

1 **Distributed Flashiness-Intensity-Duration-Frequency products over the**
2 **conterminous US**

3 Zhi Li^{1,*}, Shang Gao², Mengye Chen¹, Jiaqi Zhang¹, Jonathan J. Gourley³, Humberto Vergara⁴,
4 Siyu Zhu¹, Sebastian Ferraro¹, Yixin Wen⁵, Tiantian Yang¹, Yang Hong^{1,*}

5 ¹ School of Civil Engineering and Environmental Science, University of Oklahoma, Norman,
6 OK, USA

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

8 ³ NOAA/National Severe Storms Laboratory, Norman, OK, USA

9 ⁴ College of Engineering, University of Iowa, Iowa City, IA, USA

10 ⁵ Department of Geography, University of Florida, Gainesville, FL, USA

11

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

13

14 **Key Points:**

- 15 • We developed distributed flashiness-intensity-duration-frequency products with machine
16 learning and hydrologic simulation
- 17 • Both products can identify flash flood-prone regions in the CONUS
- 18 • We cross-compared both products over the CONUS and highlight their strengths and
19 limitations
- 20 • The utility of the two products is discussed with their synergistic use by decision makers
21

22 **Abstract**

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

43 **1. Introduction**

44 **1.1 Background**

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

52 Forecasting flash flood qualitatively is difficult, forecasting and quantifying the specific
53 magnitudes of flash flooding at a specific location is much more challenging. Due to these
54 challenges, operational forecasting flash floods on a national scale was not feasible until the
55 1980s (Georgakakos, 1986). Two types of threshold-based guidance products have emerged and
56 are currently being utilized by forecasters at the National Weather Service (NWS). The Flash
57 Flood Guidance (FFG), implemented after a deadly 1969 flash flood in Ohio, has become a
58 national standard for weather forecasters henceforth (Clark et al., 2014). Taking quantitative
59 precipitation estimates (QPE) as inputs, FFG determines if the amount of rain will produce bank-
60 full conditions on streams. However, FFG does not account for the land cover and routing in
61 simulating pluvial flash flooding. Hydrologic models, on the other hand, simulate the rainfall-
62 runoff processes to predict the occurrence of flash floods with unit streamflow values (Gourley et
63 al., 2017). With increasing available computational resources, flash flood forecast products
64 derived from hydrologic models are beginning to play a more prominent role in predictive storm
65 warning and disaster management. Gourley & Vergara (2021) found the equitable threat score
66 generally increases with the sophistication of flash flood forecast products, particularly
67 highlighting the importance of land cover and surface routing process.

68 **1.2 Problem statement**

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

82 flow. These two factors hinder effective communication between decision-makers and the
83 public, consequently placing vulnerable communities at increased risk.

84 **1.3 New promises**

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

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

112 **2. Data and Methods**

113 **2.1 Data for hydrologic simulation**

114 We use the CREST hydrologic model to simulate sub-hourly streamflow from 2001 to
115 2012. The model inputs include precipitation and potential evapotranspiration as forcings and a
116 set of a-priori parameters at a desired resolution (i.e., 1 km). We use the Multi-Radar Multi-
117 Sensor reanalysis product at 10-min time intervals over the CONUS to provide precipitation data
118 (Zhang & Gourley, 2018) and the USGS monthly climatological potential evapotranspiration for
119 the model (Allen et al., 1998). The MRMS is a radar-gauge merged quantitative precipitation
120 estimation (QPE) product by merging 180 operational radars and creating a 3D radar mosaic
121 over the CONUS (Zhang et al., 2016). A set of calibrated a-priori model parameters are accessed
122 from <https://github.com/chrimerss/EF5-US-Parameters>, and the model performance with such
123 data is evaluated by Vergara et al. (2016) and Flamig et al. (2020).

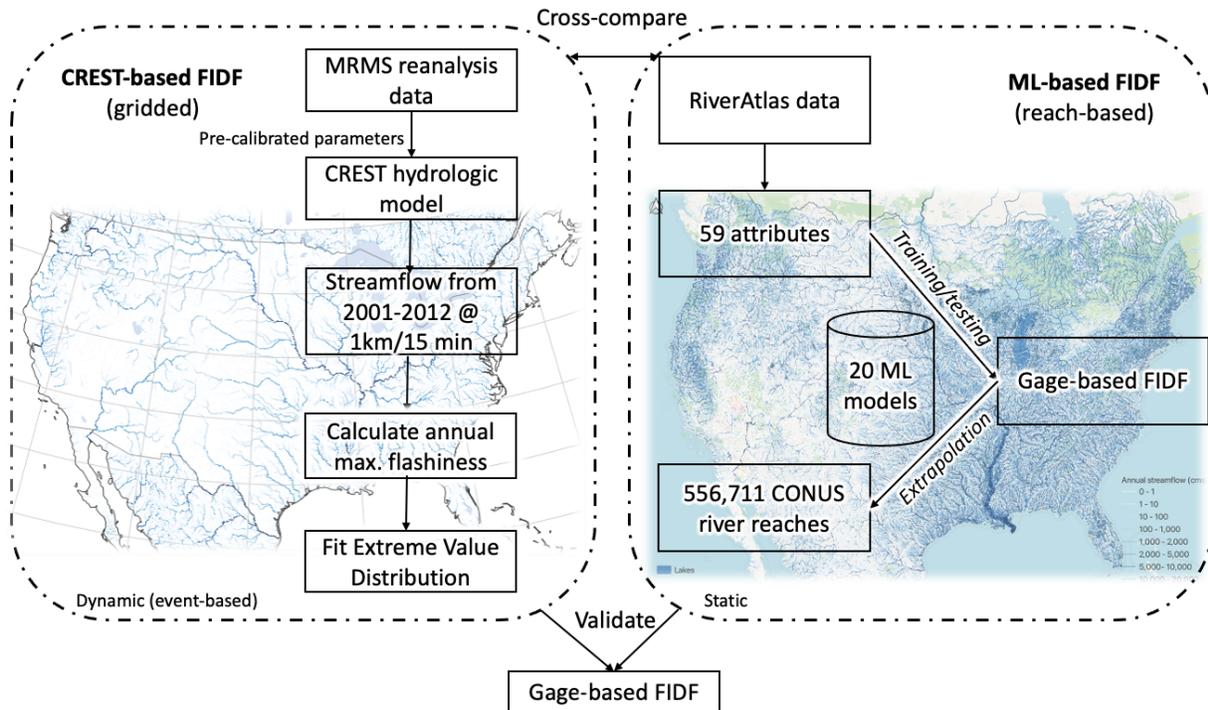
124 **2.2 RiverAtlas data**

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

136 **2.3 Framework**

137 Figure 1 depicts the overall framework used in this study to produce distributed F-IDF
138 values over the CONUS. This framework intends to produce two distributed F-IDF products
139 covering the CONUS. One is CREST-based F-IDF that is generated by the CREST hydrologic
140 model and fits an Extreme Value Distribution (EVD). A counterpart is machine-learning (ML)-
141 based F-IDF that is an extrapolation of gage-based F-IDF values over the CONUS, which was

142 conducted by Li et al. (2023). Another distinct feature of these two approaches is their spatial
 143 representativeness. The CREST-based F-IDF product is gridded, with the cell size the same as
 144 the distributed hydrologic model (i.e., 1 km). The ML-based F-IDF product is river reach-based
 145 since the hydrologic attributes are aggregated in hydrologic response units (i.e., sub-basins) and
 146 assigned to corresponding river reaches. The methods for calculating CREST-based and ML-
 147 based F-IDF are articulated in Sections 2.5 and 2.6, respectively.



