Response to Reviewer 1

Comment1： The manuscript is well-organized, with a clear structure that guides readers through the methodology, results, and implications. However, the study could be strengthened by addressing some scientific gaps, such as the mechanisms behind machine learning's temporal downscaling, the reliability of results in data limited regions like the southwest, and the lack of comprehensive uncertainty quantification. Additionally, minor typographical errors and inconsistent figure formatting slightly detract from the presentation. Overall, this is a high quality study with significant contributions to hydrology and climate adaptation, but it requires minor revisions to enhance clarity, rigor, and practical applicability.

Response: We sincerely thank you for your careful review of our manuscript and for providing valuable suggestions for its improvement. We have thoughtfully considered each of these suggestions and have provided point-by-point responses below. We hope that these responses have adequately addressed the reviewer's concerns, and enhanced the quality of our work.

Comment2: The manuscript contains numerous acronyms and would benefit from a consolidated list or table of definitions.

Response: We have compiled a table for important but uncommon acronyms. This table will be added as an appendix to serve as a convenient reference for readers.

Table R1. List of acronyms and their definitions

Acronym	Definition
ET	Extremely Randomized Trees
GB	Gradient Boosting
GEV	Generalized Extreme Value
IDF	Intensity-Duration-Frequency
IDW	Inverse Distance Weighting
KED	Kriging with External Drift

KED_AP	Kriging with External Drift using Mean Annual Precipitation
KED_DEM	Kriging with External Drift using Elevation
KGE	Kling-Gupta Efficiency
LDLR	Long Duration-Large Return Period
LDSR	Long Duration-Small Return Period
LR	Linear Regression
ML	Machine Learning
MLP	Multilayer Perceptron
NE	Northeastern Monsoon Region
NSE	Nash-Sutcliffe Efficiency
NW	Northwestern Arid Region
OK	Ordinary Kriging
PBIAS	Percent Bias
RF	Random Forest
RMSE	Root Mean Square Error
SDLR	Short Duration-Large Return Period
SDSR	Short Duration-Small Return Period
SE	Southeastern Monsoon Region
SW	Southwestern Tibetan Plateau Region

Comment 3: The study demonstrates that ML models, like gradient boosting, can estimate sub-daily intensities from daily gridded data with accuracy comparable to interpolation methods using hourly data. However, the manuscript lacks a detailed explanation of how ML achieves this temporal downscaling. What specific features or model structures enable this capability? For example, are statistical features like daily extreme precipitation or skewness critical? A discussion or sensitivity analysis of key input variables (Table 1) would clarify this process.

Response: We agree that a more detailed explanation of the machine learning model's performance would strengthen the manuscript. Conceptually, the gradient boosting (GB)

model does not disaggregate rainfall temporally in a mechanistic sense. Rather, it makes attempts and learns a cross-scale statistical mapping. To appreciate this mapping, we have conducted a feature importance analysis using Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017). This method, grounded in cooperative game theory, allows us to attribute the model's prediction to each input feature, providing a quantitative explanation of the model's behavior.

The feature analysis shows that daily-scale features, particularly those that summarize the right-tail characteristics of the daily precipitation distribution, carry information about the local storm climatology (Figure R1). The GB model leverages this information, in conjunction with static covariates like latitude, longitude, and altitude, to predict the expected intensity of sub-daily extreme events at each location. In all four cases we examined, the dominant drivers were features derived from the gridded precipitation data characterizing daily extremes, specifically the gridded daily precipitation of return periods and the average annual maximum daily precipitation. Collectively, these features contributed approximately half to two-thirds of the total feature importance.

In addition, the model preferentially relies on daily-scale tail metrics that are closest in frequency to the sub-daily target frequency. For instance, for the SDLR case, the daily 100-year feature was the most influential. For the LDSR case, the daily 5-year feature was the most significant. This finding is reasonable. When an extreme precipitation event occurs at a station, regardless of its duration, it is expected that the corresponding daily gridded precipitation would also be large with strong spatial correlations. The machine learning model, with its ability to capture non-linear relationships, leverages this common pattern observed across a vast number of samples to infer the station's extreme precipitation at various scales (including sub-daily scales) from the daily extreme data. For the LDLR case, yet, the daily 50-year feature was more informative than the 100-year feature. A possible explanation is that the extremeness of precipitation is somewhat moderated over longer durations. This may allow the model to rely on daily precipitation metrics that are slightly less extreme. Because 50-year events are more common than 100-year events, they offer a more stable signal of local

tail behavior with less sampling noise and reduced difficulty in extrapolation. This stability makes the 50-year feature a more reliable predictor for the model to utilize.

Furthermore, geographic variables such as altitude, latitude, and longitude consistently rank as highly important secondary features. This indicates that the model is not simply performing a statistical scaling but is also effectively learning and incorporating fundamental climatological and topographical controls on precipitation. Given the general trend of precipitation decreasing from the lower-altitude southeastern coastal areas to the higher-altitude northwestern regions in mainland China, a pattern driven by monsoonal weakening with inland distance and continental topographic gradients, these geographic features provide the model with a crucial climatological baseline. They establish the large-scale spatial context for extreme precipitation intensity, which the model then refines using the daily precipitation data. In short, the model's ability to leverage these spatial features in conjunction with daily extreme precipitation statistics is central to its capacity to produce spatially heterogeneous IDF estimations.

