Distributed Flashiness-Intensity-Duration-Frequency products over the conterminous US

Zhi Li[1,*], Shang Gao[2], Mengye Chen[1], Jiaqi Zhang[1], Jonathan J. Gourley[3], Humberto Vergara[4], Siyu Zhu[1], Sebastian Ferraro[1], Yixin Wen[5], Tiantian Yang[1], Yang Hong[1,*]

[1] School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK, USA

[2] School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, USA

[3] NOAA/National Severe Storms Laboratory, Norman, OK, USA

[4] College of Engineering, University of Iowa, Iowa City, IA, USA

[5] Department of Geography, University of Florida, Gainesville, FL, USA

Corresponding to: Zhi Li (li1995@ou.edu), Yang Hong (yanghong@ou.edu)

Key Points:

- We developed distributed flashiness-intensity-duration-frequency products with machine learning and hydrologic simulation
- Both products can identify flash flood-prone regions in the CONUS
- We cross-compared both products over the CONUS and highlight their strengths and limitations
- The utility of the two products is discussed with their synergistic use by decision makers

Abstract

Effective flash flood forecasting and risk communication are imperative for mitigating the impacts of flash floods. However, the current forecasting of flash flood occurrence and magnitude largely depends on forecasters' expertise. An emerging flashiness-intensity-duration-frequency (F-IDF) product is anticipated to facilitate forecasters by quantifying the frequency and magnitude of an imminent flash flood event. To make this concept usable, we develop two distributed F-IDF products across the contiguous US, utilizing both a Machine Learning (ML) approach and a physics-based hydrologic simulation approach that can be applied at ungaged pixels. Specifically, we explored 20 common ML methods and interpreted their predictions using the Shapley Additive exPlanations method. For the hydrologic simulation, we applied the operational flash flood forecast framework – EF5/CREST. It is found that: (1) both CREST and ML depict similar flash flood hot spots across the CONUS; (2) The ML approach outperforms the CREST-based approach, with the drainage area, air temperature, channel slope, potential evaporation, soil erosion identified as the five most important factors; (3) The CREST-based approach exhibits high model bias in regions characterized by dam/reservoir regulation, urbanization, or mild slopes. We discuss two application use cases for these two products. The CREST-based approach, with its dynamic streamflow predictions, can be integrated into the existing real-time flash flood forecast system to provide event-based forecasts of the frequency and intensity of floods at multiple durations. On the other hand, the ML-based approach, which is a static measure, can be integrated into a flash flood risk assessment framework for urban planners.

1. Introduction

1.1 Background

Flash floods are a type of flooding that occur rapidly, often within a few minutes or hours of the onset of rainfall (Hong et al., 2013). Flash floods are oftentimes a weather phenomenon, which is closely tied to storms (e.g., convective system, squall lines, supercells) in the US (Doswell et al., 1996; Maddox et al., 1979). Forecasting flash floods is perceived as one of the grand challenges within the hydrology community. Weather forecasting inherently carries significant challenges. When considering flash flood forecasting, an additional uncertainty arises due to the impact of land surface that can both act as a buffer or even exacerbate flooding.

Forecasting flash flood qualitatively is difficult, forecasting and quantifying the specific magnitudes of flash flooding at a specific location is much more challenging. Due to these challenges, operational forecasting flash floods on a national scale was not feasible until the 1980s (Georgakakos, 1986). Two types of threshold-based guidance products have emerged and are currently being utilized by forecasters at the National Weather Service (NWS). The Flash Flood Guidance (FFG), implemented after a deadly 1969 flash flood in Ohio, has become a national standard for weather forecasters henceforth (Clark et al., 2014). Taking quantitative precipitation estimates (QPE) as inputs, FFG determines if the amount of rain will produce bank-full conditions on streams. However, FFG does not account for the land cover and routing in simulating pluvial flash flooding. Hydrologic models, on the other hand, simulate the rainfall-runoff processes to predict the occurrence of flash floods with unit streamflow values (Gourley et al., 2017). With increasing available computational resources, flash flood forecast products derived from hydrologic models are beginning to play a more prominent role in predictive storm warning and disaster management. Gourley & Vergara (2021) found the equitable threat score generally increases with the sophistication of flash flood forecast products, particularly highlighting the importance of land cover and surface routing process.

1.2 Problem statement

Previously developed flash flooding methods present several challenges related to forecast ability and risk communication. First and foremost, the threshold-based system, as previously discussed, is a form of subjective guidance that necessitates the incorporation of past experience. For instance, the best predictors of flash flood occurrence were with 1- and 3-h rainfall that exceeded FFG by ratios greater than 100% (Clark et al., 2014). For the unit streamflow simulated by a hydrologic model, this threshold is subject to different model simulations and configuration (Gourley et al., 2021). There is an absence of a comprehensive, objective reference system to support decision-making process (Morss et al., 2016). Second, the severity of a flash flood event is still challenging to describe to the public, with respect to risk communication. Despite its frequent misuse in the news press, the terms such as '100-year flood' often used in frequentist statistics, provide the public with a perception of flood risk. However, such frequency associations are not available for flash floods, primarily because they require a quantifiable measure to describe their nature – specifically, the speed and depth of the water flow. These two factors hinder effective communication between decision-makers and the public, consequently placing vulnerable communities at increased risk.

1.3 New promises

In light of these issues, Li et al. (2023) first proposed a new metric called Flashiness-Intensity-Duration-Frequency (F-IDF), analogous to rainfall IDF in a way that attempts to quantify a flash flood event by its duration and return periods. This not only allows us to determine the likelihood of a flash flood event but also enables us to quantify its severity (such as a 100-year flash flood event). As a proof-of-concept, our previous study was conducted only at 3,722 stream gage sites across the contiguous US (CONUS), but we recognize the pressing need to be generalized to ungaged areas. This study aims to develop a distributed F-IDF product that addresses the data gap of ungaged basins, particularly in urban areas. In pursuing this goal, we employ two methods. The first is a traditional approach that relies on a distributed hydrologic model, which resolves the rainfall-runoff process at a flash flood scale (i.e., 1 km and 10 minutes) over the CONUS. The second is an emerging statistical approach that uses Machine Learning (ML) to construct the correlation between basin attributes and F-IDF quantities. Albeit with the same end product, these two methods are distinct in the way that they are developed. The hydrologic simulation, despite being less accurate than ML models as demonstrated by many studies (Kim et al., 2021; Ouyang et al., 2021), provides an interpretable framework that enhances our understanding of hydrologic processes (Clark et al., 2008). Conversely, while ML models may offer superior solutions (because of targeted training), they present challenges in interpreting the underlying hydrologic processes (Shen, 2018). This study advocates the synergistic application of these two approaches for decision making and risk management to mitigate flash flood risks. The objectives of this study are threefold: (1) To develop first-of-its-kind distributed F-IDF products over the CONUS based on both a physics-based model and an ML model; (2) To cross-compare the advantages and limitations of each approach; (3) To discuss the utility of both products and benefits of their synergistic use.

The rest of this paper is organized as follows. Section 2 introduces the data used in this study and the framework we propose for this work. Section 3 elucidates the results of this study regarding model verification, cross comparison, and presents a case study. In Section 4, we discuss the limitations of the model simulation and the utility of the F-IDF products.

2. Data and Methods

2.1 Data for hydrologic simulation

We use the CREST hydrologic model to simulate sub-hourly streamflow from 2001 to 2012. The model inputs include precipitation and potential evapotranspiration as forcings and a set of a-priori parameters at a desired resolution (i.e., 1 km). We use the Multi-Radar Multi-Sensor reanalysis product at 10-min time intervals over the CONUS to provide precipitation data (Zhang & Gourley, 2018) and the USGS monthly climatological potential evapotranspiration for the model (Allen et al., 1998). The MRMS is a radar-gauge merged quantitative precipitation estimation (QPE) product by merging 180 operational radars and creating a 3D radar mosaic over the CONUS (Zhang et al., 2016). A set of calibrated a-priori model parameters are accessed from https://github.com/chrimerss/EF5-US-Parameters, and the model performance with such data is evaluated by Vergara et al. (2016) and Flamig et al. (2020).

2.2 RiverAtlas data

The training features for the ML-based model arise from the RiverAtlas v10 dataset, hosted on the hydrosheds website (https://www.hydrosheds.org/hydroatlas) (Lehner et al., 2022). The RiverAtlas data are a compilation of river attributes, spanning eight sections: (1) Hydrology (e.g., annual runoff, natural discharge, groundwater table), (2) Physiography (e.g., channel slope, basin slope, elevation, drainage area), (3) Climate (e.g., annual precipitation, actual evaporation, climate moisture index, aridity index), (4) Soils & Geology (e.g., soil water content, clay fraction, silt fraction, karst fraction), (5) Anthropogenic (e.g., road density, urban density, population), (6) Land cover (e.g., area extent of trees, shrubs, herbaceous), (7) Natural vegetation (e.g., evergreen, deciduous, savanna), and (8) Wetland (e.g., peatland, river). Overall, 59 river attributes are used as training features, and a detailed table of these attributes can be found in Supplementary Table 1.

2.3 Framework

Figure 1 depicts the overall framework used in this study to produce distributed F-IDF values over the CONUS. This framework intends to produce two distributed F-IDF products covering the CONUS. One is CREST-based F-IDF that is generated by the CREST hydrologic model and fits an Extreme Value Distribution (EVD). A counterpart is machine-learning (ML)-based F-IDF that is an extrapolation of gage-based F-IDF values over the CONUS, which was conducted by Li et al. (2023). Another distinct feature of these two approaches is their spatial representativeness. The CREST-based F-IDF product is gridded, with the cell size the same as the distributed hydrologic model (i.e., 1 km). The ML-based F-IDF product is river reach-based since the hydrologic attributes are aggregated in hydrologic response units (i.e., sub-basins) and assigned to corresponding river reaches. The methods for calculating CREST-based and ML-based F-IDF are articulated in Sections 2.5 and 2.6, respectively.

