Consequence entanglement, measured: a falsifiable operationalization of the polycrisis severity claim

Paper E of the Fenocrisis founding program (PROMPTS.md P-6.3; the severity pivot). The consequence-propagation companion to Paper A: where Paper A tested entanglement in the timing *of crisis onset and found it absent, or fragile under a variance-maximal shared-modulation absorption bound, this tests entanglement in crisis severity — the form the field's own mechanism vocabulary makes plausible — behind a pre-registered decidability atlas, against a null that holds the count data identical, and with the effect size below which its negative says nothing stated in the abstract rather than buried in the limitations. Aimed at the polycrisis field's home venue (*Global Sustainability*) or a marked-point-process methods venue; the venue call is an owner action (Owner queue).*

Author. Evan Tabak Atlas (ORCID 0009-0007-7374-2338).

Status. Complete draft, gated on owner review before submission (Owner queue). The records this paper reads are committed and regenerable under `out/`: `out/v6` (the single-stream marked atlas), `out/v11` (the marked panel atlas), `out/v9` (the converged multi-seed severity tournament, together with its registered-tier EM-DAT addendum), and `out/v12` (the severity-excess calibration record — the known-null false-excess ceiling, the drift-equivalence curve, and the recovery-bootstrap counter-cell). A registry of headline numbers is machine-bound to the sealed record JSON each is read from by the repository's number-anchor gate (`check_citations.py`), and the build fails if any of them drifts. That registry is a hand-registered spot check on a chosen subset, not a completeness guarantee: it verifies that the registered literals are consistent, not that every number in the paper is bound. An earlier draft of this paragraph said "every quantitative statement traces to a committed, regenerable record", and §5.3 documents the concrete counterexample that sentence had — a known-null ruler that existed only as prose. It is withdrawn as written and replaced by the sentence above. The EM-DAT-derived numbers — §5.3, §5.4 and the bound §5.5 states, since the disaster half of that comparison is EM-DAT's — are *derived statistics* computed from a licensed file that is never committed; they are labelled **[registered-tier]** wherever they appear and are not machine-bound. This is a working draft, not a submission. **Scope: this tests entanglement in crisis *severity*. It does not reopen the onset-time verdicts (`out/v1`, `out/v3`, `out/v5`), which stand unchanged; nothing here forecasts, warns, or ranks (F7).**

Abstract

The polycrisis field asserts that crises are causally entangled across domains, but it has never written down a quantitative claim that could be shown to be false. Its synthesis papers develop mechanism vocabularies; its one quantitative corpus is analyzed with

counts and co-occurrence plots and drops event magnitudes explicitly; its formal apparatus is cross-impact matrices and causal-loop diagrams whose edges are elicited rather than estimated. There is, at present, nothing in the literature for a disagreeing analyst to refute. This paper supplies the missing object for the half of the claim the field's own mechanisms make most plausible — **consequence propagation: that crises make each other *worse*, not merely more frequent** — and then tests it.

We formalize a crisis stream as a **marked** point process: per month a count `N_t` and, for each event, a mark `y = log(1 + magnitude)`. Two coupling channels are kept structurally separate — an *intensity* channel in which a latent regime modulates the event **rate** (onset timing, the channel Paper A tested) and a *mark* channel in which it modulates event **severity**. Each has an explicit null the field never wrote down. Within a domain it is **random labelling**: the observed marks permuted across the stream, holding the counts and the marginal severity distribution *exactly* fixed. The counts are therefore **identical data** between the observed stream and every surrogate, so the entire difference in maximised likelihood is attributable to the temporal *assignment* of marks. (We previously wrote that the count channel "cancels exactly for the arms whose likelihood factors". That is a stronger mathematical claim than the estimator supports: the likelihood factors into a count part and a mark part only at *fixed* parameters, and one decay `β` is shared between the count and mark kernels, so the profile likelihoods do not separate parameter by parameter. The excess remains a well-defined statistic — identical count data on both sides — which is what the argument needs.) Across domains it is a circular-shift null on each domain's detrended monthly mean log-severity. Because a null result is only informative where the design could have seen a positive, both tests are gated by decidability atlases computed on known-truth simulation: a single-stream atlas (`out/v6`, 65 cells) **sealed before any real mark was read**, and a cross-domain marked *panel* atlas (`out/v11`, 65 cells) built **after** the cross-domain read and therefore queried retrospectively — a provenance difference stated here rather than in the limitations, because it changes what the cross-domain gate is worth.

