

Supplementary Material 4

A standardized rainfall extremity index for equitable flash-flood warning in South Korea

Misook Kwak, Hyunjun Ahn, Sanghoo Yoon

This Supplementary Material collects the robustness and validation analyses supporting the main text. Sections are ordered by their first citation in the main text: the non-stationarity diagnostics of Sec. 4.4 (S4.1-S4.5), the event-level validation of Sec. 4.7 (S4.6-S4.8), and the warning-equity analysis of Sec. 4.8 (S4.9).

S4.1. Sensitivity to the Regularity-Filter Boundary

The adoption rule for non-stationary fits requires a regular shape estimate, ξ in (b, 1) with baseline $b = -1.2$ (main text Sec. 2.4). Re-applying the full adoption rule with $b = -1.1$ and $b = -1.3$ (Table S50) leaves the GPD selection completely unchanged at all three bounds, including the 25/33 sign asymmetry. The GEV selection is unchanged when the bound is relaxed to -1.3 ; tightening it to -1.1 removes five combinations whose shape estimates lie in $[-1.18, -1.10]$ (Gangneung CV 80; Tongyeong CV 40, 70; Yeongdeok CV 40, 60), reducing the GEV count from 18 to 13. Note that the single BH-surviving GEV combination of S4.1 (Yeongdeok, CV 40 mm; shape -1.18) is among the removed cells, underlining that individual-cell evidence for non-stationarity is sensitive to analyst choices, whereas the aggregate GPD-level signal is not. The two stationary-side exclusions at CV = 90 mm (Jeju, ID 184; Namwon, ID 247) carry no stored shape estimate in the parameter files; their potential re-entry under $b = -1.3$ would require a single refit and does not affect the non-stationary adoption counts.

Table S50. Non-stationary adoption counts as a function of the lower regularity bound.

Lower bound b	GPD NS adopted	GPD positive trends	Sign-test p	GEV NS adopted	GEV positive trends
-1.1	33	25	0.005	13	10
-1.2 (baseline)	33	25	0.005	18	12
-1.3	33	25	0.005	18	12

S4.2. False-Discovery-Rate Adjustment of the Non-Stationarity Tests

The Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995) was applied to the likelihood-ratio test (LRT) p -values of the non-stationary GEV and GPD fits (main text Sec. 4.4). The family of tests comprises the LRTs actually performed: 269 station–CV combinations for the GPD model (the cells at which the non-stationary MLE converged) and 132 for the GEV model. For the GPD model no combination survives at $q = 0.05$ or $q = 0.10$ (smallest adjusted $p = 0.124$), although the raw rejection rate ($33/269 = 12.3\%$) exceeds the nominal 5% level—the signature of many weak trends rather than a few strong ones. For the GEV model four combinations survive at $q = 0.05$, but three of these are numerically unstable fits (degenerate or exploding trend SE) that the adoption pipeline had already discarded; the single regular survivor is Yeongdeok (ID 277) at CV = 40 mm (Table S52). The aggregate sign asymmetry of the retained GPD cells (25/33 positive trends; two-sided sign test $p = 0.005$) is unaffected by the adjustment.

Table S51. Summary of the BH adjustment of the non-stationarity LRT p -values.

Model	Tests performed (m)	Raw $p < 0.05$	BH $q = 0.05$ survivors	BH $q = 0.10$ survivors	Smallest adjusted p
GPD (declustered)	269	33	0	0	0.124
GEV	132	21	4 (1 regular)	9 (2 regular)	< 0.001

Table S52. GEV combinations surviving BH at $q = 0.05$. The three numerically unstable fits were not adopted by the model selection procedure (main text Sec. 4.4); the trend SE values shown are the bootstrap standard errors of the non-stationary location trend.

