



TailWeather: from tail to extremes, a global climatological dataset for machine-learning weather forecasting

Zhi-Song Liu^{1,2}, Michael Boy^{1,2,3}, and Risto Makkonen⁴

¹Department of Computational Engineering, LUT University, Finland

²Atmospheric Modelling Center, Lahti (AMC-Lahti), Finland

³Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, Finland

⁴Finnish Meteorological Institute (FMI), Finland

Correspondence: Zhi-Song Liu (zhisong.liu@lut.fi) and Michael Boy (michael.boy@helsinki.fi)

Abstract. Weather forecasting models are usually judged by their average skill, yet average skill says little about the rare conditions that matter most. Testing forecasts on those conditions requires a target defined consistently everywhere. Disaster catalogs record harm, but they are sparse and shaped by reporting; climatological tails can be defined everywhere, but rarity does not imply harm. We present TailWeather, a global, land-only, 0.25° dataset of climatological tail events derived from ERA5, covering 1981–2022 and early January 2023. It labels heatwaves, cold waves, heavy precipitation, and extreme wind daily, and meteorological drought monthly. Every label carries an ordinal severity tier and an intensity score relative to the local 1991–2020 climate, so users can move thresholds, within the score’s resolution, without recomputing the climatology. Compared with the EM-DAT disaster catalog, stricter tails are more strongly enriched in recorded impacts for all five hazards, by 1.6 (extreme wind) to 9.3 times (heatwave) at the best-balanced threshold. Yet 80.1 % of the world’s land population lives where the catalog is too sparse to test this link locally for at least one hazard. In forecast examples, leading AI systems that predict 2 m temperature to within about 1 K at day 1 reach a critical success index below 0.5 on heatwave tails, and their tail skill is lowest for extreme wind. TailWeather offers a reusable physical target for studying and evaluating weather extremes, one that complements rather than replaces disaster catalogs. The dataset is available at <https://doi.org/10.57967/hf/10661>.

1 Introduction

Between 1996 and 2023, the EM-DAT disaster database lists more than 250 000 deaths from heatwaves and more than 220 000 from storms (Centre for Research on the Epidemiology of Disasters (CRED), 2024). A forecast helps against such events only if it catches the rare conditions behind them. AI weather models now match or beat operational numerical weather prediction on standard scores, at a fraction of the computational cost (Bi et al., 2023; Lam et al., 2023; Price et al., 2025; Rasp et al., 2024). Those scores are averages, and averages are dominated by ordinary weather. They cannot tell us whether a forecast finds the rare, hazardous conditions at a given place and time. Figure 1 shows the gap: systems that forecast 2 m temperature to within about 1 K at day 1 still reach a critical success index below 0.5 for heatwave tails. Measuring this gap everywhere requires a target defined at every land cell and on every day.



Two kinds of data describe extremes, and neither can stand in for the other. Disaster catalogs such as EM-DAT (Centre for Research on the Epidemiology of Disasters (CRED), 2024) and the NOAA Storm Events Database (NOAA National Centers for Environmental Information, 2024) record what happened to people. What enters them depends on exposure, vulnerability, and reporting practice as much as on the weather, so they are sparse and uneven, and cannot score a forecast everywhere. For heatwaves, EM-DAT is too sparse to support a local comparison over 78.1 % of global land (Sect. 4.2). Other archives and benchmarks answer neighbouring questions: IBTrACS compiles tropical-cyclone tracks and characteristics (Knapp et al., 2010); HR-Extreme provides high-resolution regional extreme-weather cases (Ran et al., 2025); Extreme Weather Bench pairs curated high-impact cases with observations and evaluation tools (McGovern et al., 2026); and ExEBench spans several event categories, tasks, and data sources (Zhao et al., 2025).

Climatological tails answer a different question: is this value rare for this place and season? The idea underlies the indices of the Expert Team on Climate Change Detection, percentile-and-persistence heatwave definitions, the Standardized Precipitation Index, and extreme-value analysis (Zhang et al., 2011; Perkins and Alexander, 2013; McKee et al., 1993; Coles, 2001). A tail can be defined the same way everywhere, but rarity is not harm. A hot day over an empty ice sheet can be statistically extreme and hurt no one; a deadly tornado can be too small for a 0.25° reanalysis to see. We therefore keep the two views apart, and measure how they relate. Throughout the paper, we call the first *recorded impacts* and the second *physical tails*.

This paper presents TailWeather, a global, land-only dataset of climatological tail events covering 1981–2022 and early January 2023. It labels heatwaves, cold waves, heavy precipitation, and extreme wind daily, and meteorological drought monthly. Each label carries a severity tier and an intensity score. Within the score’s stored resolution and valid range, users can choose their own thresholds or build compound tails without recomputing the reference climatology. TailWeather is meant as a dense physical target for weather models; impact records serve as an independent check on its societal relevance, not as its definition.

We make four contributions. First, TailWeather provides one consistent, global target for five hazards. Second, we show that stricter tails are more often associated with recorded impacts, but that the most informative threshold differs by hazard and that impact records cannot assess much of the world. Third, we show that compound tails do not consistently strengthen this association. Fourth, we use TailWeather to expose weaknesses in AI extreme-event prediction that average scores hide, and we test it as a training signal; that final experiment is an illustration, not a proposed training method. TailWeather does not resolve convective hazards, nor does it measure impacts directly. Its strengths lie elsewhere: a long, global record and definitions applied identically everywhere. Section 2 describes the labels, Sect. 3 evaluates forecasts against them, Sects. 4 and 5 compare them with documented disasters, Sect. 6 tests them as a training signal, and Sect. 7 discusses their limits.

2 The TailWeather dataset

All labels are derived from ERA5 (Hersbach et al., 2020), read from the analysis-ready WeatherBench2 archive (Rasp et al., 2024), specifically `1959–2023_01_10-wb13-6h-1440x721_with_derived_variables.zarr`. A single source keeps provenance consistent across hazards; the price is that every label shares ERA5’s variable-specific biases.

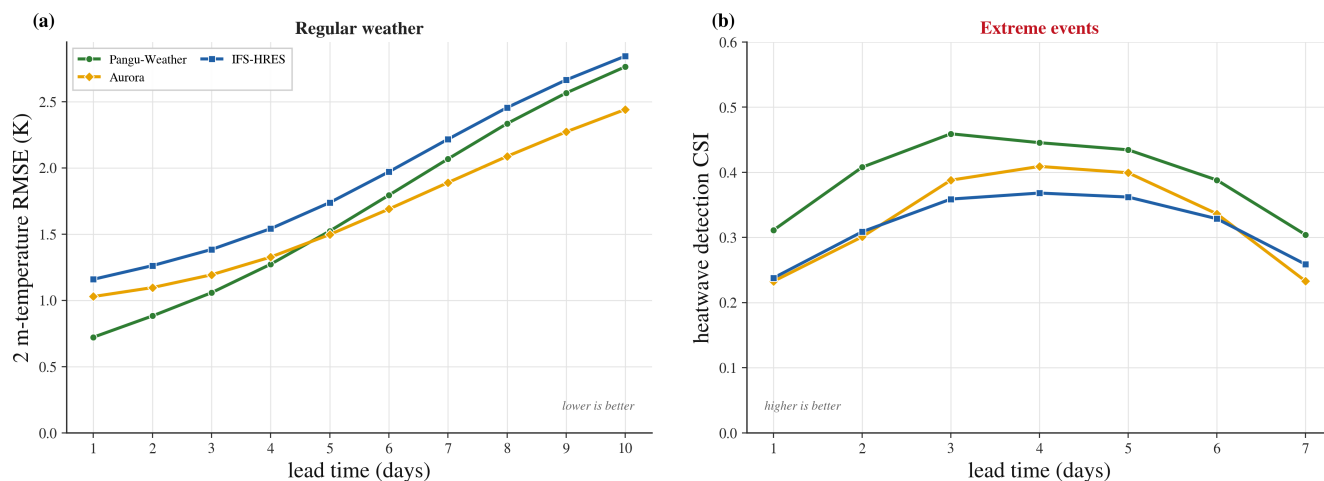


Figure 1. Strong average skill does not imply strong extreme skill. Three leading forecast systems evaluated on WeatherBench2 for 2022. (a) Regular-weather skill: 2 m-temperature root-mean-square error (RMSE) against ERA5 grows smoothly with lead time, from about 1 K at day 1 to 3 K at day 10, close to the state of the art. (b) Extreme-event skill: heatwave-detection critical success index (CSI) against the tail label, for the same systems, peaks below 0.5 near day 3 and then declines. Each panel uses the standard metric of its regime: RMSE for field regression, CSI for event detection.

55 **Grid and period.** Labels lie on a regular 0.25° grid (721×1440). The reference climate is 1991–2020, the current World Meteorological Organization (WMO) standard normal. The record runs from 1981 to early January 2023; the final year is incomplete, and Appendix C lists hazard-specific end dates and processing conventions.

Land mask. Land cells are those with an ERA5 land–sea mask value of at least 0.5: 351 848 cells, ice sheets included. They make up 33.9 % of grid points, which is not 33.9 % of Earth’s surface, because latitude–longitude cells shrink towards the poles. Over the ocean, labels are zero and continuous fields are missing, so we always apply the land and validity masks.

Severity and intensity.

The severity tier $s \in \{0, 1, 2, 3\}$ (none/moderate/severe/extreme) gives the score an operational reading. For the four score-based hazards, s bins I at 0.90/0.95/0.99, but only on cell-days where the hazard’s own persistence rule is already satisfied: a day that scores 0.97 in isolation, without forming part of a qualifying run, is tier 0, not tier 2 – severity marks confirmed events, not merely high-scoring days. Extreme wind departs from this scheme and bins on the raw wind speed instead, using World Meteorological Organization (WMO) Beaufort force boundaries (20.8 m s^{-1} for severe, 28.5 m s^{-1} for extreme), because its hazard scale is already physically anchored; drought bins on SPI depth ($-1.5, -2.0$) rather than on its own $[0, 1]$ score, for the same reason. A given tier therefore marks a fixed position on each hazard’s own scale, not a common percentile across hazards – Table 1 lists which basis applies to which hazard, and Appendix A gives the exact boundaries.

70 **Empirical versus parametric tails.** Temperature, precipitation, and wind thresholds are empirical: for each cell and calendar day, a local reference sample is pooled from the centred ± 15 -day window described above, and both the threshold and the inten-

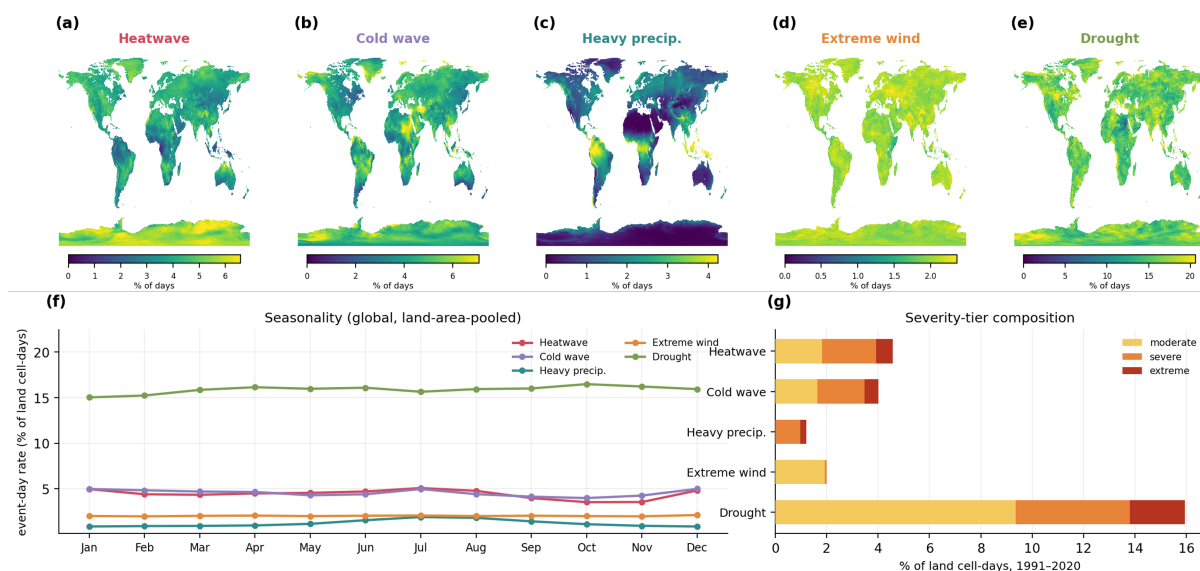


Figure 2. Event rates follow the persistence rules and the local climate, not only the percentile threshold. (a–e) Per-cell event-day rate (% of days flagged, $s \geq 1$) for each hazard, 1981–2022. Unlike the maps of Sect. 4.6, these describe the labels alone, independent of any impact record. Each panel is scaled to its own 99th-percentile land value, because overall prevalence differs by an order of magnitude between hazards (Table 1); colour therefore shows each hazard’s spatial pattern, not a cross-hazard comparison. (f) Event-day rate by calendar month, pooled over land area. The pooled curves mix the two hemispheres’ seasons and are weighted towards the Northern Hemisphere, which holds most of the land. (g) Severity-tier composition for each hazard, 1991–2020 (values in Table 1). Heavy precipitation has no moderate tier by construction (Sect. A3), and extreme wind is almost entirely moderate because its 20.8 m s^{-1} escalation threshold is rarely met.

sity score are read from the resulting empirical CDF by linear interpolation between the nearest order statistics (Appendix C). No distributional shape is assumed, so the climatology is exactly as noisy, and exactly as representative of the true long-run climate, as the pooled reference sample itself. Drought instead uses a parametric tail: monthly precipitation is accumulated over three months, a mixed distribution, a point mass at zero, since many cell-months receive no rain at all, plus a gamma fit to the non-zero accumulations, is fitted separately for each cell and target month, and the cumulative probability under that fit is converted to a standard-normal SPI value via the probability integral transform (Appendix A5). This construction lets SPI extrapolate below the smallest accumulation seen in the reference record, which the empirical hazards cannot do, at the cost of depending on the gamma fit’s adequacy, which degrades where zero-precipitation months dominate. Both approaches inherit sampling uncertainty from a finite 30-year reference, and neither is immune to a cell whose reference sample happens to be unrepresentative of its true climate.

