https://doi.org/10.5194/egusphere-2026-528
Preprint. Discussion started: 10 March 2026

A temporally continuous, probabilistic framework for observed multi-decadal flood susceptibility evolution in Canadian watersheds

Heather McGrath[1]

[1]Canada Centre for Mapping and Earth Observation, Natural Resources Canada, Ottawa, K1A 1G5 Canada

Correspondence to: Heather McGrath (heather.mcgrath@nrcan-rncan.gc.ca)

Abstract: This study advances flood susceptibility analysis by introducing a temporally continuous, uncertainty-aware framework that moves beyond static or snapshot-based mapping. We leverage outputs from a machine learning model trained on a multi-decadal record of historic flood events which generated 24 annual flood susceptibility (FS) maps spanning 2000–2023. Annual watershed scores are derived from normalized pixel proportions and thresholds. Generalized Extreme Value (GEV) distributions fitted to these score series define watershed-specific tails of wetness and dryness, with uncertainty quantified via moving-block-bootstrap. Extreme years are refined using neighbour-year expansion to capture short-term hydroclimatic regimes and validated through change-point detection and Mann–Kendall trend analysis. Pixelwise envelopes are generated by aggregating FS values across selected extreme years and spatial smoothing for coherence. National-scale analysis reveals a clear increase in flood susceptibility in the 2020s across many watersheds, with notable clusters of extreme wet years from 2017–2023 in Atlantic Canada and the St. Lawrence River basin. The 2000s serve as a baseline period, the 2010s represent a transition decade with rising FS, and the 2020s demonstrate the strongest increase in wet extremes and spatial clustering. By explicitly treating flood susceptibility as a temporally evolving, stochastic process, this framework provides probabilistic bounds and diagnostic insights that extend beyond conventional static mapping, offering a robust basis for adaptive flood susceptibility assessment and long-term planning under changing hydroclimatic conditions.

1. Introduction

Flood susceptibility (FS) mapping is widely used to support flood screening, inform land-use planning and disaster mitigation, yet most existing approaches provide static representations of flooding or rely on event-based analyses (Tsumita et al. 2025; Chihi et al. 2025; S et al. 2025; Grosso et al. 2015). Such representations implicitly assume stationarity in the spatial patterns of FS, despite growing evidence that hydroclimatic variability and long-term change are altering flood-generating processes across many regions (Tramblay et al. 2025; Wyżga et al. 2016). Consequently, there is limited understanding of how FS evolves over time and how these changes manifest spatially at hydrologically meaningful scales. Addressing this gap requires frameworks that move beyond single-period assessments to explicitly quantify temporal trends and uncertainty in FS using observed data. In this study, we introduce a temporally continuous, uncertainty-aware framework that leverages multi-decadal records of historic flood events to characterize changes in FS across Canadian watersheds.

Flood mapping has advanced considerably over the past two decades, with approaches ranging from physically based hydrologic and hydraulic models to empirical and data-driven techniques (Schumann et al. 2015; Fivos Sargentis et al. 2025; Chandran et al. 2024; Bentivoglio et al. 2025). Machine learning (ML) methods have been increasingly adopted due to their ability to capture nonlinear relationships between flood occurrence and a wide range of physiographic, hydrometeorological, and land surface predictors. Numerous studies have demonstrated the effectiveness of such models in producing accurate FS maps at local to regional scales (Wang et al. 2025; Tsumita et al. 2025; Chihi et al. 2025; Aghenda et al. 2025). However, these applications are most often conducted for a single reference period or individual flood events, with model outputs interpreted as static representations of flood-prone areas (Pourzangbar et al. 2025). Even when multiple periods or scenarios are considered (Khodaei et al. 2025; Al-Ruzouq et al. 2024), analyses typically focus on comparative snapshots rather than explicitly characterizing the temporal evolution of FS as a continuous process. Recent work has explored similar questions using station-based hydrologic observations. For example, (Ibebuchi and Abu 2025) applied an explainable AI framework to predict FS and analyze multi-decadal trends in extreme streamflow across USGS monitoring stations in the western United States. While their approach provides valuable insights into drivers of susceptibility at the gauge scale, such methods are constrained by the spatial distribution of monitoring stations and may not translate easily to regions like Canada, where gauge density is lower and climate variability is pronounced.

The Generalized Extreme Value (GEV) distribution remains a cornerstone of flood frequency analysis (FFA) for estimating rare design quantiles, with recent advances showing gains from augmenting short records using regional or local counterfactual events (Voit et al. 2025; Kochanek et al. 2014). While such studies operate on discharge-based peaks, we extend the same rationale to temporally evolving FS indices, using GEV as a quantile-thresholding device rather than a peak-flow model.

Despite the widespread use of FS mapping, methods that explicitly quantify how susceptibility evolves over time using observed data remain scarce. Instead, existing approaches often rely on discrete snapshots or scenario-based analyses, which fail to capture the continuity and variability of susceptibility patterns, and uncertainty is typically addressed through ensemble or hypothetical modelling rather than climatologically plausible extremes derived from historical variability. To address these limitations—and recognizing the relatively sparse gauge network and pronounced climate variability in Canada—we introduce a temporally continuous, uncertainty-aware framework that leverages a 24-year sequence of calibrated FS maps to characterize watershed-specific extremes and their short-term hydroclimatic regimes. Our contributions include (i) moving beyond static mapping by exploiting annual susceptibility outputs as a spatiotemporal dataset, (ii) applying GEV-based tail modelling with block-bootstrap uncertainty quantification, and (iii) deriving spatial envelopes of susceptibility through neighbour-year expansion and pixel-level validation against exceedance frequencies. To our knowledge, no prior study combines these elements to provide probabilistic bounds and trend diagnostics for FS evolution at the watershed scale, offering a robust basis for adaptive flood management under changing hydroclimatic conditions.

2. Data

The data used in this study are outputs from an inference run of an XGBoost classification model. The model was trained using historic major flood events mapped by Natural Resources Canada's (NRCan) Emergency Geomatics Services (EGS) group for the years 2005, 2006, 2008, 2011, 2013–2015, and 2017–2023 across Canada. The XGBoost ensemble consists of 10 independently trained models, forming a total of 600 decision trees with a maximum depth of 20. Model performance on the validation set yielded an overall accuracy of 0.945, with true positive and true negative rates of 0.95 and 0.94, respectively, (McGrath 2025; McGrath et al. 2026; McGrath and Alhassan 2026; Dunbar et al. 2025) for full model details. Two categories of predictors are included, static and dynamic. The static predictors describing landscape characteristics, terrain, lithology, and hydrography and dynamic predictors including precipitation and temperatures preceding the spring freshet, land use/land cover from the North American Land Change Monitoring System and mean Normalized Difference Vegetation Index (NDVI) for Julian weeks 16–23 to capture typical early-spring vegetation conditions. The training/test/validation dataset was balanced and had 268,049 samples.

To convert the predicted probabilities into binary wet and dry classes, an optimal probability threshold was determined by using a hold-out validation set. Thresholds were scanned across the range of predicted probabilities, and the cutoff maximizing the F1-optimal threshold was selected. This resulted in a wet/dry cutoff $\theta_{raw} \approx 0.383$. To improve probability reliability, an isotonic regression calibration model $f: [0,1] \rightarrow [0,1]$ was fit to the validation set, providing monotonic mapping from raw model scores to calibrated probabilities. Because probability calibration preserves ranking, but does not enforce 0.5 as an optimal cutoff, the threshold was recomputed on calibrated probabilities by scanning the precision-recall curve and selecting the F1-optimal threshold, $\theta_{cal.}$

The calibrated XGBoost model was applied to annual predictor datasets for the period 2000–2023, producing a temporal ensemble of FS maps at 30-m spatial resolution covering all of Canada in EPSG:3979. Outputs were rescaled to the range 0–100 and stored as 8-bit raster files to reduce storage requirements. These 24 yearly rasters serve as the input dataset for this work.

The National Hydrographic Network, specifically the feature named 'Work Unit' (WU) is used as the processing units. These WU were created based on the Water Survey of Canada Sub-Sub-Drainage Areas (Canada, n.d.), where these represent natural drainage boundaries and may vary in size.

3. Method

3.1. Data and Processing Units

A watershed-specific, uncertainty-aware pipeline to identify extreme wet and dry years and neighbour-years from annual FS rasters and construct probabilistic spatial envelopes of FS is developed, as illustrated in Figure 1.

https://doi.org/10.5194/egusphere-2026-528
Preprint. Discussion started: 10 March 2026

There are 1338 WU in the dataset, and each WU has a two digit prefix to associate it with a major drainage basin, Figure 2. Three WU were selected to explore and describe in this analysis. Each contains labeled data that was used to train the ML model and are spread across the country and have varying latitudes. 01AL000, in Atlantic Canada which drains to the Saint John River, 05OG000 in central Canada, where terrain is relatively flat and agricultural land use, and 09AB000 in the Yukon, 95 the most northern site.

