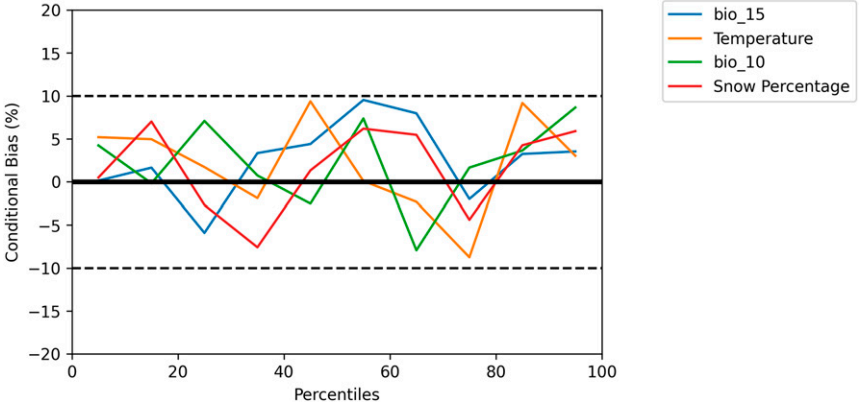

FIG. 6. Conditional bias for climatological predictors.

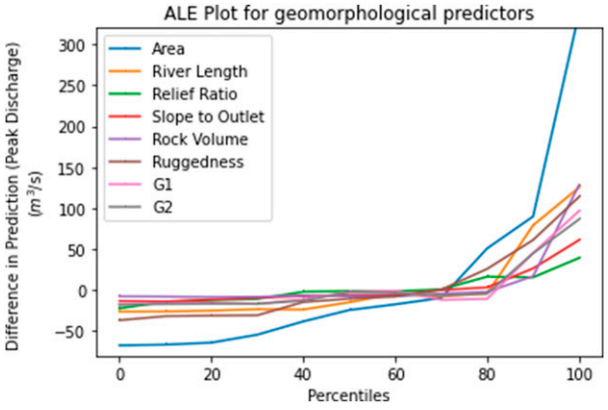

FIG. 7. ALE analysis for geomorphological predictors.

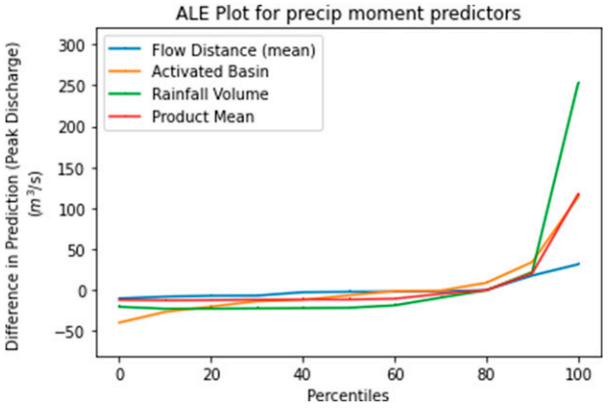

FIG. 8. ALE analysis for precipitation spatiotemporal variability predictors.

generate higher differences in predicted peak discharge), while the climatological predictors have less impact.

As expected, the basin area ALE shows the largest range and a positive trend in the range $[-65; 300]$ $m^3 s^{-1}$, again highlighting the positive relation between the surface collecting precipitation and the peak discharge. The relation is not linear, and the impact of area increases with area itself. Similarly, the peak discharge increases with relief ratio, although to a lesser extent. Higher terrain slope is expected to accelerate the conversion of precipitation into runoff, leading to higher peak discharge values.

For precipitation moments, rainfall volume has the most impact on the model outputs as it relates to the volume of water contributing to the peak discharge. Interestingly, the impact is limited before the 85th percentile and significant over the last 15% percentiles of rainfall volume. This may reflect a competition between the rainfall forcing and hydrological processes on the flood response, and that the forcing overcomes the latter through extreme rain rates. Activated basin also relays similar information and hence is also shown to be important in the prediction. The mean flow distance primarily relates to water routing and to the delay between the rainfall forcing and the flood occurrence rather than to the peak discharge, hence the weaker impact on the model outputs. In a general, precipitation moments have a significant impact on peak discharge.