The five hazards in brief. Each hazard applies a percentile-and-persistence rule to one ERA5 variable, judged against its local 1991–2020 climatology. Table 1 lists the variable, threshold, and persistence for each, and Appendix A gives the full definitions. In outline, a heatwave (cold wave) is a run of at least three consecutive days above the local 90th (below the 10th) percentile of daily maximum (minimum) 2 m temperature (Perkins and Alexander, 2013). Heavy precipitation is a wet day



above the local 95th wet-day percentile, with a separate long-wet-spell diagnostic (Zolina et al., 2010). Extreme wind is a day above the local 98th percentile of daily-maximum 10 m wind speed, with Beaufort-anchored severity tiers. Meteorological drought is a 3-month Standardized Precipitation Index at or below -1 , the dataset's only parametric tail (McKee et al., 1993).

Prevalence and data split. Table 1 gives each hazard's prevalence, the fraction of land cell-days with $s \geq 1$. Prevalence

90 follows from three things together, not the percentile alone: the threshold's strictness, the persistence rule, and how strongly the underlying variable persists from day to day. A P90 threshold alone would flag close to 10 % of days almost anywhere; temperature's three-day persistence rule removes short, isolated exceedances and still leaves heatwave and cold-wave rates near 4–4.5 %, because temperature anomalies themselves persist for days, whereas wind's one-day rule leaves its rate close to its nominal 2 %, because individual storms decorrelate faster. Drought's much higher rate (15.9 %) reflects its coarser monthly
 95 resolution and its threshold's lower nominal strictness, $\text{SPI} \leq -1$ is close to the 16th percentile of a standard normal, not a weaker definition. These order-of-magnitude differences in base rate matter for machine learning: a model trained jointly on all five hazards sees drought positives roughly eight times more often than wind positives at the same severity tier, so per-hazard loss weighting or evaluation is advisable rather than pooling raw counts across hazards. Annual comparisons should use complete years only, since a partial year's rate is not comparable to a full one (Appendix B).

100 For machine learning, we recommend a temporal split: 1981–2020 for training, 2021 for validation, and 2022 plus the available January 2023 days for testing. This keeps the entire 1991–2020 reference climatology inside the training period, so no validation or test date ever contributes to the threshold it is itself evaluated against; computing the climatology from all years, validation and test included, would instead leak each of those years' own distribution into the labels it receives. The training split holds about 5.13×10^9 labelled land cell-days; validation and test are each roughly one-fortieth of that, a single
 105 year against forty, enough for stable aggregate metrics but not for rebuilding a cell-level climatology for diagnostic purposes. Figure 2 summarises where, when, and at what severity the labels occur; Appendix B adds their year-by-year evolution.

Spatial and seasonal patterns. Because every threshold is local, a percentile rule alone would give each land cell roughly the same event rate. The departures in Fig. 2a–e therefore reveal what the persistence rules and local climates add. Extreme wind, a one-day exceedance of the 98th percentile, stays close to the 2 % its threshold implies almost everywhere. Heatwaves and
 110 cold waves vary more widely, likely because the three-day rule favours places where temperature anomalies persist. Heavy precipitation flags about 5 % of wet days, so its rate scales with how often it rains: it almost vanishes over the Sahara, the Arabian Peninsula, and the ice sheets. Drought is flagged on 15–20 % of days over most land, but less often in parts of the hyper-arid subtropics. This is consistent with the mixed distribution of Appendix A5: where more than about 16 % of three-month totals are exactly zero, even a zero total has an SPI above -1 , so the drought threshold cannot be reached. Through the
 115 year (Fig. 2f), heatwaves and cold waves vary modestly, and extreme wind and drought hardly at all. Heavy precipitation is the exception, peaking in June–August, when the Northern Hemisphere, which holds most of the land, is in its wet season.

Data records. Labels are stored as NetCDF files, one file per hazard and year (e.g. `heatwave_2020.nc`). Files use the coordinates `time` (daily, UTC calendar day; monthly for the native drought file), `latitude` (0.25° spacing, -90 to 90° , ascending), and `longitude` (0.25° spacing, 0 to 359.75°). Every file carries `binary_label` (the persistence-filtered event
 120 flag), `intensity_score`, and an ordinal severity field, plus one or two variables specific to that hazard; Table 2 lists them.



Table 1. Overview of the five labelled hazards. All labels derive from ERA5 at 0.25° over land, reference period 1991–2020, with the ordinal tier convention {0 none, 1 moderate, 2 severe, 3 extreme}. [†]Heavy precipitation has a separate long-wet-spell diagnostic; in the released files, its severity follows the heavy-day flag. [‡]Drought is computed monthly and optionally broadcast to daily. [§]Empty by construction: the R95p criterion places every flagged day at or above the 95th wet-day percentile, so the moderate bin $[0.90, 0.95)$ cannot be populated (Sect. A3).

Property	Heatwave	Cold wave	Heavy precip.	Extreme wind	Drought
Source variable	TX (2 m max)	TN (2 m min)	total precip.	10 m wind	precip. accum.
Threshold method	90th-percentile climatology	10th-percentile climatology	95th-percentile wet day; spell diagnostic	98th-percentile climatology	SPI-3 gamma
Tail type	empirical	empirical	empirical	empirical	parametric
Min. duration	3 days	3 days	1 day [†]	1 day	monthly
Native temporal res.	daily	daily	daily	daily	monthly [‡]
Severity basis	score bins	score bins	score bins	Beaufort bands	SPI bands
Reference	Perkins and Alexander (2013)	Perkins and Alexander (2013)	Zolina et al. (2010)	WMO Beaufort	McKee et al. (1993)
<i>Prevalence (land cell-days, 1991–2020)</i>					
Event-day rate (%)	4.566	4.008	1.209	1.996	15.934
% moderate ($s=1$)	1.815	1.643	0.000 [§]	1.931	9.359
% severe ($s=2$)	2.104	1.833	0.964	0.062	4.448
% extreme ($s=3$)	0.648	0.532	0.246	0.003	2.127

Heatwave and cold-wave files add a running streak-length field (`hw_duration/cw_duration`); heavy-precipitation files separate the plain heavy-day flag (`heavy_precip`) from the long-wet-spell diagnostic (`heavy_long_spell`) and also store the raw daily total (`precip_mm`); wind files store the raw daily-maximum wind speed (`wind_speed`); and drought files add the underlying `spi3` field, since `intensity_score` alone cannot be inverted back to SPI outside drought. Each hazard except drought also has a companion `<hazard>_threshold.nc` file giving the per-(day-of-year, cell) threshold actually used to build its labels (e.g. `wind_threshold.nc` stores the P_{98} wind speed in ms^{-1}), so a user can inspect or re-derive the climatology without rerunning the labelling code. For continuous fields, NaN marks both ocean cells and, for the four non-drought hazards, cell-days where the field is not defined (e.g. `intensity_score` off-event); binary fields are 0 over the ocean rather than missing. Appendix C gives the complete per-variable dictionary, including data types and the handling of the drought daily-broadcast variant.

3 Example application: TailWeather reveals an extreme-skill gap

Our first application uses TailWeather as an evaluation target. Consistent with WeatherBench2 (Rasp et al., 2024), three leading systems forecast 2 m temperature to within about 1 K at day 1 (Fig. 1a). On the tail, however, the picture changes: their critical success index (CSI) for heatwave tails stays below 0.5 even at its day-3 peak (Fig. 1b). Grid-point error, dominated by ordinary weather, cannot reveal this difference.



Table 2. Core variables in each hazard’s NetCDF files, all variables share the time, latitude, longitude grid unless noted. [†]The monthly, native drought file; a second file, `drought_daily_<year>.nc`, repeats each month’s values on every day of that month for convenience (Sect. A5).

Hazard	File(s)	Hazard-specific variables
Heatwave / cold wave	<code>heatwave_<year>.nc</code> / <code>coldwave_<year>.nc</code>	<code>hw_duration</code> / <code>cw_duration</code> (streak length, days)
	<code>heatwave_threshold.nc</code> / <code>coldwave_threshold.nc</code>	T_{90}/T_{10} (°C), by day-of-year
Heavy precipitation	<code>precip_<year>.nc</code>	<code>wet_day</code> , <code>heavy_precip</code> , <code>heavy_long_spell</code> , <code>precip_mm</code> (raw daily total), <code>wet_spell_len</code>
	<code>precip_p95_wetday.nc</code> / <code>precip_p75_spell.nc</code>	R_{95} (mm) / spell-length threshold (days)
Extreme wind	<code>wind_<year>.nc</code>	<code>wind_speed</code> (raw daily maximum, ms^{-1})
	<code>wind_threshold.nc</code>	W_{98} (ms^{-1}), by day-of-year
Drought (SPI-3)	<code>drought_<year>.nc</code> [†]	<code>spi3</code> (dimensionless)

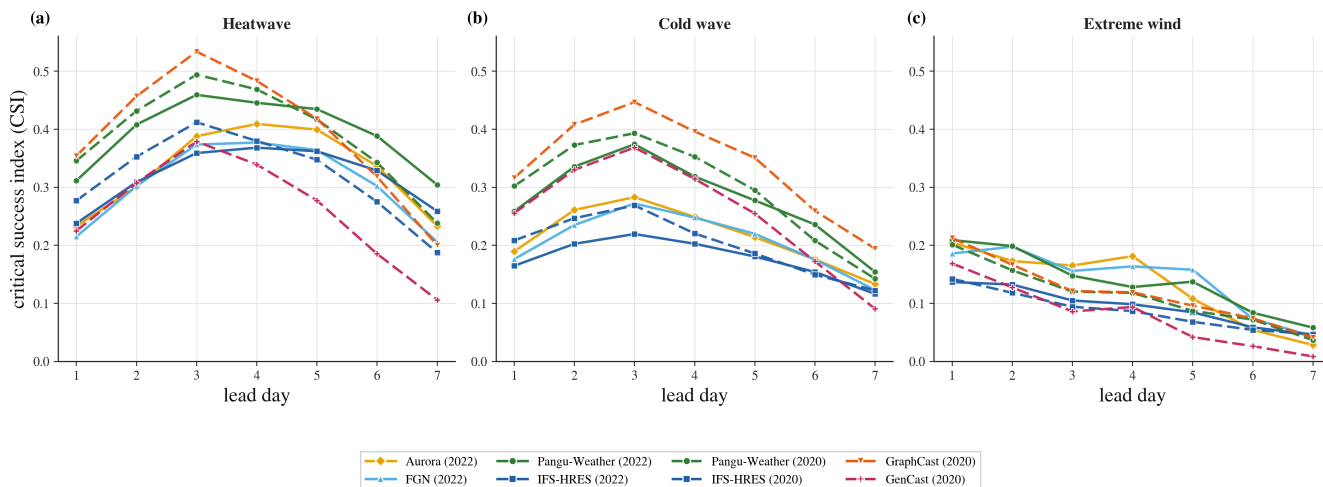


Figure 3. Tail-event skill is highest for temperature and lowest for extreme wind. Critical success index (CSI; higher is better) of six forecast systems, scored against the tail labels for the three day-resolved hazards. Solid lines show the 2022 evaluation (Aurora, FGN, Pangu-Weather, IFS-HRES) and dashed lines the 2020 evaluation (GraphCast, GenCast, Pangu-Weather, IFS-HRES); Pangu-Weather and IFS-HRES appear in both as cross-year anchors. CSI peaks near day 3 for the temperature hazards, and is lowest and decays fastest for extreme wind.