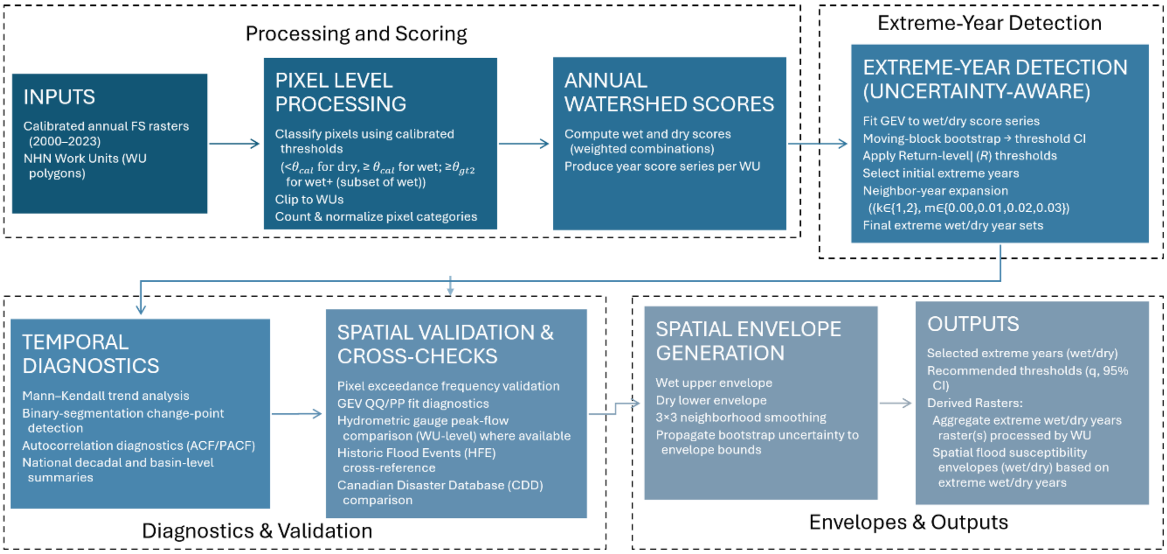
Processing and Scoring
INPUTS
Calibrated annual FS rasters (2000–2023)
NHN Work Units (WU polygons)
PIXEL LEVEL PROCESSING
Classify pixels using calibrated thresholds
(<θ_{cal} for dry, ≥ θ_{cal} for wet; ≥θ_{gt2} for wet+ (subset of wet))
Clip to WUs
Count & normalize pixel categories
ANNUAL WATERSHED SCORES
Compute wet and dry scores (weighted combinations)
Produce year score series per WU
Extreme-Year Detection
EXTREME-YEAR DETECTION (UNCERTAINTY-AWARE)
Fit GEV to wet/dry score series
Moving-block bootstrap → threshold CI
Apply Return-level| (R) thresholds
Select initial extreme years
Neighbor-year expansion ((k∈{1,2}, m∈{0.00,0.01,0.02,0.03})
Final extreme wet/dry year sets
TEMPORAL DIAGNOSTICS
Mann–Kendall trend analysis
Binary-segmentation change-point detection
Autocorrelation diagnostics (ACF/PACF)
National decadal and basin-level summaries
SPATIAL VALIDATION & CROSS-CHECKS
Pixel exceedance frequency validation
GEV QQ/PP fit diagnostics
Hydrometric gauge peak-flow comparison (WU-level) where available
Historic Flood Events (HFE) cross-reference
Canadian Disaster Database (CDD) comparison
Diagnostics & Validation
SPATIAL ENVELOPE GENERATION
Wet upper envelope
Dry lower envelope
3×3 neighborhood smoothing
Propagate bootstrap uncertainty to envelope bounds
OUTPUTS
Selected extreme years (wet/dry)
Recommended thresholds (q, 95% CI)
Derived Rasters:
Aggregate extreme wet/dry years raster(s) processed by WU
Spatial flood susceptibility envelopes (wet/dry) based on extreme wet/dry years
Envelopes & Outputs

Figure 1 Overview of the temporally continuous, uncertainty-aware workflow.

https://doi.org/10.5194/egusphere-2026-528
Preprint. Discussion started: 10 March 2026

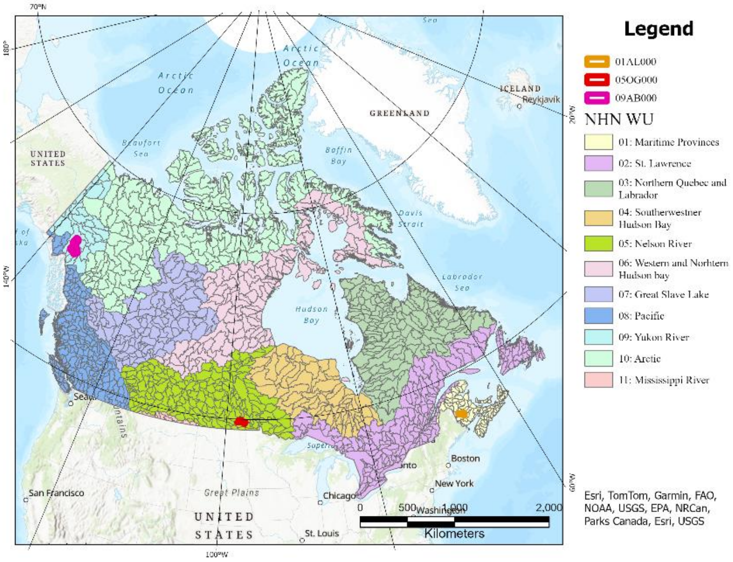

Figure 2 Study area and identification of all the NHN Work Units (WU), coloured by major drainage basins. Three WU which are analysed in the manuscript are identified using solid colours: (01AL000) in the Maritime Provinces (orange), 05OG000 in Nelson River (red) and 09AB000 in Yukon River (pink).

3.2. Wet and Dry scores

The byte-scaled susceptibility FS (0–100) data is processed by WU i and year t. Three classes are counted, *dry*, *wet* and *wet+*. The dry and wet boundary is determined by previous work, $\theta_{\text{cal}} = 0.383 * 100$. The higher threshold, $\theta_{\text{wet+}} = 88$ defines a class which is a strict subset of the wet class. This higher cutoff is chosen because is corresponds to the 25th percentile ($q25$) of all flooded pixel values across the full multi-year, multi-region dataset, ensuring that wet+ represents the upper three-quarters of observed wetness conditions, rather than marginal or low-intensity wet areas. The counts are normalized to ensure comparability across WUs of different sizes, Eq. 1, 2.

$$N_{i,t} = dry_{i,t} + wet_{i,t}, \tag{1}$$

$$\%\text{dry}_{i,t} = \frac{dry_{i,t}}{N_{i,t}}, \quad \%\text{wet}_{i,t} = \frac{wet_{i,t}}{N_{i,t}}, \quad \%\text{wet+}_{i,t} = \frac{wet+_{i,t}}{N_{i,t}}. \tag{2}$$

Additionally, if $wet\ +_{i,t} > wet_{i,t}$ due to upstream rounding or artifacts, we clip $wet\ +_{i,t} < wet_{i,t}$ and flag the record. All percentages are bounded to [0,1].

Two annual WU scores are computed: Wet Score, Eq. 3, and Dry Score, Eq. 4. In the Wet Score (S^w) wet and wet+ are favored while dry is penalized. For the Dry Score (S^d), the inverse, where dry is favored and wet+ is penalized more than wet.

$$S_{i,t}^{(w)} \;=\; w_{\text{wet}} \cdot \%\text{wet}_{i,t} \;+\; w_{\text{wet+}} \cdot \%\text{wet+}_{i,t} \;-\; w_{\text{dry}} \cdot \%\text{dry}_{i,t}. \tag{3}$$

$$S_{i,t}^{(d)} \;=\; \beta \cdot \%\text{dry}_{i,t} \;-\; \alpha \cdot \%\text{wet+}_{i,t} \;-\; \beta \cdot \%\text{wet}_{i,t}. \tag{4}$$

https://doi.org/10.5194/egusphere-2026-528
Preprint. Discussion started: 10 March 2026

Weights, constrained by $\alpha \geq \beta$, were set to: $w_{\text{wet}} = 1.0, w_{\text{wet+}} = 1.5, w_{\text{dry}} = 1.0, \alpha = 2.0, \beta = 1.0$. Additional weights were tested and results were found to be insensitive to reasonable weight changes.

To evaluate the temporal change from the Wet and Dry scores, Mann-Kendall's (MK) τ was computed separately for the annual wet and dry scores, τ_w *and* τ_d. Statistical significance was assessed using two-sided MK tests, with p-values adjusted across all water units (WUs) using the Benjamini–Hochberg false discovery rate (FDR) to get the dry and wet q values, q_d *and* q_w. WUs were subsequently classified into trend categories (strong or moderate wetting/drying, mixed significant, or inconclusive) based on the sign of Kendall's τ and the FDR-adjusted significance levels, Table 1.

Table 1 Trend classification rules as applied to Wet and Dry scores

Trend Class	Logical Condition	Expression
Mixed significant	If both wet and dry trends are significant at the strict FDR level	$q_w < 0.05$ and $q_d < 0.05$
Strong wetting	If the wet trend is significantly positive, and there is no significantly positive dry trend (strict FDR)	$\tau_w > 0$ and $q_w < 0.05$ AND NOT ($\tau_d > 0$ and $q_d < 0.05$)
Strong drying	If either the wet trend is significantly negative OR the dry trend is significantly positive (strict FDR)	($\tau_w < 0$ and $q_w < 0.05$) OR ($\tau_d > 0$ and $q_d < 0.05$)
Moderate wetting	If the wet trend is positive and significant at the loose FDR level, and no positive dry trend is significant at the same level.	$\tau_w > 0$ and $q_w < 0.10$ AND NOT ($\tau_d > 0$ and $q_d < 0.10$)
Moderate drying	If either the wet trend is negative OR the dry trend is positive, significant at the loose FDR level.	($\tau_w < 0$ and $q_w < 0.10$) OR ($\tau_d > 0$ and $q_d < 0.10$)
Inconclusive	All remaining cases not captured by the rules above	

3.3. Generalized Extreme Value (GEV) modeling

We use the stationary Generalized Extreme Value (GEV) distribution as a quantile-thresholding device on bounded, unitless wet/dry score series to obtain high-quantile cutoffs ("Return levels" (RL)) that delineate extreme years (Coles, 2001). For each WU, i, the analysis fits Generalized Extreme Value (GEV) distributions to the annual wet and dry score series, t = 2000–2023, fitted by maximum likelihood estimation (MLE). Fits are obtained with scipy.stats.genextreme.fit (The SciPy community 2008).

For a nominal return period R, the corresponding exceedance probability is

$$p = 1 - \frac{1}{R}. \qquad (5)$$

The return level X_R was obtained by inverting the fitted GEV cumulative distribution function, Eq. 6:

$$X_R = \mathrm{GEV}^{-1}(p), \tag{6}$$

consistent with the percent-point function used in the *SciPy* genextreme implementation. Letting μ, $\sigma > 0$, and ξdenote the GEV location, scale, and shape parameters, respectively, the return level for $\xi \neq 0$is (Coles, 2001, Ch. 3–4)

$$X_R = \mu + \frac{\sigma}{\xi}\,[\,(-\ln\, p)^{-\xi} - 1\,] \tag{6a}$$

with the Gumbel limit

$$X_R = \mu - \sigma \ln\,[-\ln\,(p)], (\xi = 0), \tag{6b}$$

which is the standard inversion of the GEV CDF for return-level estimation (Coles, 2001, Ch. 3).