Climatological ALE plots are flat, indicating that climatological predictors have limited impact on the model outputs. Bio_10 (mean temperature of warmest quarter) shows a positive impact with extreme quantiles. As large values of peak discharge are associated with high precipitation rates, it may identify regions with thunderstorms and convection that generate extreme precipitation. These processes tend to occur in warm areas and seasons.

5. Conclusions

Peak discharge, a key characteristic of floods and flash floods, was successfully modeled at the hydrological event scale using a robust machine-learning method applied on a dataset of 21 143 flooding events that gathers a large variety of basin physiographical, precipitation variability, and climatological characteristics across the United States. Multifold cross validation and independent evaluation results (sections 4b and 4d) demonstrate that a machine-learning approach can capture both local, regional, and event-specific hydrologic behaviors, by learning universal similarities and differences from the combined data from diverse hydrologic events.

The approach differs from traditional hydrological modeling in which the best results are usually obtained with models independently calibrated for each basin. The ability of XGBoost to simultaneously learn the interdependence between basin physiography, precipitation variability, climatology, and their impact on peak discharge, allows to bypass challenges associated with the estimation and transfer of hydrologic model parameters. An innovation of this study is to demonstrate regional modeling at the event scale, including the flash flood scale, with promising comparison with respect to a state-of-the art hydrologic modeling framework (section 4b). In addition, it implies that the information gathered in the dataset appears sufficient to characterize diverse event-specific hydrologic behaviors across catchments and precipitation events to a reasonable extent. This has proved challenging to demonstrate with traditional

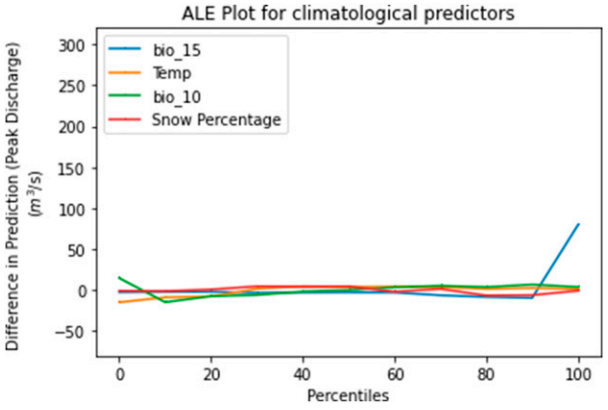

FIG. 9. ALE analysis for climatological predictors.

approaches that struggle to maintain hydrologic modeling skills from a single basin to multiple basins together.

Multivariate relationships between the peak discharge and a large number of explanatory variables such as the moments of rainfall and flow distance, spatial variability indices, geomorphologic and climatologic factors were modeled using a multidimensional framework. Along with the relationships, the significance of these factors was established by objective selection through the training procedure, and the relative influences of these factors on peak discharge were also assessed, thereby, yielding an improved understanding of these dependencies (sections 4c and 4e). The basin physiography is confirmed to have a higher impact than precipitation variability, while climatology has a lower impact.

In terms of perspectives, the conditional bias analysis reveals that while peak discharge predictions display good performances overall, they overestimate the peak discharge observed at catchments characterized by low geomorphological attributes (e.g., small catchments, flat terrain). This should be examined in a future study by including additional predictors to enhance the information content on event-specific hydrologic behaviors. Other characteristics of floods and flash floods can be considered in future analyses such as flashiness, flood duration, flood threshold exceedance levels. The method outlined in this study can be used to create efficient machine-learning models that can be compared with existing flood forecasting systems such as EF5.

Acknowledgments. This work was supported by the NOAA GOES-R Risk Reduction Science Program Award NA16OAR4320115 and the NOAA Joint Technology Transfer Initiative Award NA17OAR4590170. The authors also thank the three anonymous reviewers, along with the editors, for providing useful comments, which greatly improved the paper.

Data availability statement. This reanalysis was performed on the raw, publicly available NEXRAD data archive available from Amazon Web Services (https://aws.amazon.com/public-datasets/nexrad/). The Flood Database is publicly available (https://blog.nssl.noaa.gov/flash/database/).

REFERENCES

Alfieri, L., P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, and F. Pappenberger, 2013: GloFAS—Global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.*, **17**, 1161–1175, https://doi.org/10.5194/hess-17-1161-2013.