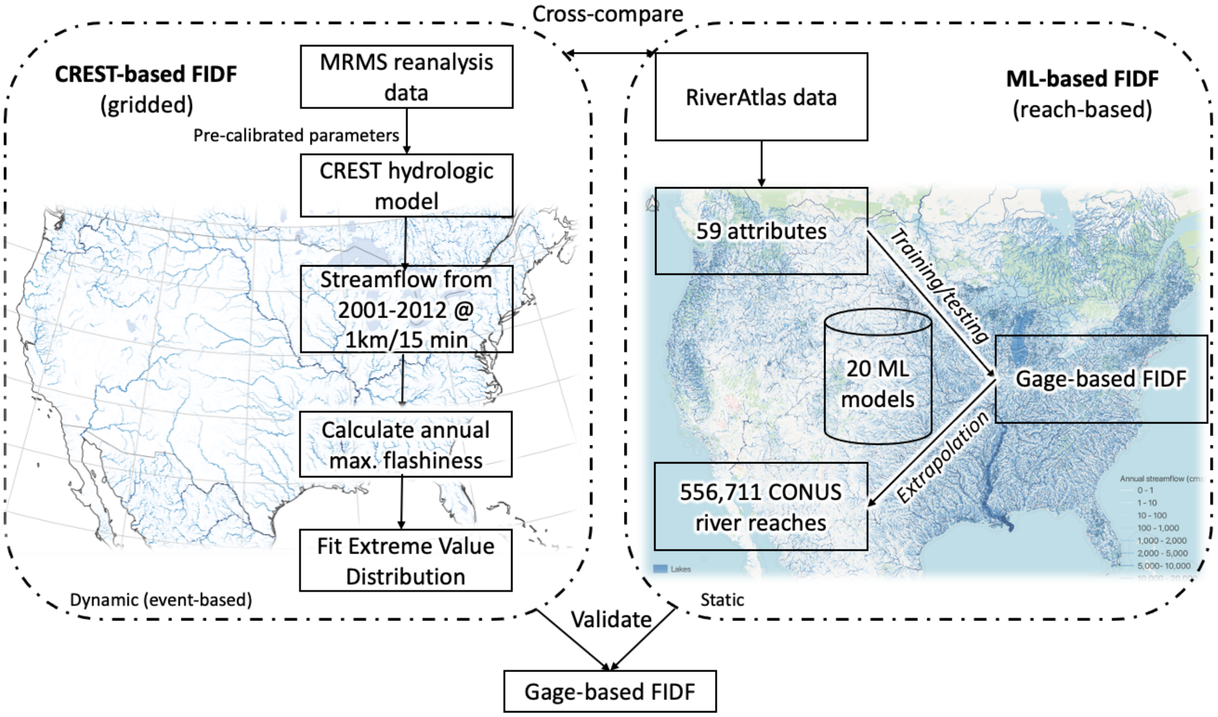

Figure 1. A schematic framework of the two approaches.

2.4 Definition of Flashiness-Intensity-Duration-Frequency

We have introduced the definition of F-IDF in Li et al. (2023) and reiterate the core concept here. The rationale for proposing a new metric is three-fold. First, this new metric quantifies the severity of a flash flood event with return periods (e.g., a 100-year flash flood event). Second, flash flood events are multi-dimensional, meaning that the duration of the event impacts the severity of the event. Third, the F-IDF is a tailored metric that can assist decision-makers in planning for and mitigating flash flood risks. The calculation of the F-IDF is as follows. First, we compute the flashiness index (Eq. 1), which is the slope of a hydrograph over a moving window that represents the duration of an event. Then, the annual maximum flashiness index is extracted by aggregating the time series. Lastly, we fit the annual maximum values into GEV and extract flashiness values for desired flash flood return periods. The flashiness values, in principle, reflect the speed at which the flood rises and the magnitude of the flood peak. Although the definitions for the flashiness index are variable, we see similarities in different methods from identified flash flood hot spots (Li et al., 2023). In addition, our method is fairly simple and reproducible compared to others (Gannon et al., 2022; Saharia et al., 2017; Smith & Smith, 2015).

$$F = \frac{\max\{Q_t - Q_{t-1}, Q_t - Q_{t-2}, \ldots\ Q_t - Q_{t-d}\}}{FAC \times d}, \tag{1}$$

where Q_t is the streamflow time series at time t, d is the duration from 1 hour to 6 hours, FAC is the drainage area (km^2). By transforming the streamflow to unit streamflow, we account for streamflow generally increasing with drainage basin size. The unit of F is dependent on the streamflow units and modeling frequency but is generally expressed in units of [L/T^2]. We standardize the unit of flashiness value to be measured in mm/h^2. In this study, we use the simulated streamflow at a 10-minute time interval, so a conversion factor of 21.6 is applied to convert $m^3/s/km^2/10\text{-min}$ to mm/h^2.

2.5 The CREST-based approach

In this study, we leverage the Coupled Routing and Excess STorage (CREST) model for its strength in flood prediction. The CREST model was jointly developed by the University of Oklahoma and NASA (Wang et al., 2011), as the first hydrologic model operated by NASA for global flood forecast during the Tropical Rainfall Measuring Mission era (Wu et al., 2012). Since its inception in 2011, the CREST model has primarily served as a flood-centric distributed hydrologic model that encapsulates a suite of remote sensing products (Chen et al., 2022; Wang et al., 2011; Li et al., 2023). As a component of the Ensemble Framework For Flash Flood Forecast (EF5) framework, EF5/CREST has been an operational setup for real-time flash flood forecast by NOAA/NSSL since 2016 and provides critical and timely information for weather forecasters in the continental US (http://flash.ou.edu/; Gourley et al., 2017). While we concentrate on the application of F-IDF using CREST in this study, F-IDF values can be generated using any distributed hydrologic model.

We simulate the 11-year streamflow using CREST from 2001 to 2011, with the first year reserved for warming up the model states. The MRMS precipitation reanalysis data at a 10-min interval and 1-km spatial resolution are used to drive the model. The model setup, such as grid resolution (1km) and a-priori parameters, are the same as the operational one, and its performance has been assessed by Flamig et al. (2020). The output streamflow is produced every 10 minutes to capture the nature of flash floods. With the streamflow values at each 1km grid cell, we extract the ten-year time series (10 years x 365 days/year x 24 hours/day x 6 10-minute/hour=525,600 time steps) and follow the F-IDF calculation as detailed in Section 2.4. We repeat this process for 4 million grid cells that have flow accumulation values greater than 1 km^2 over the CONUS to generate a distributed F-IDF product.

2.6 Machine learning based approach

Given the nature of how river attributes are aggregated, we perform the ML model at a river reach level over the CONUS using the riverATLAS dataset. Fifty-nine river attributes are fed into a suite of ML models for training on 3,722 USGS streamgage sites and then applied for 556,771 river reaches. To build the gage-based F-IDF product for ML, we extract the 15-minute streamflow time series from 1950 to 2020. These time series were fed into the F-IDF calculation as described in Section 2.4 (Li et al., 2023). With no prior information on ML model performance, we selected 20 commonly used ML models including linear, tree-based, kernel-based, and instance-based models. They are Light Gradient Boosting Machine, Random Forest Regressor, Gradient Boosting Regressor, Extra Trees Regressor, Extreme Gradient Boosting, K Neighbors Regressor, Ridge Regression, Linear Regression, Elastic Net, Lasso Least Angle Regression, Lasso Regression, Decision Tree Regressor, Bayesian Ridge, Least Angle Regression, Huber Regressor, Orthogonal Matching Pursuit, Dummy Regressor, and Passive Aggressive Regressor. A table of detailed descriptions for each model is listed in Supplementary Table 2. We use the pycaret package in Python to benchmark and automate workflows (Ali, 2020).

To split the training-testing samples, we adhere to the 70-30 principle, in which 70% of the samples are used for training, and the rest is for testing. Beyond that, we perform a 10-fold cross-validation to select the best-performing ML model out of 20 models for each return period and duration. Given six return periods (i.e., 2-yr, 5-yr, 10-yr, 25-yr, 50-yr, and 100-yr) and six durations (i.e., 1-hr, 2-hr, 3-hr, 4-hr, 5-hr, and 6-hr) of flashiness values, 36 ML models are retained for further evaluation. Because the distribution of flashiness values is positively skewed, meaning that a large number of samples are concentrated on the low end, we transform the flashiness data to resemble a Gaussian-like distribution using the Box-Cox transformation (Eq.2).

$$F' = \begin{cases} log(F), \; if \; \lambda = 0 \\ (F^{\lambda} - 1)/\lambda, \quad otherwise \end{cases} \tag{2}$$

where F' is the transformed flashiness values, F is the original flashiness values before transformation, and λ is the parameter chosen so that the distribution approximates a normal distribution. The optimal λ can be calibrated by maximizing the log-likelihood function.

2.7 Explainable Machine Learning

The Shapley Additive exPlanations (SHAP) values are used in this study to interpret the contribution of each feature to the overall prediction of flashiness values. Based on the concept of cooperative game theory, the SHAP estimates the contribution of each feature to the prediction for every instance (i.e., feature present or not) (Lundberg & Lee, 2017). Put differently, the SHAP value can be considered as the average marginal contribution of a feature value across all possible coalitions. Eq. 3 shows the mathematical expression of a shapley value given a prediction model f and an instance x:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{3}$$

where N is the set of all features; S is a subset of N that does not include feature i; $|S|$ is the number of elements in S; $|N|$ is the total number of features; $f(S \cup \{i\})$ is the prediction of the model with features in S and i; $f(S)$ is the prediction of the model with features in S only.