For each WU i, upper-tail thresholds for wetness and dryness, denoted $\mathrm{RL}_i^{(w)}$and $\mathrm{RL}_i^{(d)}$, were estimated by fitting separate GEV models to the annual Wet Score and Dry Score series. Return levels for period R were computed using Eq. 6 (Coles, 2001).

To quantify uncertainty in the return-level estimates while preserving short-run temporal dependence, we applied a moving block bootstrap (MBB) with overlapping blocks of size $b = 3$and $n_{\mathrm{boot}} = 300$ resamples per WU and per score type (wet, dry), resampling to the original length n. Each bootstrap replicate b resampled blocks to reconstruct a series of length n, yielding, Eq. 7:

$$y^{*(b)} = \{\, y_t^{*(b)} \,\}_{t=1}^{n}. \tag{7}$$

A GEV was refit by maximum likelihood to each resampled series, and a bootstrap return level $\mathrm{RL}_R^{(b)}$was computed via Eq. 6. For each bootstrap replicate, we evaluated the empirical percentile of the original (non-bootstrap) series $y = \{y_t\}_{t=1}^{n}$relative to the bootstrap threshold, Eq. 8:

$$q^{(b)} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{\, y_t \leq \mathrm{RL}_R^{(b)} \,\} \tag{8}$$

which quantifies how often the observed values fall below a threshold perturbed by sampling variability and temporal dependence. The final percentile estimate was summarized as the median across bootstrap replicates, Eq. 9:

$$\hat{q} = \mathrm{median}_{\,b} \left(q^{(b)}\right) \tag{9}$$

providing a robust, dependence-aware measure of threshold exceedance probability and uncertainty.

We report 95%percentile-bootstrap confidence intervals using the $[2.5\%, 97.5\%]$quantiles of $\{q^{(b)}\}$. Sensitivity analyses varying block size $b \in \{3,5,10\}$ and the number of resamples (100–1000) showed that $b = 3$ balances short-run autocorrelation preservation and computational efficiency, and $n_{\mathrm{boot}} = 300$ stabilizes percentile estimates at national scale.

We model annual extremes of the bounded FS score using block-maxima GEV distribution. Because FS ∈ [0,1], a finite right endpoint is plausible; the GEV Weibull domain ($\xi < 0$) provides an appropriate representation of such bounded upper tails (Coles 2001). Tail suitability was evaluated using GEV diagnostic tools, including QQ and PP plots, Anderson–Darling statistics for the upper tail, and interval estimates for the shape parameter ξ. As a robustness check, we repeated fits on (i) a logit-transformed FS series and (ii) a peaks-over-threshold (POT) approach with Generalized Pareto (GPD) tails, using threshold selected from mean residual-life and stability plots. Both alternatives formulations preserved the rank ordering of extreme years and yielded consistent return-level estimates (Katz et al. 2002; Davison and Smith 2018).

We treat the annual score series as stationary over 2000–2023 for tail fitting; RLs are therefore conditional on this assumption. Because FS scores lie in $[0, 1]$, a finite right endpoint is plausible and the Weibull domain ($\xi < 0$) is appropriate in many WUs (Coles 2001); nonetheless, we assess fit with QQ/PP diagnostics and upper-tail statistics, and quantify threshold uncertainty via a MBB to preserve short-run dependence. Stationary GEVs are adopted network-wide for reproducibility and parsimony; targeted non-stationary fits at three focal WUs (allowing $\mu(t)$, and where relevant $\sigma(t)$) to vary, improved information criteria only modestly and did not materially alter the identification of extreme years. We therefore report stationary thresholds in maps and tables and provide non-stationary overlays and AIC/BIC comparisons in the Supplement.

3.4. Selection of Extreme Years

Wet extreme years, Eq. 10.

$$\mathcal{Y}_i^{(w)} = \{\, t : S_{i,t}^{(w)} \geq \mathrm{RL}_i^{(w)} \,\} \tag{10}$$

Dry extreme years, Eq.11:

$$\mathcal{Y}_i^{(d)} = \{\, t : S_{i,t}^{(d)} \geq \mathrm{RL}_i^{(d)} \,\} \tag{11}$$

Wet and dry years are first selected against their respective GEV return levels. For each selected year t, include $t \pm 1$, only if the neighbour-year also satisfies the same threshold criterion, margin = 0.02. This captures short-term hydroclimatic regimes but avoids overexpansion. If a year is simultaneously wet and dry by thresholds, wet is prioritized, removing it from the dry set to avoid contradictory classification.

3.4. Temporal diagnostics

Binary segmentation is applied to each score series to detect mean shifts using a Sum of Squared Errors (SSE) cost function and a penalty term, Eq 12.:

$$\mathrm{y} = \lambda \log\, n, \tag{12}$$

with a minimum segment length of 3 years. Detected change years provide insight into potential structural regime shifts. Mann-Kendall (MK) statistics are computed on the annual percentage of wet/dry conditions to assess long-term susceptibility trends. A secondary diagnostic compares extreme-year selections against three-year centered rolling means of scores to visualize regime coherence.

3.5. Sensitivity analysis (weights)

A constrained random sampling over the weight space $w_{\mathrm{wet}}, w_{\mathrm{wet+}}, w_{\mathrm{dry}}, \alpha, \beta$, was performed, enforcing $w_{\mathrm{wet+}} \geq w_{\mathrm{wet}}$and $\alpha \geq \beta$. For each weight set, and each WU, we recomputed wet and dry scores, fitted return-level thresholds and identified extreme years. We quantified sensitivity by (i) Jaccard overlap of extreme-year sets relative to the baseline weights, (ii) rank-order stability (Spearman ρ and decile stability) and (ii) changes in the number of extreme wet and ry years.

3.6. Validation

The recommended percentiles and bootstrap CIs for each WU (wet/dry), along with GEV parameters, change-points, and MK statistics are recorded. These tables serve as calibration artifacts underpinning the chosen thresholds. Independent of scores, for each selected year we compute the fraction of WU pixels exceeding a fixed FS threshold (consistent with the calibrated wet/dry cutoff) and rank years by exceedance frequency. We confirm that years selected via GEV RL thresholds occupy the top tail of this distribution, indicating that the selections correspond to anomalous susceptibility conditions in the rasters themselves.

3.7. Spatial envelope generation and smoothing

From the final sets of selected wet and dry neighbour-years, we construct spatial envelopes representing the upper and lower extremes of FS values. Each WU is processed independently using its identified wet and dry neighbour-years. For every pixel, we generate upper- and lower-envelope rasters by aggregating FS values from the selected years. To enhance spatial coherence and reduce fine-scale noise—particularly noise not explicitly controlled by the machine-learning model—we apply a 3×3 neighbourhood mean filter to the stacked FS fields prior to envelope extraction. This smoothing step approximates the spatial continuity typical of hydrological processes and yields more interpretable, contiguous susceptibility bounds at the pixel scale. Finally, the upper and lower envelopes are thresholded at the calibrated wet/dry boundary $\theta_{\mathrm{cal}}$and combined to produce an FS envelope that spans the full range of wet and dry extremes.

4. Results

The analysis is completed across Canada using the modelled and calibrated FS maps across Canada for 2000 to 2023. While each of the 1338 NHN WU are processed individually, the results are discussed both individually and at the major drainage basin level, Figure 2.

4.1. Annual FS Maps

From the 24 annual calibrated FS maps, pixels were counted as *dry*, *wet*, and *wet+* for each NHN WU. An analysis of wet pixel counts over the time series reveals a clear national-scale upward trend in FS, Figure 3a. Several years, most prominently 2017 and 2021, show pronounced increases in wet pixel counts, which correspond to years with elevated precipitation and widespread flooding events (Environment and Climate Change Canada 2017). Although individual WU and major drainage basins exhibit heterogeneous temporal patterns the aggregated national profile indicates a consistent increase in wet conditions across the study period. This suggests a gradual intensification of susceptibility, with interannual variability driven by episodic extremes rather than random noise. Decadal aggregation of wet pixel counts by major drainage basins, (see Figure 3b and Supplement, 4.1), further highlights regionally distinct but generally increasing trajectories. The strongest decadal increases occur in Atlantic Canada (WU 01), Northern Quebec and Labrador (WU 03), Southwest Hudson Bay (WU 04), and the Yukon

River basin (WU 09). In contrast, the Nelson River (WU 05) and Great Slave Lake (WU 07) regions peak during the 2010s, with slightly lower but still elevated values in the 2020s. A marked peak in the 2010s, with low value in 2020s is also observed in the Mississippi River basin (WU 11). Because WU 11 represents the smallest geographic unit in Canada, however, its contribution to the national-scale statistics remains comparatively small despite its internal extremes.