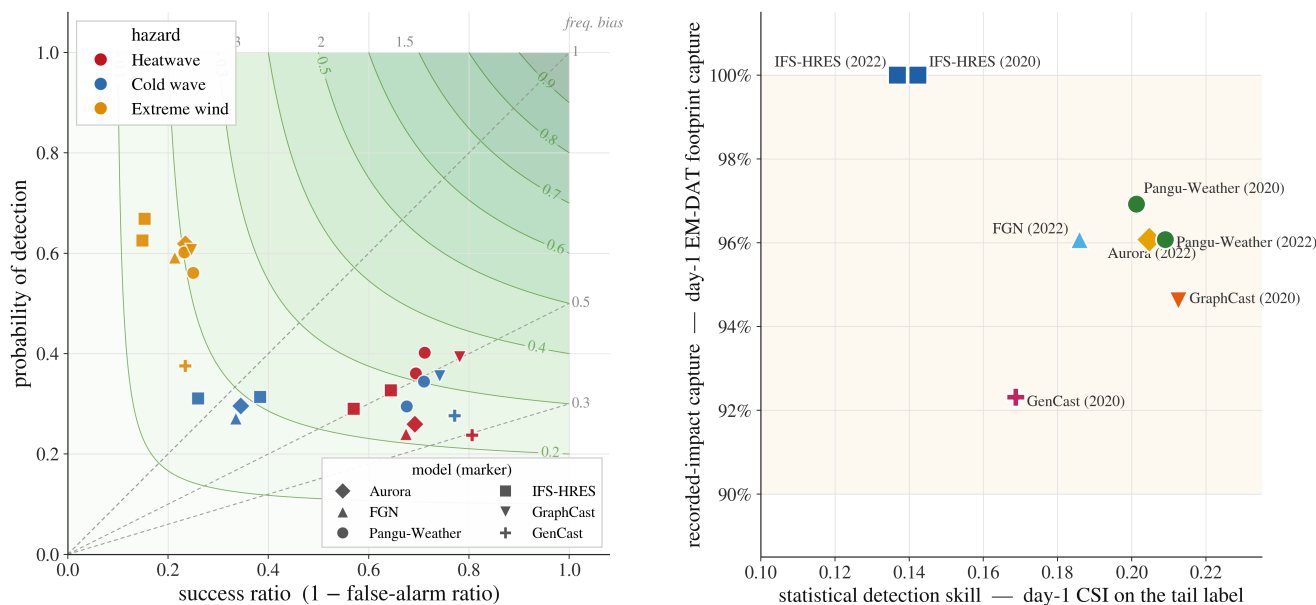


Figure 4. Models fail in opposite directions for temperature and wind, and tail skill and disaster capture rank them differently. (a) Day-1 detection performance against the tail labels, shown as probability of detection and success ratio (green CSI isolines; grey frequency-bias rays). Models tend to under-predict temperature tails and over-predict extreme-wind tails. (b) Tail-label detection skill and recorded-impact capture do not rank the six systems in the same order for extreme wind, because they measure different targets.

We evaluate six published systems, namely Pangu-Weather, GraphCast, GenCast, Aurora, FGN, and the high-resolution Integrated Forecasting System (IFS-HRES), against the three day-resolved temperature and wind labels (Bi et al., 2023; Lam et al., 2023; Price et al., 2025; Bodnar et al., 2025; Alet et al., 2025; European Centre for Medium-Range Weather Forecasts, 2026). The available forecast archives span two test years, 2020 and 2022; Pangu-Weather and IFS-HRES appear in both and anchor the comparison. We omit heavy precipitation, because the archives lack a comparable precipitation forecast, and drought, because it is monthly. Every forecast is scored against the same ERA5-derived labels with the CSI, the number of hits divided by the sum of hits, misses, and false alarms, which runs from 0 (no skill) to 1 (perfect). Because CSI counts false alarms on non-event cell-days, a model cannot score well by forecasting extremes everywhere; evaluating only on days when extremes occurred would reward exactly that behaviour. Scoring against ERA5 may slightly disadvantage IFS-HRES at short lead times because it starts from ECMWF’s operational analysis rather than from ERA5 (Rasp et al., 2024).

Tail-event skill depends strongly on the hazard (Fig. 3). It is highest for the temperature hazards and lowest for extreme wind, for every model at every lead time. The performance diagram shows how the models fail (Fig. 4a): they miss temperature tails but over-predict wind tails, two opposite failure modes that a single score blurs together. Comparing wind-tail skill with the capture of recorded wind-disaster footprints (Fig. 4b) shows that the two do not rank the models in the same order. Detecting a statistical tail and capturing a disaster are different, and TailWeather makes the first one answerable at every land cell.



4 Relationship between tail events and documented disasters

TailWeather labels describe unusual weather, not the damage it causes. Their definitions are applied consistently, but nothing guarantees that they line up with documented disasters. Here we measure how far they do. This is a plausibility check of societal relevance, not a validation of ERA5's meteorological accuracy. We ask three questions in turn. Where is the impact record dense enough for a comparison (Sect. 4.2)? Do the labels find recorded disasters (Sect. 4.3)? And does a stricter tail mean a stronger link (Sects. 4.4–4.6)?

4.1 Design

We compare the labels with EM-DAT (Centre for Research on the Epidemiology of Disasters (CRED), 2024), an independently compiled global disaster catalog, over 1996–2023. Each record is rasterised onto the TailWeather grid, using sub-national boundaries from the Geocoded Disasters dataset (GDIS) where available and national boundaries otherwise (Rosvold and Buhaug, 2021). A multi-day record covers its full reported period, widened by ± 2 days to absorb the offset between local reporting dates and UTC days. Heat, cold, wind, and drought records are matched to the like-named labels. Flood records are matched to heavy precipitation, a looser pairing, because floods also depend on upstream rainfall, soil moisture, and river routing. The comparison rests on 219 heatwave, 288 cold-wave, 2495 storm, 3966 flood, and 402 drought records, so sample sizes differ by more than an order of magnitude between hazards. Most records carry sub-national geometry (73–93 % for four hazards), but only 52 % of heatwave records do: much of the heat record is the 2022 European heatwave, entered as one row per affected country and falling after GDIS coverage ends in 2018. Appendix D gives the matching rules and their limitations.

For each threshold q , we report two complementary quantities. Catalog precision is the fraction of tail cell-days that overlap a disaster record; catalog recall is the fraction of recorded-impact cell-days that are tails. An unmatched cell-day is not proof that no impact occurred, only that none was recorded. We estimate uncertainty with a 15-day block bootstrap, which respects short-term temporal dependence. Appendix D gives the full definitions and the limits of this procedure.

4.2 Documented disaster records cover only part of the world

Before asking how well tails and records agree, we ask where the question can be answered at all. For each hazard, we set aside land cells with fewer than five recorded impact days over 1996–2023. This is a simple screen, since five days may come from a single disaster, and we apply it again to the per-cell maps in Sect. 4.6. Figure 5 shows what the screen leaves out for heat, together with a few telling events.

Much of the world fails the screen. For heat (78.1 %), cold (77.0 %), and drought (66.6 %), most land has too few records for a local comparison; for wind, the share falls just short of half (45.8 %). Heavy precipitation, with the densest reporting of the five, still leaves 35.0 % of land unassessable (Fig. 6), and 34.4 % of land fails the screen for all five hazards at once. These gaps do not mark quieter weather. For heat, cold, and wind, the excluded areas have tail-event rates equal to or higher than those of assessable areas (ratios 1.03–1.30), and drought is close to parity (0.89). Only heavy precipitation differs markedly: its excluded areas really do have fewer tail events (ratio 0.34). The gap also reaches the world's land population, 80.1 % live where

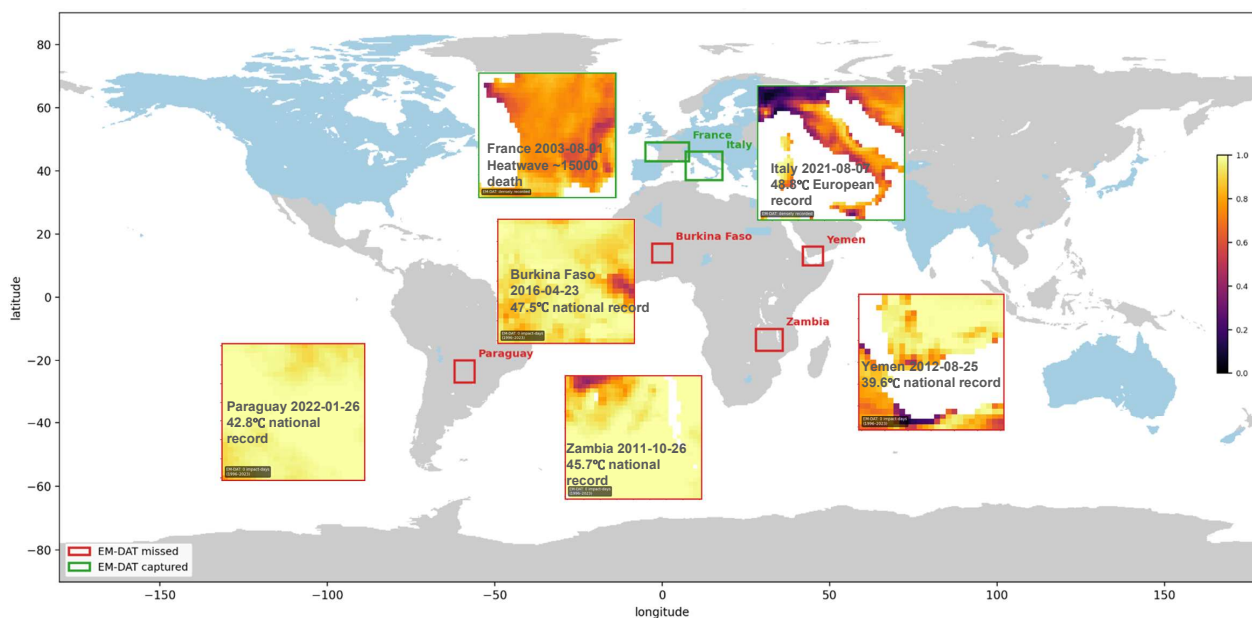


Figure 5. The heat-impact catalog can assess only part of the land surface. Grey cells have fewer than five recorded impact days in the 1996–2023 comparison period; blue cells pass this coverage screen. Red boxes mark four station or national heat records (Paraguay, Zambia, Burkina Faso, and Yemen) with no matching EM-DAT entry in the comparison; green boxes mark the 2003 France and 2021 Italy heatwaves.

the catalog cannot support a local estimate for at least one hazard (Lloyd et al., 2017). A benchmark built only on reported disasters would leave most people unassessed for at least one hazard. TailWeather is computed identically wherever ERA5 has land data, which is why it complements impact records.

4.3 Tail labels overlap documented disasters

Do the labels find documented disasters? We begin with six well-documented European heatwaves of the ERA5 era. All six appear in the heatwave label at the maximum intensity score, with flagged durations that match their reported lengths (Fig. 7a).

We then test every EM-DAT record, counting a disaster as captured if the persistence-filtered label fires anywhere inside its reported area and dates. Capture ranges from 57 % for heat to about 94 % for heavy precipitation and drought, with cold and wind in between (Appendix G). These raw rates are not detection skill: a large or long-lived footprint simply offers more chances to overlap a tail by coincidence. An illustrative chance calculation in Appendix G shows how strongly area, duration, and tail prevalence shape the result. Heat has the lowest capture. Records dated only to the month are captured less often than day-precise records (Fig. 7b), and wider matching windows raise capture. The heat result is therefore sensitive to how the catalog dates its events, although dating alone need not explain every miss.

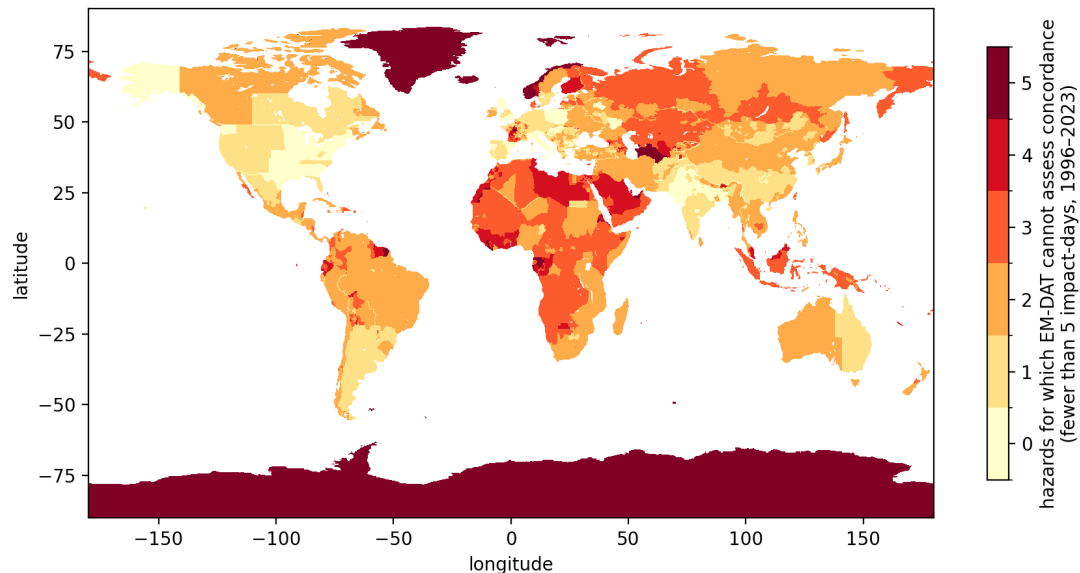


Figure 6. Where documented disaster records cannot evaluate a tail-event dataset. Per land cell, the number of hazards (0–5) for which EM-DAT records fewer than five impact days over 1996–2023 and so cannot support a local estimate (Sect. 4.2). Large contiguous regions, concentrated in Africa, South America, and interior Asia, are uninformative for three or more hazards at once; only 4.9 % of land is informative for all five.