148
 149 **Figure 1.** A schematic framework of the two approaches.

150 **2.4 Definition of Flashiness-Intensity-Duration-Frequency**

151 We have introduced the definition of F-IDF in Li et al. (2023) and reiterate the core
 152 concept here. The rationale for proposing a new metric is three-fold. First, this new metric
 153 quantifies the severity of a flash flood event with return periods (e.g., a 100-year flash flood
 154 event). Second, flash flood events are multi-dimensional, meaning that the duration of the event
 155 impacts the severity of the event. Third, the F-IDF is a tailored metric that can assist decision-
 156 makers in planning for and mitigating flash flood risks. The calculation of the F-IDF is as
 157 follows. First, we compute the flashiness index (Eq. 1), which is the slope of a hydrograph over a
 158 moving window that represents the duration of an event. Then, the annual maximum flashiness

159 index is extracted by aggregating the time series. Lastly, we fit the annual maximum values into
160 GEV and extract flashiness values for desired flash flood return periods. The flashiness values, in
161 principle, reflect the speed at which the flood rises and the magnitude of the flood peak.
162 Although the definitions for the flashiness index are variable, we see similarities in different
163 methods from identified flash flood hot spots (Li et al., 2023). In addition, our method is fairly
164 simple and reproducible compared to others (Gannon et al., 2022; Saharia et al., 2017; Smith &
165 Smith, 2015).

$$166 \quad F = \frac{\max \{Q_t - Q_{t-1}, Q_t - Q_{t-2}, \dots, Q_t - Q_{t-d}\}}{FAC \times d}, \quad (1)$$

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

174 **2.5 The CREST-based approach**

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

187 We simulate the 11-year streamflow using CREST from 2001 to 2011, with the first year
188 reserved for warming up the model states. The MRMS precipitation reanalysis data at a 10-min
189 interval and 1-km spatial resolution are used to drive the model. The model setup, such as grid
190 resolution (1km) and a-priori parameters, are the same as the operational one, and its
191 performance has been assessed by Flamig et al. (2020). The output streamflow is produced every
192 10 minutes to capture the nature of flash floods. With the streamflow values at each 1km grid
193 cell, we extract the ten-year time series (10 years x 365 days/year x 24 hours/day x 6 10-
194 minute/hour=525,600 time steps) and follow the F-IDF calculation as detailed in Section 2.4. We
195 repeat this process for 4 million grid cells that have flow accumulation values greater than 1 km²
196 over the CONUS to generate a distributed F-IDF product.

197 **2.6 Machine learning based approach**

198 Given the nature of how river attributes are aggregated, we perform the ML model at a
199 river reach level over the CONUS using the riverATLAS dataset. Fifty-nine river attributes are
200 fed into a suite of ML models for training on 3,722 USGS streamgage sites and then applied for
201 556,771 river reaches. To build the gage-based F-IDF product for ML, we extract the 15-minute
202 streamflow time series from 1950 to 2020. These time series were fed into the F-IDF calculation
203 as described in Section 2.4 (Li et al., 2023). With no prior information on ML model
204 performance, we selected 20 commonly used ML models including linear, tree-based, kernel-
205 based, and instance-based models. They are Light Gradient Boosting Machine, Random Forest
206 Regressor, Gradient Boosting Regressor, Extra Trees Regressor, Extreme Gradient Boosting, K
207 Neighbors Regressor, Ridge Regression, Linear Regression, Elastic Net, Lasso Least Angle
208 Regression, Lasso Regression, Decision Tree Regressor, Bayesian Ridge, Least Angle
209 Regression, Huber Regressor, Orthogonal Matching Pursuit, Dummy Regressor, and Passive
210 Aggressive Regressor. A table of detailed descriptions for each model is listed in Supplementary
211 Table 2. We use the pycaret package in Python to benchmark and automate workflows (Ali,
212 2020).

213 To split the training-testing samples, we adhere to the 70-30 principle, in which 70% of
214 the samples are used for training, and the rest is for testing. Beyond that, we perform a 10-fold
215 cross-validation to select the best-performing ML model out of 20 models for each return period
216 and duration. Given six return periods (i.e., 2-yr, 5-yr, 10-yr, 25-yr, 50-yr, and 100-yr) and six
217 durations (i.e., 1-hr, 2-hr, 3-hr, 4-hr, 5-hr, and 6-hr) of flashiness values, 36 ML models are

218 retained for further evaluation. Because the distribution of flashiness values is positively skewed,
 219 meaning that a large number of samples are concentrated on the low end, we transform the
 220 flashiness data to resemble a Gaussian-like distribution using the Box-Cox transformation
 221 (Eq.2).

$$222 \quad F' = \begin{cases} \log(F), & \text{if } \lambda = 0 \\ (F^\lambda - 1)/\lambda, & \text{otherwise} \end{cases} \quad (2)$$

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

226 **2.7 Explainable Machine Learning**

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

$$234 \quad \phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (3)$$

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

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

246 marginal contribution of that feature for that combination. Third, we calculate the average of its
247 marginal contributions across all combinations.

248 **3. Results**

249 **3.1 Model verification**

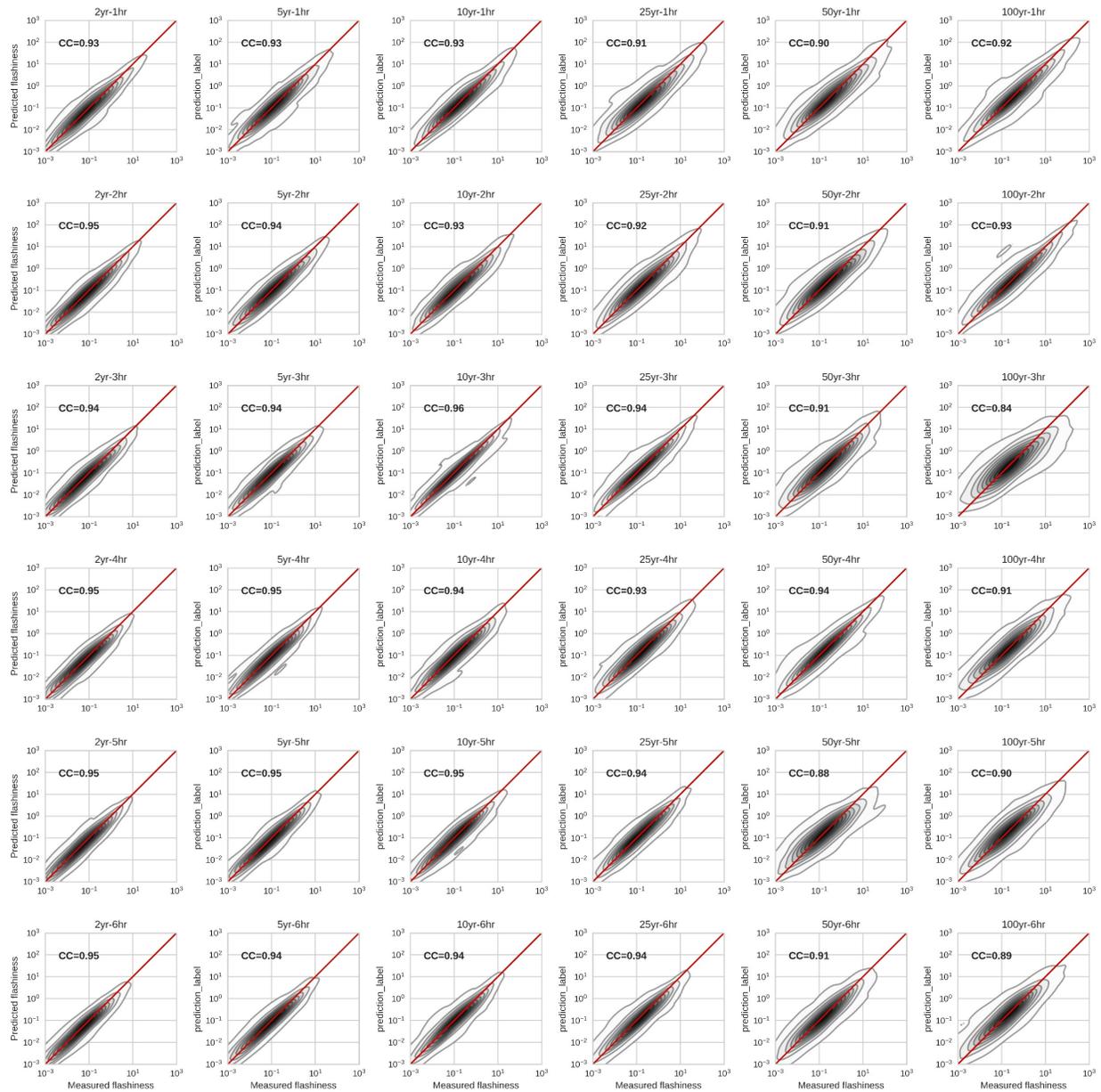
250 We evaluate the model performance with respect to calculated flashiness values for the
251 testing samples used by the ML approach. The Spearman correlation coefficient (CC) is used to
252 depict the goodness-of-fit of predicted flashiness values and target values.