In practice, the SHAP values are generally calculated through the following steps. First, we enumerate all possible combinations of features. For a given instance to be predicted, we consider all possible combinations of input features. Given 59 features in our case, computing all combinations is infeasible, as the total number of combinations is $2^{59} - 1 > 5 \times 10^{17}$. We decided to select only the 20 best features, so the number of combinations becomes 1,048,575. The selection criterion is based on the univariate statistical tests – the F-statistic in this case – to measure the general significance of the explanatory factor in regression analysis. Second, we calculate the prediction with and without a particular feature and record the difference as the marginal contribution of that feature for that combination. Third, we calculate the average of its marginal contributions across all combinations.

3. Results

3.1 Model verification

We evaluate the model performance with respect to calculated flashiness values for the testing samples used by the ML approach. The Spearman correlation coefficient (CC) is used to depict the goodness-of-fit of predicted flashiness values and target values.

3.1.1 ML-based approach

The ML-based approach depicts an overall good fit (mean CC>0.9) between predicted flashiness and target flashiness values (processed from USGS streamgages), indicating that the 59 hydrologic attributes adequately explain the variability of flashiness values over the CONUS (Fig. 2). Among 36 enumerations (6 frequencies x 6 durations), the Light Gradient Boosting Machine model tops in 33 combinations, except for the 25yr-3hr, 100yr-2hr, and 100-6hr, which are best predicted by Gradient Boosting Machine, Random Forest, and Gradient Boosting Machine, respectively. In general, tree-based machine learning models perform better than linear models, instance-based models (i.e., k Neighbors Regressor), and kernel-based models (i.e., Support Vector Machine); and ensemble-based models perform better than deterministic models. The tree-based models resemble human decision-making processes and have been widely applied in flood attribution and for identifying flood-generating mechanisms (Kemter et al., 2023; Stein et al., 2021).

Figure 2 also indicates that ML model performance deteriorates with increasing return periods (column-wise comparison), but improves with longer durations (row-wise comparison). When referring to performance improvement (or deterioration), we mean not only the increase (or decrease) in CC but also the decrease (or increase) in the uncertainty spread, as indicated by the contour area. This is expected for two reasons. First, for rare events (e.g., 1-in-100-year), static hydrologic signatures become less impactful while it depends more on the event characteristics such as event rainfall, antecedent soil moisture, channel routing, etc. In other words, as the rainfall event magnitude increases, it overshadows underlying climatological characteristics. For instance, rainfall spatiotemporal variability is found to determine heavier streamflow tails (Wang et al., 2022). Second, the rare event dynamics involve more hydrologic processes and thus need more variables to describe. In other words, in the occurrences of extreme runoff events, nonlinear hydrological responses start to dominate (Basso et al., 2023). Under these circumstances, the ML model becomes less effective due to a lack of training samples.

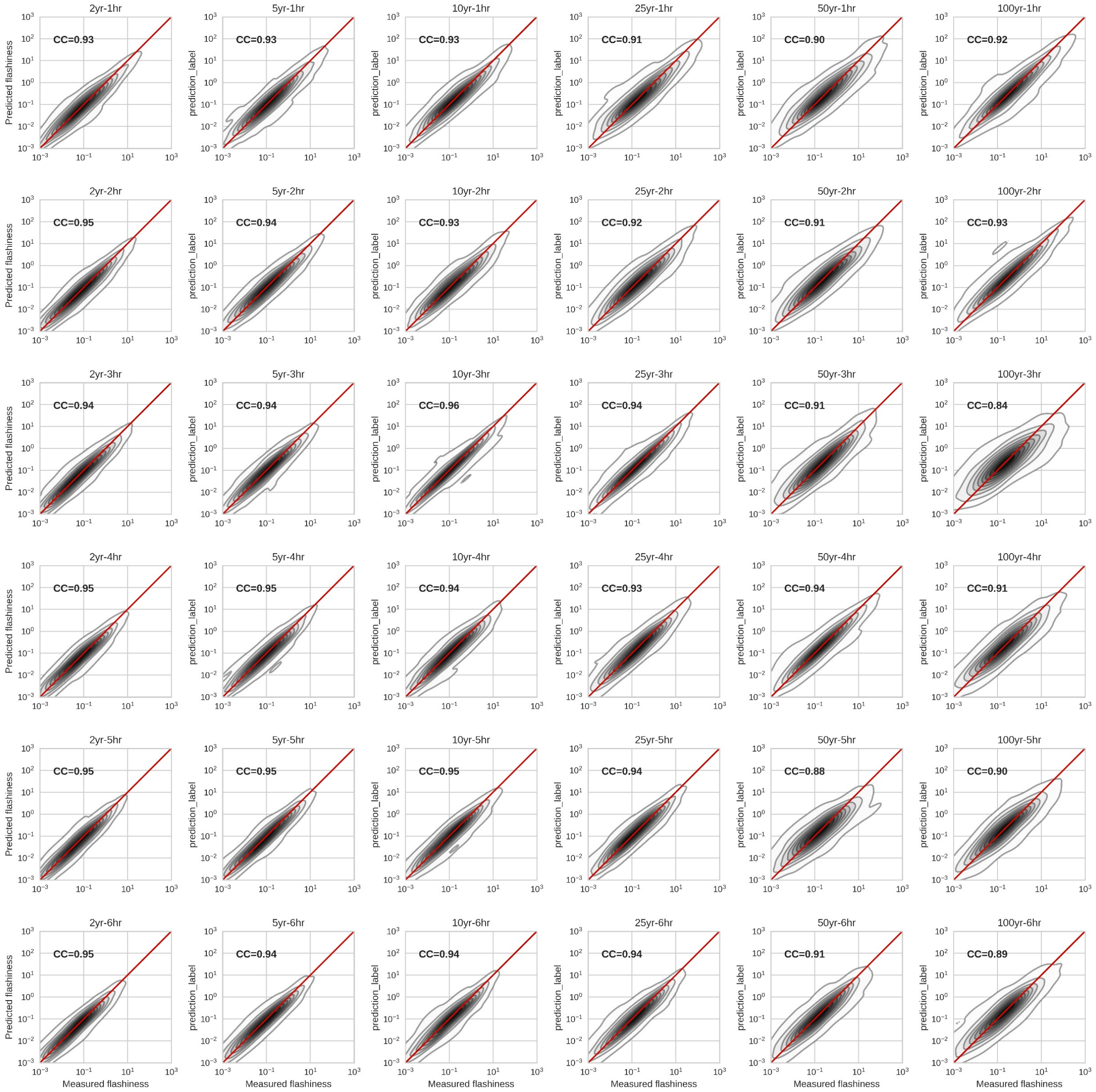

Figure 2. Density plot of the predicted flashiness values by the Machine Learning model versus target data for the testing data (processed from USGS streamgages). The red line is a 1:1 line showing the bias of the prediction – the model is overestimating (underestimating) if it is above (below) the 1:1 line.

The important factors ranked by the SHAP values are shown in Fig. 3. The drainage area is the most important factor in the ML prediction methods. We note that the flow accumulation (a proxy for drainage area) appears in the denominator of Eq. 1 and thus normalizes the streamflow values into unit streamflow. Even following the normalization, the drainage area values contribute positively to the model prediction. Put simply, including drainage areas in the ML model can improve ML prediction skills in small drainage basins. Smaller basins are more susceptible to being below the scale of the contributing storm scale and thus completely covered by the causative rainfall. Conversely, the ML model is less skillful in large drainage basins to predict flashiness values, as we can expect, larger basins have spatially heterogeneous attributes such as spatial rainfall variability and soil classes, which complicate the prediction. Air temperature is ranked as the second important factor, and higher temperature positively impacts the model prediction. The spatial distribution of the SHAP values suggests that air temperature exerts its most positive influence on model predictions only to the south of 30°N, especially for southern Texas and central Florida (Fig. S1). The channel slope factor, as expected, improves model predictions when its values are high. On the contrary, basin slope impacts less on model predictions, probably because the time scale of a hillslope routing is beyond the flash flood time scale for large basins. The comparison of spatial SHAP values is presented in Fig. S2a, where one can see higher SHAP values of channel slope across the Appalachians, Intermountain West, and Missouri Valley. In these regions, the importance of channel slope outweighs basin slope (Fig. S2b). The potential evapotranspiration factor is similar to the air temperature because higher temperature leads to higher saturated water vapor and thus requires less energy to evaporate (Thornthwaite, 1948). The spatial distribution of the annual runoff variable (Fig. S2c) corresponds better with flash flood hotspots (e.g., West Coast) than that of annual rainfall (Fig. S2d). Despite the Southeast receiving abundant annual rainfall, the SHAP values in this region are negative. This implies that rainfall, in this context, acts more as a confounder than as a contributor to predicting flashiness. Related to soil variables, soil water content and clay soil fraction are the two leading variables to improve model prediction. They have similar behavior – higher soil water content or higher clay soil fraction leads to positive model performance. That is, regions with higher soil water content and/or clay soil fractions are more susceptible to flash flooding. For human impacts, densely populated regions and higher road density enhance model predictability by taking into account the fast flow generation process (Yang et al., 2011). The

SHAP method assists us in retracing significant contributing factors for flash flood prediction and in identifying hydrologic processes through data mining. These processes should be incorporated into hydrologic model development to better simulate rapid runoff generation.