While this analysis clarifies which features are most critical for this cross-scale statistical prediction, we acknowledge that it does not fully elucidate the specific internal mechanisms of the downscaling process. we agree that a deeper mechanistic exploration is an interesting and important challenge for future studies. We believe our interpretability provides a valuable first step in the right direction and a foundation for such future investigations.

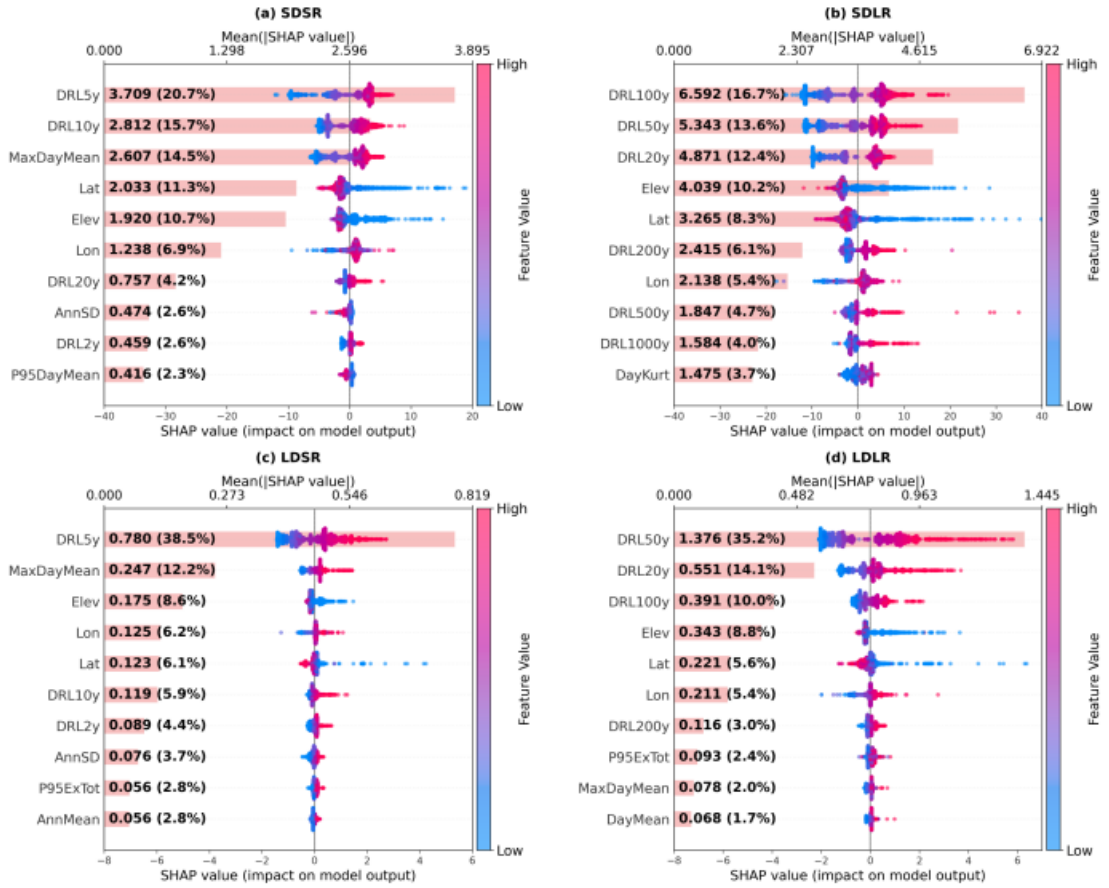

Figure R1. SHAP feature importance for GB-based IDF regionalization in mainland China (four target cases). Horizontal bars show global importance as mean absolute SHAP values (mean|SHAP|) with percentages; beeswarm points show sample-level signed SHAP values; point color encodes the standardized feature value (z-score) from low to high. Lat: latitude; Lon: longitude; Elev: elevation; AnnMean: mean annual precipitation; AnnSD: standard deviation of annual precipitation; DayMean: mean daily precipitation; DayKurt: kurtosis of daily precipitation; MaxDayMean: multi-year mean of the annual maximum daily precipitation; P95DayMean: multi-year mean of the annual 95th percentile of daily precipitation; P95ExTot: multi-year mean annual total precipitation from days exceeding the 95th percentile; DRL2y-DRL1000y: daily return level for a 2-1000 year return period.

Comment 4: Table 1 lists geographic coordinates, elevation, and precipitation statistics as independent variables for ML. Why were these variables chosen, and were other meteorological variables, such as temperature or humidity, tested? Given their potential

influence on extreme precipitation, justifying their exclusion or inclusion would enhance the robustness of the ML approach.