Ashley, S. T., and W. S. Ashley, 2008: Flood fatalities in the United States. *J. Appl. Meteor. Climatol.*, **47**, 805–818, https://doi.org/10.1175/2007JAMC1611.1.

Blöschl, G., and M. Sivapalan, 1995: Scale issues in hydrological modelling: A review. *Hydrol. Processes*, **9**, 251–290, https://doi.org/10.1002/hyp.3360090305.

——, ——, T. Wagener, A. Viglione, and H. Savenije, Eds., 2013: *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*. Cambridge University Press, 490 pp.

Beven, K., 2001: How far can we go in distributed hydrological modelling? *Hydrol. Earth Syst. Sci.*, **5**, 1–12, https://doi.org/10.5194/hess-5-1-2001.

——, and A. Binley, 1992: The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Processes*, **6**, 279–298, https://doi.org/10.1002/hyp.3360060305.

Brownlee, J., 2016: *XGBoost with Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*. Machine Learning Mastery, 115 pp., https://www.goodreads.com/book/show/50621772-xgboost-with-python.

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

Emmanuel, I., H. Andrieu, E. Leblois, N. Janey, and O. Payrastre, 2015: Influence of rainfall spatial variability on rainfall-runoff modelling: Benefit of a simulation approach? *J. Hydrol.*, **531**, 337–348, https://doi.org/10.1016/j.jhydrol.2015.04.058.

Fernández-Delgado, M., M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, 2019: An extensive experimental survey of regression methods. *Neural Networks*, **111**, 11–34, https://doi.org/10.1016/j.neunet.2018.12.010.

Flamig, Z. L., H. Vergara, and J. J. Gourley, 2020: The Ensemble Framework For Flash Flood Forecasting (EF5) v1.2: Description and case study. *Geosci. Model Dev.*, **13**, 4943–4958, https://doi.org/10.5194/gmd-13-4943-2020.

Fry, J. A., and Coauthors, 2011: Completion of the 2006 National Land Cover Database for the conterminous United States. *Photogramm. Eng. Remote Sens.*, **77**, 858–864.

Gauch, M., J. Mai, S. Gharari, and J. Lin, 2019: Data-driven vs. physically-based streamflow prediction models. *Proc. Ninth Int. Workshop on Climate Informatics*, Paris, France, École Normale Supérieure, 5 pp., https://cs.uwaterloo.ca/~jimmylin/publications/Gauch_etal_CI2019.pdf.

Gourley, J. J., and Coauthors, 2013: A unified flash flood database across the United States. *Bull. Amer. Meteor. Soc.*, **94**, 799–805, https://doi.org/10.1175/BAMS-D-12-00198.1.

——, and Coauthors, 2017: The FLASH project: Improving the tools for flash flood monitoring and prediction across the United States. *Bull. Amer. Meteor. Soc.*, **98**, 361–372, https://doi.org/10.1175/BAMS-D-15-00247.1.

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.

Hu, R., F. Fang, C. C. Pain, and I. M. Navon, 2019: Rapid spatio-temporal flood prediction and uncertainty quantification using a deep learning method. *J. Hydrol.*, **575**, 911–920, https://doi.org/10.1016/j.jhydrol.2019.05.087.

Kirstetter, P.-E., H. Andrieu, G. Delrieu, and B. Boudevillain, 2010: Identification of vertical profiles of reflectivity for correction of volumetric radar data using rainfall classification. *J. Appl. Meteor. Climatol.*, **49**, 2167–2180, https://doi.org/10.1175/2010JAMC2369.1.

——, ——, B. Boudevillain, and G. Delrieu, 2013: A physically based identification of vertical profiles of reflectivity from volume scan radar data. *J. Appl. Meteor. Climatol.*, **52**, 1645–1663, https://doi.org/10.1175/JAMC-D-12-0228.1.

Kratzert, F., D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, 2018: Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.*, **22**, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018.

——, ——, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing, 2019: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.*, **23**, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019.

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.

Molnar, C., 2019: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com, 320 pp., https://christophm.github.io/interpretable-ml-book/.

Mosavi, A., O. Pinar, and C. Kwok-wing, 2018: Flood prediction using machine learning models: Literature review. *Water*, **10**, 1536, https://doi.org/10.3390/w10111536.

Ni, L., D. Wang, J. Wu, Y. Wang, Y. Tao, J. Zhang, and J. Liu, 2020: Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. *J. Hydrol.*, **586**, 12 4901, https://doi.org/10.1016/j.jhydrol.2020.124901.