The result is a split verdict, and the positive half is narrower than the design set out to establish. **Within conflict** (UCDP GED `best` fatalities, N = 347,109 events over T = 420 months, 1990–2024) the excess over the mark-shuffle null survives decisively and seed-robustly — ΔBIC excess = 7,654, clearing the surrogate ensemble's maximum by more than 7,600 BIC, in all five pre-declared cuts. What that establishes is that **the monthly allocation of conflict severity is not exchangeable given the counts** — and that is all it establishes. **The magnitude is not diagnostic**: a sealed calibration record built for this paper (`out/v12`) shows that an *exogenous* monthly-mean log-severity drift of about 0.13 log-units, with no severity dynamics of any kind, reproduces an excess of the same size, as does a literal two-population composition model, with a fitted mark coefficient of the same order. We therefore report the rank verdict and withdraw every effect-size comparison built on the 7,654. We resist three further tempting extensions of it. The instrument sees only per-bin sufficient statistics, so the claim is about a monthly series and not about event-to-event dependence. The best-fitting arm is the one in which monthly mean severity depends on its own recent history, but the sign of that dependence is *not* recoverable from the sealed record — the coefficient is unconstrained in sign and

only its absolute value was stored — so we decline to call it self-excitation. And two alternative explanations are named and neither is excluded: the null is exchangeability against a *single global* severity law, which a slowly-changing mixture of heterogeneous conflicts rejects with no dynamics at all, and a recording process whose magnitude completeness varies with recent activity is the seismological objection to every result of this shape, whose laboratory control has no crisis analogue. Composition is the front-runner and, unlike recording, is checkable now on open data.

Within disasters [registered-tier] the same excess is decidable by rank but *near-negligible* — 33.6 BIC, under 0.5 % of conflict's. The ruler an earlier draft measured it against ("about 1.4× the false-excess ceiling") was an unregenerable prose figure and is withdrawn; against `out/v12`'s measured false-excess ceiling at the disaster configuration, +4.02 BIC, the cell sits well clear of a true-null baseline, which makes it a *presence*, not a magnitude. **Across domains** [registered-tier] there is no shared severity regime: the detrended contemporaneous correlation of conflict and disaster monthly mean log-severity is −0.1147 against a circular-shift null (one-sided positive-coupling p = 0.9492; two-sided p = 0.1358), and no lag in a ±6-month scan turns positive. So on these data the most plausible form of polycrisis fails in the same direction as the least plausible one: whatever severity structure exists is **domain-local**.

We state what would refute all of it, and what bounds the negative. The panel atlas's weakest mapped shared-regime strength is θ = 0.30, recoverable from D = 2 domains at its reference cell and sitting exactly on the decidability threshold there — so the negative is a bound at θ ≳ 0.3 *for the panel-likelihood instrument the atlas certifies, at the one (σ, N/domain) point where that strength is mapped*. The correlation actually computed is lower-powered, so the operative floor is at least that and is not itself mapped. It is an effect-size floor published with its coordinates rather than implied. We also report in full a house correction: the first pass at this tournament (`out/v7`) read its verdict off a marked-Hawkes fit that never converged; its optimizer audit withdrew a σ-tier and cost four of five cuts their published tiers; the converged multi-seed re-run reported here (`out/v9`) resolves every affected cut to the same decisive positive and *raises* the headline magnitude. Every number is a retrospective model-comparison on public, aggregate data. **Scope: this is the severity channel. The onset-time verdicts are not reopened and this result does not speak to them.**

---

1. The gap: a field with nothing to refute

The polycrisis literature is conceptual, elicited, or descriptive. Its causal-mechanism synthesis (Lawrence, Homer-Dixon, Janzwood, Rockström, Renn & Donges 2024, *Global Sustainability* 7, "Global polycrisis: the causal mechanisms of crisis entanglement") builds a vocabulary of common stresses, domino effects, and inter-systemic feedbacks, and develops a research agenda; it writes down no estimable model. Its companion consensus roadmap (Lawrence et al. 2024, *Polycrisis Research and Action Roadmap*, Cascade Institute) says so directly: "Polycrisis analysis also needs formal models of crisis

interaction and early-warning indicators of crisis emergence and escalation." (This paper supplies neither as an instrument — it is a retrospective model-comparison and nothing in it is an early-warning indicator, F7 — but the first half names the gap it addresses.) A 2025 systematic review (Rakowski et al., *Annual Review of Environment and Resources* 50) found **one dynamic model among 59 analyzed polycrisis publications**. The field's own most searching internal critique asks whether "polycrisis" is a concept or a meme (Lawrence 2022) — and the honest answer turns on whether anyone can state the claim in a form that could come out false.