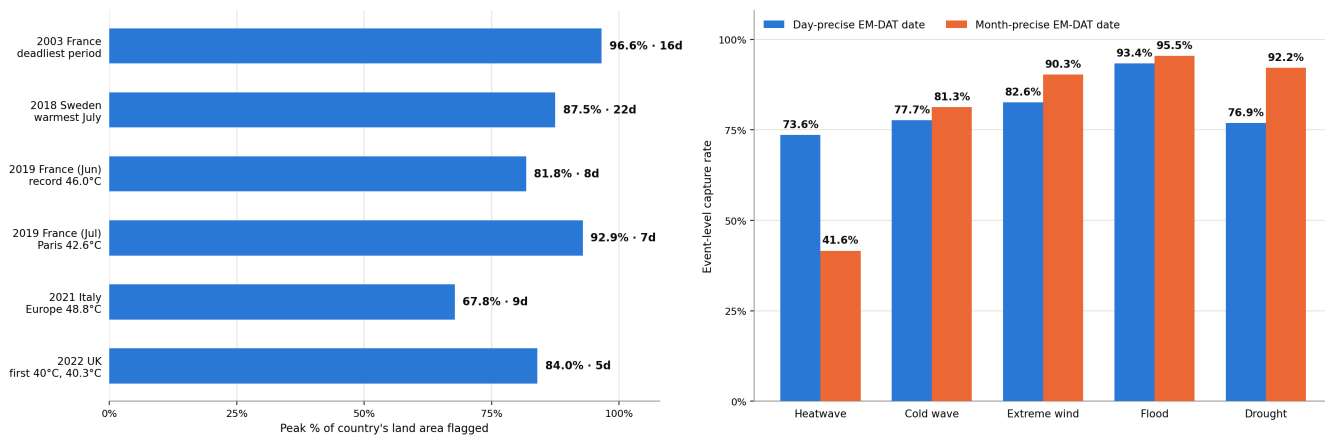


Figure 7. The heatwave label recovers six documented European heatwaves, and heat capture depends on catalog date precision. (a) Six European heatwaves compared with TailWeather using externally reported dates and approximate national bounding boxes. Bars show the peak flagged fraction of land area and the flagged duration; each event reaches the label's maximum intensity score of 1.0. (b) Event-level capture stratified by EM-DAT date precision. Differences between groups show sensitivity to catalog construction; they do not attribute every unmatched event to dating errors.



Table 3. Catalog-overlap summary, 1996–2023, global land. (a) Cell-level recall $P(T_q | X)$ (%) at each hazard’s F1-optimal threshold q^* and at the common reference threshold $q = 0.90$ (P90); the last column gives their ratio. (b) Association between tail labels and recorded impacts. Precision, Recall, and Lift are evaluated at q^* . Tail(%) and Impact(%) are base rates at $q = 0.90$: the fraction of land cell-days flagged as tail and the fraction with a recorded impact. On this shared scale, $q = 0.90$ coincides with the conventional label threshold only for heatwave and cold wave. For extreme wind (P98), heavy precipitation (P95), and drought (scaled SPI, where $q = 0.90$ is a much stricter cut), Tail(%) should not be compared with the event-day rates of Table 1. Even for heatwave the two differ (4.92 versus 4.566 %), because they cover different periods (1996–2023 versus 1991–2020).

(a) Recall at q^* and P90				(b) Tail–impact association						
Hazard	q^*	P90	$\times$	Hazard	q^*	Tail (%)	Impact (%)	Precision (%)	Recall (%)	Lift ($\times$)
Heatwave	1.82	15.07	8.3	Heatwave	0.990	4.92	0.112	1.04	1.82	9.3
Cold wave	2.54	9.43	3.7	Cold wave	0.975	3.67	0.190	0.79	2.54	4.2
Extreme wind	4.73	13.01	2.8	Extreme wind	0.975	10.04	0.749	1.17	4.73	1.6
Heavy precip.	23.77	7.33	0.3	Heavy precip.	0.600	2.43	1.688	4.18	23.77	2.5
Drought	2.49	1.30	0.5	Drought	0.800	0.38	3.546	9.99	2.49	2.8

4.4 Stricter tails are more often linked to recorded impacts

The event-level test asks whether a disaster overlaps any tail at all. We now sweep the threshold and compare individual cell-days, which asks something different: as the physical definition tightens, how does the balance shift between finding recorded impacts and flagging cells without one? We summarise each sweep by q^* , the threshold that maximises F_1 , the harmonic mean of precision and recall. The value q^* summarises this comparison; it does not replace each hazard’s label definition.

For heat, cold, and wind, q^* is far stricter than the label threshold. At the common reference threshold $q = 0.90$, recall is 15 % for heat, 9.4 % for cold, and 13 % for wind (Table 3); the F_1 optimum gives up recall in exchange for precision. For heavy precipitation and drought, the optimum lies below 0.90. For drought, q is a scaled SPI score rather than a percentile, and the released drought definition remains that of Appendix A5. One threshold cannot serve every hazard. Precision must also be read against very different base rates. At $q = 0.90$, heatwave tails cover 4.92 % of land cell-days, whereas recorded heat impacts cover only 0.112 %. Even a useful label therefore has low raw precision. We report instead the enrichment, or lift, $P(X | T_q)/P(X)$: how many times more likely a recorded impact is on a tail cell-day than on an average cell-day. A lift of 1 means no association; a lift of 9 means a recorded impact is nine times more likely on a tail cell-day. Base rates do not tell the whole story, however. At $q = 0.99$, heat tails and heat-impact cells are similar in number, yet precision is still only 1.04 %.

Tails and recorded impacts are related, but they are not interchangeable. As the threshold tightens, recall falls and precision rises (Fig. 8). More importantly, enrichment rises for all five hazards: the rarer the weather, the more likely it coincides with a recorded impact. The intensity score thus carries information about impacts, even though it is not itself a measure of harm.



Tail-vs-extreme concordance: precision and recall vs. threshold, all 5 hazards

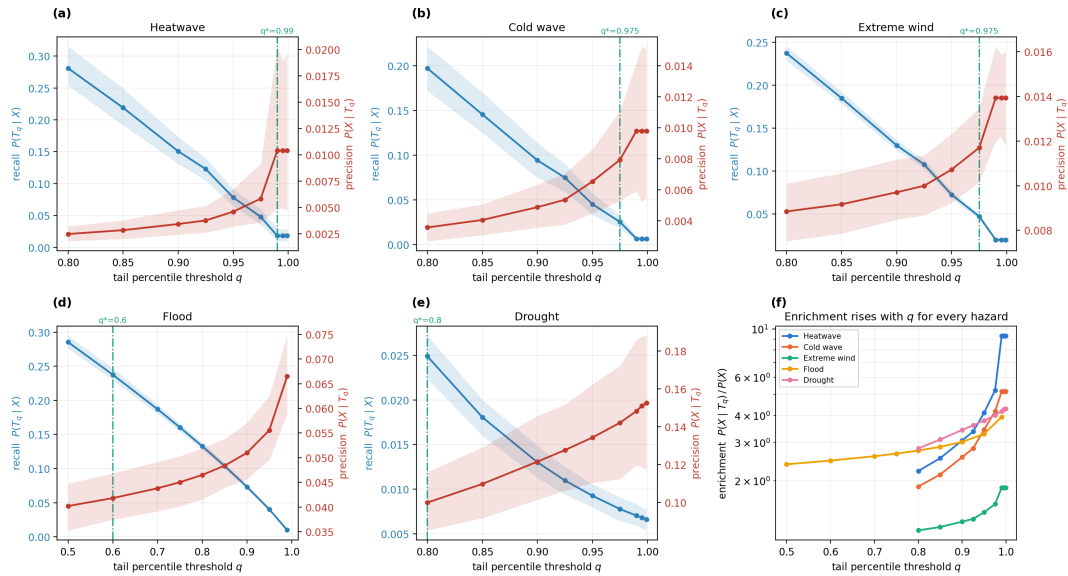


Figure 8. Enrichment rises with tail strictness for every hazard. Panels (a–e): recall $P(T_q | X)$ (blue, left axis) and precision $P(X | T_q)$ (red, right axis) as functions of the tail threshold q for each hazard. The two conditionals are drawn on independent axes; the dashed line marks each hazard’s F_1 -optimal q^* . Recall falls and precision rises monotonically in every panel. Panel (f) overlays the resulting enrichment $P(X | T_q)/P(X)$ for all five hazards on a common logarithmic axis: it increases monotonically with q in every case, showing that the intensity score carries impact information (Sect. 4.4).

4.5 One threshold does not fit all hazards

The F_1 -optimal threshold runs from 0.99 for heat to 0.60 for heavy precipitation (Fig. 9). Optimal threshold for drought (0.80) lies on its own scaled-SPI axis and cannot be compared directly. These values summarise one catalog comparison. They are not universal hazard thresholds. Users should choose a threshold for their own target, weighing missed events against unmatched detections and bearing in mind the differences in score scaling and precision described in Appendix F.

4.6 The relationship differs by hazard and place

The association is weakest for extreme wind, with a lift of only 1.6 at q^* (Table 3). Six-hourly ERA5 fields on a 0.25° grid do not resolve short-lived gusts, and this mismatch of scales plausibly contributes to the weak link. The wind label is a useful physical target, but an imperfect proxy for wind damage. Figures 10 maps the local association wherever at least five impact days exist. Where it can be estimated, it is mostly positive. Yet most of Africa and South America stay blank: the record there is too thin. Reporting follows population, institutional capacity, and media attention, and a disaster-only benchmark inherits that uneven geography. A tail dataset defined against the local climate evaluates the physical event in the same way in well-reported and poorly reported regions alike.

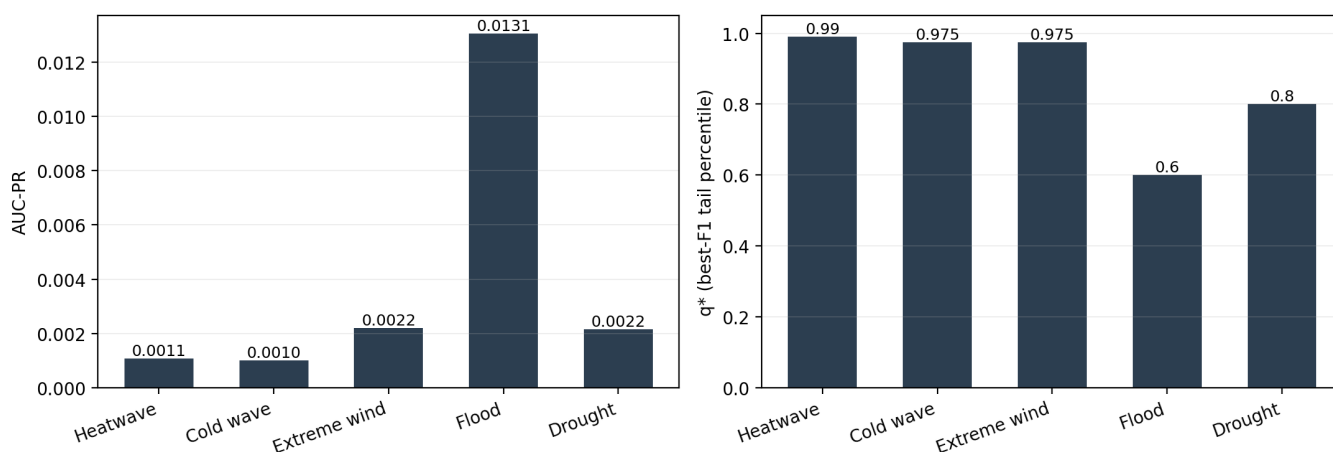


Figure 9. No universal “right” percentile. (a) Partial area under the precision–recall curve per hazard. This quantity is not comparable across hazards, since it tracks each hazard’s impact base rate more than its predictability, and it is integrated only over the recall range the sweep spans (Appendix F). (b) F1-optimal score threshold q^* for each hazard. The dotted line marks $q = 0.90$; drought uses a scaled SPI score and should not be read on a common percentile scale.