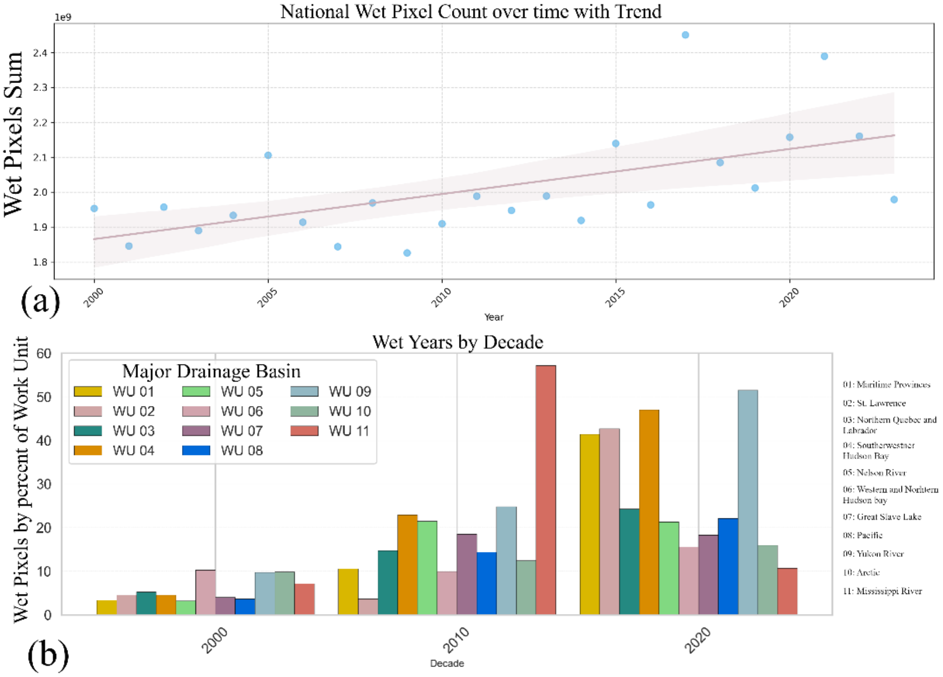

Figure 3 (a) top: National wet pixel count over the time series with fitted trendline. The shaded region represents the 95% confidence interval for the trend estimate, (b) bottom: wet pixel count by decade per major drainage basin.

4.2. Wet and Dry Scores

Figure 4 summarized the relationship between wet and dry trends, and the resulting trend classifications across Canada, by WU. Figure 4a shows the Kendall's τ estimates for wet vs dry scores. The strong negative association indicates that wet and dry τ values are largely discordant: WUs with positive wet-season trends tend to show negative dry-season trends and vice-versa. This pattern reflects diverging hydroclimatic behaviour across years. The marginal distributions of τ, Figure 4b, exhibit pronounced asymmetry. Wet year τ values are strongly right-skewed, with only a small proportion of years yielding negative or non-zero trends. Conversely, dry year τ values are left-skewed, with only a small number of Wus showing positive trends. This asymmetry arises from an upward shift in wet conditions in the most recent decade. High impact wet years, notably 2017 and 2021, expand the upper tail threshold of the fitted GEV model. Such patterns are consistent with a non-stationary hydroclimatic regime characterized by sustained wetting.

False Discovery Rate (FDR) control was applied separately to wet and dry year tests. Using the Benjamini–Hochberg (BH) procedure, wet *p*-values and dry *p*-values were each adjusted independently across all WUs. The resulting *q*-values were used uniformly for (i) assessing significance and (ii) assigning trend classes (per Table 1). No joint pooling across wet and dry p-values was performed, thereby avoiding an inflated effective family size and preserving the physical interpretability of yearly trends. As BH FDR may be mildly liberal under positive spatial dependence, q-value significance is interpreted as a screening criterion rather than strict error control; conclusions were corroborated using q-value maps, see Supplement 4.2.

The spatial distribution of the Net Trend ($\tau_w - \tau_d$.) is shown in Figure 4d. Many WUs, especially in southern Ontario and along the Saint Lawrence River, exhibit a net positive trend. Negative trends are found mainly in northern parts of the Country and mostly coastal WUs. Figure 4e illustrates the trend classes assigned to each WU. Many WUs, 743, are categorized as inconclusive, reflecting weak or non-monotonic seasonal trends. Nonetheless, coherent regional patterns emerge. Mixed-significant, those showing statistically significant but opposing wet and dry trends are common, 395 WUs, and cluster prominently across southern Ontario and southern Quebec, from James Bay to Lake Ontario, the suggesting complex seasonal hydroclimatic dynamics. Additionally, in the west, along the Rocky Mounty range is a cluster of Mixed Significant WUs. Moderate wetting is observed in 156 WUs, distributed more diffusely, but consistent with national-scale wetting tendencies. After BH correction, 0.90% of the WUs show significant net wetting, 0.15 show significant net drying and 29.74% fall into the mixed-significant category. The timing of the detected change points span 2009 – 2020, with a median occurrence year of 2017, aligning with the national rise in wet susceptibility.

Parameter-sensitivity experiments for the weighting coefficients $w_{\text{wet}}, w_{\text{wet+}}, w_{\text{dry}}, \alpha, and\ \beta$ shows that annual wet and dry scores are robust to reasonable weight variation. Across 500 constrained random weight sets, Spearman rank correlations with the baseline scores remained near unity at stations 01AL000, 05OG000 and 09AB000, median ρ ≈ 0.99–1.00; 5th percentile ≥ 0.98). Year-to-year rankings stayed within ±1 decile of the baseline, and the top-3 extreme year sets exhibited high Jaccard similarity (≥ 0.8–1.0), with substitutions occurring only under near-ties. Spatial persistence patterns were consistent with known hydroclimatic behaviour.

https://doi.org/10.5194/egusphere-2026-528
Preprint. Discussion started: 10 March 2026

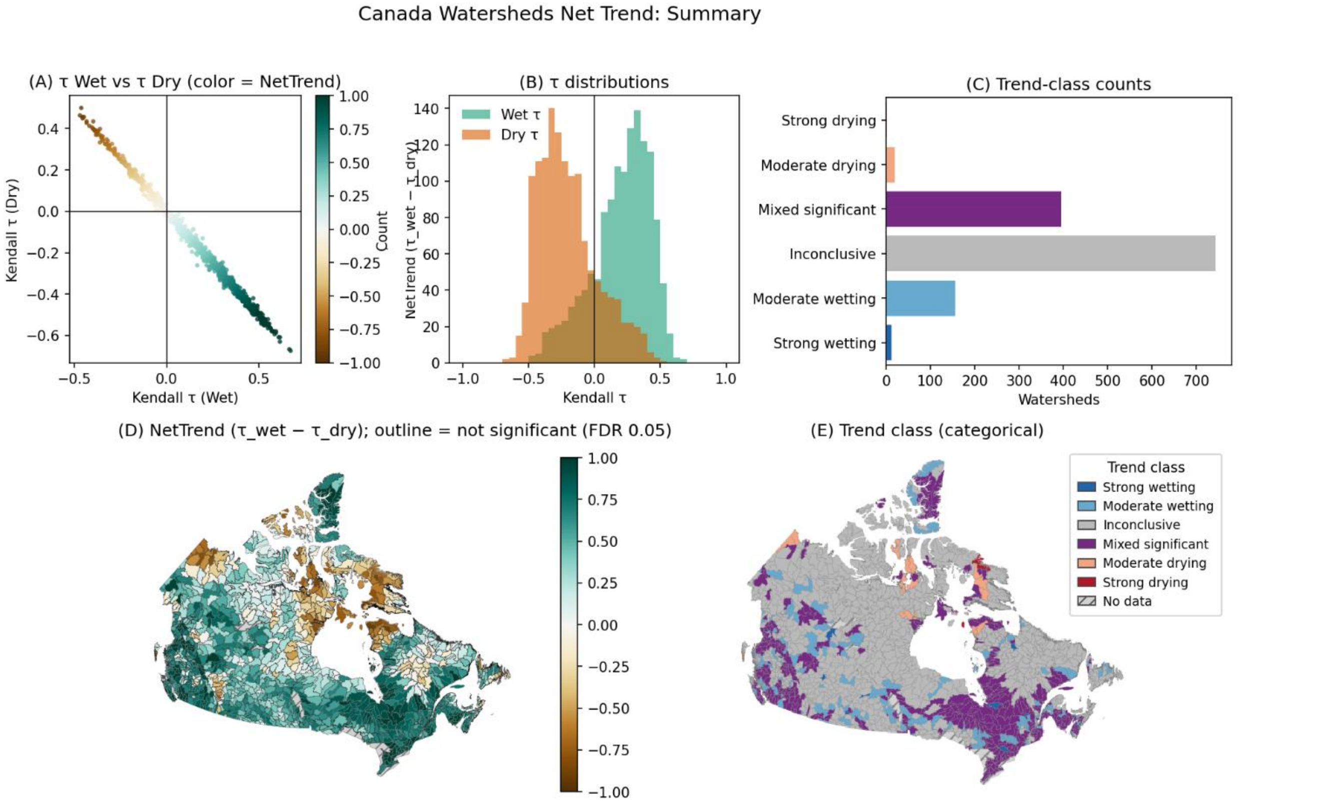

Figure 4 Kendall's τ computed on annual wet and dry scores across Canada, (a) τ Wet vs τ Dry, (b) Distributions of Kendall's τ for wet and dry scores, (c) trend class counts, (d) net trend map, (e) categorized trend class map.

4.3. Identification of Extreme Wet and Dry Years

From the GEV analysis, extreme wet and dry years and their corresponding neighbour-expanded wet and dry years - were identified. Figure 5 summarizes the spatial distribution of these extremes across Canadian WUs using two complementary approaches: bivariate decade maps (a and b) and neighbour-year frequency maps (c and d).

Across all WUs, the average number of extreme wet and dry years is 3 and 2, respectively, with observed ranges of 1–8 years (wet) and 1–10 years (dry). Nationally, the most frequent wet years are 2021 (600 WUs) and 2017 (528 WUs), while 2009 and 2007 represent the most widespread dry-extreme years (513 and 444 WUs, respectively), aligning with earlier national-scale findings, Figure 3. Filling out the rest of a 'top 5' list of wet years, across the entire country is 2020, 2022 and 2015, while dry years are 2006, 2010 and 2001. Of the top 5 wet years, only 2015 can be found in the top 20 list of both wet and dry years, occurring wet in 255 WUs and dry in 107 WUs, see Supplement 4.3.

Figure 5a presents the bivariate decade classification for wet extremes, showing both the decade containing the greatest number of wet-extreme years and the relative position of the second-most-frequent decade. For example, a WU shaded in orange indicates that its dominant wet-extreme decade is the 2020s. The shading intensity reflects alignment with the secondary decade: the lightest shade denotes that both primary and secondary wet-extreme years occur within the 2020s; medium shading indicates an adjacent decade (e.g. 2010s) whereas the darkest shading a more distant decade (e.g. 2000s).