Response: Our study set out to provide a comprehensive comparison between site-observation-based and gridded-precipitation-based IDF regionalization methods across mainland China. For the interpolation methods, the foundational data set consists of the IDF curves derived from station-level precipitation records. Advanced geostatistical techniques like Kriging with External Drift (KED) can incorporate covariates to improve accuracy, and previous studies have shown that variables such as elevation and mean annual precipitation are common and effective choices (Van De Vyver, 2012; Yin et al., 2018; Zou et al., 2021). To ensure a methodologically fair and direct comparison between the two approaches, we constrained the machine learning models to use a set of predictors in parallel. This approach can also show the potential inherent within precipitation data itself, without introducing additional meteorological variables.

Our findings demonstrate that high levels of accuracy were achieved using only these features. For instance, the best-performing interpolation method, KED with mean annual precipitation, yielded KGE values greater than 0.96 for 1-hour-5-year storms and greater than 0.84 for 1-hour-100-year storms. The top machine learning model, Gradient Boosting, achieved comparable accuracy with KGE values exceeding 0.94 and 0.87 for the same respective storm events. This high performance underscores that the predictive ability of our models stems from the precipitation data rather than the simple aggregation of numerous meteorological variables. An evident advantage of this methodology is its enhanced applicability to regions where other meteorological data, such as temperature and humidity, may be unavailable, less reliable, or of coarser resolution. We do, however, acknowledge that our models showed reduced performance in certain complex regions. The inclusion of additional meteorological variables could indeed have the potential to improve accuracy in these specific areas, and this represents a valuable and promising direction for our future research.

Comment 5: Line 291, you mention it was repeated five times. Clarify if this was with or without replacement.

Response: The five-fold cross-validation was performed using sampling without replacement, ensuring that each data point was used for validation once only. We will clarify this in the revised manuscript.

Comment 6: Section 2.1, the division of mainland China into four regions (NE, SE, NW, SW) is based on climate and topography, with the Eastern Monsoon region split along the Qinling-Huaihe line due to its heterogeneity. Was this subdivision sufficient to capture regional variability, particularly in the SE region with extreme precipitation? Could further sub-regionalization or alternative regionalization schemes improve model performance?

Response: The primary purpose of our initial four-region division was to assess whether the methods, despite their strong performance at the national scale, would exhibit significant performance degradation for distinct climatic and topographical zones. This objective was successfully met by demonstrating that model accuracy varies regionally, most notably highlighting the need for caution when applying the dataset in the SW region. However, we agree with the reviewer that a more granular subdivision could capture regional variability in greater detail and provide more specific guidance for users, thereby increasing the practical value of our work. We have conducted an additional analysis using an alternative regionalization scheme based on the nine major river basins of China (Figure R2).

The results from this basin-level analysis corroborate the findings presented in our manuscript (Tables R2-R9). For instance, both the GB and KED_AP methods consistently show the poorest performance in the Southwest Basin across all four test cases, reaffirming the challenges posed by its complex topography and the sparse station network. The Southeast Basin also exhibited reduced predictive accuracy, a result likely attributable to the combined effects of extreme precipitation and hilly topography, which present inherent challenges to spatial modeling. In contrast, basins with denser gauge networks and more stable precipitation patterns such as the Huaihe and Haihe basins retain high precision. This basin-based perspective provides a useful alternative view for readers who focus on these specific hydrologic regions. We will

include these results as the supplementary materials for reference.

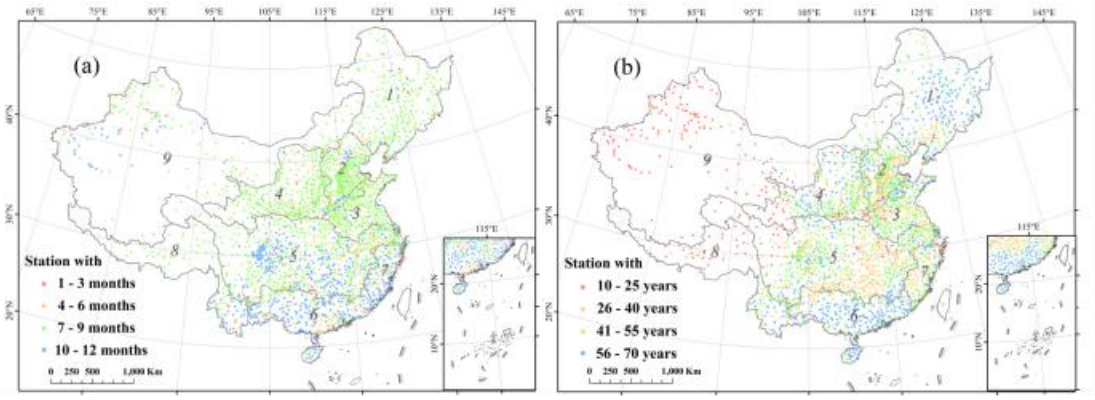

Figure R2. Similar to Figure 1 in the main text, yet this map illustrates the spatial distribution of the nine major river basins in mainland China, which were used as an alternative regionalization scheme for performance evaluation. The numbered basins are: 1: Songhua and Liaohe River Basin; 2: Haihe River Basin; 3: Huaihe River Basin; 4: Yellow River Basin; 5: Yangtze River Basin; 6: Pearl River Basin; 7: Southeast Basin; 8: Southwest Basin; 9: Continental Basin.