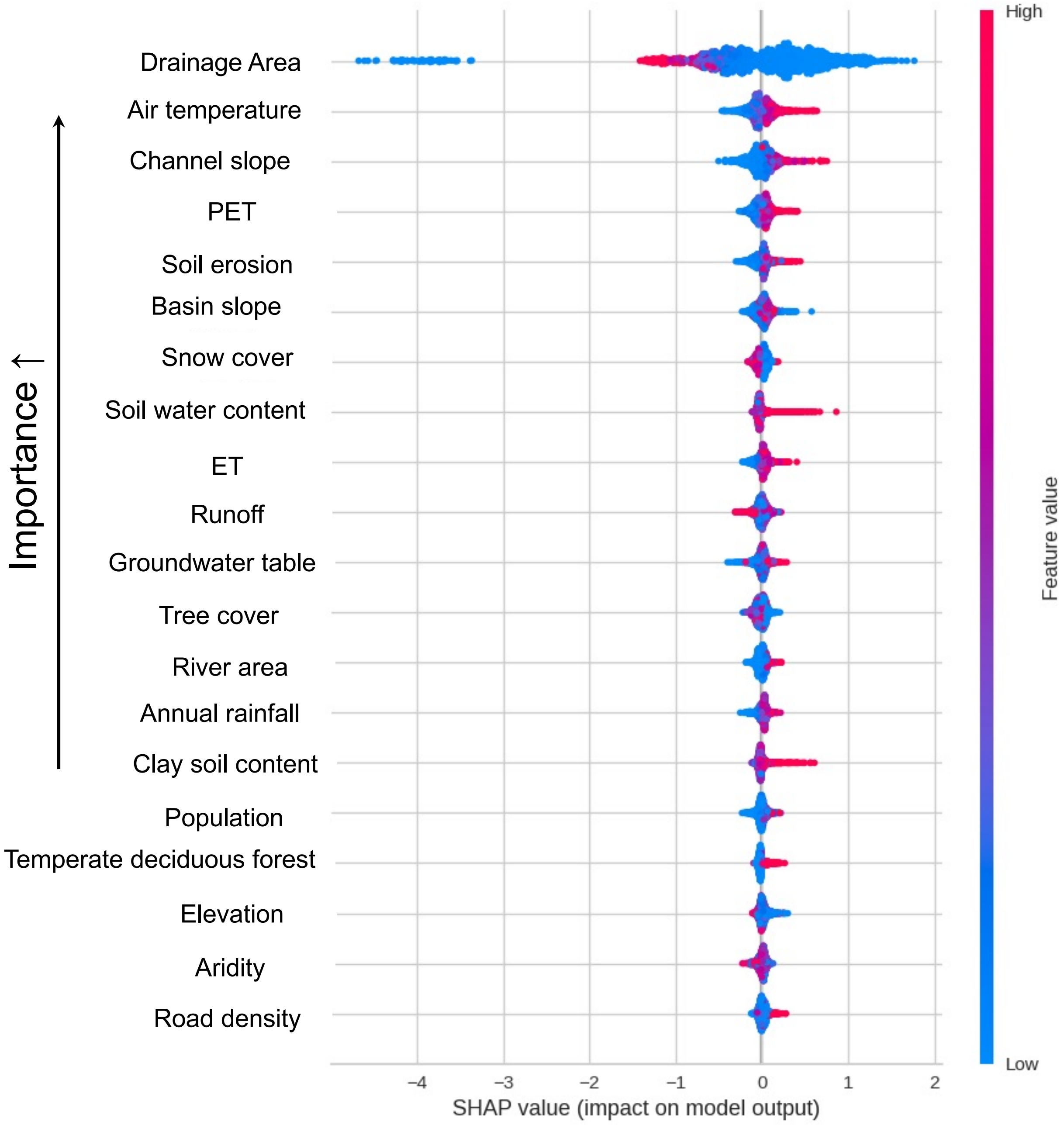

Figure 3. Important features are ranked by the SHAP values (increase from bottom to top). The color of the dots shows the feature values, and locations show the SHAP values for 2-year and 1-hour flash flood events. Positive SHAP values indicate that the inclusion of this factor can improve the model prediction. Likewise, negative values mean that this factor does not

contribute to the model performance improvement. Take the drainage area as an example, we see that low drainage area values contribute positively to the model prediction.

3.1.2 CREST-based approach

Generally, the performance of the CREST-based approach falls short of the ML-based approach, as it is not specifically designed for flashiness simulation. The highest CC value in Fig. 4 among the 36 combinations is 0.6, occurring in the 2-year and 6-hour event, as compared to 0.95 for the ML-based model. Similar to the results from the ML model, CC values increase with event duration and decrease with return periods. Conversely, the uncertainty range decreases with event duration and increases with return periods. Different from the ML model, CREST model tends to overestimate the flashiness values, as indicated by the core density region lying above the 1:1 line. The overestimation could be attributed to a positive bias of streamflow and faster flood rising with the kinematic wave parameterization (Flamig et al., 2020; Vergara et al., 2016). In short, the CREST model routes overland runoff and in-channel flood water through a simplified shallow water equation – kinematic wave model, and a-priori kinematic wave model parameters were derived based on statistical relationships with physiography, precipitation, and soil parameters (Vergara et al., 2016). However, at the higher end (with a flashiness index greater than 10), the CREST-based approach exhibits an underestimation across 36 combinations. We explore possible reasons for the bias in Section 3.1.3.

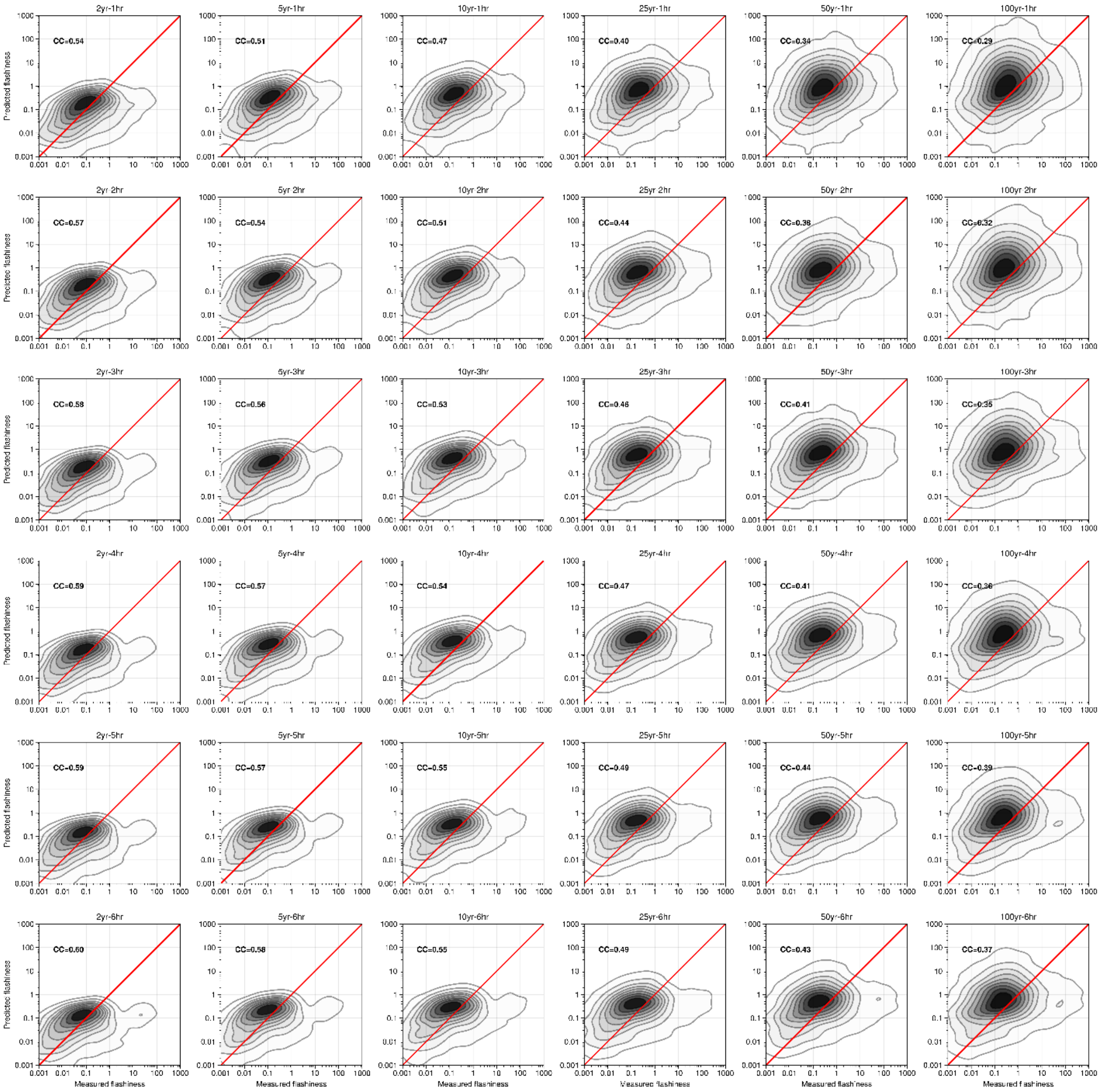

Figure 4. Similar to Fig. 2, but for the CREST-based results.

3.1.3 Comparing CREST- and ML-based approaches at all gages

The spatial distribution of the flashiness bias is shown in Fig. 5 for CREST (Fig. 5a) and ML (Fig. 5b). At a first glance, CREST-simulated flashiness values exhibit higher biases than those of ML, which is expected and has been demonstrated in Figs. 2 and 4. CREST model tends to underestimate flash flood hotspot regions, such as the Appalachians, the Southwest, and the Flash Flood Alley in Texas. It corroborates with the observation from the density plot – the CREST model exhibits an underestimation at high flashiness values. For other regions, the

CREST model demonstrates a high positive bias, probably falling within the flashiness range of 0.1 to 1 in the density plot (Fig. 4). For the ML model, it shows a sporadic spatial distribution of flashiness biases, which are the random errors.