A strong clustering of wet extremes is evident in eastern Canada—including Ontario, southern Quebec, and the Atlantic provinces—where most WUs exhibit the 2020s as both their dominant and secondary wet-extreme decade. Western and northern regions display greater temporal heterogeneity, with some Arctic WUs and isolated areas in Newfoundland and the Pacific region showing dominant wet extremes in the 2000s or 2010s. Across the Prairie region (Alberta and Saskatchewan), a southwest band of WUs indicates dominant wet extremes in the 2010s, with fewer transitioning into the 2020s.

Figure 5b shows the bivariate decade map for dry extremes. Much of the country records its driest years in the 2000s, with secondary decades aligning more frequently with the 2010s than the 2020s. Very few WUs exhibit dominant dry extremes in the 2020s, though small clusters occur in northern Quebec along the Hudson Bay shoreline. Overall, dry-extreme conditions exhibit greater spatial consistency than wet extremes, with the 2000s emerging as the nationally driest decade and the 2020s the wettest.

After applying neighbour-expansion, the average number of extreme wet years remains unchanged, while extreme dry years increase to an average of 4 years. For wet extremes, 65% of WUs show no change under neighbour-expansion, 17% add one adjacent year, and 10% add two. In contrast, dry extremes exhibit greater sensitivity: only 35% of WUs remain unchanged, while 24% add one adjacent dry year, and 19% and 10% add two and three neighbour years, respectively. A visualization of the neighbour-expanded counts are shown in Figure 5c (wet) and 6d (dry), where lighter shades correspond to WUs with the fewest extreme years. Notably, 203 WUs have a single wet-extreme year, and 129 (63.5%) of these remain unchanged after neighbour-expansion; most are found in the Nelson River basin (WU 05) and correspond to the year 2017.

The patterns in this figure suggests increasing temporal clustering of wet extremes—particularly in recent decades—consistent with broader regional climate signals.

https://doi.org/10.5194/egusphere-2026-528
Preprint. Discussion started: 10 March 2026

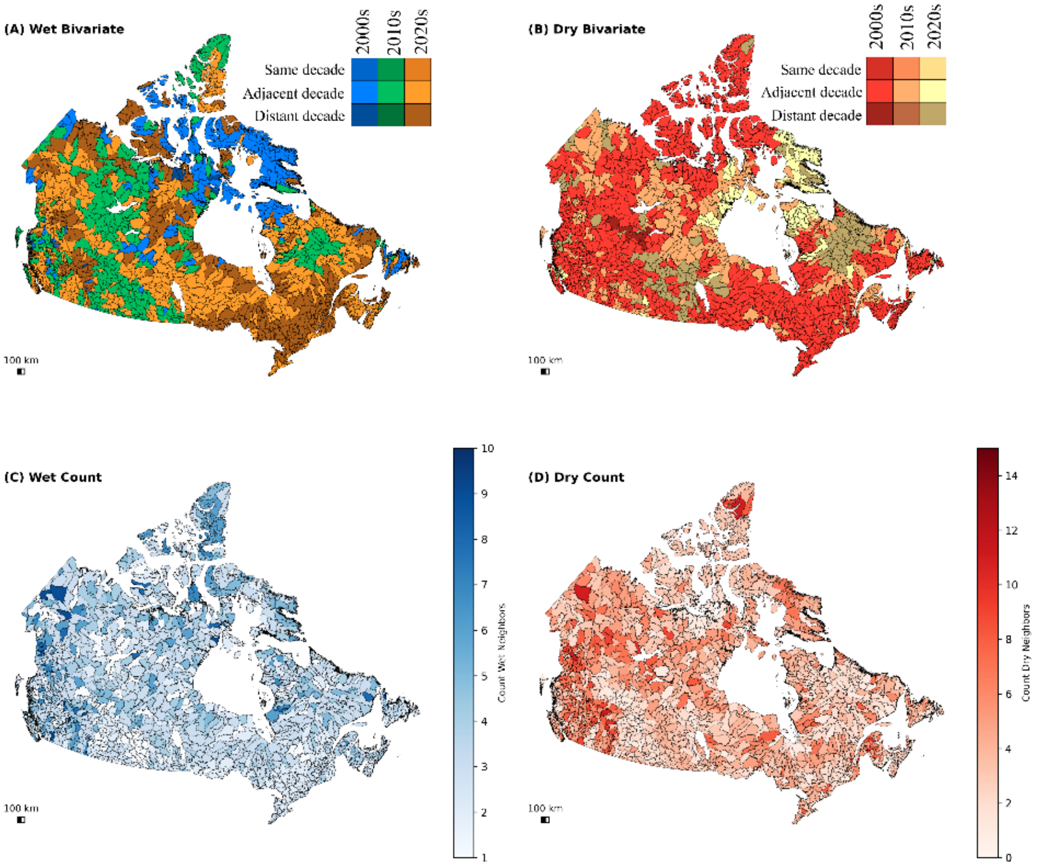

Figure 5 Extreme Hydroclimatic Years Across Canadian WUs, (a) Dominant decade of wet extremes: Colours represent the decade in which each WU experienced the greatest number of wet-extreme years—2000s (blue), 2010s (green), and 2020s (orange). Shading indicates how closely the second-most-frequent decade aligns with the dominant decade: lighter shades denote that wet extremes also occurred in an adjacent decade, whereas darker shades indicate that secondary extremes occurred in a more distant decade. (b) Dominant decade of dry extremes: Colours represent the dominant decade for dry-extreme years—2000s (red), 2010s (orange), and 2020s (yellow). As in (a), shading intensity reflects whether the secondary decade is adjacent to, or distant from, the dominant decade. (c) Total number of wet-extreme years per WU, shown with a sequential blue gradient (darker shades indicate higher frequencies). (d) Total number of dry-extreme years per WU, shown with a sequential red gradient (darker shades indicate higher frequencies).

4.3.1. Neighbour-year expansion sensitivity and justification

To assess the robustness of the neighbour-year expansion procedure, we performed a sensitivity analysis across a full factorial grid of quantile margins $m \in \{0.0,0.01,0.02,0.03\}$and temporal windows $k \in \{1,2,3,5\}$, comparing GEV-defined baseline extreme-year sets with expanded selections for every WU. Across the domain, extreme-year identification proved effectively invariant: for all WUs and nearly all (m, k)combinations, Jaccard similarity, mean run-length, singleton rate, and wet–dry separation metrics were identical to baseline values (Jaccard = 1.0, Δ mean-run = 0, Δ separation = 0), indicating global methodological stability rather than sensitivity confined to specific regions or parameter ranges. Only a small number of isolated cases (e.g., WU 01DS000 under the most permissive expansion, $k = 5, m = 0.03$) exhibited minor deviations in run-length or Jaccard similarity, attributable to local time-series idiosyncrasies such as ties or threshold-adjacent years and showing no geographic clustering. Cross-WU Jaccard analysis reveals physically interpretable spatial structure: wet-year similarity is moderate on average ($J_{\mathrm{wet}} \approx 0.26$), whereas dry-year similarity is lower ($J_{\mathrm{dry}} \approx 0.18$), reflecting more localized drought behaviour. Basin-level aggregation highlights regions of strong hydroclimatic coherence (e.g., Southwestern Hudson Bay with $J_{\mathrm{wet}} = 0.414$, $J_{\mathrm{dry}} = 0.313$, and similar behaviour in the Nelson River), contrasted with eastern basins such as the St. Lawrence, where wet-year agreement remains moderate ($J_{\mathrm{wet}} = 0.320$) but dry-year concordance is very low ($J_{\mathrm{dry}} = 0.117$). The near-zero sensitivity across all parameter combinations demonstrates that extreme-year detection is dominated by the underlying data-driven hydrological variability captured by the FS/XGBoost anomaly structure, rather than by thresholding or neighbourhood choices, supporting the use of fixed expansion parameters and reinforcing that observed inter-basin differences reflect genuine hydroclimatic controls rather than methodological artefacts.

4.4. Threshold Uncertainty and Recommended Quantiles

MBB sizes 3, 5 and 10, were tested to quantify uncertainty in GEV RL thresholds by resampling blocks of consecutive years (to preserve temporal dependence) and recalculate thresholds. The block-bootstrap is designed to retain short-range temporal dependence while estimating sampling variability in threshold quantiles. The distribution of bootstrap confidence-interval (CI) widths (hi–lo) for wet and dry conditions across the three block sizes can be found in 4.4 in the Supplement. The wet condition exhibits a stable median CI width of 0.25 for all block sizes (3, 5, and 10), indicating that increasing the block length does not materially change the central tendency of uncertainty for wet thresholds in our dataset. By contrast, the dry condition shows decreasing median CI width with longer blocks: 0.208 for blocks 3 and 5, and 0.169 for block 10. Thus, for dry thresholds, larger blocks yield smaller typical CI widths, consistent with greater smoothing of short-term variability, see Supplement 4.4. Using a width ≤ 0.150 threshold as "narrow," the share of narrow intervals increases monotonically with block size for both conditions, but more markedly for dry thresholds. This pattern reinforces the interpretation that longer blocks reduce CI width most clearly for the dry case, whereas the wet CI width remains centered around ~0.25 with only modest gains in the narrow fraction. The shorter blocks (3–5 years) preserve more fine-scale dependence and local variability, which yields peaky or irregular CI-width distributions for wet conditions. In contrast, longer blocks (10 years) and the dry condition tend to smooth short-term fluctuations and aggregate variability, producing rounder kernel density curves with fewer small-scale undulations. Practically, this means short blocks in wet analyses capture granular temporal structure at the cost of rougher uncertainty profiles, whereas longer blocks (and dry analyses) favor stability and narrower intervals, especially in the central mass of the distribution.

The datasets median recommended quantiles are $\mathrm{q}_{med,wet} = \mathrm{q}_{med,dry} = 0.917$, indicating both recommended thresholds are anchored at high quantile levels (≈0.92), Table 2. This finding contradicts the common assumption that dry thresholds are selected from much lower quantiles (e.g., 0.2–0.4) and suggests our dry definition relies on upper-tail behaviour in a dry-condition index or on symmetrically defined exceedance measures rather than the lower tail of the same variable. The consistency of *q* across conditions implies a shared decision rule for thresholding extremes within this study. From these results, the block size of 3 was selected, to emphasize short range dependence and site-level heterogeneity.