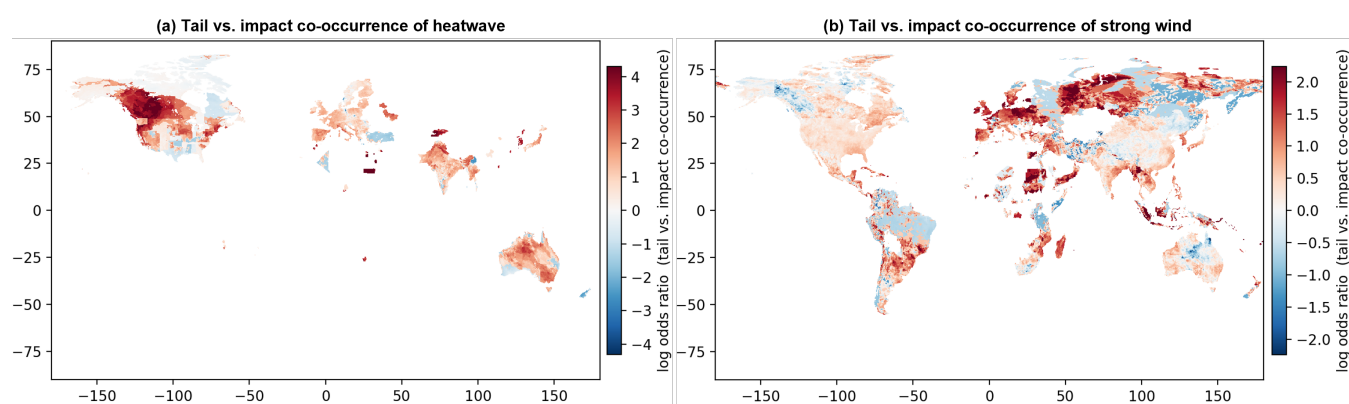


Figure 10. Co-currence of tail vs. recorded events. (a) Heat tails and recorded heat impacts co-occur wherever the record allows a test. Per-cell log odds ratio between heatwave tail labels and recorded impacts at $q = 0.90$. Cells with fewer than five impact days are masked, and the diverging scale is clipped symmetrically at ± 4 so that a small number of saturated coastal cells (a rasterisation edge effect) do not dominate the colour range. The association is positive nearly everywhere it can be estimated; the blank areas over Africa and South America reflect absent recorded impacts rather than absent tail events (Sect. 4.6). (b) The wind tail–impact association is weaker and patchier than for heat. The pattern is consistent with the weaker pooled relationship between the wind-tail proxy and reported impacts (Sect. 4.6)

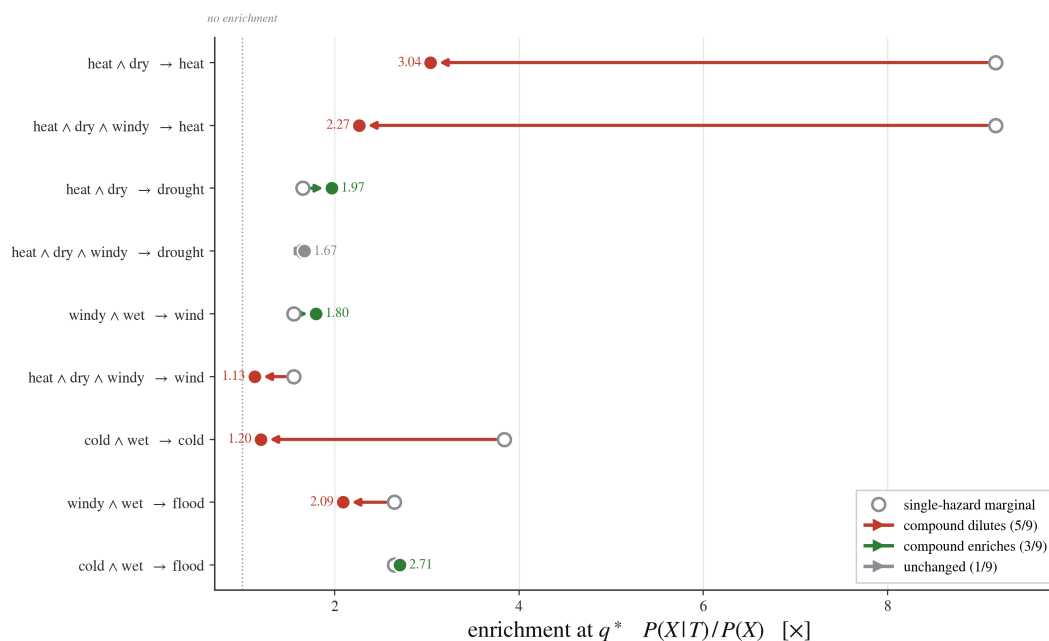


Figure 11. Compounding does not systematically sharpen the tail–impact correspondence. Each row is one compound label compared against the EM-DAT impact record of one of its constituents (written *compound* $\rightarrow$ *constituent*). The hollow marker is the constituent single-hazard enrichment at its own q^* ; the filled marker is the compound’s enrichment at its own q^* ; the arrow runs from the former to the latter. Red arrows (five of nine) show compounding *diluting* the correspondence, green (three) *strengthening* it, grey (one) leaving it unchanged.

5 Compound tail events

Since every label shares one grid, compound tails are easy to build. Compound events, such as heat coinciding with drought, can be more damaging than either hazard alone (Zscheischler et al., 2018), so one might expect compound tails to track recorded impacts more closely than single hazards do. We test this expectation directly. We take the compound intensity as the lower of its constituent scores, so a hot–dry cell-day must be both hot and dry at the chosen threshold. We test four compounds, hot–dry, hot–dry–windy, windy–wet, and cold–wet, against the recorded impacts of each constituent.

The expectation does not hold in general: compounding is not automatically more informative (Fig. 11). It weakens the association in five comparisons, strengthens it in three, and leaves it unchanged in one. Adding drought and wind to heat, for example, lowers the association with recorded heat impacts, consistent with some heat disasters lacking all three conditions. A hot–dry label, by contrast, tracks recorded drought impacts slightly better than drought alone. Each comparison uses the label’s own fitted threshold, so these results do not isolate the effect of compounding alone.

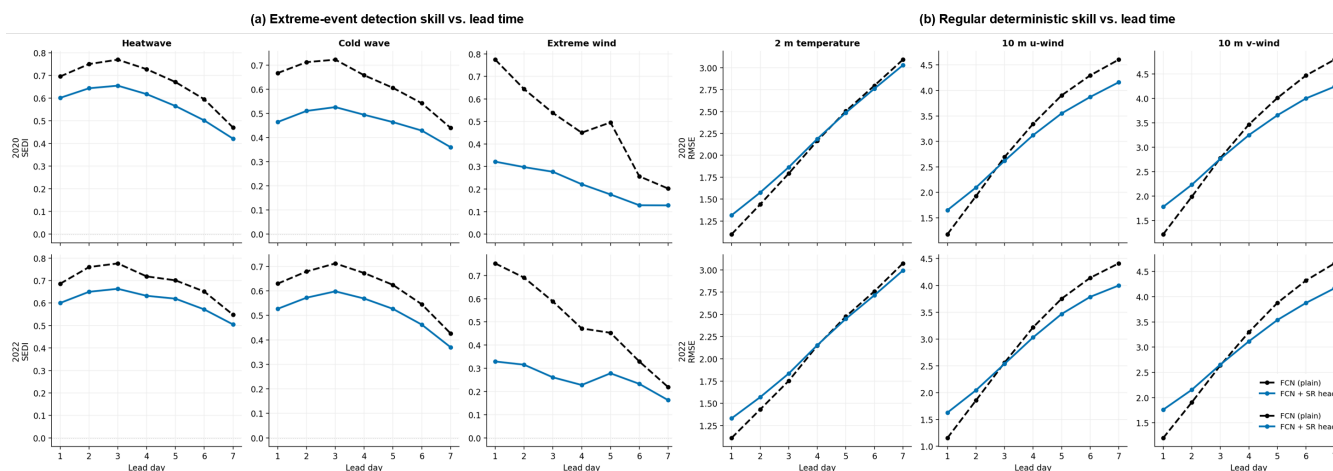


Figure 12. A tail-aware correction improves regular skill but degrades extreme-event detection. Frozen FourCastNet (FCN) forecasts with (solid) and without (dashed) a trainable sharpening head applied as post-processing. (a) Extreme-event detection skill (SEDI; higher is better) falls at every lead day. (b) Regular deterministic skill (latitude-weighted RMSE) tells the opposite story at longer lead times: the correction lowers 10 m wind RMSE by 30–55 % by day 7, and 2 m temperature RMSE more modestly.

6 Example application: AI model training with TailWeather

Our second application asks whether TailWeather can also be used for AI optimization. In a deliberately limited experiment, we attach a small trainable sharpening head to frozen FourCastNet (FCN) forecasts. The head is trained to weight cells and days
 240 by their TailWeather intensity score, and it corrects only the final forecast; the FCN rollout itself is unchanged (Appendix E).

The result cuts both ways. The correction lowers 10 m wind RMSE by 30–55 % at day 7, so the signal can improve an ordinary pointwise forecast. Yet extreme-event detection gets worse at every lead time, for every hazard, and in both test years. We measure detection with the Symmetric Extremal Dependence Index (SEDI), a score designed for rare binary events that runs from -1 to 1 , with 0 for no skill and 1 for perfect detection (Ferro and Stephenson, 2011). Its largest drop, 0.45 , is for
 245 extreme wind at day 1 (Fig. 12). Better pointwise accuracy does not guarantee better event detection. The experiment poses this interesting trade-off question for future exploration.

7 Discussion

TailWeather and disaster catalogs answer different questions. A TailWeather label marks weather that is rare for its place and season; a disaster record marks reported harm, which also depends on exposure, vulnerability, and reporting. The two are
 250 linked, since stricter tails coincide more often with recorded impacts, but they are not interchangeable. TailWeather supplies a global physical target, while impact catalogs remain indispensable whenever the question is societal consequence.



The dataset makes a hidden forecasting question measurable. AI systems can score well on average while missing tail events, most clearly for extreme wind. TailWeather can also serve as a training signal, but our sharpening experiment shows that the signal alone is not enough: the correction improved pointwise accuracy and degraded event detection. Thresholds and compound labels should be chosen for the hazard and the question.

TailWeather has four main limitations. First, every label inherits ERA5's biases and the limits of 0.25° , 6 h sampling (Hersbach et al., 2020). Convective gusts, tornadoes, hail, and downbursts are out of reach. ERA5 precipitation, a short-range forecast field, is also less tightly constrained by observations than temperature, which likely makes the heavy-precipitation and drought labels the least certain of the five. Second, all thresholds refer to a fixed 1991–2020 climate. They describe rarity relative to that baseline, not to a shifting one. However, our code support a moving-reference analysis to refit the climatology. Third, the empirical intensity score saturates near 0.99, so the farthest tail cannot be graded (Appendix F). Fourth, the comparison with recorded impacts inherits the catalog's coarse geolocation, inclusion criteria, and uneven reporting, so its numbers describe recorded disasters rather than all harmful weather. Appendices C and F provide more details.

Future versions can address several of these limits. A fitted parametric upper tail, such as a generalised Pareto distribution (Coles, 2001), would grade the far tail beyond the reach of the empirical rank. Extending the record past January 2023, and using a gust field where one is available, would lengthen the test period and sharpen the wind label. Scoring models separately on moderate, severe, and extreme cells, which the tiers already support, would show where forecast skill breaks down. The impact comparison would gain from dependence-preserving null models and from stratification by footprint geometry and date precision. TailWeather does not replace the careful reading of disaster records. It gives the physical side of that story a consistent, global form, so that forecasts can be tested wherever the weather is extreme, not only where extremes are reported.

8 Code and data availability

9 Data availability

TailWeather is available at <https://doi.org/10.57967/hf/10661> (Liu et al., 2026). Illustrative subsets and an interactive demonstration are available at <https://huggingface.co/spaces/zhisong-liu/TailWeather>. The dataset is licensed under CC BY 4.0.

10 Code availability

The label files and processing logs are available in the dataset repository at <https://huggingface.co/datasets/zhisong-liu/TailWeather>.

11 Conclusions

We have presented TailWeather, a 42-year daily record of five climatological hazards over global land, in which every event carries a severity tier and an intensity score. Three findings stand out. First, the labels are physically meaningful: they recover



documented European heatwaves, and for every hazard, stricter tails coincide more often with recorded disasters, with lifts of 1.6 to 9.3. Second, disaster records alone cannot support a fair global assessment: they are too sparse for at least one hazard where 80.1 % of the world's land population lives, whereas TailWeather applies the same definition everywhere. Third, the labels expose a gap that average scores hide: leading AI forecasts with errors of about 1 K score below 0.5 at detecting heatwaves and worst on extreme wind, and a tail-aware correction improved RMSE while degrading detection. TailWeather therefore complements, rather than replaces, disaster records. Its common grid, tiers, and scores make it a ready basis for global studies of extremes that do not depend on where reporting is dense, for tracing compound and cascading events, although compounds must be chosen with care, and for training and testing forecasts wherever extremes occur.

Appendix A: Event definitions

Each hazard is flagged by one reproducible percentile-and-persistence rule, applied to a single ERA5 variable against the local 1991–2020 climatology. Section 2 summarises the rules and Table 1 lists their parameters; the full definitions follow.

A1 Heatwave

We follow the percentile-and-persistence definition of Perkins and Alexander (2013). Let $TX(c, d)$ denote the daily maximum 2 m temperature at land cell c on date d , with day-of-year $j = \text{doy}(d)$. The seasonally varying threshold is the local 90th percentile of the reference climatology within a centred ± 15 -day window,

$$T_{90}(c, j) = P_{90}\{TX(c, d') : d' \in \text{ref. period}, |\text{doy}(d') - j| \leq 15 \text{ (circular)}\}. \quad (\text{A1})$$

A day is an exceedance if $E(c, d) = 1[TX(c, d) > T_{90}(c, j)]$, and a heatwave is flagged where exceedances persist for at least three consecutive days,

$$HW(c, d) = 1[\exists \text{ a run of } \geq 3 \text{ consecutive days with } E = 1 \text{ that includes } d]. \quad (\text{A2})$$

The intensity score $I(c, d)$ is the empirical percentile rank of $TX(c, d)$ in the local climatological cumulative distribution function (CDF); severity tiers bin it at 0.90/0.95/0.99. *Parameters:* P_{90} , ± 15 -day window, ≥ 3 days, reference 1991–2020.