That is the gap this paper addresses, and it is a different kind of gap from a missing result. There is no published quantitative claim about cross-domain crisis interaction that a disagreeing analyst could refute: no null, no test statistic, no stated sample size at which the question becomes answerable, and therefore no way for the field's central assertion to lose. The gap is not an inference of ours; the field states it. The most quantitative polycrisis study in existence (Delannoy et al. 2025) analyses a multi-category national-shock database for trends, spatial distribution, and cross-category co-occurrence — and disclaims severity explicitly, recording that the authors found "the magnitude and impact of shocks (casualties, U.S. dollars, etc.) inconsistent across datasets, and as such, excluded them from the present analysis (only the number of shocks is presented)". It reports no null model, no surrogate, and no significance test. Its immediate neighbours are franker still: a 2025 crisis- transmission paper in the same journal describes itself as "a hypothesis- proposing study" that "does not offer empirical testing which is left for future research", and states that "future empirical testing needs to demonstrate which of the presented conditions are most relevant, need adjustment, or can be abandoned" (Brosig 2025). This paper is an attempt at exactly that.

Paper A of this program supplied one such refutable object for the *timing* of crisis onset and reported its failure. This paper supplies the object for the half of the claim the field's own mechanism vocabulary makes far more plausible — and which the program canon's onset-time scope clause (§6.5) explicitly reserves as *not* addressed by any onset-timing verdict:

> An onset-timing tournament tests onset-time entanglement only. A negative does not refute consequence propagation (a marked-process question), systemic-risk interdependence, longer-horizon chains, or coupling below the decidability threshold.

Why severity is the plausible form. Almost every mechanism the field names is a mechanism about *consequences*, not about clocks. A drought does not make a coup more punctual; it makes a food-price shock bite harder. A pandemic does not schedule a debt crisis; it deepens one already under way. "Crises interact" read as *onset-time entanglement* is the strong and brittle version of the claim — it asserts that the arrival of one crisis shifts the hazard of another's arrival — and this program has now measured it three ways and found it fragile or absent each time (`out/v1`, `out/v3`, `out/v5`). Read as *consequence propagation*, the same sentence says something much weaker and much more likely: that magnitudes co-move even where arrivals do not. That claim has never

been tested, because testing it requires marks, and the field's harmonized corpus (Delannoy et al. 2025) deliberately excludes event magnitudes.

Why the machinery exists but has not been used here. Marked point processes are a mature apparatus, and the question "are the marks independent of the history, or do they cluster?" has been fought over for two decades in seismology, where it is the question of whether earthquake magnitudes are correlated. **That debate is live and unsettled, and it is worth stating which way each side cuts**, because a one-sided reading of it would misrepresent the only literature that has taken this question seriously. Corral (2006) found recurrence times to depend on seismicity history while magnitudes did *not* — the mark channel memoryless where the count channel is not. Lippiello, de Arcangelis & Godano (2008) and Lippiello, Godano & de Arcangelis (2012) found the opposite: magnitudes conditioned on preceding seismicity, with the 2012 paper arguing the correlation is not a catalogue-incompleteness artifact. Davidsen & Green (2011) argued that it largely is. Taroni (2024), re-examining Southern California with incompleteness removed, did not find the correlation surviving. Petrillo & Zhuang's (2022) meta-analysis of twenty-nine papers concludes that the question is not yet answerable at the required standard. Xiong et al. (2023) report magnitude clustering as prevalent across field *and laboratory* catalogues — the laboratory arm being their answer to the incompleteness objection, and an answer no crisis dataset can supply.