Table R2. Accuracy metrics for the KED_AP interpolation method across the nine major river basins for the SDSR

Region	NSE	PBIAS (%)	RMSE	KGE
1.Songhua and Liaohe River Basin	0.85	0.38	2.68	0.88
2.Haihe River Basin	0.9	-0.41	3.19	0.93
3.Huaihe River Basin	0.96	-0.17	3.9	0.97
4.Yellow River Basin	0.89	-0.03	3.28	0.92
5.Yangtze River Basin	0.82	0.13	4.03	0.86
6.Pearl River Basin	0.72	-0.02	5.25	0.8
7.Southeast Basin	0.58	0.36	3.62	0.7
8.Southwest Basin	0.09	-0.72	3.44	0.25
9.Continental Basin	0.72	-0.06	3	0.83
Mainland China	0.94	0.03	3.79	0.96

Table R3. Accuracy metrics for the KED_AP interpolation method across the nine major river basins for the SDLR

Region	NSE	PBIAS (%)	RMSE	KGE
1.Songhua and Liaohe River Basin	0.66	0.64	8.24	0.73
2.Haihe River Basin	0.73	-0.4	14.99	0.78
3.Huaihe River Basin	0.91	0.49	12.27	0.92
4.Yellow River Basin	0.65	-0.27	15.26	0.72
5.Yangtze River Basin	0.57	0.09	15.22	0.65
6.Pearl River Basin	0.52	-0.07	18.33	0.62
7.Southeast Basin	0.43	0.68	15.53	0.48
8.Southwest Basin	0.22	-2.5	14.42	0.25
9.Continental Basin	0.27	-0.48	11	0.46
Mainland China	0.79	-0.01	14.72	0.84

Table R4. Accuracy metrics for the KED_AP interpolation method across the nine major river basins for the LDSR

Region	NSE	PBIAS (%)	RMSE	KGE
1.Songhua and Liaohe River Basin	0.9	0.14	0.19	0.92
2.Haihe River Basin	0.88	0.11	0.27	0.92
3.Huaihe River Basin	0.96	0.15	0.42	0.96
4.Yellow River Basin	0.87	-0.19	0.34	0.89
5.Yangtze River Basin	0.86	-0.1	0.44	0.89
6.Pearl River Basin	0.86	-0.02	0.67	0.89
7.Southeast Basin	0.61	0.93	0.73	0.69
8.Southwest Basin	0.42	0.84	0.55	0.56
9.Continental Basin	0.78	0.18	0.28	0.85

Mainland China	0.95	0.07	0.45	0.96

178

179 Table R5. Accuracy metrics for the KED_AP interpolation method across the nine
180 major river basins for the LDLR

Region	NSE	PBIAS (%)	RMSE	KGE
1.Songhua and Liaohe River Basin	0.71	0.97	0.66	0.77
2.Haihe River Basin	0.67	-0.56	1.41	0.73
3.Huaihe River Basin	0.9	0.35	1.36	0.9
4.Yellow River Basin	0.56	-0.15	1.58	0.65
5.Yangtze River Basin	0.64	-0.04	1.64	0.71
6.Pearl River Basin	0.72	0.37	2.12	0.77
7.Southeast Basin	0.48	0.75	2.16	0.52
8.Southwest Basin	0.2	-0.69	1.48	0.21
9.Continental Basin	0.31	-0.63	1.09	0.48
Mainland China	0.82	0.05	1.6	0.87

181

182 Table R6. Accuracy metrics for the GB machine learning method across the nine
183 major river basins for the SDSR

Region	NSE	PBIAS (%)	RMSE	KGE
1.Songhua and Liaohe River Basin	0.77	-0.78	3.26	0.8
2.Haihe River Basin	0.87	-0.23	3.52	0.91
3.Huaihe River Basin	0.94	-1.36	4.74	0.94
4.Yellow River Basin	0.82	-0.14	4.26	0.86
5.Yangtze River Basin	0.78	0.51	4.55	0.82
6.Pearl River Basin	0.69	-1.02	5.72	0.81
7.Southeast Basin	0.39	0.86	4.33	0.49
8.Southwest Basin	-0.23	-1.72	3.99	0.18

9.Continental Basin	0.75	1.43	2.86	0.82
Mainland China	0.92	-0.01	4.34	0.94

184

185 Table R7. Accuracy metrics for the GB machine learning method across the nine
186 major river basins for the SDLR

Region	NSE	PBIAS (%)	RMSE	KGE
1.Songhua and Liaohe River Basin	0.68	-1.18	7.9	0.71
2.Haihe River Basin	0.79	0.48	13.18	0.82
3.Huaihe River Basin	0.91	-0.45	12.24	0.91
4.Yellow River Basin	0.67	-0.36	14.87	0.74
5.Yangtze River Basin	0.66	1.07	13.57	0.75
6.Pearl River Basin	0.68	-1.54	15.23	0.77
7.Southeast Basin	0.57	0.66	13.51	0.63
8.Southwest Basin	0.27	-4.19	14.7	0.12
9.Continental Basin	0.4	0.49	9.93	0.52
Mainland China	0.83	0.1	13.3	0.87