We further dissect the bias based on four factors – annual rainfall, degrees of regulation, urban extent, and channel slope, as depicted in Fig. 6. The annual rainfall has the least impact on the CREST model bias among the four factors, largely because it has been incorporated when developing the kinematic wave parameters as a proxy (Vergara et al., 2016). The highest bias is associated with the regulation factor, as the CREST model has not yet considered any human controls in the streamflow generation process. The model biases are positive across various degrees of regulation, but they peak between 0 and 10, where the drainage area is relatively small compared to regions with higher degrees of regulation. For the urban extent, the CREST model bias transitions from positive to negative with increasing urbanization. In a highly urbanized region, which is more prone to flash floods, the CREST model tends to underpredict the flashiness values. Given the fact that CREST has incorporated urban imperviousness as a land surface parameter, the error term should originate from this parameterization or perhaps the kinematic wave parameterization. Lastly, the channel slope presents a similar pattern as the urban extent, where CREST model results have a positive bias over regions with mild slopes yet a slight negative bias over steeper terrain.

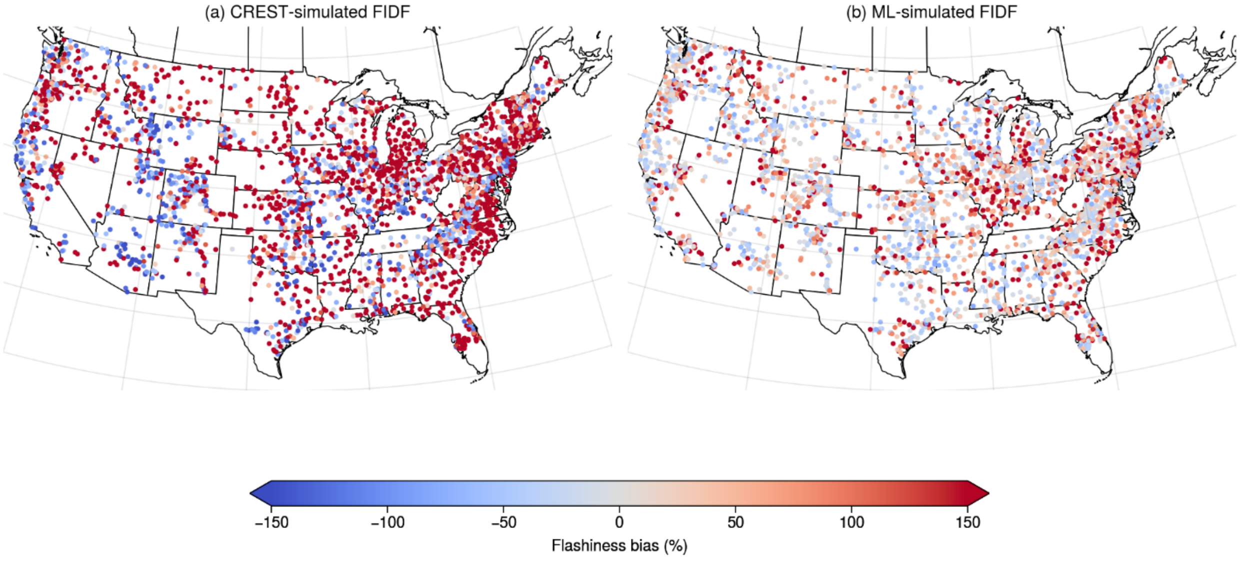

Figure 5. Maps of the flashiness bias by (a) CREST-simulated FIDF and (b) ML-simulated FIDF. It shows the 2-yr and 1-hr flashiness biases and others have a similar pattern.

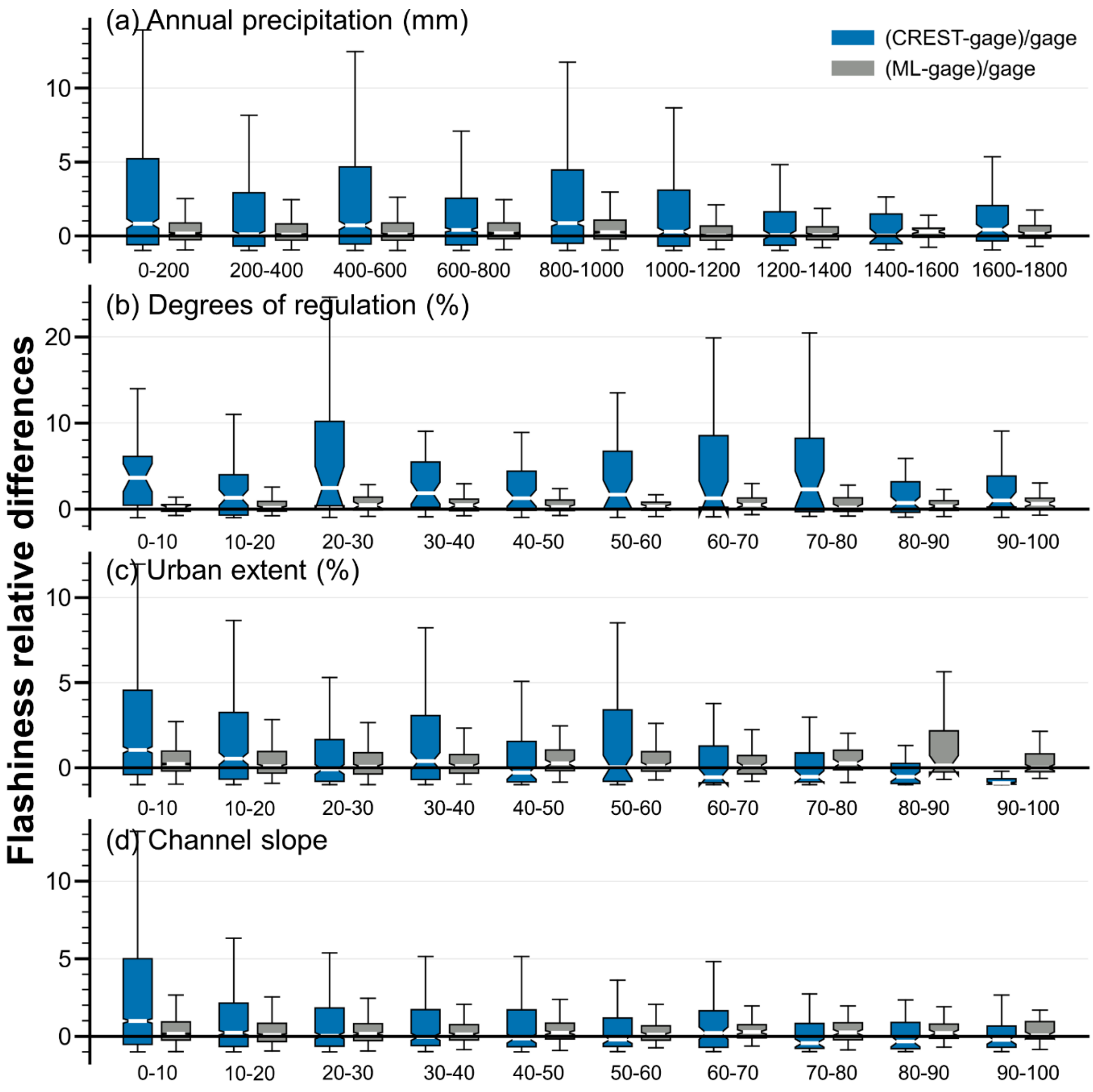

Figure 6. The plot of conditional bias of CREST-predicted and ML-predicted flashiness values based on (a) annual precipitation, (b) degrees of regulation, (c) urban extent, and (d) channel slope.

3.2 CONUS-wide distributed FIDF

After verifying our model at gaged locations, we have a certain confidence to produce a distributed product. Figures 7 and 8 show the CONUS-wide distributed F-IDF curves for the CREST and ML simulations, respectively. The CREST-simulated results have some voids over the Intermountain West. Some of these voids correspond to gaps in the NEXRAD radar coverage, which are the basis of the precipitation inputs. Notably, the CREST model generates gridded outputs, whereas the ML model generates reach-based outputs (in a vector format). A common feature of both products is that large rivers, such as the Mississippi River, appear in a dim color, indicating that flash flooding is not a disastrous concern due to the nature of their slow-rising flow. In contrast, rivers in headwater catchments, urbanized regions, and complex terrain exhibit high flashiness values. In particular, regions such as the Missouri Valley, Appalachians, Flash Flood Alley in Texas, and the Southwest are identified as flash flood hotspots. However, the results simulated by the CREST model appear more fragmented than those simulated by the ML model. This is because each grid cell extracts its own streamflow time series and fits into the GEV, making it independent from others. On the contrary, the ML model uses a single model to interpolate/extrapolate the flashiness values in space, which serves to smooth out any speckles.