Three things follow for this paper. First, the **instrument** is settled even where the seismological verdict is not: it is **random labelling** — hold the event times fixed, permute the marks, ask whether the observed mark sequence beats the permuted one — a textbook null for marked point patterns (Diggle 2013; Illian et al. 2008), of which Xiong et al.'s shuffled-magnitude test is the closest sibling to the design used here. Second, the **standard objection** is settled too, and we adopt it as our own leading alternative explanation (§5.6): apparent mark clustering can be manufactured by a detection process whose completeness varies with recent activity. Third, the **contribution of this paper is not a new statistic**. The mark-shuffle surrogate is a random-labelling test in a lineage decades old; the marked Hawkes and marked Cox arms are standard; the novelty is entirely in the *application* to social crisis domains, in the mechanism discrimination raced behind it, and in what is built around it — a pre-registered decidability gate, an explicit effect-size floor, and a claim stated so that it can fail.

And the null-hypothesis template is not new either. The compound- and multivariate-extremes literature in climate science routinely tests dependence between drivers against an explicit independence null and quantifies how that dependence shifts joint-exceedance probabilities (Zscheischler & Seneviratne 2017). That work is cross-*variable* within one domain rather than cross-domain across social systems, but it is the template this paper imports, and the debt is acknowledged rather than elided.

Either verdict is a deliverable. A confirmation would be the first quantitative evidence that crisis severity propagates across domains; a failure is the first honest test of the most plausible form of the field's central claim. Both arrived: a positive within one domain and a negative across domains, in the same tournament, on the same protocol.

2. The claim as the field states it — and the two nulls it never wrote down

2.1 The formal object

Let a crisis domain `d` emit events at times `t_i` with magnitudes `m_i` (deaths, fatalities, affected). Bin to months. A **marked stream** is then, per bin `t`, a count `N_t` together with, for each event in the bin, a **mark**

`y = log(1 + magnitude)`

— log-severity, the `+1` offset admitting the zero-magnitude events that crisis corpora genuinely contain (in the conflict stream used here, 7.9 % of events carry a `best` estimate of zero). Conditional on a latent state the marks are i.i.d. Gaussian, so a bin is fully summarized by its sufficient statistics `(N_t, Σy, Σy²)` and the whole competitor roster scores the mark channel in closed form.

A latent regime can now act through **two structurally distinct channels**:

- the **intensity channel** — the regime modulates the event *rate*, `log λ = log λ₀ + κ · log(m_hi) · net_t`. This is onset timing: the exact unmarked MSM-Cox channel Paper A tested.
- the **mark channel** — the regime modulates event *severity*, `m_t = m₀ + θ · net_t`. A hot regime makes events bigger.

The cell `κ = 0, θ > 0` is precisely "severity propagates even when timing does not". Keeping the two channels separate is the whole design: it is what makes "consequence entanglement" a distinct hypothesis rather than a re-description of a timing result.

2.2 The within-domain claim, and the null that cancels the count channel

Claim C1 (severity is not exchangeable). *Within a domain, the allocation of event severity across months is not exchangeable given the counts: the recent history of monthly mean log-severity carries information about the next month's.*

We state C1 at monthly resolution deliberately, because that is the resolution the instrument has. A marked stream is reduced to its per-bin sufficient statistics `(N_t, Σy_t, Σy_t²)`; every arm scores the mark channel through those, and the marked-Hawkes arm's mark driver is `d_t = ȳ_t − m₀`, a monthly bin mean. Within-month ordering is invisible to the observed fit and is destroyed identically by the surrogate. At roughly 826 conflict events per month this is a claim about a coarse monthly series, **not** about event-to-event dependence, and we do not make the event-level claim anywhere.

The null is **random labelling**: the observed marks, permuted uniformly at random across the events of the stream. This surrogate holds

- the **counts** `N_t` in every bin **exactly** fixed, and
- the **marginal severity distribution** — every observed magnitude, with its multiplicity — **exactly** fixed,

and destroys only the temporal *assignment* of severity to events. Because the counts are bit-identical between the observed stream and every surrogate, any advantage an arm enjoys purely from the clustering of *arrivals* is present in identical measure in the surrogates and cancels in the difference. We call the quantity that survives the **severity excess**:

`excess = ΔBIC(arm vs null)_observed − mean_surrogates ΔBIC(arm vs null)`

with the verdict read off the observed statistic's **rank** in the surrogate ensemble, never off a z-score (§6 explains why the z was retired). Under C1's null the observed stream is exchangeable with its surrogates and its rank is uniform.

Three properties of that quantity should be stated before any number is read off it, because two of them are weaker than the design's summary suggests.