187

188 Table R8. Accuracy metrics for the GB machine learning method across the nine
189 major river basins for the LDSR

Region	NSE	PBIAS (%)	RMSE	KGE
1.Songhua and Liaohe River Basin	0.87	2.39	0.22	0.92
2.Haihe River Basin	0.92	-0.07	0.22	0.96
3.Huaihe River Basin	0.96	-0.7	0.39	0.94
4.Yellow River Basin	0.83	0.28	0.39	0.91
5.Yangtze River Basin	0.88	0.08	0.42	0.93
6.Pearl River Basin	0.83	-0.83	0.77	0.89

7.Southeast Basin	0.7	-0.39	0.63	0.75
8.Southwest Basin	0.46	-1.64	0.53	0.74
9.Continental Basin	0.86	0.74	0.23	0.9
Mainland China	0.95	-0.08	0.44	0.96

Table R9. Accuracy metrics for the GB machine learning method across the nine major river basins for the LDLR

Region	NSE	PBIAS (%)	RMSE	KGE
1.Songhua and Liaohe River Basin	0.68	3.94	0.69	0.81
2.Haihe River Basin	0.77	-0.17	1.16	0.82
3.Huaihe River Basin	0.92	-0.09	1.23	0.91
4.Yellow River Basin	0.61	-0.06	1.48	0.73
5.Yangtze River Basin	0.77	0.47	1.3	0.85
6.Pearl River Basin	0.79	-1.11	1.88	0.85
7.Southeast Basin	0.68	-0.85	1.7	0.75
8.Southwest Basin	0.08	-3.37	1.72	0.39
9.Continental Basin	0.5	-0.31	0.93	0.63
Mainland China	0.88	-0.05	1.36	0.91

Comment 7: The study interpolates missing hourly data for gaps <12 hours and assigns zero for gaps ⩾12 hours (beginning on line 157). How was the impact of this imputation strategy assessed, and what are its implications for IDF curve accuracy in regions with frequent missing data?

Response: The core of our methodology for developing the station-level IDF curves involves identifying the annual maximum series for various durations and subsequently fitting a Generalized Extreme Value (GEV) distribution to these series. A crucial aspect of this process is that it relies on the peak rainfall values within each year, rather than

the entire continuous record. We first applied a quality control filter, ensuring that only stations with less than 10% missing data per year were included in the analysis. This initial screening reduces the probability that a gap in the data record would coincide with the annual maximum rainfall event for that year, which by its nature represents a very small fraction of the total time in any given year. Furthermore, the statistical fitting of the GEV distribution across a long-term series of annual maxima provides an additional layer of robustness. Even if the maximum value for a small number of years in the series were slightly affected by missing data, the overall integrity of the GEV distribution is largely maintained by the majority of years when the record was complete.

To quantitatively quantify the effect of our data-filling strategy, we conducted a comparative analysis. We recalculated the IDF curves for all stations from the raw data, treating the missing periods as having no precipitation, and compared these results to the IDF curves generated using the data-filling procedure described in the manuscript. For the four representative cases of SDSR, SDLR, LDSR, and LDLR, the Pearson correlation coefficients between the two sets of results were 0.9996, 0.9995, 1.0000, and 0.9998, respectively. This high level of consistency confirms that our approach for managing missing data has a negligible effect on the derived IDF curves. While we acknowledge the small chance that a significant number of extreme precipitation events could occur during periods of missing data and thereby affect the results, we believe that the chosen method has been a practical and effective solution given the available data, ensuring the high quality of the station-level IDF curves that form the basis of our regionalization study.

Comment 8: The SW region shows significantly lower accuracy (KGE as low as 0.31 for KED_AP and 0.14 for GB), attributed to sparse station density and complex topography. Given the lack of validation stations in parts of the NW and SW regions, how reliable are the IDF curves in these areas? Should users be explicitly cautioned against using these curves without further validation?

Response: We agree that the predictive accuracy in the NW and SW regions is lower, and we appreciate the reviewer highlighting this point. In fact, a key reason for dividing

the study area into distinct regions was to demonstrate that a model performing well at the national level may have reduced accuracy at a regional scale, which warrants caution during application. In our revised manuscript, we will place emphasis on this point as a caution about the limitations of reliability when applying the IDF curves in these specific regions.

Comment 9: The manuscript notes that hyperparameter tuning via grid search did not significantly improve ML performance, so default settings were used. Why do you think tuning was ineffective? Were the default parameters near-optimal, or were the tuning ranges too narrow? Clarifying this would help readers assess the robustness of the ML models.

Response: We acknowledge that our initial grid search, which used manually selected nodes with relatively large intervals and a limited scope, might have overlooked hyperparameter combinations that could enhance model accuracy. To address this limitation, we employed Bayesian optimization, a more advanced and efficient method for hyperparameter tuning (Mockus, 1998; Snoek et al., 2012). This technique is particularly suitable for exploring large and complex parameter spaces because it uses the results from previous iterations to inform which set of hyperparameters to test next. For this new analysis, we also expanded the search range for the hyperparameters of our machine learning models to ensure a more thorough exploration (Tables R10-R13).