253 **3.1.1 ML-based approach**

254 The ML-based approach depicts an overall good fit (mean $CC > 0.9$) between predicted
255 flashiness and target flashiness values (processed from USGS streamgages), indicating that the
256 59 hydrologic attributes adequately explain the variability of flashiness values over the CONUS
257 (Fig. 2). Among 36 enumerations (6 frequencies x 6 durations), the Light Gradient Boosting
258 Machine model tops in 33 combinations, except for the 25yr-3hr, 100yr-2hr, and 100-6hr, which
259 are best predicted by Gradient Boosting Machine, Random Forest, and Gradient Boosting
260 Machine, respectively. In general, tree-based machine learning models perform better than linear
261 models, instance-based models (i.e., k Neighbors Regressor), and kernel-based models (i.e.,
262 Support Vector Machine); and ensemble-based models perform better than deterministic models.
263 The tree-based models resemble human decision-making processes and have been widely
264 applied in flood attribution and for identifying flood-generating mechanisms (Kemter et al.,
265 2023; Stein et al., 2021).

266 Figure 2 also indicates that ML model performance deteriorates with increasing return
267 periods (column-wise comparison), but improves with longer durations (row-wise comparison).
268 When referring to performance improvement (or deterioration), we mean not only the increase
269 (or decrease) in CC but also the decrease (or increase) in the uncertainty spread, as indicated by
270 the contour area. This is expected for two reasons. First, for rare events (e.g., 1-in-100-year),
271 static hydrologic signatures become less impactful while it depends more on the event
272 characteristics such as event rainfall, antecedent soil moisture, channel routing, etc. In other
273 words, as the rainfall event magnitude increases, it overshadows underlying climatological
274 characteristics. For instance, rainfall spatiotemporal variability is found to determine heavier
275 streamflow tails (Wang et al., 2022). Second, the rare event dynamics involve more hydrologic

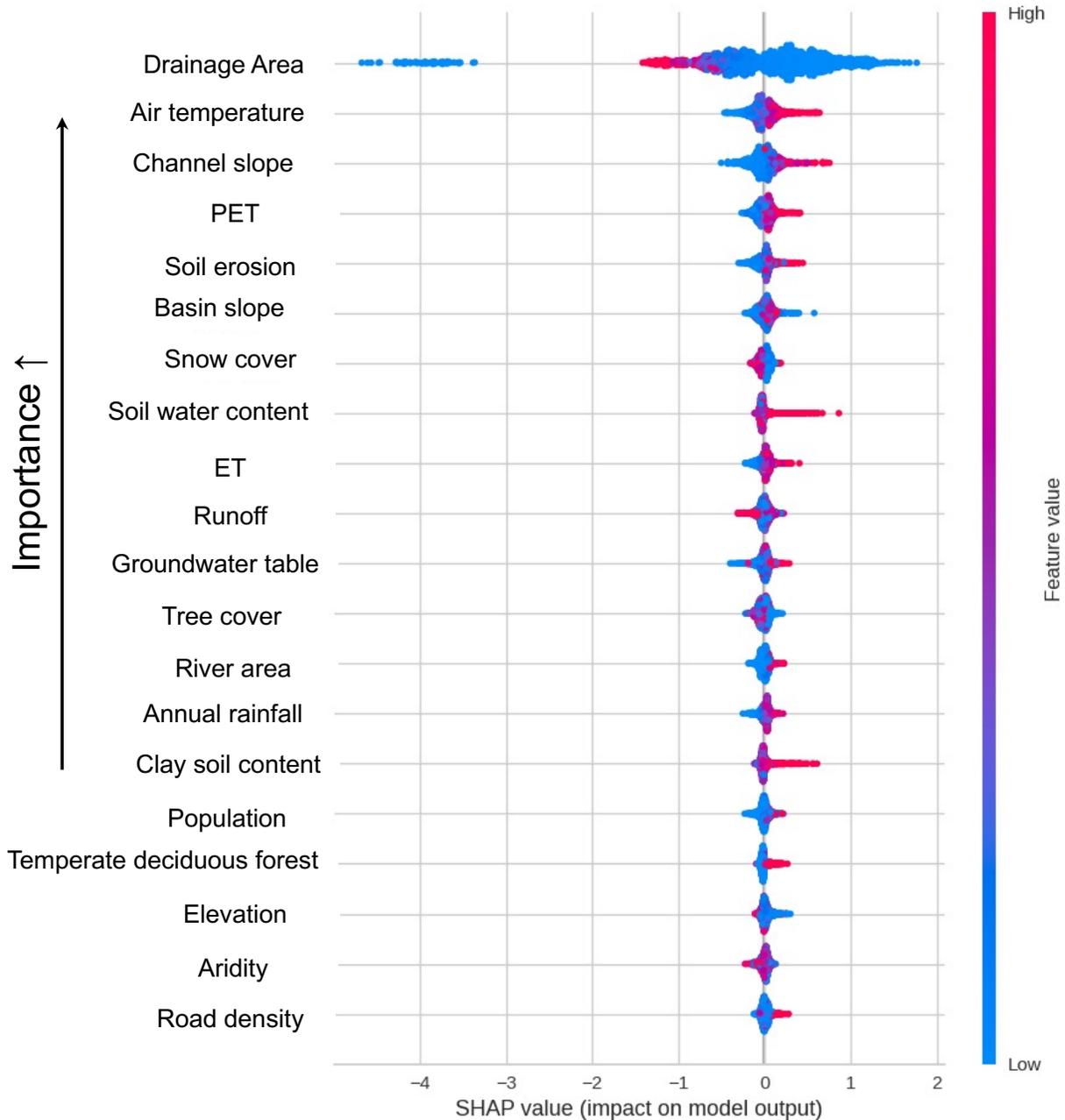
276 processes and thus need more variables to describe. In other words, in the occurrences of
 277 extreme runoff events, nonlinear hydrological responses start to dominate (Basso et al., 2023).
 278 Under these circumstances, the ML model becomes less effective due to a lack of training
 279 samples.