The cancellation is exact for two arms and first-order for the third. The marked null and the marked Hawkes likelihoods factor as `count_ll + mark_ll`, so for them the count term is literally identical between observed and surrogate and subtracts out. The marked MSM-Cox likelihood does not factor: a single latent path is inferred jointly from counts and marks, so shuffling the marks moves the inferred path and with it the count-emission term. The implementation says so in its own docstring, and we repeat it rather than smoothing it: for the MSM-Cox row of §5.1 the "count channel" column is not exactly the count channel and the residual is a mark-plus-interaction quantity. **The headline excess is read off the Hawkes arm, where the cancellation is exact.**

The excess is a likelihood contrast, not a model-selection quantity. Because the marked null's maximum-likelihood fit depends only on the totals `Σy` and `Σy²`, it is exactly shuffle-invariant; and every fit in the expression carries the same parameter count and the same `n_obs`. The null term and the entire BIC penalty therefore cancel algebraically, leaving `excess ≡ 2·(ℓ_obs − mean ℓ_surr)` — twice a log-likelihood contrast for **one** model on two datasets. It is reported in BIC units for comparability with the tournament table, but nothing in it is a penalised between-model comparison.

So the "95 % is counts" figure is a ratio of two different things, and we label it as such. In the primary conflict cell the Hawkes arm's raw advantage over the marked null is 165,211 BIC — that *is* a penalised between-model comparison — of which 157,557 BIC is reproduced by the mark-shuffle surrogates. The correct reading of the ratio is: *the surrogates reproduce roughly 95 % of the arm's advantage over a homogeneous-count null, and the severity claim rests on the residual 4.6 %.* That residual is a within-domain arrival-clustering term being differenced away — **within-domain overdispersion in conflict arrivals, which is not the cross-domain onset-time quantity `out/v1`, `out/v3` and `out/v5` tested** and should not be attributed to them. What the ratio does establish is that any version of this test which skipped the surrogate would have reported a vast "severity" effect that is almost entirely arrivals.

One calibration caveat belongs here rather than in the limitations, because it conditions how the surrogate should be read. Random-labelling envelopes are known to be **miscalibrated when the ground process is inhomogeneous or clustered**

(Grabarnik, Myllymäki & Stoyan 2011), and a crisis event stream is emphatically both. The specific danger is not the counts — those are held identical — but the mark law: permuting marks globally imposes **one** severity distribution on every month of a 35-year, 190-country stream. Any departure from that single global law — a drift in what is recorded, or simply a slowly-changing mixture of which conflicts are running where — registers as an excess. §5.6 treats the two resulting alternative explanations as co-leading, and neither is excluded here.

2.3 The cross-domain claim, and its null

Claim C2 (a shared severity regime). *Across domains, severity co-moves: hot months are systematically deadlier in more than one domain at once, driven by a shared latent regime.*

This is the polycrisis headline in the consequence channel, and it is a different statistic from C1. Its estimable form is the panel model in which `D` domains' mark **means** share one latent cascade,

`m_{d,t} = m₀_d + θ · net_t`,

with per-domain counts left independently Poisson so that the *only* cross-domain coupling in the model lives in the severity means. This arm nests the independent-panel null at `θ = 0` ($\Delta k = 3$), so `BIC(regime) < BIC(null)` is a clean shared-severity-cascade excess. The reduced statistic actually computed on real data is the detrended contemporaneous correlation of the two domains' monthly mean log-severity, tested against a **circular-shift** null which preserves each domain's own autocorrelation while destroying alignment between them, plus a ±6-month lead–lag scan (the severity analogue of the directional design of `out/v5`). Both series are linearly detrended first, so a shared *linear* reporting ramp cannot manufacture co-movement — the qualifier matters, because the detrend removes a single straight line fitted over the full 420-month window and is then applied to the 288 common-support months where the correlation is taken. A nonlinear ramp survives it, and the sparse, noisy pre-2000 period influences the slope subtracted from the later window. Fitting the detrend on the common support, and testing a spline or first-difference alternative, belong with the gated re-run.

2.4 What makes these claims falsifiable

Three things, none of which the field's existing formalisms supply:

1. **A null that is computable from the observed data alone**, with no free parameters to tune — the permutation and the shift are both exact.
2. **A pre-registered decidability region** (§3): the sample sizes and effect sizes at which each test can distinguish its alternative from its null, mapped on known-truth simulation *before* any real mark is read, so that a negative can be reported as a negative rather than as a shrug.
3. **A stated effect-size floor** below which the negative is silent. A test whose power is unmapped cannot be refuted either — it can only be doubted.