Station	CV (mm)	Raw p	Adjusted p	Trend (h/yr)	Trend SE	Status
Uljin (130)	20	2.0E-27	< 0.001	+0.000	0.000	unstable; not adopted
Tongyeong (162)	30	2.7E-13	< 0.001	-5.217	12.33	unstable; not adopted
Icheon (203)	30	6.5E-10	< 0.001	+1.228	13.92	unstable; not adopted
Yeongdeok (277)	40	6.6E-5	0.002	+0.014	0.016	regular; adopted NS

S4.3. Non-Stationary Parameter Trajectories

Figure S1 shows the 1990–2024 trajectories of the time-varying parameters at four representative CVs (20, 40, 60, 80 mm), referenced in main text Sec. 4.4. The two models detect complementary aspects of change: GEV adoption peaks at intermediate CV (40–60 mm), where the location parameter of the annual extreme shifts, whereas GPD adoption concentrates at the lowest CVs, where the POT log-scale broadens. Both signs coexist within every panel, so the retained combinations do not support a single secular direction; the record exhibits spatially heterogeneous change rather than Korea-wide intensification. Of the 48 (station, CV) combinations in which either model adopts non-stationarity, only three are detected by both models at the same CV (Sokcho, CV 30 mm; Seosan, CV 20 mm; Yeosu, CV 50 mm), all with agreeing positive trends. Per-CV adoption counts are given in Table 4 of the main text (GPD) and Supplementary Table S4 (GEV).

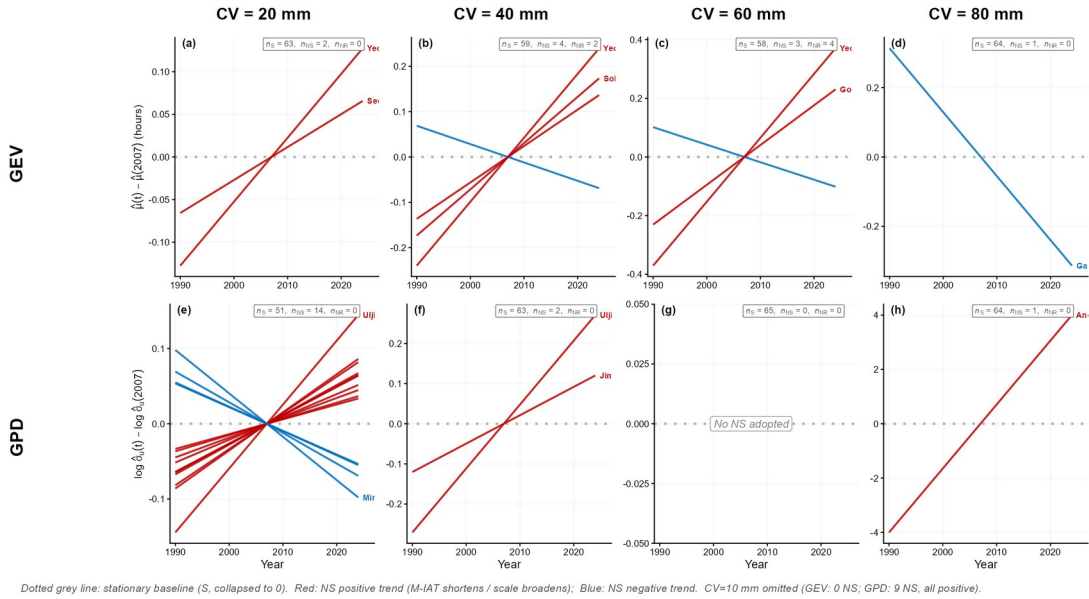


Figure S1. Non-stationary trajectories of the time-varying parameter at CV = 20, 40, 60, and 80 mm. Top row (a-d): GEV model, $\mu(t) - \mu(2007)$ in hours; bottom row (e-h): GPD model, $\log \sigma_u(t) - \log \sigma_u(2007)$. Stationary combinations (S): dotted baseline at zero; non-stationary combinations (NS): solid trajectories, red for positive and blue for negative trends; non-regular combinations (NR) omitted. Per-panel boxes report the counts n_S , n_{NS} , n_{NR} .