280
 281 **Figure 2.** Density plot of the predicted flashiness values by the Machine Learning model versus
 282 target data for the testing data (processed from USGS streamgages). The red line is a 1:1 line
 283 showing the bias of the prediction – the model is overestimating (underestimating) if it is above
 284 (below) the 1:1 line.

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

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

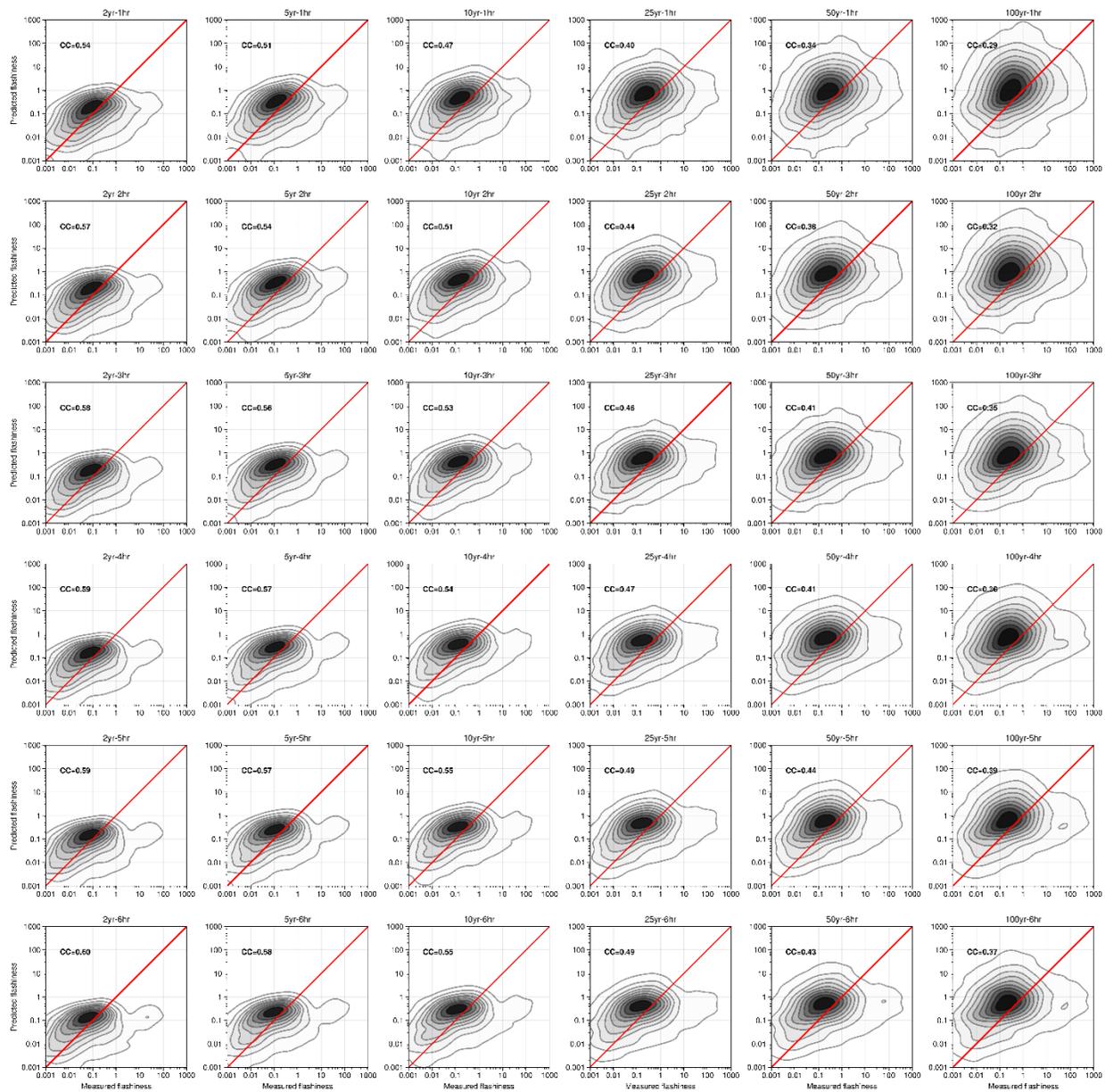


319
320 **Figure 3.** Important features are ranked by the SHAP values (increase from bottom to top). The
321 color of the dots shows the feature values, and locations show the SHAP values for 2-year and 1-
322 hour flash flood events. Positive SHAP values indicate that the inclusion of this factor can
323 improve the model prediction. Likewise, negative values mean that this factor does not

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

326 **3.1.2 CREST-based approach**

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



342

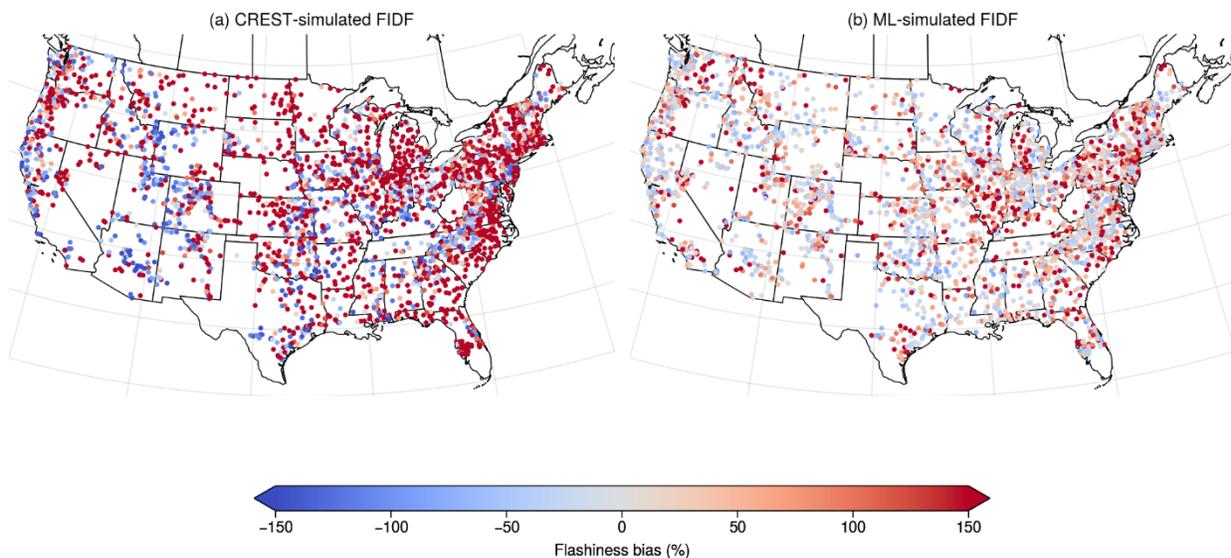
343 **Figure 4.** Similar to Fig. 2, but for the CREST-based results.