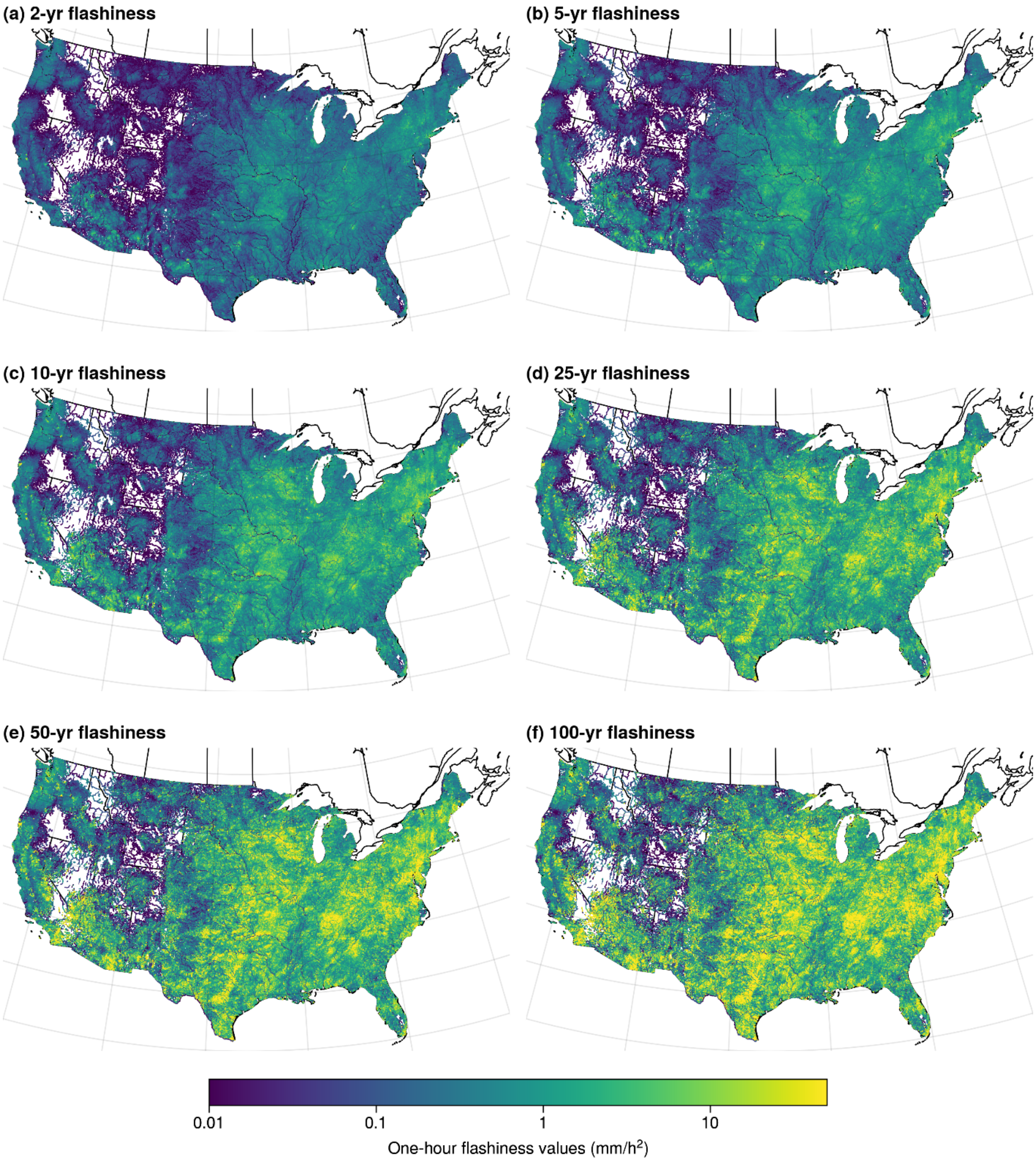

Figure 7. A grid-based F-IDF map over the CONUS by the CREST model.

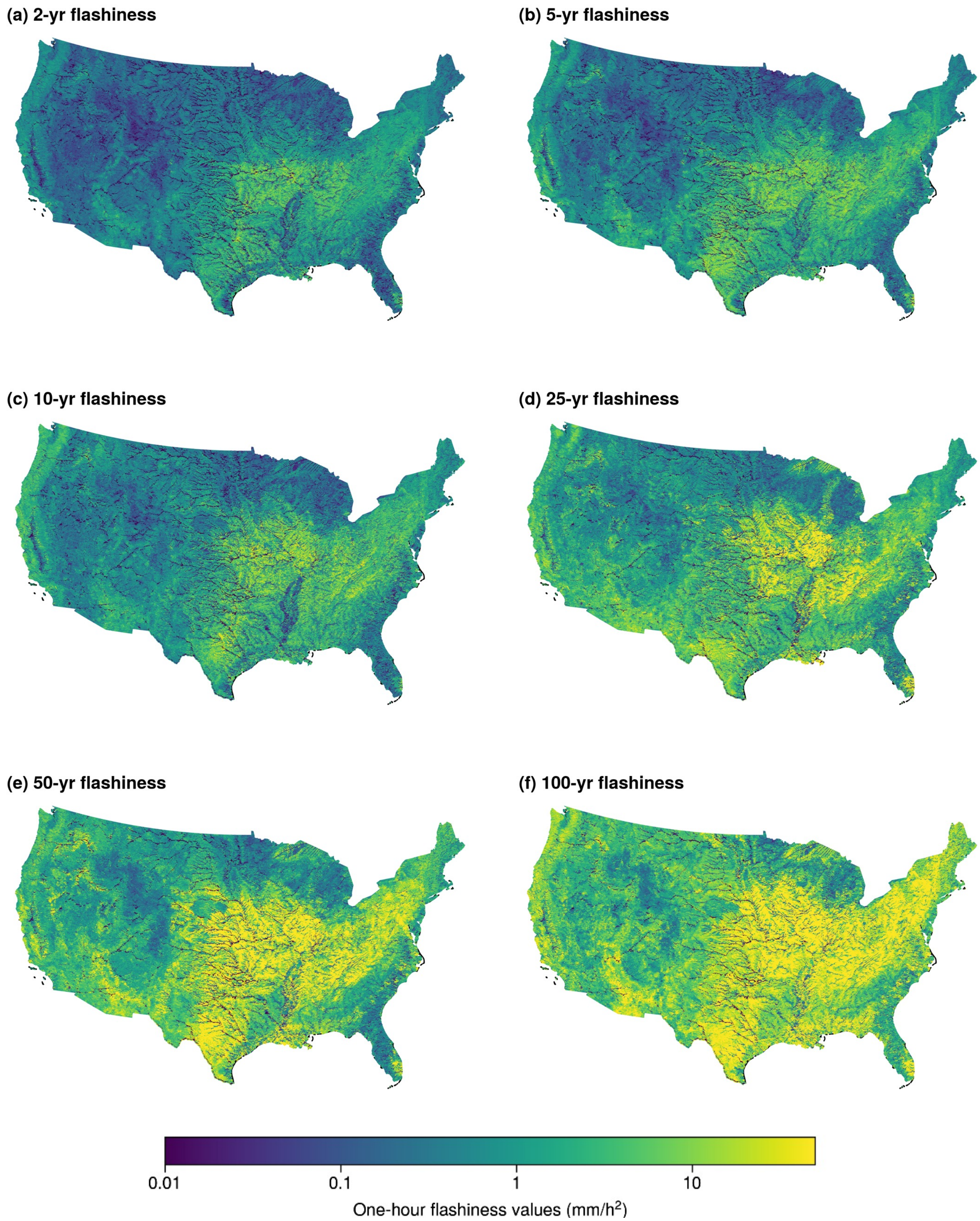

Figure 8. Similar to Fig. 7, but for the ML-based prediction.

3.3 Event-based analysis

To illustrate the utility of the distributed F-IDF products, we showcase their performance for a real flash flood event – the 2006 Louisville flash flooding event. On September 22 and 23, 2006, a slow-moving storm system passed through, resulting in up to 10 inches (254 mm) of rain in the Louisville region within a 24-hour period. The northwestern region suffered the most and six people lost their lives during this event (https://louisvillemsd.org/programs/programs-and-projects/floodplain-management/flooding-history-louisville#:~:text=September%202006,-A%20slow%2Dmoving&text=Up%20to%2010%20inches%20of,since%20the%20March%201997%20flood). Because the city of Louisville is surrounded by mountains, it is susceptible to flash flooding and has long been known as a flash flood hot spot in the Missouri Valley.

We extracted the time series of streamflow simulation over this region, calculated the event flashiness values, and then compared them to the CREST-simulated F-IDF curves to plot the gridded return periods (Fig.9). The results of return periods are also compared with those by streamgages with the same approach except using its own F-IDF values. The CREST and streamgage values have agreement on the flash flood core region, as highlighted by the ellipse. For a 1 (2/3/4/5/6) hour event, 4 (5/4/5/5/5) out of 7 gages in the highlighted region classifies this as a 100-year flash flood event. Since it is a slow-moving event, event frequency becomes rarer with higher event duration. However, the CREST simulation tends to overestimate the magnitude of this event, especially on a dichotomous metric – streamgages that did not recognize this as a flash flood event (with return periods < 2 years) were incorrectly predicted by CREST as an event (return periods >= 2 years. There is a generally good agreement between the CREST model and streamgage values when considering high-end events (return periods >= 50 years). This demonstrates the utility of the CREST-simulated F-IDF product, which can quantify the frequency of an impending flash flood event coupled with a weather forecast model or radar-based precipitation inputs. It not only enables us to define the extent of a flash flood warning but also to gauge the severity of the event for effective emergency communication.

Unlike the dynamic hydrologic model, the ML-based prediction does not directly generate streamflow time series, so event-based analyses, such as determining event return periods, are not feasible. Figure 10 provides a close-up view of the flashiness values in this region instead. One can observe that streamgages identified as flash flood events (return periods >= 2 years) are located in smaller drainage basins, and their flashiness values range between 1 and 10. While the ML-based F-IDF product cannot function on a forecast basis due to its limitations, it still possesses significant value in risk management. For instance, certain influential factors determining flashiness values, such as regulation or land use, can be engineered. Therefore, this tool could be effectively integrated into flash flood risk management strategies.