S4.4. Exceedance-Frequency Trend

Table S53. Exceedance-frequency trend by CV. Annual counts of rainy-cluster events whose M-IAT falls below the fitted GPD threshold (network fixed at 65 stations, 1990–2024) are modelled by a Poisson trend in year; z and p are overdispersion-adjusted (quasi-Poisson; dispersion $\phi \approx 9$ –19). Trends are positive at every CV but none is significant at the 5% level. Across the 583 station–CV combinations the per-station trend sign is positive in 399 (68%; 69 nominally

significant increases versus 18 decreases), and no individual trend survives the Benjamini–Hochberg FDR at $q = 0.05$.

CV (mm)	Annual exceedances	Trend (%/decade)	z (quasi-Poisson)	p
10	12830	+5.2	1.32	0.185
20	8707	+7.7	1.80	0.072
30	5964	+6.5	1.33	0.185
40	5329	+6.9	1.39	0.164
50	4310	+6.8	1.35	0.176
60	3564	+4.6	0.84	0.399
70	3000	+5.2	0.89	0.373
80	2553	+5.6	0.85	0.395
90	2132	+3.5	0.54	0.588

S4.5. Fifty-Year (1975–2024) Exceedance-Frequency Extension and Homogeneity Screening

To distinguish limited statistical power from a genuine absence of change, the exceedance-frequency analysis of Sec. S4.4 was repeated on the longer 1975–2024 record available at 60 of the 65 stations (the network is complete over this fifty-year window). The counting rule and the overdispersion-adjusted quasi-Poisson trend model are identical to those of Table S53, with the GPD thresholds held fixed at their full-record values. In contrast with the 35-year analysis (0/9 CVs significant), the frequency of fast-accumulation events increases significantly at all nine critical values ($p < 0.02$ throughout), with trends of +7.7 to +10.7% per decade; the 35-year non-detection reported in the main text is therefore attributable to limited power rather than to a stationary occurrence process.

Table S54. Fifty-year (1975–2024) exceedance-frequency trend by CV, 60 stations. Annual counts of rainy-cluster events with M-IAT below the fitted GPD threshold are modelled by a Poisson trend in year; z and p are overdispersion-adjusted (quasi-Poisson dispersion ϕ). All nine CVs are significant at the 5% level.

CV (mm)	Total exceedances	Trend (%/decade)	z (quasi-Poisson)	p	ϕ
10	16014	+7.7	3.3	0.001	16.7
20	10933	+9	3.67	2e-04	12.6
30	7218	+9.5	3.3	0.001	11.2
40	6538	+9.6	3.25	0.0011	10.7
50	5301	+10	3.18	0.0015	9.9
60	4321	+10.4	2.92	0.0035	10.3
70	3662	+10.7	2.82	0.0048	9.7
80	3118	+10.5	2.56	0.0106	9.7
90	2601	+9.9	2.39	0.0169	8.4

Because this extension draws on the pre-1990 segment, which lies outside the primary study period, the extended series were screened for non-climatic inhomogeneity: the Pettitt change-point test (Pettitt, 1979) and the Standard Normal Homogeneity Test (Alexandersson, 1986) were applied to each station's annual-precipitation record, with two-sided p-values controlled across stations by the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). Sixty of the sixty-five gauges provided complete, homogeneous coverage over 1975–2024 and were retained; the five omissions comprise four short records beginning only in the mid-1980s and one gauge with an early-period observation gap. The single station flagged for an abrupt shift—an isolated East Sea island site with a step in the late 1990s—was confirmed to correspond to a regional precipitation regime change rather than to any relocation or instrument replacement. No retained gauge therefore carries an identifiable instrumental discontinuity, and, consistent with

this, the frequency increase is gradual and spread across critical values and stations rather than concentrated at a break date or a single gauge; the 50-year signal is thus unlikely to be a non-climatic artefact.

S4.6. Leave-One-Out Jackknife of the Event-Level Severity Correlations

A leave-one-out jackknife assesses whether the event-level Spearman correlations between event-peak GPD-STRI and recorded severity (main-text Sec. 4.7) are driven by any single event: each of the 27 events is removed in turn and the correlation recomputed. The overall correlations ($\rho = 0.563$ casualties, 0.464 property) change by at most 0.064 and 0.067 respectively, with no sign change. Only Typhoon Soulik—a concordant low-STRI, low-severity case—reduces ρ by more than 0.05 ; the events whose removal instead raises ρ are discordant high-damage/low-STRI cases, so the reported correlations are, if anything, conservative.