344 **3.1.3 Comparing CREST- and ML-based approaches at all gages**

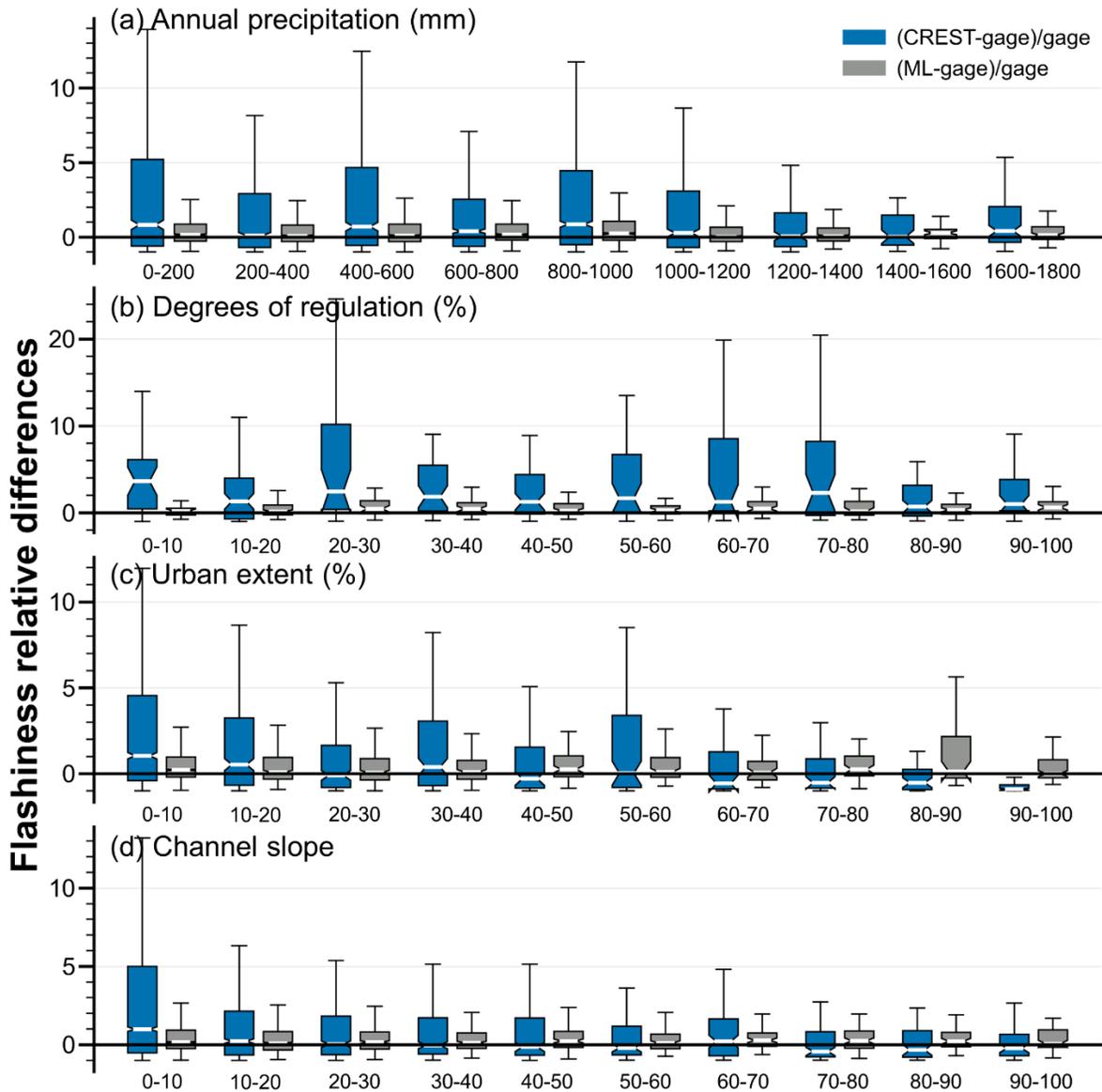
345 The spatial distribution of the flashiness bias is shown in Fig. 5 for CREST (Fig. 5a) and
 346 ML (Fig. 5b). At a first glance, CREST-simulated flashiness values exhibit higher biases than
 347 those of ML, which is expected and has been demonstrated in Figs. 2 and 4. CREST model tends
 348 to underestimate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

351 CREST model demonstrates a high positive bias, probably falling within the flashiness range of
352 0.1 to 1 in the density plot (Fig. 4). For the ML model, it shows a sporadic spatial distribution of
353 flashiness biases, which are the random errors.