Table 2 Median bootstrap-recommended RL quantiles and CI widths per WU, summarized across block sizes

	Wet	**Dry**
Median CI width	0.250	0.195
Median recommended quantiles	0.917	0.917

4.5. Sample WU Analysis

The three example WUs (Figure 2) illustrate contrasting hydroclimatic behaviours and trend sensitivities across Canada, see Supplement 4.5 for more details and figures:

- 01AL000 (Nashwaak River near Fredericton, NB): In this WU, the wet score (blue) remains near zero through the 2000s, then rises sharply around 2019 and remains elevated, approaching the wet RL target (dashed blue line). The dry score shows the opposite pattern: a gradual decline through time, moving steadily away from its RL target. Together, these indicate a regime shift toward wetter conditions in recent years. The MK trend tests confirm this behaviour, showing significant increasing (wet) and decreasing (dry) monotonic trends. GEV results are consistent with this pattern—wet tails tend to be heavier (higher q) than dry tails—although uncertainty is large across bootstrap block sizes.
- 05OG000 (La Salle River between Portage la Prairie and Winnipeg, MB): Both wet and dry scores oscillate around zero, with short wet and dry episodes but no sustained exceedance of their RL targets. This indicates strong interannual variability but no directional hydroclimatic shift. The MK tests show no significant trends, providing no evidence of monotonic wetting or drying. Likewise, the GEV exceedance quantiles show wet and dry q medians that are close together, with 95% confidence intervals that strongly overlap across all block sizes—suggesting no robust separation of wet and dry tails.
- 09AB000 (Headwaters Yukon, near Whitehorse, YK): This northern WU shows a steadily increasing wet score across the record, with frequent wet-episode flags, particularly during the 2010s and early 2020s, and values that approach the wet RL target in recent years. This indicates more frequent and intense wet conditions, accompanied by fewer dry episodes. The MK trends highlight a strong wetting signal, with significant increases in wet scores (τ ≈ +0.529, p ≈ 0.00032) and significant decreases in dry scores (τ ≈ –0.547, p ≈ 0.00020). GEV results support this interpretation: wet *q* medians exceed dry *q* medians across all block sizes. Although confidence intervals widen and sometimes

overlap, the directional ordering is consistent, and block-size sensitivity remains qualitatively stable. Overall, this WU shows a robust shift toward wetter extremes.

Collectively, these three examples highlight the method's ability to differentiate stable hydroclimatic regimes, transitional systems undergoing directional change, and high-variability WUs without coherent trends. The bootstrap-based GEV analysis further provides a nuanced quantification of uncertainty, reflecting differences in local variability, persistence, and sample size across sites.

4.6. Validation

4.6.1. Against Pixel Exceedance Frequencies

To verify that GEV-selected extreme years correspond to genuinely anomalous hydrologic conditions on the ground, we computed, the fraction of pixels exceeding calibrated wet/dry raster thresholds, three WUs are discussed.

- 01AL000: Regime transition to wet starting in 2016 the wet area fraction rises, peaking ≳ 0.7–0.8 in 2018–2021 while the dry fraction collapses in those years, Figure 6. The GEV picks capture the spatial shift into a wetter regime: when the point series flags an extreme wet year, a large share of pixels simultaneously exceed the wet threshold and the WU severity median jumps. This is strong cross-validation: point-based extremes are coherent across the raster, not artifacts of single-pixel behaviour.
- 05OG000: both wet and dry area fractions swing widely year to year with many years surpassing the 0.6 threshold line. The GEV matches spatial peaks, dry year (2006), occurs at a dry fraction near 1.0 with a wet fraction trough. The wet years (2012, 2017 and 2018) coincide with wet fraction ≥ 0.9 and FS median spikes ≈ 70 – 80+. The GEV picks do correspond to WU anomalies, but the basin is noise-dominated, large, spatially coherent wet or dry exceedances appear frequently, not just in extreme years, see Supplement 4.6.
- 09AB000: Consistently high wet fraction, most years sit at ~ 0.7 – 0.9. GEV wet years 2014-205 and 2022 land on local maxima of wet fraction (0.9-0.95) and coincide with median peaks ≈ 85 – 90. The FS median trend rises into the late 2010s/early 2020s, echoing the upward wet fraction—consistent with the station's wetting signal, see Supplement 4.6.

Across these three WU, GEV-selected extreme years identified from the point series coincide with large, spatially coherent exceedances in the underlying raster fields. In the figures for 01AL000 and 09AB000, wet 'stars' occur when ≥ 60% of pixels exceed calibrated wet thresholds and the FS median rises well above its wet guide, confirming basin-scale wet anomalies and recent wetting regimes. In contrast, the La Salle (05OG000) exhibits frequent basin-wide exceedances for both wet and dry thresholds, indicating a high-variability system where GEV 'stars' still mark the largest anomalies but separation from typical years is weaker; here, tighter thresholds and/or persistence criteria are recommended.

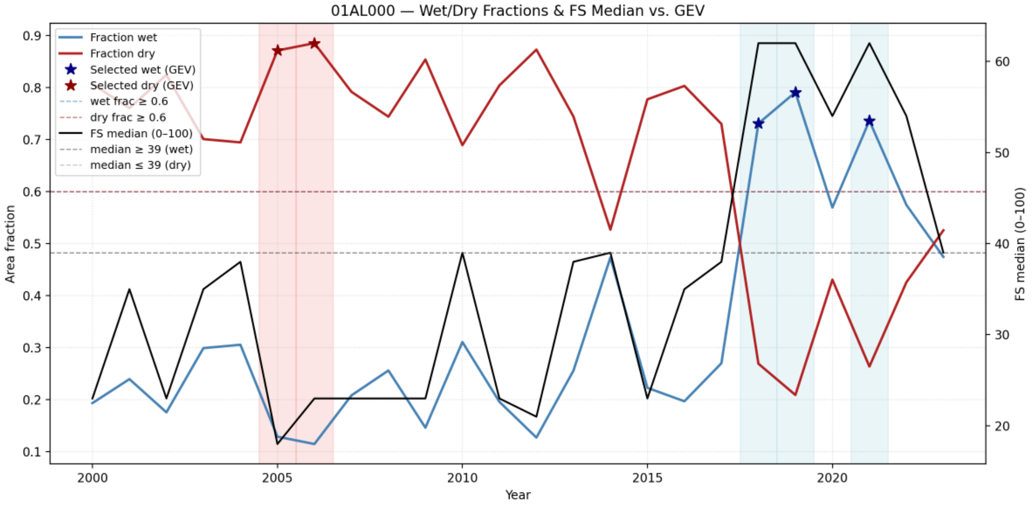

Figure 6 Temporal Patterns of Wet/Dry Area Fractions and FS Median for Tile 01AL000 Compared to GEV selected Extremes, wet years only (wet year neighbours not included), wet data are coloured in blue, dry in red, stars indicated GEV selected extreme years, area fraction on the left y-axis and FS median value, range 0 to 100 are on the secondary y-axis.

4.6.2. Validation & Diagnostics

The goodness-of-fit of the GEV model to annual wet-score series using Quantile–Quantile (QQ) and Probability–Probability (PP) plots were accessed and plotted, see Supplement 4.6.

- 01AL000: These diagnostics flag upper-tail misfit under a stationary GEV (expected under a recent wetting regime), but the raster validation strongly supports that the selected extremes correspond to real, spatially widespread wet anomalies in the Work Unit.
- 05OG000: The GEV looks statistically well-behaved, but the Work Unit shows broad, frequent spatial anomalies. This is a noise-dominated system; extremes from the point record do not cleanly isolate unique, basin-wide anomaly years in the rasters.
- 09AB000: Diagnostics indicate a sound GEV fit and clear spatial validation of wet episodes—consistent with the strong wetting found in the score and MK analyses.

Results from these three WU illustrate in the wetting regime, 01AL000, a non-stationary GEV (time-varying location/scale) would likely improve calibration, in 05OG000 a good stationary fit, appropriate for a site where interannual variability dominates and no monotonic trend is present, and 09AB000, adequate stationary fit with minor curvature; still compatible with a gradual wetting signal. FDR diagnostics confirm that 01AL000 and 09AB000 remain significant after multiple-testing control, while 05OG000 does not—exactly matching the station-level MK results. In 01AL000 & 09AB000, the years flagged as wet by the time-series method, the fraction of WU pixels exceeding RL-based thresholds spikes (and rises over time in 09AB000) shows that GEV-selected extremes are not just point anomalies; they coincide with basin-scale positive anomalies in the underlying rasters. While 05OG000 exhibits frequent, high-amplitude exceedance fractions in many years, diluting the diagnostic power of any single threshold. This is consistent with a high-variability hydrometeorological regime (e.g., convective rainfall or local noise) where extremes are spatially widespread but not trend-bearing. Finally, the recommended *q* values (≈0.917–0.941) generally work well where a directional signal exists.

4.6.3. Hydrometric Data Comparison

Across the three WUs with available hydrometric records (01AL000, 05OG000, 09AB000), correspondence between classified wet/dry years and observed discharge varies by site; Figure 7 shows representative gauge records with the best annual coverage. At 01AL000, wet-classified years (2018–2019, 2021), coincide with elevated spring freshet flows but are not exceptional relative to earlier peaks (e.g., ~2010), indicating above-average seasonal conditions rather than unique hydrologic extremes. At 09AB000, year-by-year alignment between GEV-selected extreme years and gauge maxima is inconsistent, with interpretation further limited by sparse late-record coverage. At 05OG000 (La Salle River), wet-extreme years (2011, 2017) generally align with higher stages across multiple gauges, but are not absolute maxima, and a counterexample occurs in 2009—a dry-classified year with the highest recorded level at station 05OG804—suggesting local decoupling likely related to routing, storage, or operational effects. Overall, hydrometric agreement is heterogeneous, reflecting scale mismatch between spatially aggregated WU-level susceptibility and point-based river gauges, along with timing differences and gauge representativeness; accordingly, the gauge comparison is interpreted as contextual support rather than a strict validation metric.