After 100 iterations, the results of this extensive tuning process revealed that the performance of our models is stable and not highly sensitive to hyperparameter variations within these broad ranges (Table R14 and Table 3 in the manuscript). While the Bayesian optimization did yield some minor improvements in accuracy, the overall gains were marginal and did not alter the fundamental conclusions of our study. This indicates that the default hyperparameter settings already provide near-optimal performance for this model in the context of our study. Additionally, in our study, the main findings are based on the outputs of the GB model, given its superior performance. However, a comparison of the performance metrics before and after optimization shows that GB exhibited only a slight increase in accuracy. For any other model to challenge

the primary findings, it would require a substantial increase in accuracy from hyperparameter tuning, enough to surpass the performance of GB. Our comprehensive search demonstrated that such a significant improvement was not attainable. Therefore, our original choice to proceed with default hyperparameters for simplicity and consistency is further supported by this new analysis.

Nevertheless, we agree that no search can be truly exhaustive, and it remains possible that parameter combinations outside the already extensive ranges we tested could provide incremental benefits. We will state this point in the Discussion section of our manuscript to acknowledge this inherent limitation.

Table R10. Hyperparameter search ranges for the Random Forest model using Bayesian optimization.

Hyperparameter	Search range for Bayesian optimization	SDSR	SDLR	LDSR	LDLR
n_estimators	50 - 1000	914	173	943	1000
min_samples_split	2 - 50	2	2	2	2
max_features	0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, None, “sqrt”, “log2”	None	None	0.90	None

Table R11. Hyperparameter search ranges for the Gradient Boosting model using Bayesian optimization.

Hyperparameter	Search range for Bayesian optimization	SDSR	SDLR	LDSR	LDLR
n_estimators	50 - 1000	826	1000	640	607
max_depth	2 - 50	4	2	2	6
max_features	0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, None, “sqrt”, “log2”	0.90	0.40	0.20	0.10

subsample	0.10 - 1.00	0.3771	0.8657	0.9867	0.2546
learning_rate	0.01 - 0.50	0.0791	0.1145	0.2807	0.0100

Table R12. Hyperparameter search ranges for the Extremely Randomized Trees model using Bayesian optimization.

Hyperparameter	Search range for Bayesian optimization	SDSR	SDLR	LDSR	LDLR
n_estimators	50 - 1000	938	899	72	1000
min_samples_split	2 - 50	2	12	4	11
max_features	0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, None, “sqrt”, “log2”	None	None	None	0.10

Table R13. Hyperparameter search ranges for the Multilayer Perceptron model using Bayesian optimization.

Hyperparameter	Search range for Bayesian optimization	SDSR	SDLR	LDSR	LDLR
hidden_layer_sizes	(50,), (64,), (100,), (128,), (256,), (500,), (750,), (1000,), (64, 64), (128, 64), (64, 128), (150, 150), (200, 200), (250, 250), (15, 30, 45), (30, 45, 30),	(128,64)	(150,150)	(150,150)	(30,45,30)

	(45, 30, 15), (10, 10, 10, 10), (15, 15, 15, 15)				
activation	"logistic", "tanh", "relu", "identity"	"tanh"	"relu"	"relu"	"relu"
alpha	(1e-6) - 1	0.7277	0.01525	(1e-6)	0.7825

Table R14. Accuracy metrics of machine learning methods after hyperparameter tuning via Bayesian optimization.

	SDSR				SDLR				LDSR				LDLR			
	NSE	PBIAS (%)	RMSE	KGE	NSE	PBIAS (%)	RMSE	KGE	NSE	PBIAS (%)	RMSE	KGE	NSE	PBIAS (%)	RMSE	KGE
RF	0.92	0.11	4.30	0.94	0.83	0.33	13.40	0.87	0.94	0.00	0.47	0.96	0.88	0.18	1.37	0.91
GB	0.93	0.19	4.18	0.95	0.84	-0.06	13.16	0.89	0.95	0.00	0.44	0.97	0.88	0.07	1.36	0.90
ET	0.92	0.08	4.48	0.93	0.82	0.24	13.82	0.86	0.94	0.01	0.49	0.95	0.86	0.02	1.42	0.89
MLP	0.92	-0.35	4.33	0.93	0.83	-0.16	13.38	0.88	0.95	-0.15	0.43	0.97	0.88	-0.61	1.34	0.92

Comment 10: Line 291, The introduction references non-stationarity in IDF curves due to climate change, but the methodology does not account for it (for example, different RCP scenarios). Were tests conducted to evaluate the impact of non-stationarity, particularly for long return periods, 100 or 1000 years? A brief discussion or analysis of this issue would align the study with current climate research trends.

Response: We agree that in the context of global warming, the distribution functions used to establish IDF curves are expected to change with the observed increase in extreme rainfall events. Since the estimates for cases with large return periods are calculated from the tail of the distribution, they would be more sensitive to such changes, which could indeed alter the results for these cases. Given that our current study has already achieved good results using stationary IDF curves based on historical

observations, it is a worthwhile endeavor to explore how these methods would perform with non-stationary IDF curves in a warming world. We will add a brief discussion in the discussion section to acknowledge this point and suggest a further exploration in the future.