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



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

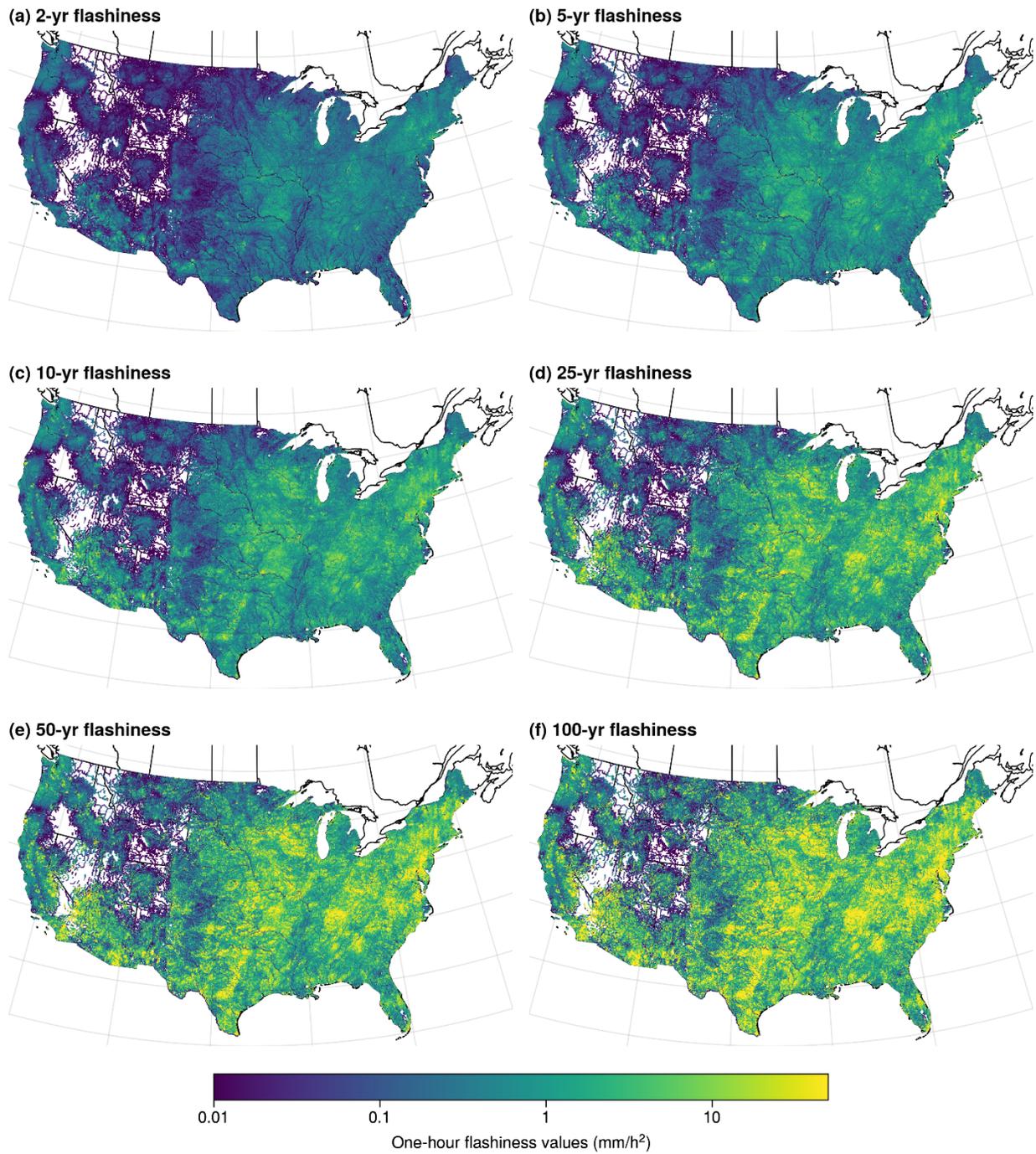


372
 373 **Figure 6.** The plot of conditional bias of CREST-predicted and ML-predicted flashiness values
 374 based on (a) annual precipitation, (b) degrees of regulation, (c) urban extent, and (d) channel
 375 slope.

376 **3.2 CONUS-wide distributed FIDF**

377 After verifying our model at gaged locations, we have a certain confidence to produce a
 378 distributed product. Figures 7 and 8 show the CONUS-wide distributed F-IDF curves for the
 379 CREST and ML simulations, respectively. The CREST-simulated results have some voids over

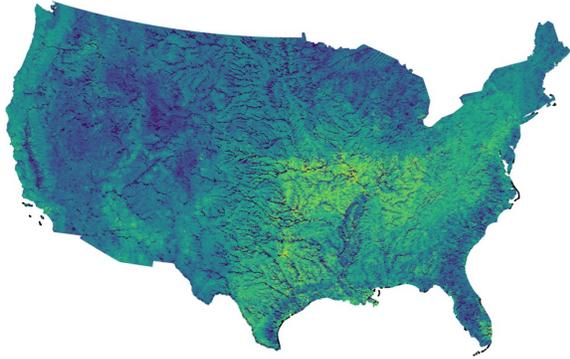
380 the Intermountain West. Some of these voids correspond to gaps in the NEXRAD radar
381 coverage, which are the basis of the precipitation inputs. Notably, the CREST model generates
382 gridded outputs, whereas the ML model generates reach-based outputs (in a vector format). A
383 common feature of both products is that large rivers, such as the Mississippi River, appear in a
384 dim color, indicating that flash flooding is not a disastrous concern due to the nature of their
385 slow-rising flow. In contrast, rivers in headwater catchments, urbanized regions, and complex
386 terrain exhibit high flashiness values. In particular, regions such as the Missouri Valley,
387 Appalachians, Flash Flood Alley in Texas, and the Southwest are identified as flash flood
388 hotspots. However, the results simulated by the CREST model appear more fragmented than
389 those simulated by the ML model. This is because each grid cell extracts its own streamflow
390 time series and fits into the GEV, making it independent from others. On the contrary, the ML
391 model uses a single model to interpolate/extrapolate the flashiness values in space, which serves
392 to smooth out any speckles.



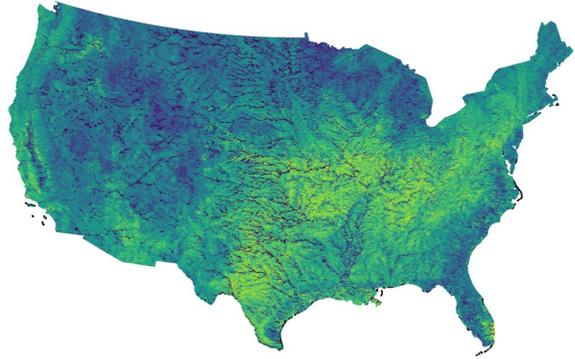
393

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

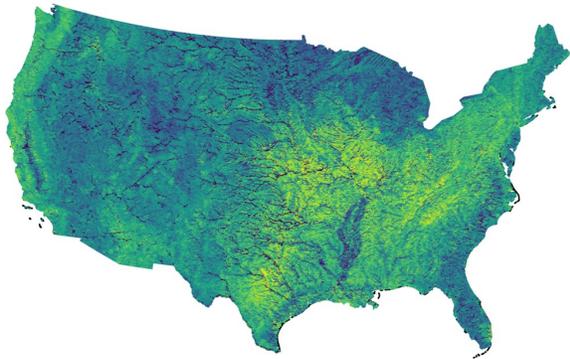
(a) 2-yr flashiness



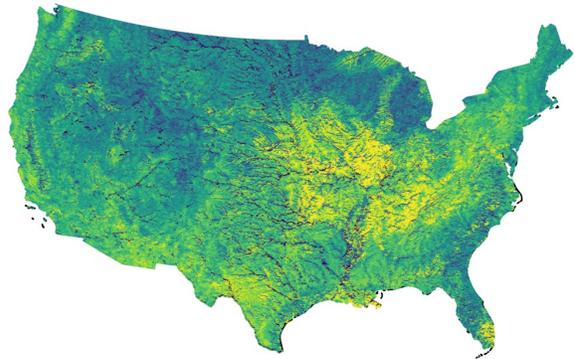
(b) 5-yr flashiness



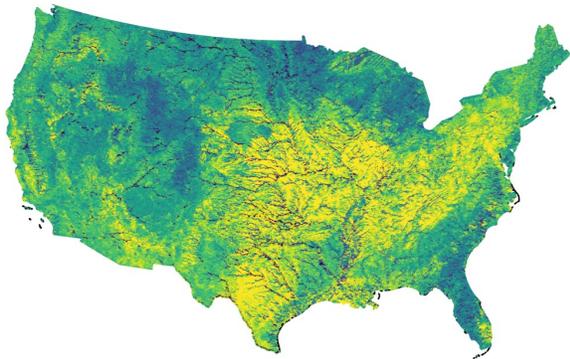
(c) 10-yr flashiness



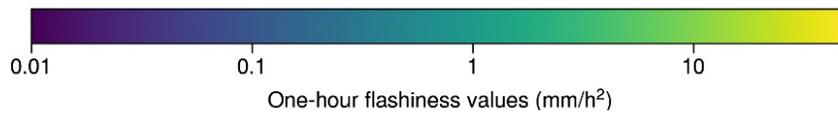
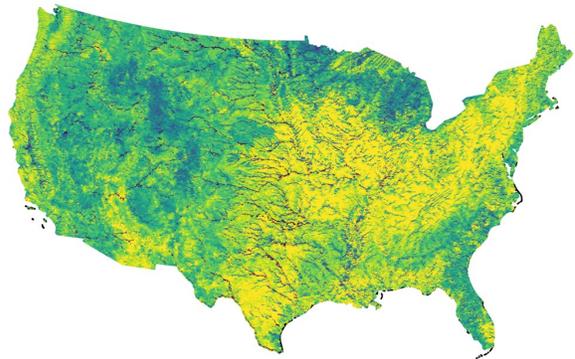
(d) 25-yr flashiness