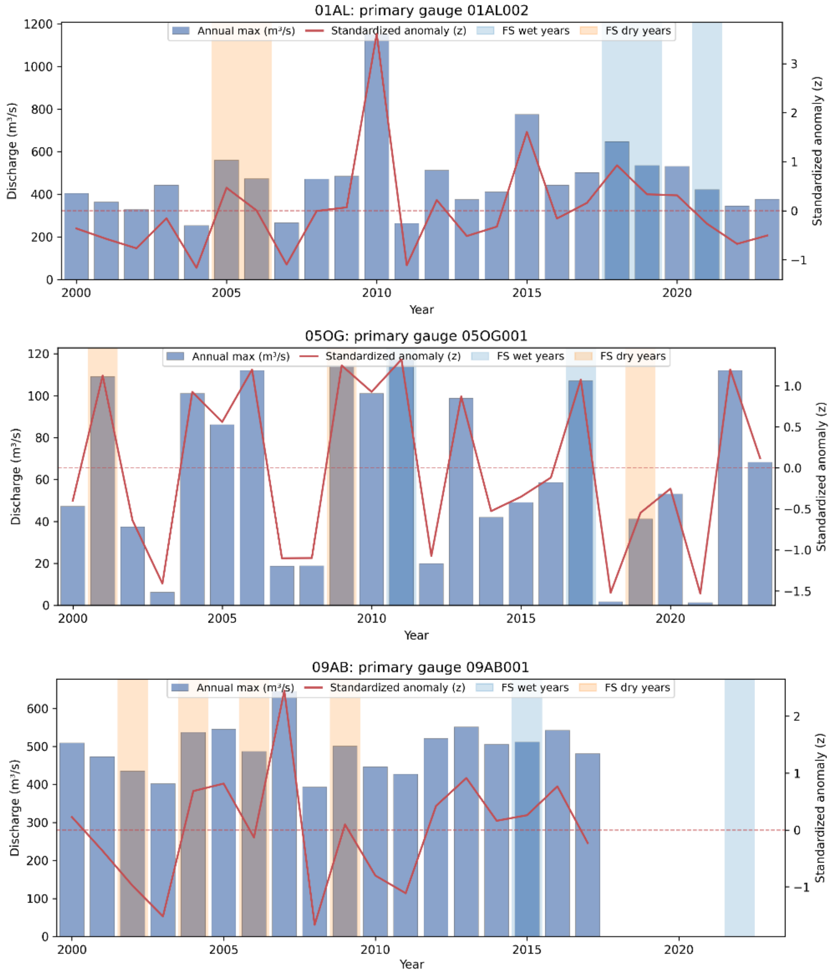

Figure 7 Annual maximum discharge (bars), standardized anomaly (red), and FS wet/dry year shading (blue/orange) for the three validation WUs using the best-coverage gauge per site. Given the WU-scale (spatially aggregated) FS versus point-scale gauges, this provides context, not strict validation

4.6.4. Historic Flood Record

Two national sources were consulted for historical context—the Canadian Disaster Database (CDD) and the Historic Flood Events (HFE) points—but their coarse/ambiguous geocoding and patchy, inconsistently updated records (sparse post-2020 entries) limit reliable matching to specific WUs. Consequently, these datasets provide only supplementary, anecdotal context rather than definitive validation, and detailed site-by-site comparisons are presented in the Supplement 4.6.

Environment and Climate Change Canada tracks and summarizes seasonal and long term precipitation and temperature trends in their Climate Trends and Variation Bulletins (CTVB) (Environment and Climate Change Canada 2017). While they haven't published bulletins for all years explored in this work, we do find relatively good agreement in the GEV extreme years and their bulletins. For example, according to the 2017 Spring CTVB, this season ranked as the third-wettest spring on record, with total precipitation 19.2% above the baseline average. Wetter-than-average conditions were most pronounced across southern British Columbia, central Alberta, central Saskatchewan, southeastern Ontario, southern Quebec, and parts of southern and northern Nunavut. Conversely, southern Saskatchewan, Manitoba, northern Quebec, and western Northwest Territories experienced drier anomalies. This largely lines up with the results of our analysis with WU with GEV determined wet years that include 2017, see Supplement 4.6, Fig4.6.4-1.

4.7. Spatial Envelopes of Flood Susceptibility

We aggregated the per-year susceptibility rasters using the selected extreme neighbour-years to derive spatial envelopes. This produced an upper envelope (wet) and a lower envelope (dry) that bound the range of flood susceptibility through time. To improve spatial coherence consistent with hydrological connectivity, we applied a 3×3 neighbourhood mean filter to each pixel and used the filtered values as the final flood-susceptibility (FS) estimate. The extreme-year analysis therefore outputs three rasters:

- Mean of wet extreme-neighbour years - FS values scaled 0–100
- Mean of dry extreme-neighbour years - FS values scaled 0–100
- Flood envelope (categorical) - three discrete classes indicating non-flooded, lower (dry) boundary, and upper (wet) boundary.

4.8. Limitations

The experiments demonstrate that the proposed workflow reliably identifies temporal patterns of extreme wetness and dryness; however, inference is bounded by the quality, completeness, and representativeness of the input datasets. Conclusions are strongest at the comparative and regional scales (national and basin-level patterns of FS change) and are not intended for direct translation to site-specific hazard, design floods, or regulatory delineation.

FS inputs are produced by a Canada-wide model trained on labeled flood boundaries, although performance is strong overall, the training archive is denser after ~2005 and sparser in the early record and in remote regions (e.g., the Arctic, northern Québec, Labrador). In these settings, the XGBoost model has seen fewer analogous examples, so predictions are more extrapolative and likely carry greater uncertainty. This could manifest as attenuation of early-period FS values and weaker sensitivity to local extremes. The earliest harmonized LULC inputs date to 2000 and were downsampled from 250 m to 30 m to fit with the rest of the data. As a result, FS values in the first years of the time series may be underestimated where pre-2000 land change (e.g., urban expansion, drainage modification, forestry) increased exposure. While LULC ranked among the least-influential features in both partial mutual information and partial correlation analyses (Dunbar et al., 2025), its coarser resolution nonetheless introduces uncertainty for early years. FS is a susceptibility indicator, it is not a probability of flooding, a frequency estimate, or a regulatory floodplain. Pixel-level exceedance offers strong internal validation, but external hydrometric corroboration is limited by gauge density, spatial representativeness, and non-uniform record lengths.

Extreme-year identification relies on GEV-based scoring applied to evolving hydroclimatic conditions. Although trends are treated explicitly in the analysis (e.g., MK tests), departures from stationarity and spatial dependence among WUs can influence significance and uncertainty. We partially mitigate these issues through non-parametric tests and multiple-comparison control at the mapping stage, but residual dependence may still affect local p-values and confidence. We fit stationary GEVs to the annual unitless wet/dry score series to obtain high-quantile cutoffs ("return levels") that serve as thresholds for extreme years. Because stationarity is an approximation, we performed a targeted sensitivity at the three focal WUs, fitting non-stationary GEVs with $\mu(t)$(and, where relevant, $\sigma(t)$) linear in centered time. For wet scores at 01AL000 and 09AB000, non-stationary models improved AIC/BIC and produced gently increasing $x_q(t)$, consistent with the observed wetting; the set of extreme years changed minimally. For dry scores and the wet series at 05OG000, stationary and non-stationary fits were indistinguishable by AIC/BIC and yielded identical extreme years. We therefore retain the stationary GEV for network-wide mapping and reporting and include non-stationary overlays and AIC/BIC tables in the Supplement. (Note: these "return levels" are quantile cutoffs on unitless scores, not hydrologic return periods.)

Spatial unevenness in labeled flood data and environmental covariates may produce region-dependent biases (e.g., conservative FS in data-sparse northern basins). Where training density is low, bootstrap and neighbourhood-year sensitivity tests indicate higher spread, which we interpret as lower confidence, not as absence of trend.

Independent climate fields (e.g., decadal precipitation anomalies, reanalysis moisture/temperature metrics, snowmelt indicators, permafrost state) were explored qualitatively to contextualize patterns; a systematic attribution study is beyond scope. We recommend targeted regional corroboration for operational use, especially in areas with large observed FS increases and sparse labels.

5. Conclusion

This study presents a novel, temporally continuous, and uncertainty-aware framework for FS analysis that advances beyond traditional static or event-based mapping approaches. This work leverages 24 annual FS maps spanning 2000–2023 derived from a machine learning model trained on a multi-decadal record of historic flood events and incorporating time-varying climate data, land use, and vegetation indices. These maps were synthesized into WU-scale scores reflecting wet, wet+, and dry conditions, with calibrated thresholds defining FS states.

We applied Generalized Extreme Value (GEV) distributions to these temporal score series to characterize WU-specific tails of wetness and dryness, quantifying uncertainty through a moving-block-bootstrap approach. The identification of extreme years was refined using neighbour-year expansion to capture short-term hydroclimatic regimes and validated via change-point detection and MK trend analysis. Spatial continuity was ensured by aggregating FS values across extreme years and applying neighbourhood filtering, with bootstrap confidence bounds propagated to the resulting probabilistic envelopes which represent diagnostic summaries of high and low FS years.

Our national-scale analysis reveals a significant increase in FS during the 2020s, with many WUs—particularly in Atlantic Canada and the St. Lawrence River basin—experiencing clusters of extreme wet years from 2017 to 2023. Decadal analysis suggests the 2000s represent a baseline period of FS, the 2010s mark a transition with rising susceptibility, and the 2020s demonstrate the strongest increase in wet extremes and spatial clustering. These findings underscore the evolving nature of flooding under changing hydroclimatic conditions and highlight the value of temporally continuous susceptibility mapping for adaptive FS management.