Table S55. Leave-one-out jackknife of the event-level GPD-STRI severity correlations. Delta- $\rho = \rho(\text{leave-one-out}) - \rho(\text{all } 27)$; only events with $|\Delta\rho| \geq 0.05$ in either metric are listed (overall $\rho = 0.563$ casualties, 0.464 property).

Event	$\Delta\rho$ (casualties)	$\Delta\rho$ (property)
Typhoon Soulik	−0.053	−0.061
Typhoon Gladys	+0.064	+0.018
Gyeonggi–northern Gangwon heavy rainfall	+0.053	+0.051
Typhoon Bolaven & Tembin	+0.027	+0.067

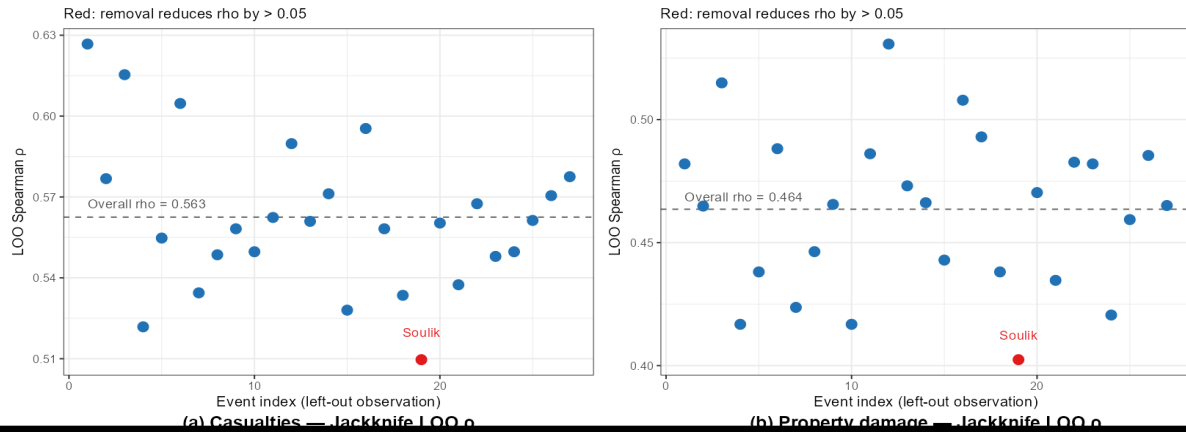


Figure S2. Leave-one-out jackknife of the event-level Spearman correlations for (left) casualties and (right) property damage. Each point is ρ recomputed with one event removed; the dashed line is the overall ρ . Red marks the single event (Typhoon Soulik) whose removal reduces ρ by more than 0.05 .

S4.7. Comparison Against Simple Precipitation Baselines

Three baselines were computed from the same hourly record used for M-IAT: the daily precipitation total (P24), the maximum running 3-hour accumulation (P3h; windows may cross midnight and are attributed to the ending day), and the maximum hourly value (P1h). Unit consistency was verified against the recorded daily precipitation of the damage register (correlation 1.000; e.g. Gangneung, 31 Aug. 2002, Typhoon Rusa: 837.5 mm as computed from the hourly ASOS series; the official KMA daily record for that date is 870.5 mm). For the event-level external validation, each baseline was reduced to an event peak exactly as for STRI: the maximum over the affected station–days of the event (main text Sec. 4.7). A same-station variant (pairing STRI with the rainfall at the single station attaining the event's peak STRI, instead of independent per-baseline maxima) leaves the intensity signal essentially unchanged: the Spearman correlation between event-peak GEV-STRI and P1h is about 0.85 either way, so the strong STRI-hourly-rainfall association is not an artefact of the peak-aggregation choice.

Table S56. Event-level Spearman rank correlation with independently recorded damage severity ($n = 27$ events).