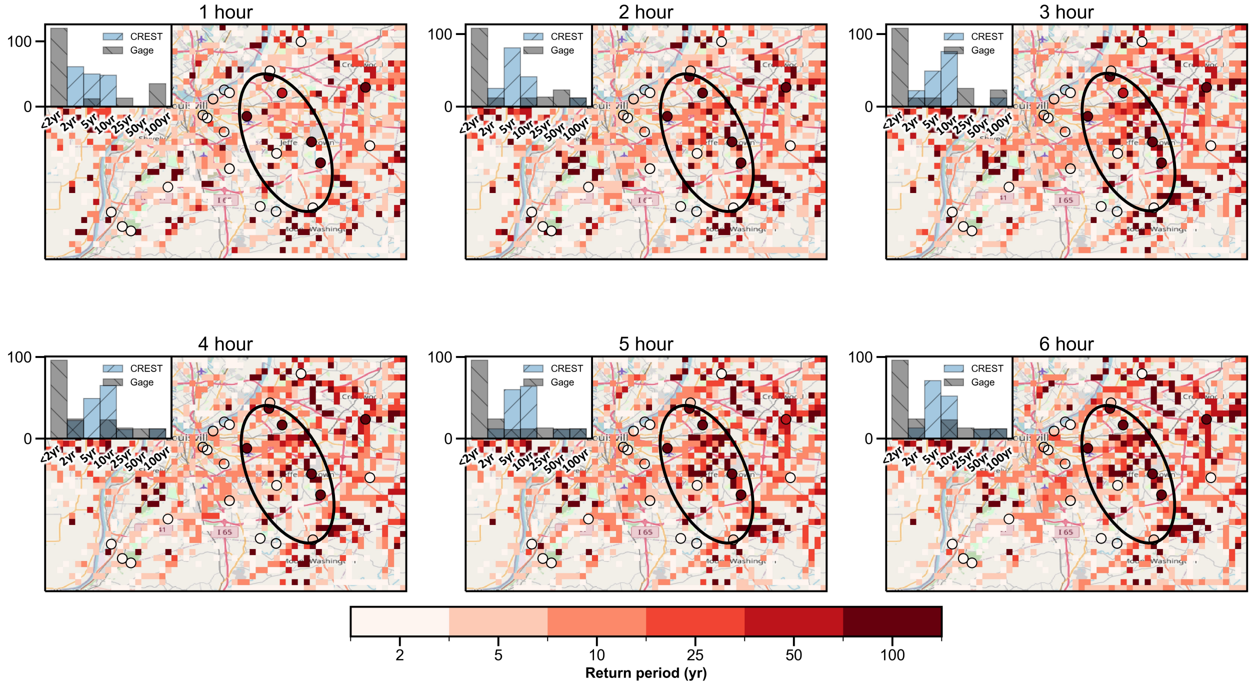

Figure 9. Maps of the return periods of flashiness values by the CREST simulation for the event, overlaid with gage-based return periods of flashiness values. The inset on the top left of each panel is the histogram of estimated return periods by CREST model and stream gages. The ellipse highlights the region with high return periods.

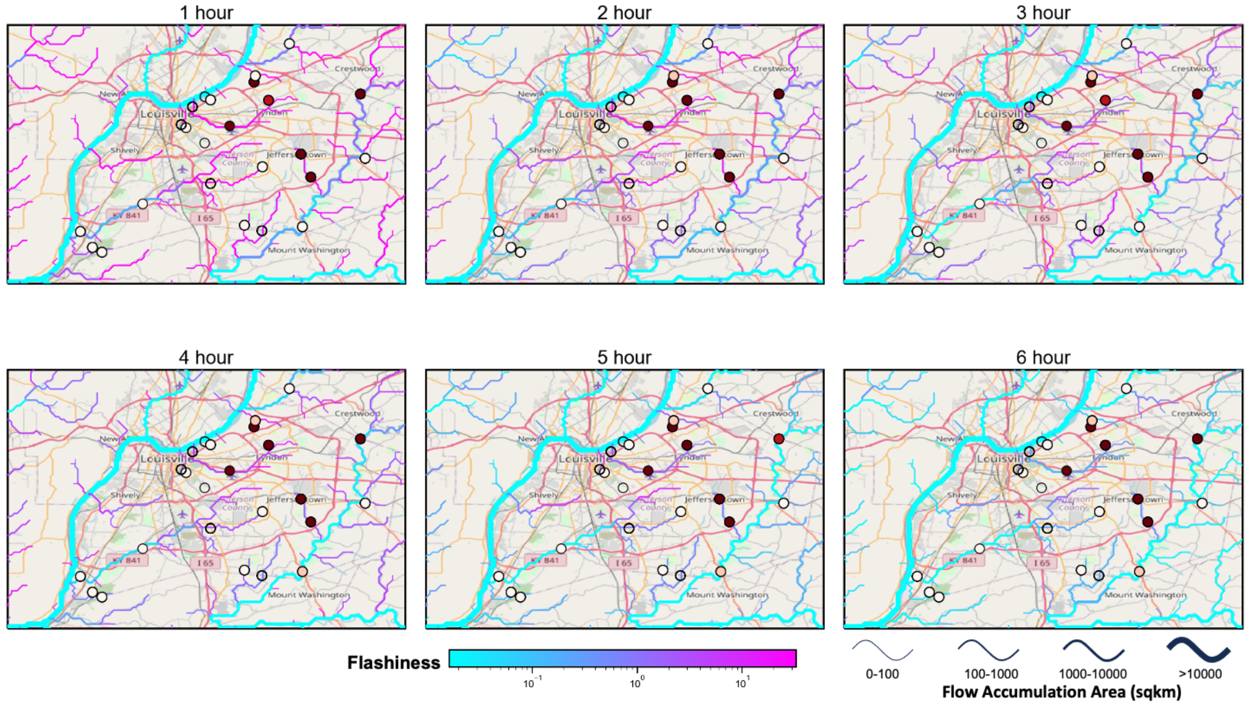

Figure 10. Maps of the flashiness values by the ML model for the event, overlaid with return periods estimated by the streamflow at gages.

4. Discussion

4.1 Uncertainties in Models

The accuracy and effectiveness of the F-IDF curves rely heavily on two models, which inevitably bear uncertainties with respect to inputs, model physics, aggregating methods, etc. We break down the uncertainties into two main categories: epistemic and aleatoric uncertainty (Beven, 2016). The epistemic uncertainty arises from a lack of knowledge about the forcing data, model structure, and model parameters. The nature of epistemic uncertainty is reducible, meaning that with the advancement in our knowledge and techniques, we can narrow down the epistemic uncertainty. However, the aleatoric uncertainty is a main result of random noise but may be structured (bias, autocorrelation, and long-term persistence). The CREST model simulation embraces major epistemic uncertainties from precipitation inputs, evapotranspiration, model parameters, and model structure. Among them, precipitation data is one of the primary uncertainty sources for flash flood prediction. In this study, we use the MRMS reanalysis data consisting of weather radar and in-situ instruments because it is so far the only available precipitation product at sub-hourly and 1 km resolution over the CONUS. One of the noticeable limitations of this product is its coverage in complex terrain such as the Rockies which is the radar “blind” zone (Zhang et al., 2016). Even within radar coverage, its quality degrades because of beam broadening issues over radar sparse regions (Zhang et al., 2012). The MRMS data can be fused with satellite precipitation data, such as the GPM IMERG to fill the gap and produce reliable F-IDF values over the Rockies. The second source of uncertainty stems from the model parameters and physics (Clark et al., 2016). Despite calibration, the performance of the CREST model is not uniformly high across different regions. For instance, the model tends to have large errors in snow-dominant regions due to its simplified conceptualization of the snow process (Flamig et al., 2020). Fortunately, flash floods are typically less influenced by snowmelt and more so by heavy rainfall. Pertaining to calculating the flashiness index, the routing parameters are arguably crucial as they have a high sensitivity to both the timing and magnitude of the flood simulation. These parameters control how water is routed through the hydrological system, effectively determining how quickly a flood rises and how high the flood peak becomes. Thus, they have a significant impact on the flashiness index and ultimately, the assessment of flash flood risk. Careful calibration of these parameters can lead to more accurate and reliable flash flood forecasts.

On the other hand, the ML model mainly suffers from aleatoric uncertainty, as its model bias tends to be random (Fig. 5b). But it still has epistemic uncertainties that are reducible, one of such being the training data length. The model is now only trained on 3,722 streamgage sites that have 15-minute time interval of streamflow observations with at least 25-years length. Increasing sample sizes can enhance its representation of tree-based models and mitigate the overfitting issue. Particularly, a lack of training samples in rare events (e.g., 100-year flash flood event) degrades model performance, as shown in Fig. 2. In parallel to increase sample sizes, including more features relevant to flash flood prediction could be beneficial. Another way of reducing epistemic uncertainty is to use Bayesian methods to encode our prior knowledge about the distribution of the model parameters and provide probabilistic outputs (Nuti et al., 2021). Also notably, the SHAP method, used in this study to unearth the interpretability of the ML model, does not elucidate any causality or correlation between each feature and flashiness. Rather, it provides insights into how a feature influences the model’s predictability.

4.2 Synergetic use of two products to mitigate flash flood impacts

The CREST-based and ML-based F-IDF products have different characteristics and can serve different purposes. In terms of prediction accuracy, the ML-based F-IDF demonstrates a closer resemblance to the observed F-IDF values derived from streamgages, whereas the performance of the CREST-based simulation is somewhat inferior. However, the ML method cannot be utilized to derive event-based statistics, a task for which the CREST simulation is well-suited.

Given its dynamic feature, the CREST simulation can be of use for operational flash flood forecasts. Currently, weather forecasters from the National Weather Service issue flash flood warnings guided by the unit streamflow variable from the CREST model amongst other information (Gourley & Vergara, 2021). This F-IDF product offers a more tangible and comprehensive approach to conceptualize the severity of flash floods. By framing the intensity of a flash flood in terms of a "100-year event," for example, we aim to facilitate more effective public communication. This approach allows the public to correlate their accumulated experience with 100-year floods, enabling a better understanding of the severity of flash flood events. Importantly, this framework is model agnostic. This means it can seamlessly integrate with any hydrologic model, such as the National Water Model, provided that the model is capable of generating timely streamflow predictions.