Despite these advances, several limitations remain. The framework focuses on FS rather than realized flood hazard and thus does not directly capture flood occurrence or severity. The representativeness of the machine learning model outputs depends on the quality and spatial coverage of input data, including climate and land use variables. While non-stationarity in FS is diagnosed through trend and change-point analyses, fully accounting for non-stationary processes in extreme value modelling remains a challenge. External validation, including comparison with river gauge data and observed flood events, is a critical next step to assess model performance and refine susceptibility estimates. Future work will also explore integrating hydrodynamic modelling and socio-economic exposure data to enhance FS assessments.

Overall, this study provides a rigorous probabilistic framework and diagnostic tools for understanding the temporal evolution of FS at WU scales, offering a robust foundation for long-term flood planning and climate adaptation strategies.

Code and data availability

Code available: https://github.com/hmcgrath/FS-extreme-years

Input raster data, to be released, ETA April 2026

Author contributions

HM designed the experiments, and carried them out. HM developed the model code and performed the simulations.

HM prepared the manuscript.

Competing interests

The author declares no competing interests

Acknowledgements

The author would like to acknowledge the support of the Flood Hazard Identification and Mapping Program (FHIMP and Charlotte Crevier for preparing the files in our DataCube for dissemination.

The author acknowledges the use of Microsoft Copilot (M365 Copilot) for limited assistance with code optimization and manuscript review (grammar, phrasing, and clarity).

Financial support

Support funds from Flood Hazard Identification and Mapping Program (FHIMP)

References

Aghenda, Mohamed, Adnane Labbaci, Lhoussaine Bouchaou, Mohammed Hssaisoune, and Yassine Ait Brahim. 2025. "Flood Prediction Using Machine Learning and Deep Learning Models: A Systematic Review." *Mediterranean Geoscience Reviews*, ahead of print, November 14. https://doi.org/10.1007/s42990-025-00201-6.

Al-Ruzouq, Rami, Abdallah Shanableh, Ratiranjan Jena, Mohammed Barakat A. Gibril, Nezar Atalla Hammouri, and Fouad Lamghari. 2024. "Flood Susceptibility Mapping Using a Novel Integration of Multi-Temporal Sentinel-1 Data and eXtreme Deep Learning Model." *Geoscience Frontiers* 15 (3): 101780. https://doi.org/10.1016/j.gsf.2024.101780.

Bentivoglio, Roberto, Sebastiaan Nicolas Jonkman, Elvin Isufi, and Riccardo Taormina. 2025. "Probabilistic Flood Hazard Mapping for Dike-Breach Floods via Graph Neural Networks." *EGUsphere*, December 19, 1–29. https://doi.org/10.5194/egusphere-2025-5582.

Canada, Natural Resources. n.d. "National Hydro Network - NHN - GeoBase Series - Open Government Portal." Accessed August 28, 2025. https://open.canada.ca/data/en/dataset/a4b190fe-e090-4e6d-881e-b87956c07977.

Chandran, Divya V, J Anitha, Divya V Chandran, Amrutha Vijayakumar, and Anjana Raveendran. 2024. "Advancements in Flood Mapping Methodologies: A Comprehensive Review." *2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 1–6. https://doi.org/10.1109/RAICS61201.2024.10690080.

Chihi, Hayet, Mohamed Amine Hammami, and Imen Mezni. 2025. "Flood Susceptibility Mapping in Data-Scarce Arid Environments: Guided by Geology-Driven Knowledge and Multi-Event Cloud-Based Validation." *Natural Hazards* 121 (18): 20855–901. https://doi.org/10.1007/s11069-025-07533-4.

Coles, Stuart. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer. https://doi.org/10.1007/978-1-4471-3675-0.

Davison, A. C., and R. L. Smith. 2018. "Models for Exceedances Over High Thresholds." *Journal of the Royal Statistical Society: Series B (Methodological)* 52 (3): 393–425. https://doi.org/10.1111/j.2517-6161.1990.tb01796.x.

Dunbar, Karen Elaine, Heather McGrath, and Usman T. Khan. 2025. *Enhancing Flood Susceptibility Modelling in Canada: Integrating Seasonal Meteorological Data, Feature Selection and Machine Learning Approaches*. Nos. EGU25-3871. EGU25. Copernicus Meetings. https://doi.org/10.5194/egusphere-egu25-3871.

Environment and Climate Change Canada. 2017. "Climate Trends and Variations Bulletin – Spring 2017." Program results. Climate Trends and Variations Bulletin – Spring 2017, December 6. https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/climate-trends-variability/trends-variations/spring-2017-bulletin.html.

Fivos Sargentis, G., Romanos Ioannidis, Matina Kougkia, et al. 2025. “The Technological Evolution in Flood Risk Estimation.” In *Proceedings of the International Conferences on Digital Technology Driven Engineering 2024*, edited by Nikos D. Lagaros, Rajai Z. Alrousan, Khairedin M. Abdalla, Marios C. Phocas, and Giuseppe Carlo Marano. Springer Nature Switzerland.

Grosso, N., L. Dias, H. P. Costa, F. D. Santos, and P. Garrett. 2015. “Continental Portuguese Territory Flood Social Susceptibility Index.” *Natural Hazards and Earth System Sciences* 15 (8): 1921–31. https://doi.org/10.5194/nhess-15-1921-2015.

Ibebuchi, Chibuike Chiedozie, and Itohan-Osa Abu. 2025. “Probabilistic Flood Susceptibility Mapping Using Explainable AI for the Western United States.” *Environmental Research Communications* 7 (10): 105008. https://doi.org/10.1088/2515-7620/ae0c5c.

Katz, Richard W., Marc B. Parlange, and Philippe Naveau. 2002. “Statistics of Extremes in Hydrology.” *Advances in Water Resources* 25 (8): 1287–304. https://doi.org/10.1016/S0309-1708(02)00056-8.

Khodaei, Hamidreza, Farzin Nasiri Saleh, Afsaneh Nobakht Dalir, and Erfan Zarei. 2025. “Future Flood Susceptibility Mapping under Climate and Land Use Change.” *Scientific Reports* 15 (1): 12394. https://doi.org/10.1038/s41598-025-97008-0.

Kochanek, K., B. Renard, P. Arnaud, et al. 2014. “A Data-Based Comparison of Flood Frequency Analysis Methods Used in France.” *Natural Hazards and Earth System Sciences* 14 (2): 295–308. https://doi.org/10.5194/nhess-14-295-2014.

McGrath, Heather. 2025. “Multi-Event Machine Learning for Annual Flood Susceptibility Prediction at a National Scale.” Paper presented at ISRPS/CSRSS, Toronto, ON. *ISPRS Annals*, November 15. Under Review.

McGrath, Heather, and Victor Alhassan. 2026. “Flood Susceptibility Mapping: Can XGBoost Match CNN Spatial Awareness?” Paper presented at 16th International Conference on Hydroinformatics, Zaragoza, Spain. *16th International Conference on Hydroinformatics*, June.

McGrath, Heather, K.E. Dunbar, and Usman. 2026. “Comparative Analysis of Feature Selection Methods in ML-Based Flood Susceptibility Mapping.” Paper presented at International Conference on Flood Management (ICFM10), London ON, Canada. *ICFM10*, May.

Pourzangbar, Ali, Peter Oberle, Andreas Kron, and Mário J. Franca. 2025. “Analysis of the Utilization of Machine Learning to Map Flood Susceptibility.” *Journal of Flood Risk Management* 18 (2): e70042. https://doi.org/10.1111/jfr3.70042.

S, Sreekala, P. Geetha, and Dhanya Madhu. 2025. “Flood Susceptibility Map of Periyar River Basin Using Geo-Spatial Technology and Machine Learning Approach.” *Remote Sensing in Earth Systems Sciences* 8 (1): 1–21. https://doi.org/10.1007/s41976-024-00101-7.

Schumann, Guy J.-P., Paul D. Bates, Jeffrey C. Neal, and Konstantinos M. Andreadis. 2015. “Chapter 2 - Measuring and Mapping Flood Processes.” In *Hydro-Meteorological Hazards, Risks and Disasters*, edited by John F. Shroder, Paolo Paron, and Giuliano Di Baldassarre. Elsevier. https://doi.org/10.1016/B978-0-12-394846-5.00002-3.

The SciPy community. 2008. “Scipy.Stats.Genextreme — SciPy v1.17.0 Manual.” Scipy.Stats.Genextreme. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.genextreme.html.

Tramblay, Y., G. Thirel, L. Strohmenger, et al. 2025. “Evolution of Flood Generating Processes under Climate Change in France.” *Hydrology and Earth System Sciences* 29 (23): 7023–39. https://doi.org/10.5194/hess-29-7023-2025.

Tsumita, Noriyasu, Suwanno Piyapong, Ratthanaporn Kaewkluengklom, Sittha Jaensirisak, and Atsushi Fukuda. 2025. “Flood Susceptibility Mapping of Urban Flood Risk: Comparing Autoencoder Multilayer Perceptron and Logistic Regression Models in Ubon Ratchathani, Thailand.” *Natural Hazards* 121 (15): 17833–67. https://doi.org/10.1007/s11069-025-07494-8.

Voit, Paul, Felix Fauer, and Maik Heistermann. 2025. “From Worst-Case Scenarios to Extreme Value Statistics: Local Counterfactuals in Flood Frequency Analysis.” *EGUsphere*, November 11, 1–18. https://doi.org/10.5194/egusphere-2025-4951.

Wang, Yongyang, Pan Zhang, Yulei Xie, Lei Chen, and Yu Li. 2025. “Toward Explainable Flood Risk Prediction: Integrating a Novel Hybrid Machine Learning Model.” *Sustainable Cities and Society* 120 (February): 106140. https://doi.org/10.1016/j.scs.2025.106140.

Wyżga, Bartłomiej, Zbigniew W. Kundzewicz, Virginia Ruiz-Villanueva, and Joanna Zawiejska. 2016. “Flood Generation Mechanisms and Changes in Principal Drivers.” In *Flood Risk in the Upper Vistula Basin*, edited by Zbigniew W. Kundzewicz, Markus Stoffel, Tadeusz Niedźwiedź, and Bartłomiej Wyżga. Springer International Publishing. https://doi.org/10.1007/978-3-319-41923-7_4.