(e) 50-yr flashiness



(f) 100-yr flashiness



395

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

397

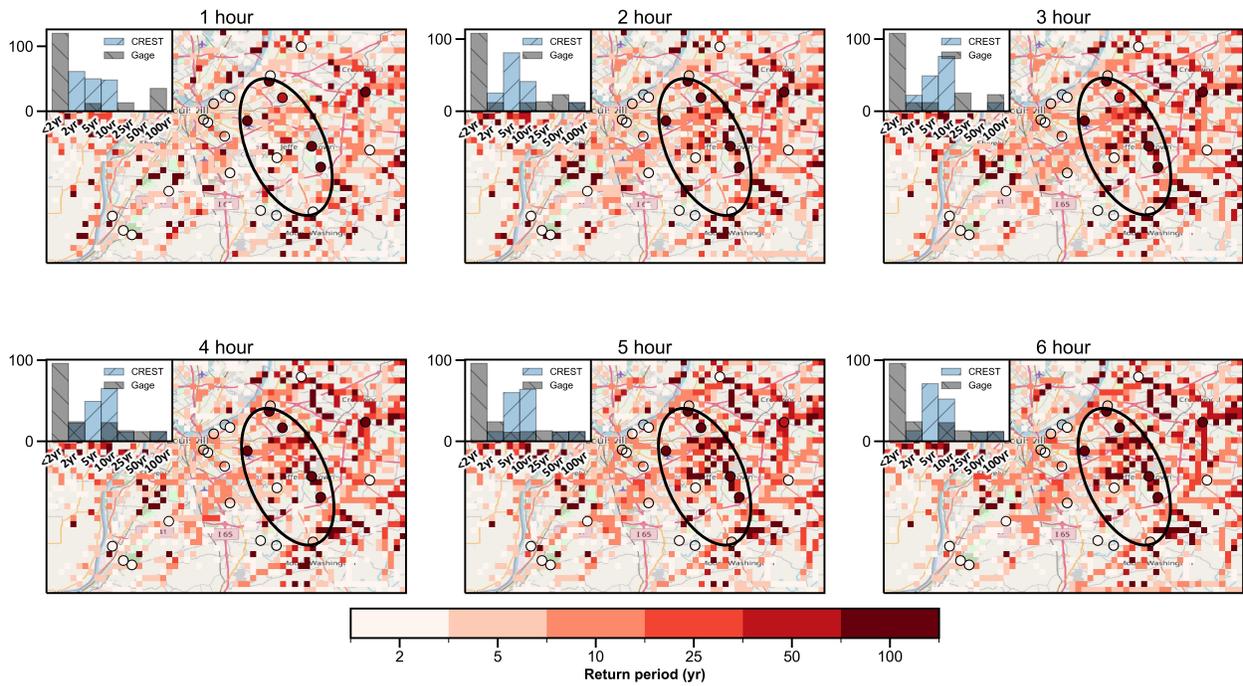
398 3.3 Event-based analysis

399 To illustrate the utility of the distributed F-IDF products, we showcase their performance
400 for a real flash flood event – the 2006 Louisville flash flooding event. On September 22 and 23,
401 2006, a slow-moving storm system passed through, resulting in up to 10 inches (254 mm) of rain
402 in the Louisville region within a 24-hour period. The northwestern region suffered the most and
403 six people lost their lives during this event (<https://louisvillemisd.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
404 flooding and has long been known as a flash flood hot spot in the Missouri Valley.

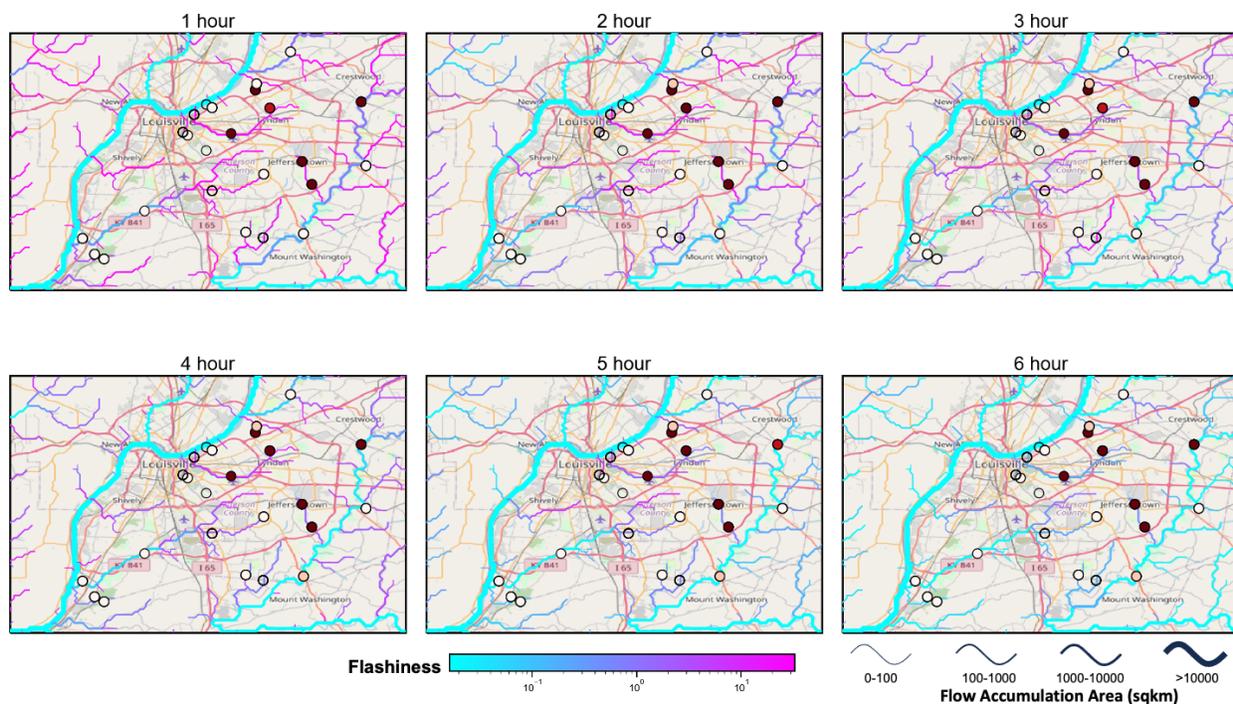
408 We extracted the time series of streamflow simulation over this region, calculated the
409 event flashiness values, and then compared them to the CREST-simulated F-IDF curves to plot
410 the gridded return periods (Fig.9). The results of return periods are also compared with those by
411 streamgages with the same approach except using its own F-IDF values. The CREST and
412 streamgage values have agreement on the flash flood core region, as highlighted by the ellipse.
413 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
414 as a 100-year flash flood event. Since it is a slow-moving event, event frequency becomes rarer
415 with higher event duration. However, the CREST simulation tends to overestimate the magnitude
416 of this event, especially on a dichotomous metric – streamgages that did not recognize this as a
417 flash flood event (with return periods < 2 years) were incorrectly predicted by CREST as an
418 event (return periods \geq 2 years. There is a generally good agreement between the CREST
419 model and streamgage values when considering high-end events (return periods \geq 50 years).
420 This demonstrates the utility of the CREST-simulated F-IDF product, which can quantify the
421 frequency of an impending flash flood event coupled with a weather forecast model or radar-
422 based precipitation inputs. It not only enables us to define the extent of a flash flood warning but
423 also to gauge the severity of the event for effective emergency communication.