A2 Cold wave

The cold wave mirrors the heatwave in the lower tail. Let $TN(c, d)$ be the daily minimum 2 m temperature, and take the local 10th percentile within the same ± 15 -day seasonal window,

$$T_{10}(c, j) = P_{10}\{TN(c, d') : d' \in \text{ref. period}, |\text{doy}(d') - j| \leq 15\}. \quad (\text{A3})$$

A day is an exceedance if $E(c, d) = 1[TN(c, d) < T_{10}(c, j)]$, and a cold wave requires at least three consecutive exceedance days. The cold-tail intensity score is $I(c, d) = 1 - (\text{percentile rank})$, so a value colder than the entire reference distribution scores ≈ 1 ; tiers use the same 0.90/0.95/0.99 bins. The three-day minimum is our choice and differs from the six-day minimum of the ETCCDI cold-spell duration index. *Parameters:* P_{10} , ± 15 -day window, ≥ 3 days, reference 1991–2020.



310 A3 Heavy precipitation

We follow the wet-day and wet-spell framework of Zolina et al. (2010): a 95th-percentile wet-day threshold (R_{95p}), complemented by a wet-spell diagnostic. Let $P(c, d)$ be daily total precipitation. A *wet day* has $P \geq 1$ mm, and a *wet spell* is a run of at least two consecutive wet days. The heavy-precipitation threshold is the 95th percentile of the *wet-day* distribution only,

$$R_{95}(c) = P_{95}\{P(c, d') : d' \in \text{ref. period}, P(c, d') \geq 1 \text{ mm}\}, \quad (\text{A4})$$

315 so that $\text{heavy}(c, d) = \mathbf{1}[P(c, d) \geq 1 \text{ mm} \wedge P(c, d) > R_{95}(c)]$. The structural diagnostic of Zolina et al. (2010) flags a heavy day embedded in an unusually long wet spell. With $L(c, d)$ the length of the wet spell containing d and $S_{75}(c)$ the 75th percentile of reference wet-spell lengths,

$$\text{heavy_long}(c, d) = \mathbf{1}[\text{heavy}(c, d) = 1 \wedge L(c, d) \geq S_{75}(c)]. \quad (\text{A5})$$

In the released files, the heavy-precipitation label and its severity follow `heavy (precip_severity > 0` exactly where `heavy_precip > 0`), while `heavy_long` is stored separately as `heavy_long_spell`. A heavy-precipitation label therefore does not require membership in a wet spell. The intensity score is the wet-day percentile rank of $P(c, d)$. Because R_{95p} already places every flagged day at or above the 95th wet-day percentile, the moderate tier $[0.90, 0.95)$ is empty by construction: heavy precipitation has no $s = 1$ events (Table 1). This follows from the definition; it is not a processing error. *Parameters:* wet day ≥ 1 mm, R_{95} over wet days; for the long-spell diagnostic, spells of ≥ 2 wet days and P_{75} spell length.

325 A4 Extreme wind

Let $W(c, d) = \max_{t \in d} \sqrt{u_{10}(c, t)^2 + v_{10}(c, t)^2}$ be the daily maximum sustained 10 m wind speed over the available six-hourly samples. The labelling code uses the ERA5 instantaneous 10 m gust when it is available, since it is the more impact-relevant, WMO-defined quantity. The archive used here contains no gust field, however, so every released wind label uses the sustained-speed fallback above, as the file metadata record. TailWeather wind labels are therefore not peak-gust extremes. With the local
 330 98th percentile within the ± 15 -day window,

$$W_{98}(c, j) = P_{98}\{W(c, d') : d' \in \text{ref. period}, |\text{doy}(d') - j| \leq 15\}, \quad (\text{A6})$$

a day is flagged if $W(c, d) > W_{98}(c, j)$. Wind extremes are transient synoptic storms lasting hours to about a day, so the minimum duration is one day. Severity tiers use speed boundaries motivated by the Beaufort scale: a flagged day is at least moderate, becomes severe at 20.8 ms^{-1} (Beaufort forces 9–10), and extreme at 28.5 ms^{-1} (force 11 and above). ERA5 at
 335 0.25° and 6 h resolves synoptic-scale wind extremes but not convective gusts. The label is thus the storm analogue that a reanalysis can provide; Sect. 4.6 discusses how this scale mismatch weakens the link to recorded wind damage. *Parameters:* P_{98} , ± 15 -day window, ≥ 1 day, Beaufort tiers at $20.8/28.5 \text{ ms}^{-1}$.



A5 Meteorological drought (SPI-3)

Drought is defined by the Standardized Precipitation Index (SPI) at a 3-month accumulation (McKee et al., 1993), standardised
 340 through a fitted gamma distribution (Lloyd-Hughes and Saunders, 2002). It is the only parametric tail in the dataset. Daily
 precipitation is aggregated to monthly totals and summed over three months, $P_3(c, m) = \sum_{k=0}^2 P_{\text{month}}(c, m - k)$. For each cell
 and target month, a mixed distribution is fitted to the reference-period accumulations to accommodate exact zeros,

$$H(x) = q + (1 - q) G(x; \alpha, \beta), \quad (\text{A7})$$

where $q = \Pr(P_3 = 0)$ and $G(\cdot; \alpha, \beta)$ is the gamma CDF fitted by maximum likelihood to the non-zero values. The index is the
 345 standard-normal quantile of the cumulative probability,

$$\text{SPI}(c, m) = \Phi^{-1}(H(P_3(c, m))). \quad (\text{A8})$$

Negative values indicate drier-than-reference conditions. The index is only approximately standard normal: this depends on the
 fit and is limited by the discrete probability mass at zero. A cell is in drought when $\text{SPI} \leq -1.0$, with tiers at $-1.0/-1.5/-2.0$
 following McKee et al. (1993).

350 SPI-3 is inherently monthly. A daily-broadcast variant repeats each month's label on every day of that month; this adds
 no temporal information and creates within-month dependence, although it does not by itself restrict the range of thresholds
 that can be evaluated. Because a monthly total uses the whole month's precipitation, the broadcast field must not be used as a
 contemporaneous daily predictor. For the released drought field, a score threshold of $q = 0.80$ (drought's F_1 -optimal threshold,
 Table 3) corresponds to approximately $\text{SPI} \leq -2.4$ on finite-score cells, and $q = 0.90$ to $\text{SPI} \leq -2.7$; neither is the 80th or 90th
 355 climatological percentile. Because the drought score is missing outside drought, conditions milder than the default threshold
 cannot be recovered from it; the stored `spi3` field is needed for that. *Parameters*: 3-month scale, gamma maximum-likelihood
 fit, thresholds $-1.0/-1.5/-2.0$.

Appendix B: Dataset characteristics

This appendix summarises the descriptive statistics of the labelled dataset: annual frequency (Fig. B1) and a comparative
 360 profile of the five hazards (Fig. B2). Severity-tier distributions are shown in Fig. 2g. Since thresholds are fixed at 1991–2020,
 frequencies in different years are measured against the same baseline. Heatwave frequency rises from 3.7 % of land cell-days
 (1981–1990) to 5.9 % (2014–2023), while cold-wave frequency falls from 6.4 % to 3.1 %, as expected in a warming climate.
 Neither result depends on the partial final year: restricting the recent decade to the nine complete years 2014–2022 changes
 heat from 5.85 % to 5.95 % and cold from 3.05 % to 2.87 %. The drought series reaches 33.8 % in 2023 (118 968 of 351 848
 365 land cells with finite SPI), but this value comes from a single, partial January. It is not an annual rate and should not be
 compared with the complete years earlier on the same curve. A fixed reference does not smooth over an incomplete final
 period, as the drought endpoint shows. The stored scores measure rarity relative to the fixed 1991–2020 baseline throughout; a
 moving-reference analysis would require refitting the climatology from the underlying ERA5 fields for each target period.

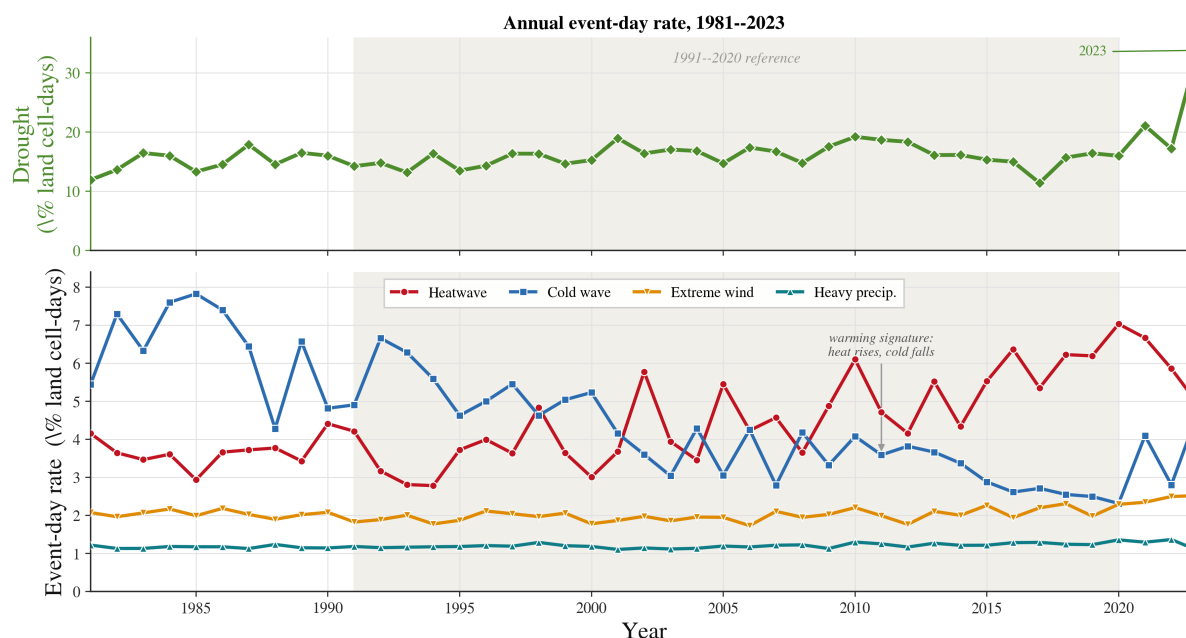


Figure B1. Against the fixed 1991–2020 baseline, heatwaves become more frequent and cold waves rarer. Annual event-day rate for each hazard, 1981–2022 plus a partial 2023, as a percentage of land cell-days, with the 1991–2020 reference period shaded. Drought (top) has its own scale because it is four to eight times more prevalent than the other four hazards (bottom). Because thresholds are held fixed at the reference climatology, the curves show how each hazard’s frequency has evolved relative to a stationary baseline: the heatwave and cold-wave series cross as heat rises and cold falls, the expected warming signature. The 2023 drought value (33.8 %) comes from a single, partial January.

Appendix C: Processing details

370 This appendix records the construction details needed to reproduce the labels; the per-hazard rules are given in Appendix A.
Terminal coverage and sampling. The WeatherBench2 archive ends at 18:00 UTC on 10 January 2023 (WeatherBench 2 contributors, 2026). The 2023 label files therefore cover 1–10 January for heat, cold, and wind; 1–9 January for precipitation; and a single January for monthly drought. These partial periods must not be treated as a full year. Daily extremes computed from six-hourly samples also differ from those computed from hourly data. Monthly precipitation totals and SPI need complete
 375 accumulation windows, so incomplete terminal months should be excluded.

Climatological reference distribution. For temperature and wind, the seasonal reference sample pools values within a circular ± 15 -day window over 1991–2020, giving approximately $30 \times 31 = 930$ values before missing data and leap days are accounted for. The precipitation threshold (Appendix A3) instead pools wet days across the full reference period, without a seasonal window. SPI is fitted separately for each calendar month. Pooling increases the sample size, but it does not make consecutive
 380 weather observations independent.

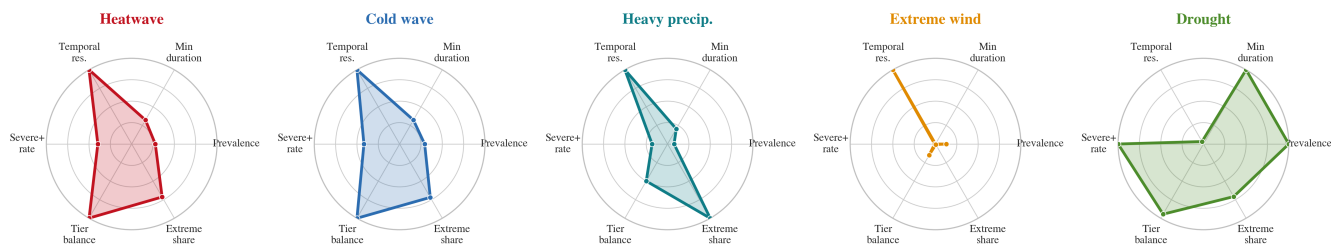


Figure B2. The five hazards differ in prevalence, duration, and tier balance. Comparative profile of the five hazards across six dataset properties: prevalence (event-day rate), minimum duration, temporal resolution, the severe-or-extreme event rate, tier balance (the normalised entropy of the moderate/severe/extreme mix), and extreme-tier share. Each axis is normalised to its cross-hazard maximum. Drought is prevalent, long-lasting, and coarse in time, whereas extreme wind is rare, short, and strongly concentrated in its lowest severity tier.