Comment 11: The manuscript cites a high-resolution IDF dataset in the introduction for the Qinghai-Tibet Plateau (Ren et al., 2025). A quantitative comparison with this dataset in the SW region would benchmark the study's results and highlight its unique contributions.

Response: Thank you for your constructive suggestion. We have conducted this quantitative comparison analysis. In our analysis, we extracted the gridded IDF values from the Ren et al. (2025) product at the locations of our observation stations within the SW region as examples and compared them against the predictions from our GB models (Figure R3). The comparison shows a moderate degree of correlation for cases with smaller return periods, but a weaker relationship for cases with larger return periods. We also observed that our model's estimates for short-duration extreme precipitation tended to be lower than those from the Ren et al. dataset. We attribute these discrepancies primarily to the differences in spatial resolution between the two studies. Events with short durations and large return periods represent the most intense and often localized rainfall phenomena. As discussed in our paper, the SW region is characterized by exceptionally complex topography and high regional heterogeneity. These characteristics exert a strong influence on extreme precipitation, leading to significant local variations. Our model, which was trained on a national scale with a spatial resolution of 0.1 degrees, is designed for broad applicability and generalizability across mainland China. This coarser resolution may inherently smooth over the sharp and localized peaks of extreme rainfall that a high-resolution model is better equipped to capture. The study by Ren et al., by focusing its training and prediction specifically on samples from the SW region, was able to achieve a finer spatial resolution (1/30°) that is likely more sensitive to these localized extremes. However, as the study by Ren et al. does not provide an explicit accuracy assessment against IDF curves from ground

stations, it is difficult to evaluate quantitatively which set of predictions is more accurate for this specific region.

Despite this challenge in direct benchmarking in terms of accuracy, we believe that both studies using machine learning hold distinct and important value. Our study successfully demonstrates the robustness and general applicability of a machine learning framework on a larger spatial scale by regionalizing IDF curves across mainland China, an area encompassing diverse climates and topographies. The work by Ren et al., in contrast, serves as a valuable complement, providing a high-resolution analysis tailored specifically to one of the nation's most complex regions.

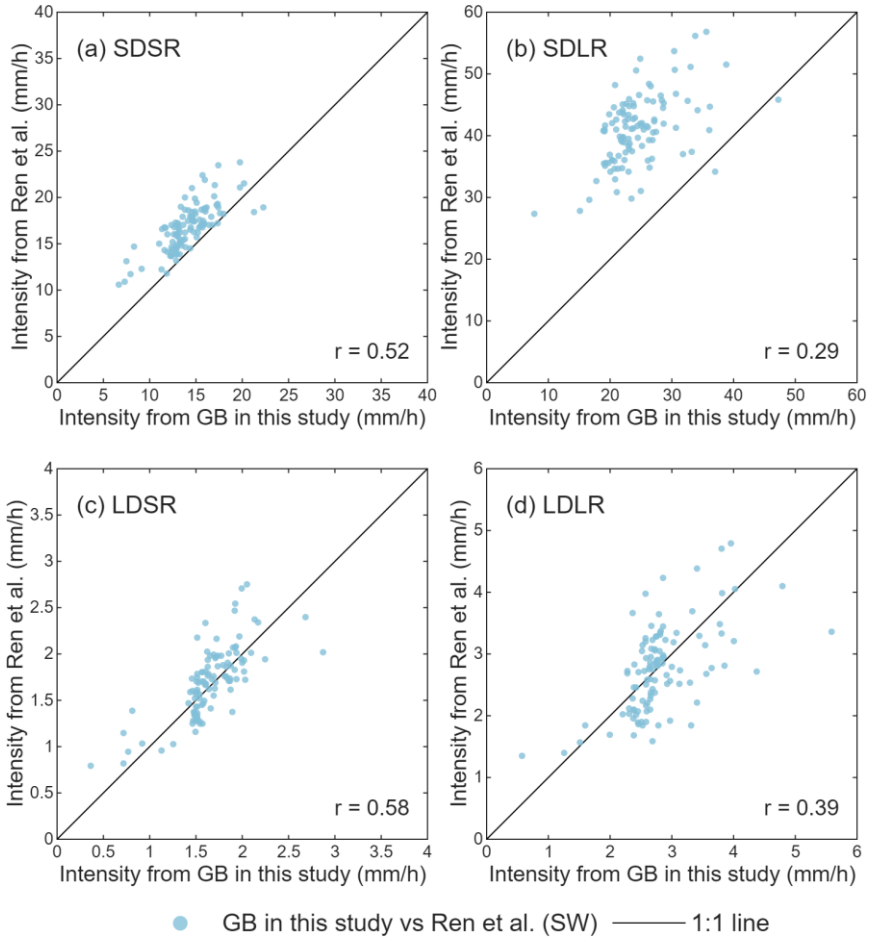

Figure R3. Scatterplots comparing IDF estimates from this study's GB model against the Ren et al dataset for four representative cases in the SW region.