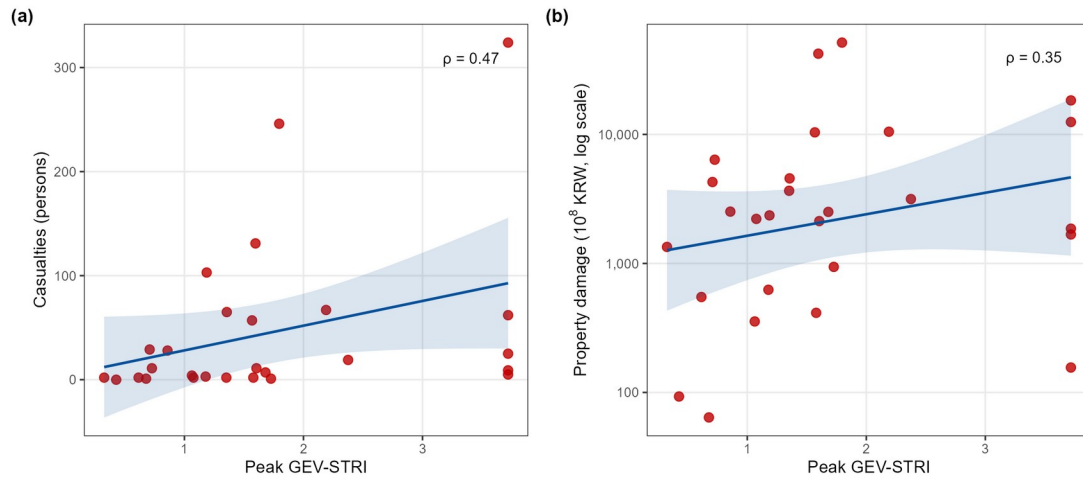


Figure S3. Event-level peak GEV-STRI versus (a) recorded casualties and (b) recorded property damage (log scale) for 27 documented flood events ($n = 27$). Spearman rank correlation ρ is shown in each panel. Compare with the GPD-STRI version in main text Figure 13.

Event peak index	Casualties rho	p	Property damage rho	p
GPD-STRI	+0.56	0.002	+0.46	0.015
GEV-STRI	+0.47	0.014	+0.35	0.073
Daily maximum precipitation (P24)	+0.72	< 0.001	+0.43	0.027
3-h maximum precipitation (P3h)	+0.53	0.005	+0.36	0.066
1-h maximum precipitation (P1h)	+0.45	0.018	+0.39	0.048

None of the pairwise differences between the baselines and GPD-STRI is statistically significant at $n = 27$ (Table S57): the bootstrap confidence intervals for the correlation differences all include zero. STRI is therefore statistically indistinguishable from raw precipitation maxima as an event-level severity correlate; STRI's value rests on the station- and climate-standardised scale and the explicit lead-time semantics of M-IAT (main text Secs. 2.1 and 5).

Table S57. Paired bootstrap (4,000 resamples over events) for the difference in Spearman correlation, $\Delta\rho = \rho(\text{baseline}) - \rho(\text{GPD-STRI})$.

Contrast	Impact measure	Mean delta-rho	95% CI
P24 – GPD-STRI	Casualties	+0.150	[−0.203, +0.493]
P3h – GPD-STRI	Casualties	−0.037	[−0.234, +0.138]
P24 – GPD-STRI	Property damage	−0.037	[−0.413, +0.323]
P3h – GPD-STRI	Property damage	−0.101	[−0.355, +0.136]

A station–day-level discrimination comparison is reported for completeness in Table S58 but is structurally uninformative: the damage-day labels embed a precipitation criterion (82.9% of damage station–days exceed the 110-mm daily representativeness filter by construction), which guarantees near-perfect AUC for the raw-precipitation scores irrespective of hydrological merit. As operational reference points on the 1,069 damage station–days: 53.9% reached the national advisory criterion (60 mm in 3 h), 21.1% reached the warning criterion (90 mm in 3 h), and 82.9% exceeded 110 mm/day.

Table S58. Station–day AUC on the per-CV alarm pools (STRI values from the main analysis; baselines do not vary with CV).