Circular day-of-year window. The ± 15 -day window wraps around the year boundary: the reference sample for early-January target days includes late-December values, and vice versa, so thresholds are continuous across 31 December–1 January rather than truncated at the calendar boundary.

Leap days. Day-of-year is the calendar's own day of year (1–365, or 1–366 in a leap year), pooled across reference years without adjustment. Because about one reference year in four is a leap year, every date from 1 March onward shifts by one day in those years, not only 29 February: the reference pool for a given target day mixes calendar dates up to one day apart. The ± 15 -day window probably dampens this offset, although we have not compared it directly with a strictly calendar-date-matched reference. Day 366 enters the pooled reference sample through the same circular day-of-year distance as other days.

Percentile estimator. Threshold percentiles are read from the sorted reference sample by linear interpolation between the two nearest order statistics, rather than by a nearest-rank rule. The rank spacing is of order $1/n$ for n reference values, but the resulting difference in physical threshold depends on the spacing of the relevant order statistics.

Intensity-score precision and far-tail saturation. Two resolutions must be kept apart: the precision with which scores are stored, and the resolution of the reference distribution. The pooled 43-year percentile tables give the 50th, 90th, 95th, and 99th score percentiles for all four daily hazards. All sixteen values are exact integer multiples of 0.0005 (e.g. heat P50 = 0.4845 = 969×0.0005), and none lies close to a multiple of $1/99 \approx 0.01$; the stored score grid is therefore 0.0005. The truncated 2023 file alone suggests a coarser spacing, but only because it contains so few days. Finer storage would not resolve the far tail, which is limited by the empirical rank itself (Appendix F).

The monthly drought score differs. Its finite values equal $\min(1, -\text{SPI}/3)$ on drought cells, and it is missing elsewhere. The smallest finite score is therefore close to $1/3$, which is why thresholds of 0.1, 0.2, and 0.3 select the same cells. Monthly broadcasting adds temporal dependence but does not cause this plateau.



Appendix D: Tail-impact comparison details

For valid comparison cell-days, define $a = \sum T_q X$, $b = \sum T_q (1 - X)$, $c = \sum (1 - T_q) X$, and $d = \sum (1 - T_q) (1 - X)$. Then

$$\text{precision} = \frac{a}{a+b}, \quad \text{recall} = \frac{a}{a+c}, \quad F_1 = \frac{2a}{2a+b+c}, \quad \text{lift} = \frac{a/(a+b)}{(a+c)/(a+b+c+d)}. \quad (\text{D1})$$

All terms use the same valid mask and weights. Per-cell log odds ratios are $\log(ad/bc)$ where defined; zero counts require an explicitly stated correction or masking convention. A missing catalog record cannot be read as a verified absence of impact. Confidence intervals come from 500 bootstrap resamples of 15-day temporal blocks. Each block's counts are pooled over the entire spatial grid before resampling, so the whole spatial field is resampled jointly and spatial as well as temporal dependence is retained. A 15-day block may still misrepresent the dependence of monthly SPI, overlapping three-month accumulations, or long disaster records. The intervals also omit uncertainty from the source reanalysis, threshold estimation, geolocation, and catalog inclusion.

Appendix E: Sharpening-head implementation

Starting from pre-trained FourCastNet (Pathak et al., 2022), we froze the backbone and attached a shallow residual head (He et al., 2016). The head reads the frozen forecast of 2 m temperature and the two 10 m wind components, proposes a post-processing correction, and predicts an attention map trained against the TailWeather intensity score (Vaswani et al., 2017). The losses give larger weight to cells and days with higher intensity scores. To expose the head to errors that grow during a rollout, we trained it with increasing rollout lengths of 1, 3, 5, and 7 days. The experiment is a simple test of whether TailWeather can supply useful supervision, not a comparison of training methods.

Appendix F: Limits of the tail-impact comparison

Cell-day and event-level quantities. EM-DAT applies inclusion criteria, so the comparison concerns recorded disasters rather than all harmful weather. A cell-day recall of 1.82 % for heat means that 1.82 % of rasterised heat-impact cell-days overlap the evaluated mask. It does not mean that only 1.82 % of heat disasters are captured; that event-level quantity is defined separately in Appendix G.

Score plateaus. Although the daily-hazard score is stored on a fine 0.0005 grid (Appendix C), the sweep still plateaus near the top. Any day whose value exceeds the entire reference sample of about 930 values receives the same maximum empirical rank, so many of the most extreme days tie at or near 1, and thresholds between about 0.99 and 1 select nearly the same cells. This is a property of an empirical rank estimated from a finite sample. For drought, finite scores begin near 1/3 and non-drought scores are missing, which explains the identical selections at 0.1, 0.2, and 0.3; the score is also capped at 1 for SPI at or below −3.

Threshold selection and partial curve area. Each q^* is selected on the comparison data and describes that sample. The partial area under the precision–recall curve integrates only over the recall range each sweep attains, which differs among hazards. It



should not be used to rank hazards or be compared with the full-curve no-skill baseline. The rise of precision with threshold is an empirical result, whereas the fall in recall follows from nested threshold masks whenever the persistence and validity rules preserve nesting.

Appendix G: Event-level capture rates

435 Section 4 compares labels and records cell-day by cell-day. A complementary, event-level view asks whether the tail label fire anywhere inside that record's footprint during its recorded dates. EM-DAT often records only the month or year of an event, so the date window adapts to the stated precision: ± 2 days for day-precise records, ± 15 days for month-precise records, and ± 182 days for the rare year-precise ones. Capture rates so defined are 57.1 % for heat ($n = 219$), 79.2 % for cold ($n = 288$), 82.8 % for wind ($n = 2495$), 93.6 % for heavy precipitation ($n = 3966$), and 93.5 % for drought ($n = 402$). A uniform ± 2 -day window
 440 would penalise records without day-level dates and understate capture for hazards whose entries are less often day-precise.

Two stratifications qualify these numbers. First, capture rises with footprint area for all five hazards, from 34 % to 86 % across area quartiles for heat and from 61 % to 90 % for wind, simply because a larger polygon over a multi-day window offers more cells that could cross the threshold, whether or not the exceedance coincides with the damage. This is the footprint confound of Appendix F, reappearing as inflated capture for large events. Second, stratifying by recorded severity gives a
 445 mixed signal: capture rises with severity for wind, heavy precipitation, and drought, but not for heat or cold.

To gauge how much capture chance alone could produce, we use an illustrative independent-cell calculation: a footprint of A cells over D days is captured by chance with probability $1 - (1 - p)^{AD}$, where p is the tail rate at the evaluated threshold (Fig. G1). The calculation uses representative areas and durations, and a constant area of about 540 km^2 per cell, a mid-latitude approximation rather than each event's actual count of rasterised cells. In the smallest-area quartile, observed capture exceeds
 450 this chance level by 22, 51, and 61 percentage points for heat, wind, and heavy precipitation.

Appendix H: Per-hazard concordance detail

This appendix resolves the pooled results of Sect. 4 by hazard. Figure H1 traces, for each hazard, the precision–recall trade-off as the tail threshold q is swept, with the F_1 -optimal q^* marked. Every curve lies well above its impact base rate, the precision a random tail would achieve, indicating a positive association in this sample. The F_1 maxima (panel f) fall at very different
 455 thresholds: no single percentile is optimal. Figure H2 completes the spatial picture of Figs. 10 with per-cell log odds ratios for the remaining three hazards.

Appendix I: Severity-tier composition through time

Figure B1 reports the total event-day rate. Fig. I1 splits it into the moderate, severe, and extreme tiers, year by year. The changes in frequency depend on the tier. The heatwave increase spans all three tiers but is proportionally largest in the extreme tier,
 460 which roughly doubles between 1981 and 2020. The 2023 drought value has large severe and extreme contributions. However,

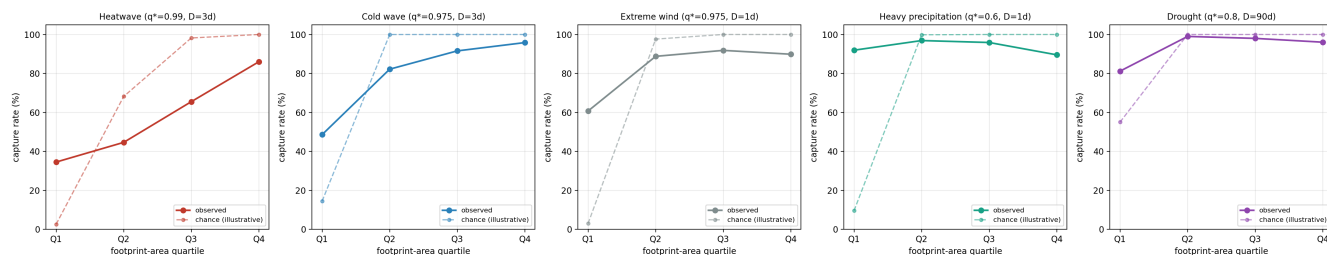


Figure G1. Small-footprint events are captured above an illustrative chance level; for large footprints, capture carries little information. Event-level capture by footprint-area quartile. Observed capture (solid, per hazard) and the binomial chance capture (dashed) for a footprint of the quartile’s representative area over a representative D -day window (Appendix G), evaluated at each hazard’s q^* . In the smallest-footprint quartile (Q1), observed capture exceeds this illustrative baseline for every hazard, most sharply for heavy precipitation and wind; in the largest (Q4), chance capture saturates near 100 %, so the gap narrows for every hazard regardless of genuine detection. Drought’s excess is largest in Q1 and falls to near parity by Q4. The representative area-to-cell conversion and per-hazard durations are illustrative and uncalibrated, and are no substitute for a dependence-preserving null model.

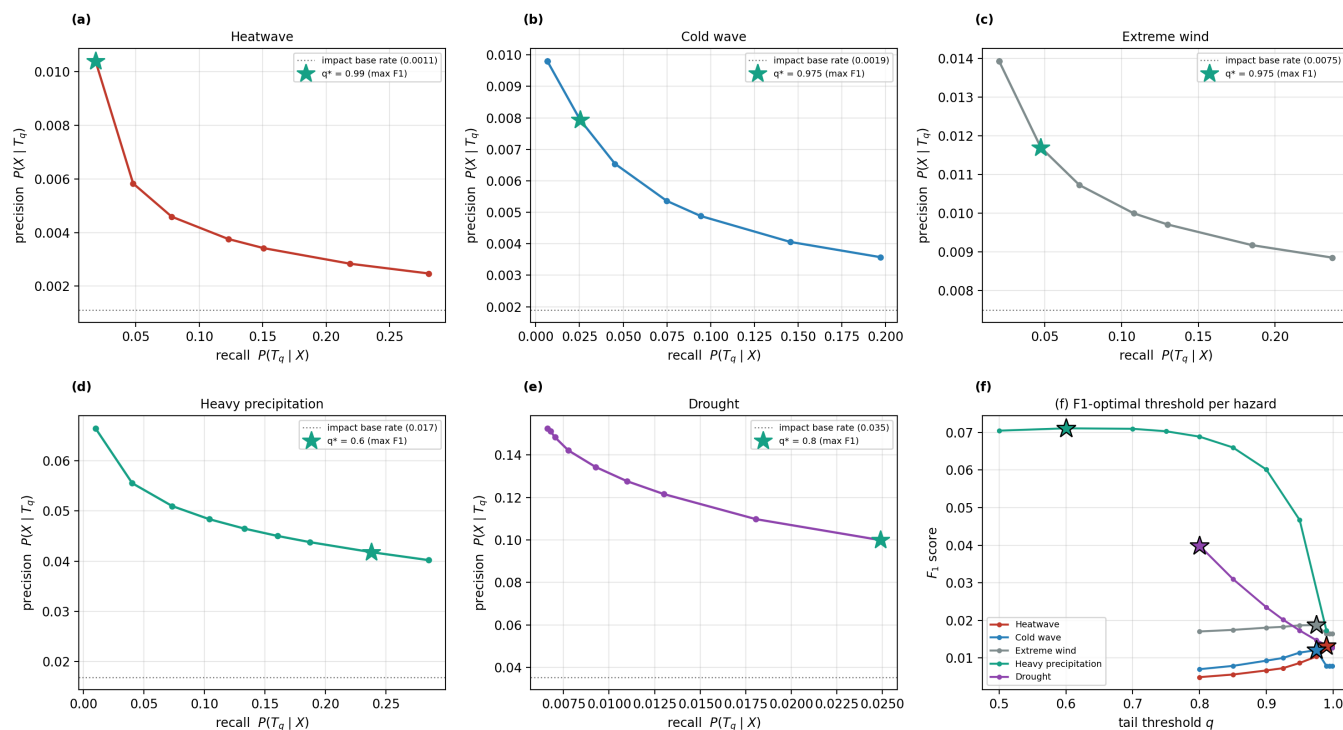


Figure H1. Every hazard’s precision–recall curve lies above its impact base rate. (a–e) Precision $P(X | T_q)$ against recall $P(T_q | X)$ as the tail threshold q is swept, one panel per hazard; the star marks the F_1 -optimal q^* and the dotted line the impact base rate (the precision a random tail would achieve). (f) The F_1 score as a function of q for all five hazards, with stars at the maxima that define q^* ; these range from $q^* = 0.60$ (heavy precipitation) to 0.99 (heat), with drought’s optimum (0.80) on its separate SPI-based scale.