Comment 12: The IDF curves are provided at 0.1° and 0.5° resolutions, but their

alignment with specific applications (such as urban drainage design, flood modeling) is unclear. Are these resolutions optimized for particular use cases, and how should users select between them? Providing guidance would enhance the dataset's practical utility.

Response: We thank the reviewer for this helpful suggestion. The high-resolution 0.1° product can better preserve the steep local gradients and high spatial variability characteristic of extreme, short-duration precipitation events. Therefore, we recommend the 0.1° dataset for detailed urban hydrology studies, the design of urban drainage networks, flood modeling in small catchments, and the design of hydraulic infrastructure such as culverts and bridges. In these contexts, the 0.5° product could lead to an underestimation of peak rainfall intensity, potentially resulting in undersized infrastructure. The 0.5° product is intended to represent areal-mean rainfall conditions over a larger area. This resolution is preferable for large-scale hydrologic modeling and basin-wide water resource assessments. This coarser resolution offers benefits in terms of computational efficiency, with fewer grid cells leading to reduced data storage requirements and faster model run times. This also aligns with the input data assumptions of some models which operate on a similar resolution.

To provide a more intuitive understanding of this scale difference, a user can consider the number of IDF curves available for a given area. For instance, for a 2500 km^2 region, our 0.1° product provides approximately 25 distinct IDF curves, whereas the 0.5° product yields approximately one, representing an averaged condition. Taking the city of Beijing (approximately 16000 km^2) as another example, an area with a relatively dense network of observation stations, the existing hourly gauges provide about 20 station-level IDF curves. Our 0.1° dataset increases this to about 170 available IDF curves, while the 0.5° dataset provides about 8 curves.

In short, the 0.1° dataset should be selected for local-scale and urban studies that demand higher spatial fidelity to capture localized extremes, while the 0.5° dataset is suitable for regional and basin-scale analyses where computational efficiency and a representation of areal-average conditions are the priority. In our revised manuscript, we will add a statement to provide this guidance, helping users select the appropriate dataset resolution for their specific application.

Comment 13: GB outperforms other ML methods, but the manuscript does not discuss its interpretability or the relative importance of input features. A feature importance analysis would provide insights into which variables drive performance, aiding future model development.

Response: To address this, we have performed an interpretability analysis using Shapley Additive Explanations (SHAP), as our response to the previous comment (Figure R1). This analysis allows for a clear quantification of the relative importance of each input feature. In the revised manuscript, we intend to include this feature importance analysis in the supplementary materials to assist the readers interested in model development.

Comment 14: Figure 7: The caption mentions 500 samples but does not explain the sampling method (for example, bootstrap or Monte Carlo). Add a brief clarification.

Response: Thank you for your suggestion. We used the Monte Carlo method to generate the samples. We will add this clarification to the caption of Figure 7 in the revised manuscript.

Comment 15:

(1) Inconsistent spacing in "machine learning" vs. "machinelearning" appears in several instances (for example lines 103, 408). Standardize to "machine learning."

(2) Inconsistent spacing before references (for example, line 108, 110, 113). Check formatting.

(3) Line 735: "Deepseek R1" clarify the tool's name and provide a citation or link for transparency.

(4) Figure 1: Include a description of the inset in the caption.

(5) Standardize color scales across panels (a–d for KED_AP, e–h for GB) to facilitate direct comparisons. Ensure units (mm/h) are explicitly labeled in the caption or legend.

(6) Table 2 and Table 3: Ensure consistent formatting of numerical values (for example, PBIAS values should all include the % symbol). Add a footnote clarifying that negative

PBIAS indicates underestimation.

Response: We appreciate the reviewer pointing out these issues with formatting and clarity. We will carefully check the entire manuscript and correct all the points mentioned.

Reference:

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).

Mockus, J., Tiesis, V., & Žilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. In L. C. W. Dixon & G. P. Szegő (Eds.), *Towards global optimisation* (Vol. 2, pp. 117–129).

Ren, Z., Sang, Y.-F., Cui, P., Chen, F., & Chen, D. (2025). A dataset of gridded precipitation intensity-duration-frequency curves in Qinghai-Tibet Plateau. *Scientific Data*, *12*(1), 3. https://doi.org/10.1038/s41597-024-04362-1

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25, pp. 2951–2959).

Van de Vyver, H. (2012). Spatial regression models for extreme precipitation in Belgium. *Water Resources Research, 48*, W09549. https://doi.org/10.1029/2011WR011707

Yin, S., Wang, Z., Zhu, Z., Zou, X., & Wang, W. (2018). Using Kriging with a heterogeneous measurement error to improve the accuracy of extreme precipitation return level estimation. *Journal of Hydrology, 562*, 518–529. https://doi.org/10.1016/j.jhydrol.2018.04.064

Zou, W., Yin, S., & Wang, W. (2021). Spatial interpolation of the extreme hourly precipitation at different return levels in the Haihe River basin. *Journal of Hydrology, 598*, 126273. https://doi.org/10.1016/j.jhydrol.2021.126273