CV (mm)	GEV-STRI	GPD-STRI	P24	P3h
10	0.955	0.926	0.995	0.986
20	0.955	0.887	0.995	0.986
30	0.811	0.865	0.995	0.986

CV (mm)	GEV-STRI	GPD-STRI	P24	P3h
40	0.902	0.887	0.995	0.986
50	0.934	0.868	0.995	0.986
60	0.910	0.856	0.995	0.986
70	0.975	0.839	0.995	0.986
80	0.970	0.828	0.995	0.986
90	0.866	0.801	0.995	0.986

Pairing and sample sensitivity of the rainfall-component analysis (main-text Fig. 12). In the main text, the event-peak GEV-STRI and the event-peak precipitation summaries are maximised independently over each event's affected stations, so the two maxima need not occur at the same gauge; the analysis is restricted to the 27 events with yearbook loss records for consistency with the severity validation above, although no loss data are required. Table S59 repeats the correlation (i) over all 36 documented events and (ii) with the precipitation taken at the station where STRI itself peaks (same-station pairing). The intensity correlation is essentially unchanged across all three designs ($\rho(\text{P1h}) = 0.83\text{--}0.85$), while the volume correlation rises moderately under same-station pairing ($0.40 \rightarrow 0.54$), reflecting that locally intense stations also tend to be climatologically wet. The intensity dominance of STRI is therefore an artefact of neither the station pairing nor the sample.

Table S59. Sensitivity of the rainfall-component correlations (Spearman ρ with event GEV-STRI) to the station pairing and the event sample.

Design	Events	$\rho(\text{P24, daily total})$	$\rho(\text{P1h, hourly max})$
Event-peak, independent maxima (main text, Fig. 12)	27	+0.40	+0.85
Event-peak, independent maxima	36	+0.47	+0.84
Same-station (precipitation at the peak-STRI station)	36	+0.54	+0.83

S4.8. Temporal-Split Cross-Validation

Stationary L-moments fits were re-estimated using only 1990–2010: 21 annual blocks for the GEV model, and for the GPD model the training-period exceedances above the full-record threshold u (held fixed across the split so that the comparison isolates parameter-estimation uncertainty), runs-declustered ($r = 1$) and fitted by L-moments. Implementation checks: the GEV refit on the full record reproduces the published parameters to three decimals (e.g. Daejeon, ID 133), and the GPD protocol reproduces the published shape estimates with correlation 0.95 (median absolute difference 0.018 over 583 cells). STRI was then recomputed for the 17 documented events of 2011–2024 from training-period parameters only. In-sample values in Tables S4.7–S4.8 use the identical stationary machinery on the full record, so the contrast isolates the time split. The out-of-sample index closely tracks its in-sample counterpart (station–day correlation $r = 0.95$ for GEV-STRI and 0.89 for GPD-STRI; event-peak correlation 0.92 and 0.85), and the external validation survives the split (Table S60). Cell-level shape estimates are considerably noisier across the split (r of about 0.65 for GEV and 0.61 for GPD between the 21-year and 35-year estimates), indicating that the out-of-sample stability of STRI derives from the fitted CDF as a whole rather than from any individual parameter.

Table S60. Event-level Spearman correlation for the 2011–2024 events ($n = 17$), in-sample (1990–2024 fits) versus out-of-sample (1990–2010 fits).

Index	Casualties ρ	p	Property damage ρ	p
GPD-STRI, in-sample	+0.69	0.002	+0.34	0.176
GPD-STRI, out-of-sample	+0.64	0.005	+0.37	0.140

Index	Casualties rho	p	Property damage rho	p
GEV-STRI, in-sample	+0.69	0.002	+0.34	0.178
GEV-STRI, out-of-sample	+0.58	0.016	+0.25	0.342

Table S61. Event-peak STRI for the 17 validation-period events (CV-median, maximised over affected station–days). Values of 3.72 correspond to the index saturation cap.