424 Unlike the dynamic hydrologic model, the ML-based prediction does not directly
425 generate streamflow time series, so event-based analyses, such as determining event return
426 periods, are not feasible. Figure 10 provides a close-up view of the flashiness values in this
427 region instead. One can observe that streamgages identified as flash flood events (return
428 periods \geq 2 years) are located in smaller drainage basins, and their flashiness values range

429 between 1 and 10. While the ML-based F-IDF product cannot function on a forecast basis due to
430 its limitations, it still possesses significant value in risk management. For instance, certain
431 influential factors determining flashiness values, such as regulation or land use, can be
432 engineered. Therefore, this tool could be effectively integrated into flash flood risk management
433 strategies.



434
435 **Figure 9.** Maps of the return periods of flashiness values by the CREST simulation for the event,
436 overlaid with gage-based return periods of flashiness values. The inset on the top left of each
437 panel is the histogram of estimated return periods by CREST model and stream gages. The
438 ellipse highlights the region with high return periods.



439

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

442 4. Discussion

443 4.1 Uncertainties in Models

444 The accuracy and effectiveness of the F-IDF curves rely heavily on two models, which
 445 inevitably bear uncertainties with respect to inputs, model physics, aggregating methods, etc. We
 446 break down the uncertainties into two main categories: epistemic and aleatoric uncertainty
 447 (Beven, 2016). The epistemic uncertainty arises from a lack of knowledge about the forcing data,
 448 model structure, and model parameters. The nature of epistemic uncertainty is reducible,
 449 meaning that with the advancement in our knowledge and techniques, we can narrow down the
 450 epistemic uncertainty. However, the aleatoric uncertainty is a main result of random noise but
 451 may be structured (bias, autocorrelation, and long-term persistence). The CREST model
 452 simulation embraces major epistemic uncertainties from precipitation inputs, evapotranspiration,
 453 model parameters, and model structure. Among them, precipitation data is one of the primary
 454 uncertainty sources for flash flood prediction. In this study, we use the MRMS reanalysis data
 455 consisting of weather radar and in-situ instruments because it is so far the only available
 456 precipitation product at sub-hourly and 1 km resolution over the CONUS. One of the noticeable

457 limitations of this product is its coverage in complex terrain such as the Rockies which is the
458 radar “blind” zone (Zhang et al., 2016). Even within radar coverage, its quality degrades because
459 of beam broadening issues over radar sparse regions (Zhang et al., 2012). The MRMS data can
460 be fused with satellite precipitation data, such as the GPM IMERG to fill the gap and produce
461 reliable F-IDF values over the Rockies. The second source of uncertainty stems from the model
462 parameters and physics (Clark et al., 2016). Despite calibration, the performance of the CREST
463 model is not uniformly high across different regions. For instance, the model tends to have large
464 errors in snow-dominant regions due to its simplified conceptualization of the snow process
465 (Flamig et al., 2020). Fortunately, flash floods are typically less influenced by snowmelt and
466 more so by heavy rainfall. Pertaining to calculating the flashiness index, the routing parameters
467 are arguably crucial as they have a high sensitivity to both the timing and magnitude of the flood
468 simulation. These parameters control how water is routed through the hydrological system,
469 effectively determining how quickly a flood rises and how high the flood peak becomes. Thus,
470 they have a significant impact on the flashiness index and ultimately, the assessment of flash
471 flood risk. Careful calibration of these parameters can lead to more accurate and reliable flash
472 flood forecasts.

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

486 **4.2 Synergetic use of two products to mitigate flash flood impacts**

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

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

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

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

521 **5. Conclusion**

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

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

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

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

552 **Data Availability**

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

556

557

558

559 **Reference**

- 560
561 Allen, R.G., L. Pereira, D. Raes, and M. Smith, 1998. Crop Evapotranspiration, Food and
562 Agriculture Organization of the United Nations, Rome, Italy. FAO publication 56. ISBN 92-5-
563 104219-5. 290p.
- 564 Chen, M., Li, Z., & Gao, S. (2022). Multisensor Remote Sensing and the Multidimensional
565 Modeling of Extreme Flood Events: A Case Study of Hurricane Harvey–Triggered Floods in
566 Houston, Texas, USA. *Remote Sensing of Water-Related Hazards*, 87-104.
- 567 Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions.
568 *Advances in neural information processing systems*, 30.
- 569 Zhang, J. & Gourley, J. (2018). *Multi-Radar Multi-Sensor Precipitation Reanalysis (Version 1.0)*.
570 Open Commons Consortium Environmental Data
571 Commons. <https://doi.org/10.25638/EDC.PRECIP.0001>
- 572 Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P., Clark, R. A., III, Argyle, E., Arthur, A.,
573 Martinaitis, S., Terti, G., Erlingis, J. M., Hong, Y., & Howard, K. W. (2017). The FLASH
574 Project: Improving the Tools for Flash Flood Monitoring and Prediction across the United
575 States, *Bulletin of the American Meteorological Society*, 98(2), 361-372.
576 doi: <https://doi.org/10.1175/BAMS-D-15-00247.1>
- 577 Kemter, M., Marwan, N., Villarini, G., & Merz, B. (2023). Controls on flood trends across the
578 United States. *Water Resources Research*, 59, e2021WR031673.
579 <https://doi.org/10.1029/2021WR031673>.
- 580 Stein, L., Clark, M. P., M. Knoben, W. J., Pianosi, F., & Woods, R. A. (2021). How Do Climate
581 and Catchment Attributes Influence Flood Generating Processes? A Large-Sample Study for 671
582 Catchments Across the Contiguous USA. *Water Resources Research*, 57(4), e2020WR028300.
583 <https://doi.org/10.1029/2020WR028300>
- 584 Wang, H., Merz, R., Yang, S., Tarasova, L., & Basso, S. (2023). Emergence of heavy tails in
585 streamflow distributions: The role of spatial rainfall variability. *Advances in Water Resources*,
586 171, 104359. <https://doi.org/10.1016/j.advwatres.2022.104359>
- 587 Thornthwaite, C. W. (1948). An Approach toward a Rational Classification of Climate.
588 *Geographical Review*, 38(1), 55–94. <https://doi.org/10.2307/210739>

589 Vergara, H., Kirstetter, P., Gourley, J. J., Flamig, Z. L., Hong, Y., Arthur, A., & Kolar, R.
590 (2016). Estimating a-priori kinematic wave model parameters based on regionalization for flash
591 flood forecasting in the Conterminous United States. *Journal of Hydrology*, 541, 421-433.
592 <https://doi.org/10.1016/j.jhydrol.2016.06.011>

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

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

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

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

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

609 Beven, K. 2016. Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood,
610 hypothesis testing, and communication, *Hydrological Sciences Journal*, 61:9, 1652-
611 1665, DOI: [10.1080/02626667.2015.1031761](https://doi.org/10.1080/02626667.2015.1031761)

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

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

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

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

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

626 Li, Z., Gao, S., Chen, M., Zhang, J., Gourley, J.J., Wen, Y., Yang, T., Hong, Y. (2023).
627 Introducing Flashiness-Intensity-Duration-Frequency (F-IDF): A New Metric to Quantify Flash
628 Flood Intensity. Preprint on Authorea. June 23, 2023. doi:
629 10.22541/essoar.168748464.41784321/v1

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

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

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

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

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

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

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

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

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

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

666

667

668