NWS, 2010: Floods: The awesome power. NOAA Publ., 16 pp., https://www.weather.gov/media/jetstream/tstorms/floods_booklet.pdf.

Orzechowski, P., W. La Cava, and J. H. Moore, 2018: Where are we now? A large benchmark study of recent symbolic regression methods. *Proc. Genetic and Evolutionary Computation Conf. (GECCO'18)*, Kyoto, Japan, Association for Computing Machinery, 1183–1190, https://doi.org/10.1145/3205455.3205539.

Razavi, T., and P. Coulibaly, 2013: Streamflow prediction in ungauged basins: Review of regionalization methods. *J. Hydrol. Eng.*, **18**, 958–975, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690.

Saharia, M., P.-E. Kirstetter, H. Vergara, J. J. Gourley, Y. Hong, and M. Giroud, 2017: Mapping flash flood severity in the United States. *J. Hydrometeor.*, **18**, 397–411, https://doi.org/10.1175/JHM-D-16-0082.1.

Salas, F. R., and Coauthors, 2018: Towards real-time continental scale streamflow simulation in continuous and discrete space. *J. Amer. Water Resour. Assoc.*, **54**, 7–27, https://doi.org/10.1111/1752-1688.12586.

Segal, M., 1988: Regression trees for censored data. *Biometrics*, **44**, 35–47, https://doi.org/10.2307/2531894.

Sillmann, J., V. V. Kharin, F. W. Zwiers, X. Zhang, and D. Bronaugh, 2013: Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *J. Geophys. Res. Atmos.*, **118**, 2473–2493, https://doi.org/10.1002/jgrd.50188.

Smith, A. B., 2020: 2010–2019: A landmark decade of U.S. billion-dollar weather and climate disasters. NOAA, https://www.climate.gov/news-features/blogs/beyond-data/2010-2019-landmark-decade-us-billion-dollar-weather-and-climate.

Smith, J. A., M. L. Baeck, K. L. Meierdiercks, P. A. Nelson, A. J. Miller, and E. J. Holland, 2005: Field studies of the storm event hydrologic response in an urbanizing watershed. *Water Resour. Res.*, **41**, W10413, https://doi.org/10.1029/2004WR003712.

Solomatine, D. P., and T. Wagener, 2011: Hydrological modeling. *Treatise on Water Science*, Vol. 2, Elsevier, 435–457, https://doi.org/10.1016/B978-0-444-53199-5.00044-0.

Strahler, A. N., 1964: Quantitative geomorphology of drainage basins and channel networks. *Handbook of Applied Hydrology*, V. Chow, Ed., McGraw Hill, 439–476.

Sweeney, T. L., 1992: Modernized areal flash flood guidance. NOAA Tech. Rep. NWS HYDRO 44, 21 pp., https://repository.library.noaa.gov/view/noaa/13498.

Vergara, H., P. E. Kirstetter, J. J. Gourley, Z. L. Flamig, Y. Hong, A. Arthur, and R. Kolar, 2016: Estimating a-priori kinematic wave model parameters based on regionalization for flash flood forecasting in the conterminous United States. *J. Hydrol.*, **541**, 421–433, https://doi.org/10.1016/j.jhydrol.2016.06.011.

Xu, Q., and Y. Liang, 2001: Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.*, **56**, 1–11, https://doi.org/10.1016/S0169-7439(00)00122-2.

Yu, X., Y. Wang, L. Wu, G. Chen, L. Wang, and H. Qin, 2020: Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. *J. Hydrol.*, **582**, 124293, https://doi.org/10.1016/j.jhydrol.2019.124293.

Zhang, J., and J. Gourley, 2018: Multi-Radar Multi-Sensor Precipitation Reanalysis (Version 1.0). Open Commons Consortium Environmental Data Commons, accessed 15 July 2019, https://doi.org/10.25638/EDC.PRECIP.0001.

——, and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zoccatelli, D., M. Borga, A. Viglione, G. B. Chirico, and G. Blöschl, 2011: Spatial moments of catchment rainfall: Rainfall spatial organisation, basin morphology, and flood response. *Hydrol. Earth Syst. Sci.*, **15**, 3767–3783, https://doi.org/10.5194/hess-15-3767-2011.