Event	Casualties	Property (1E8 KRW)	GPD in	GPD oos	GEV in	GEV oos
Typhoon Sanba (2012)	2	3,657	1.38	1.63	1.35	1.30
Typhoons Bolaven & Tembin (2012)	11	6,365	0.85	0.84	0.82	0.91
Heavy rain, 22–23 Jul. (2013)	3	625	1.65	1.55	1.39	1.16
Heavy rain, 11–15 & 18 Jul. (2013)	1	940	1.54	1.72	2.54	3.52
Heavy rain, 25 Aug. (2014)	2	1,342	0.81	0.98	0.29	0.41
Typhoon Chaba (2016)	9	1,859	3.72	3.72	3.72	3.72
Heavy rain (2017)	5	156	1.90	1.74	3.72	3.72
Typhoon Kong-rey (2018)	2	549	0.91	1.13	0.61	0.80
Typhoon Soulik (2018)	0	93	0.49	0.44	0.43	0.54
Heavy rain, 26 Aug.–1 Sep. (2018)	2	414	1.42	1.63	1.63	3.72
Typhoon Prapiroon (2018)	1	64	1.09	1.06	0.56	0.97
Typhoon Mitag (2019)	25	1,677	2.20	3.72	3.72	3.72
Typhoon Haishen (2020)	2	2,214	0.99	1.24	0.80	0.98
Heavy rain, Jul. (2020)	57	10,371	3.72	2.22	3.72	3.72
Heavy rain, Jul. (2021)	4	355	1.25	1.15	1.06	1.02
Typhoon Hinnamnor (2022)	11	2,128	2.20	1.84	1.64	1.56
Heavy rain, Aug. (2022)	19	3,155	3.72	3.72	3.72	3.72

S4.9. Warning Equity Across the Station Network

On the 1,069 damage station–days, an absolute daily-precipitation threshold and an STRI threshold were each calibrated to the same overall damage-capture rate (recall = 0.60), and per-station recall was correlated (Spearman ρ_{wet}) with station climatological wetness (95th percentile of daily precipitation); a smaller $|\rho_{\text{wet}}|$ and a smaller cross-station recall standard deviation (SD) indicate more equitable warning across the wet–dry gradient. The absolute-threshold columns do not vary with CV because daily precipitation is CV-independent. The absolute rule is consistently wet-biased ($\rho_{\text{wet}} \approx +0.32$ at every CV), whereas the standardised indices remove this bias: GEV-STRI reduces it toward zero at low CV and reverses it at high CV, and GPD-STRI stays near zero throughout. The representative main-text figure (Fig. 15, Sec. 4.8) uses CV = 30 mm, where all 65 stations are resolved. Analysis script: `equity_standardization_analysis.R`.

Table S62. Warning-equity metrics by CV. For each CV, per-station recall under an absolute daily-precipitation threshold and under an STRI threshold (both calibrated to recall = 0.60 on the 1,069 damage station–days) is related to station wetness. ρ_{wet} is the Spearman correlation between per-station recall and wetness (nearer zero = more equitable); SD is the cross-station standard deviation of recall. GEV-STRI is defined at all 65 stations for every CV; GPD-STRI omits damage days below the POT threshold. Absolute-threshold values are CV-invariant.

CV (mm)	Absolute ρ_{wet}	Absolute SD	GEV-STRI ρ_{wet}	GEV-STRI SD	GPD-STRI ρ_{wet}	GPD-STRI SD
10	+0.33	0.18	+0.14	0.13	+0.05	0.27
20	+0.32	0.18	+0.07	0.13	+0.18	0.34
30	+0.32	0.18	+0.00	0.13	+0.04	0.22
40	+0.33	0.18	−0.30	0.17	−0.06	0.19
50	+0.33	0.18	−0.30	0.15	+0.05	0.18
60	+0.32	0.18	−0.34	0.17	−0.04	0.14
70	+0.32	0.18	−0.37	0.16	+0.13	0.17
80	+0.32	0.18	−0.47	0.17	−0.02	0.18
90	+0.32	0.18	−0.54	0.18	−0.08	0.17

Supplementary References

Alexandersson, H. (1986). A homogeneity test applied to precipitation data. *Journal of Climatology*, 6(6), 661–675.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2), 126–135.