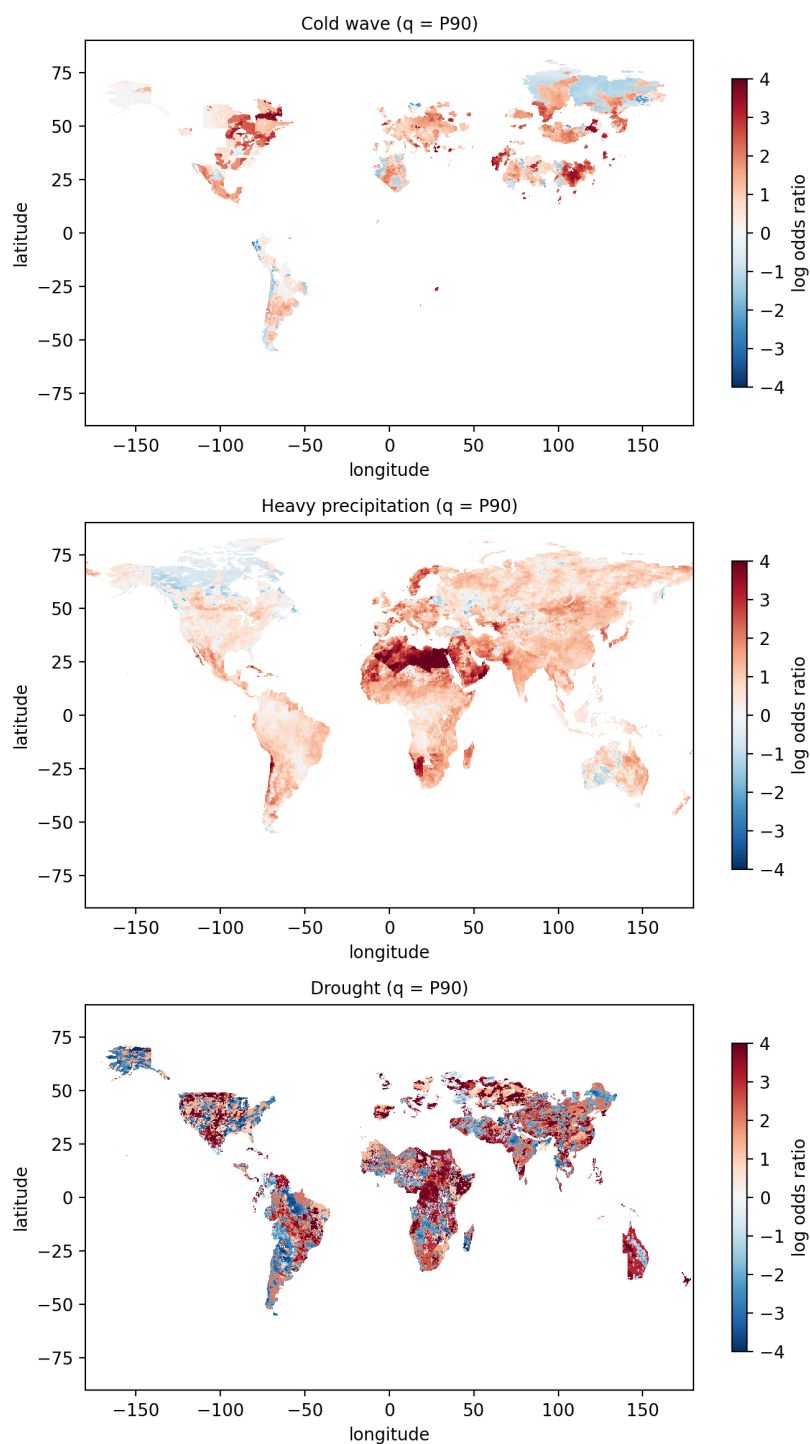


Figure H2. Heavy precipitation, with the densest impact record, can be assessed over the widest area. Per-cell log odds ratio for the remaining three hazards at $q = 0.90$, 1996–2023, as in Fig. 10: (a) cold wave, (b) heavy precipitation, (c) drought. Cells with fewer than five impact days are masked, and the diverging scale is clipped at ± 4 .

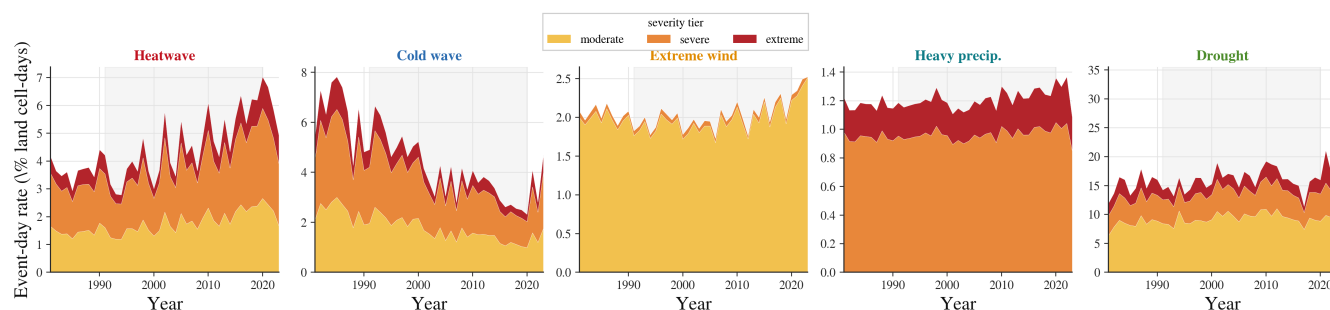


Figure 11. The heatwave increase is proportionally largest in the extreme tier. Severity-tier composition of the annual event-day rate, 1981–2022 plus a partial 2023 (see Fig. B1). Each panel stacks the moderate, severe, and extreme contributions to the event-day rate of Fig. B1 for one hazard, with the 1991–2020 reference period shaded. The vertical scale is independent for each hazard. The 2023 endpoint reflects a single partial month or ten days, not a full year, most visibly for drought (Appendix I).

it rests on a single January field, it says nothing about annual drying. Heavy precipitation has no moderate tier by construction (Sect. A3), and extreme wind is almost entirely moderate because its Beaufort escalation thresholds are seldom met.

Author contributions. Z.-S. Liu implemented the labelling pipeline, the concordance analysis, and the figures; M. Boy and R. Makkonen advised on atmospheric interpretation; all authors contributed to the manuscript.

465 *Competing interests.* The authors declare that they have no competing interests.



References

- Alet, F., Price, I., El-Kadi, A., Masters, D., Markou, S., Andersson, T. R., Stott, J., Lam, R., Willson, M., Sanchez-Gonzalez, A., and Battaglia, P.: Skillful joint probabilistic weather forecasting from marginals, arXiv [preprint], <https://doi.org/10.48550/arXiv.2506.10772>, 2025.
- Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, *Nature*, 619, 533–538, <https://doi.org/10.1038/s41586-023-06185-3>, 2023.
- Bodnar, C., Bruinsma, W. P., Lucic, A., et al.: A foundation model for the Earth system, *Nature*, 641, 1180–1187, <https://doi.org/10.1038/s41586-025-09005-y>, 2025.
- Centre for Research on the Epidemiology of Disasters (CRED): EM-DAT: The International Disaster Database, <https://www.emdat.be/>, last access: 6 October 2026, 2024.
- Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer, London, <https://doi.org/10.1007/978-1-4471-3675-0>, 2001.
- European Centre for Medium-Range Weather Forecasts: IFS Medium-range Control forecast (formerly HRES) and Analysis Data (15-days) (Set I), <https://www.ecmwf.int/en/forecasts/datasets/set-i>, last access: 6 October 2026, 2026.
- Ferro, C. A. T. and Stephenson, D. B.: Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events, *Weather and Forecasting*, 26, 699–713, <https://doi.org/10.1175/WAF-D-10-05030.1>, 2011.
- He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, <https://doi.org/10.1109/CVPR.2016.90>, 2016.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., et al.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., and Neumann, C. J.: The international best track archive for climate stewardship (IBTrACS): Unifying tropical cyclone data, *Bulletin of the American Meteorological Society*, 91, 363–376, <https://doi.org/10.1175/2009BAMS2755.1>, 2010.
- Lam, R., Sanchez-Gonzalez, A., Willson, M., et al.: Learning skillful medium-range global weather forecasting, *Science*, 382, 1416–1421, <https://doi.org/10.1126/science.adi2336>, 2023.
- Liu, Z.-S., Boy, M., and Makkonen, R.: TailWeather, Hugging Face [data set], <https://doi.org/10.57967/hf/10661>, 2026.
- Lloyd, C. T., Sorichetta, A., and Tatem, A. J.: High resolution global gridded data for use in population studies, *Scientific Data*, 4, 170 001, <https://doi.org/10.1038/sdata.2017.1>, 2017.
- Lloyd-Hughes, B. and Saunders, M. A.: A drought climatology for Europe, *International Journal of Climatology*, 22, 1571–1592, <https://doi.org/10.1002/joc.846>, 2002.
- McGovern, A., Mandelbaum, T., Rothenberg, D., Loveday, N., Potvin, C., Flora, M., Magnusson, L., Gilleland, E., and Allen, J.: Extreme Weather Bench: A framework and benchmark for evaluation of high-impact weather, arXiv [preprint], <https://doi.org/10.48550/arXiv.2605.01126>, 2026.
- McKee, T. B., Doesken, N. J., and Kleist, J.: The relationship of drought frequency and duration to time scales, in: Proceedings of the 8th Conference on Applied Climatology, pp. 179–184, American Meteorological Society, Anaheim, California, <https://api.semanticscholar.org/CorpusID:129950974>, 1993.
- NOAA National Centers for Environmental Information: Storm Events Database, <https://www.ncei.noaa.gov/stormevents/>, last access: 6 October 2026, 2024.



- Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators, arXiv [preprint], <https://doi.org/10.48550/arXiv.2202.11214>, 2022.
- 505 Perkins, S. E. and Alexander, L. V.: On the measurement of heat waves, *Journal of Climate*, 26, 4500–4517, <https://doi.org/10.1175/JCLI-D-12-00383.1>, 2013.
- Price, I., Sanchez-Gonzalez, A., Alet, F., et al.: Probabilistic weather forecasting with machine learning, *Nature*, 637, 84–90, <https://doi.org/10.1038/s41586-024-08252-9>, 2025.
- Ran, N., Xiao, P., Wang, Y., Shi, W., Lin, J., Meng, Q., and Allmendinger, R.: HR-Extreme: A high-resolution dataset for extreme weather forecasting, in: *International Conference on Learning Representations (ICLR)*, <https://doi.org/10.48550/arXiv.2409.18885>, 2025.
- 510 Rasp, S., Hoyer, S., Merose, A., Langmore, I., et al.: WeatherBench 2: A benchmark for the next generation of data-driven global weather models, *Journal of Advances in Modeling Earth Systems*, 16, e2023MS004 019, <https://doi.org/10.1029/2023MS004019>, 2024.
- Rosvold, E. L. and Buhaug, H.: GDIS, a global dataset of geocoded disaster locations, *Scientific Data*, 8, 61, <https://doi.org/10.1038/s41597-021-00846-6>, 2021.
- 515 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, in: *Advances in Neural Information Processing Systems*, vol. 30, <https://dl.acm.org/doi/10.5555/3295222.3295349>, 2017.
- WeatherBench 2 contributors: WeatherBench 2 Data Guide, <https://weatherbench2.readthedocs.io/en/latest/data-guide.html>, last access: 8 September 2026, 2026.
- Zhang, X., Alexander, L., Hegerl, G. C., et al.: Indices for monitoring changes in extremes based on daily temperature and precipitation data, 520 *Wiley Interdisciplinary Reviews: Climate Change*, 2, 851–870, <https://doi.org/10.1002/wcc.147>, 2011.
- Zhao, S., Xiong, Z., Zhao, J., and Zhu, X. X.: ExEBench: Benchmarking Foundation Models on Extreme Earth Events, arXiv [preprint], <https://doi.org/10.48550/arXiv.2505.08529>, 2025.
- Zolina, O., Simmer, C., Gulev, S. K., and Kollet, S.: Changing structure of European precipitation: Longer wet periods leading to more abundant rainfalls, *Geophysical Research Letters*, 37, L06 704, <https://doi.org/10.1029/2010GL042468>, 2010.
- 525 Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., et al.: Future climate risk from compound events, *Nature Climate Change*, 8, 469–477, <https://doi.org/10.1038/s41558-018-0156-3>, 2018.