The ML-based FIDF, on the contrary, cannot be used on an event basis because it produces static flashiness values. Yet, it can be of use to risk managers in the city with its high prediction accuracy. In regions characterized by high risks or equivalently elevated flashiness values, the implementation of protective measures is imperative to mitigate potential impacts. For instance, signage such as "potential flash flood areas" and "when flooded, turn around, don't drown" are crucial to improve driver's safety. Some flood defense measures can also be implemented to reduce the flashiness values, such as changing land use. Using the ML model, urban planners have the capacity to adjust different feature values, enabling them to identify feasible and effective strategies to decrease flashiness values. This approach offers a quantitative assessment of how flashiness changes with certain feature values, thereby supporting the decision-making process.

By integrating both these products into operational risk communication and long-term planning strategies, we anticipate a reduction in the impacts of flash floods, achieved through a blend of soft and hard measures for flood management. For model development, the important variables identified by the ML model can be incorporated into the hydrologic model, ensuring that the hydrologic processes are not overlooked. Certainly, the applications of F-IDF products are not only limited to the examples provided above.

5. Conclusion

This study presents a pioneering creation of the distributed F-IDF products over the CONUS with a physics-based hydrologic model approach and the statistics-based machine learning (ML) approach. The two products exhibit similar performance in identifying regions prone to flash floods, but their differences result in distinct applications. For the ML model, we explored its interpretability by incorporating the SHAP values for each feature to rank their importance. The conclusions are summarized as follows:

1. Both CREST and ML predict flashiness values reasonably well, with average CC values of 0.58 and 0.95, respectively, for a 2-year flash flood event;
2. The drainage area, air temperature, channel slope, potential evapotranspiration, and soil erosion features are identified as the five most important factors influencing the ML model's prediction. These factors can yield valuable insights that could inform the development of hydrologic models for better flash flood forecasting;
3. The CREST simulation exhibits high biases in regions that are characterized by dam/reservoir regulation, urbanization, or mild slopes, suggesting areas for future improvement;
4. The distributed F-IDF products, both by CREST and ML provide similar risk maps for flash flood-prone regions. However, the spatial patterns of ML-produced maps are smoother, compared to those generated by CREST. This is attributable to two primary factors. On one hand, grid cells in the CREST simulation are independent, while the ML model interpolates or extrapolates between features. On the other hand, CREST simulation benefits from radar-based rainfall inputs, a feature not available to the ML model;

5. Different yet synergistic applications for the two products are emphasized. The CREST-based simulation can provide event-based forecasts, making it suitable for operational flash flood forecasts employed by weather forecasters and emergency responders. Conversely, the ML-based simulation, which is a static feature, can be integrated into a flash flood risk assessment framework, offering a valuable tool for urban planners;

In future research, we hope to expand the study area to the globe by developing a global F-IDF product. This would enhance our ability to communicate risks associated with flash floods effectively on a worldwide scale.

Data Availability

The MRMS reanalysis data is acquired from Zhang & Gourley (2018). The RiverAtlas product is acquired from https://www.hydrosheds.org/hydroatlas. The F-IDF products generated by CREST and ML can be accessed from Li (2023).

Reference

Allen, R.G., L. Pereira, D. Raes, and M. Smith, 1998. Crop Evapotranspiration, Food and Agriculture Organization of the United Nations, Rome, Italy. FAO publication 56. ISBN 92-5-104219-5. 290p.

Chen, M., Li, Z., & Gao, S. (2022). Multisensor Remote Sensing and the Multidimensional Modeling of Extreme Flood Events: A Case Study of Hurricane Harvey–Triggered Floods in Houston, Texas, USA. *Remote Sensing of Water-Related Hazards*, 87-104.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

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

Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P., Clark, R. A., III, Argyle, E., Arthur, A., Martinaitis, S., Terti, G., Erlingis, J. M., Hong, Y., & Howard, K. W. (2017). The FLASH Project: Improving the Tools for Flash Flood Monitoring and Prediction across the United States, *Bulletin of the American Meteorological Society*, *98*(2), 361-372. doi: https://doi.org/10.1175/BAMS-D-15-00247.1

Kemter, M., Marwan, N., Villarini, G., & Merz, B. (2023). Controls on flood trends across the United States. Water Resources Research, 59, e2021WR031673. https://doi.org/10.1029/2021WR031673.

Stein, L., Clark, M. P., M. Knoben, W. J., Pianosi, F., & Woods, R. A. (2021). How Do Climate and Catchment Attributes Influence Flood Generating Processes? A Large-Sample Study for 671 Catchments Across the Contiguous USA. *Water Resources Research*, *57*(4), e2020WR028300. https://doi.org/10.1029/2020WR028300

Wang, H., Merz, R., Yang, S., Tarasova, L., & Basso, S. (2023). Emergence of heavy tails in streamflow distributions: The role of spatial rainfall variability. *Advances in Water Resources*, *171*, 104359. https://doi.org/10.1016/j.advwatres.2022.104359

Thornthwaite, C. W. (1948). An Approach toward a Rational Classification of Climate. *Geographical Review*, *38*(1), 55–94. https://doi.org/10.2307/210739

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

Flamig, Z. L., Vergara, H., and Gourley, J. J.: 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, 2020.

Zhang, J., Y. Qi, C. Langston and B. Kaney, 2012: Radar Quality Index (RQI) – a combined measure for beam blockage and VPR effects in a national network. *Weather Radar and Hydrology*, 351, 388-393.

Clark, M.P., Wilby, R.L., Gutmann, E.D. Vano, J.A., Gangopadhyav, S., Wood, A. W., Fowler, H.J., Prudhomme, C., Arnold, J.R., Brekke, L.D., 2016. Characterizing Uncertainty of the Hydrologic Impacts of Climate Change. *Curr Clim Change Rep* **2**, 55–64. https://doi.org/10.1007/s40641-016-0034-x

Nuti, G., Jiménez Rugama, L. A., & Cross, A. (2021). An Explainable Bayesian Decision Tree Algorithm. *Frontiers in Applied Mathematics and Statistics*, *7*, 598833. https://doi.org/10.3389/fams.2021.598833

Gourley, J. J., & Vergara, H. (2021). Comments on “Flash Flood Verification: Pondering Precipitation Proxies”, *Journal of Hydrometeorology*, *22*(3), 739-747. doi: https://doi.org/10.1175/JHM-D-20-0215.1

Beven, K. 2016. Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrological Sciences Journal, 61:9, 1652-1665, DOI: 10.1080/02626667.2015.1031761

Hong, Y., Adhikari, P., Gourley, J.J. (2013). Flash Flood. In: Bobrowsky, P.T. (eds) Encyclopedia of Natural Hazards. Encyclopedia of Earth Sciences Series. Springer, Dordrecht. doi:10.1007/978-1-4020-4399-4_136

Li, Zhi. (2023). Distributed F-IDF products [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8169330

Doswell , C. A., III, Brooks, H. E., & Maddox, R. A. (1996). Flash Flood Forecasting: An Ingredients-Based Methodology, *Weather and Forecasting*, *11*(4), 560-581. doi: https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2

Maddox, R. A., Chappell, C. F., & Hoxit, L. R. (1979). Synoptic and Meso-α Scale Aspects of Flash Flood Events, *Bulletin of the American Meteorological Society*, *60*(2), 115-123. doi: https://doi.org/10.1175/1520-0477-60.2.115

Clark, R. A., Gourley, J. J., Flamig, Z. L., Hong, Y., & Clark, E. (2014). CONUS-Wide Evaluation of National Weather Service Flash Flood Guidance Products, *Weather and Forecasting*, *29*(2), 377-392. doi: https://doi.org/10.1175/WAF-D-12-00124.1

Li, Z., Gao, S., Chen, M., Zhang, J., Gourley, J.J., Wen, Y., Yang, T., Hong, Y. (2023). Introducing Flashiness-Intensity-Duration-Frequency (F-IDF): A New Metric to Quantify Flash Flood Intensity. Preprint on Authorea. June 23, 2023. doi: 10.22541/essoar.168748464.41784321/v1

Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P., Clark, R. A., III, Argyle, E., Arthur, A., Martinaitis, S., Terti, G., Erlingis, J. M., Hong, Y., & Howard, K. W. (2017). The FLASH Project: Improving the Tools for Flash Flood Monitoring and Prediction across the United States, *Bulletin of the American Meteorological Society*, *98*(2), 361-372. doi: https://doi.org/10.1175/BAMS-D-15-00247.1

Morss, R. E., Mulder, K. J., Lazo, J. K., & Demuth, J. L. (2016). How do people perceive, understand, and anticipate responding to flash flood risks and warnings? Results from a public survey in Boulder, Colorado, USA. *Journal of Hydrology*, *541*, 649-664. https://doi.org/10.1016/j.jhydrol.2015.11.047

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, 2008.

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

Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., Gourley, J. J., & Hong, Y. (2021). Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study of four watersheds with different hydro-climatic regions across the CONUS. *Journal of Hydrology*, *598*, 126423. https://doi.org/10.1016/j.jhydrol.2021.126423

Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, *54*(11), 8558-8593. https://doi.org/10.1029/2018WR022643

Lehner, B., Messager, M.L., Korver, M.C., Linke, S. (2022). Global hydro-environmental lake characteristics at high spatial resolution. Scientific Data 9: 351. doi: **https://doi.org/10.1038/s41597-022-01425-z**

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., Grams, H., Wang, Y., Cocks, S., Martinaitis, S., Arthur, A., Cooper, K., Brogden, J., & Kitzmiller, D. (2016). Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621-638. https://doi.org/10.1175/BAMS-D-14-00174.1

Ali, M. (2020). PyCaret: An open source, low-code machine learning library in Python, https://www.pycaret.org.

Yang, G., Bowling, L. C., Cherkauer, K. A., & Pijanowski, B. C. (2011). The impact of urban development on hydrologic regime from catchment to basin scales. *Landscape and Urban Planning*, *103*(2), 237-247. https://doi.org/10.1016/j.landurbplan.2011.08.003