---

3. The pre-registered gate: two decidability atlases for severity

A negative on a question nobody has mapped is worth very little. This program's standing rule (F6) is that the *bidirectional recovery gate precedes every mechanism verdict*: simulate from each arm at the actual sample size, refit all arms, and verify that the generating arm is recovered — and where it is not, commit "undecidable at this N" as the finding rather than silently reporting the in-sample winner. For the severity question this required two new atlases, both known-truth simulation only. Their provenance differs and the difference matters: `out/v6` was sealed **before** any real mark was read, while `out/v11` was built **after** the cross-domain co-movement read had already run — the gap the record itself names, and the reason §5.4's certification is a retrospective query of a sealed grid rather than an in-line gate.

3.1 The single-stream marked atlas (`out/v6`)

`out/v6` races the marked trichotomy — marked null (homogeneous Poisson counts + i.i.d. marks), marked Hawkes (self-exciting counts *and* self-exciting marks: the dominoes mechanism), marked MSM-Cox (a shared latent cascade modulating marks and/or intensity: the weather mechanism) — over a grid in event count `N ∈ {100, 300, 1,000, 3,000, 10,000}`, mark dispersion `σ ∈ {0.5, 1.0, 1.8}` (light, moderate, heavy), and coupling channel (intensity-only, mark-only, both). Mark coupling is held at `θ = 0.6` across the σ axis, so a heavier mark law is a genuinely harder cell (lower regime signal-to-noise `θ/σ`) rather than a differently-parameterized one. 65 cells; replications 6/6/5/3/2 by `N`; the decidability threshold is recovery ≥ 0.8; every arm anchored against brute-force path enumeration in the repository's F1 battery before a number ships.

Four expectations were committed verbatim in the run script and printed to the log before a single cell was fit. Their answers:

M1 — severity is identifiable on its own. The mark-only channel (`κ = 0, θ > 0` — timing plain Poisson) becomes decidable at a finite event count. The mark-only recovery floors are **N = 100 (light), N = 1,000 (moderate), N = 1,000 (heavy)**. "Severity propagates even when timing does not" is a mapped, decidable region.

M2 — the mark law is a real decidability driver, and the floors are the wrong place to read it. The gridded floors for moderate and heavy marks coincide at N = 1,000, so across the top two thirds of the σ axis the floor does not move at all, and reading the driver off the floors would rest on a single light → moderate step of a three-point grid against a 3.3×-spaced `N` grid. The direction is nonetheless real, and it is carried by the **recovery curve at matched N**: mark-only recovery falls from **0.667 at σ = 1.0 to 0.000 at σ = 1.8** at N = 300, and from **1.000 to 0.800** at N = 1,000. A heavier-tailed severity law is strictly harder at matched sample size, exactly as the `θ/σ` argument predicts. *(This reading is a dated correction on the claim surface — the record's own verdict string quotes the coincident floors; the sealed JSON is unchanged and the recovery curve is read from it.)*

M3 — severity and timing are distinct decidability surfaces. The intensity-only (timing) channel recovers at **1.000 at every σ and every N in the grid, including the smallest**, so the timing floor is *at or below N = 100* — censored at the grid minimum, not a measured constant. That is enough for the structural point: the counts channel does not care how the marks are distributed, while the mark channel does, so "does timing propagate?" and "does severity propagate?" are separately decidable questions and neither answer implies the other.

M4 — specificity, and the marked face of the flagship degeneracy. On marked-null truth the coupled arms falsely win in **0.000** of replications, the maximum over all 15 null cells — no severity coupling is manufactured where there is none. Above its floor the marked trichotomy separates: on mark-only truth the shared-regime arm beats the mark-clustering Hawkes arm head-to-head in **97 %** of replications, and the mirror mechanism is itself recoverable (marked-Hawkes truth recovered at **1.000** for N ≥ 1,000). Below the floor it does not separate, and that is the marked face of this program's flagship Hawkes-versus-Cox degeneracy: *which* severity mechanism is operating is a strictly harder question than *whether* severity clusters at all, and the two are reported with different confidence throughout §5.

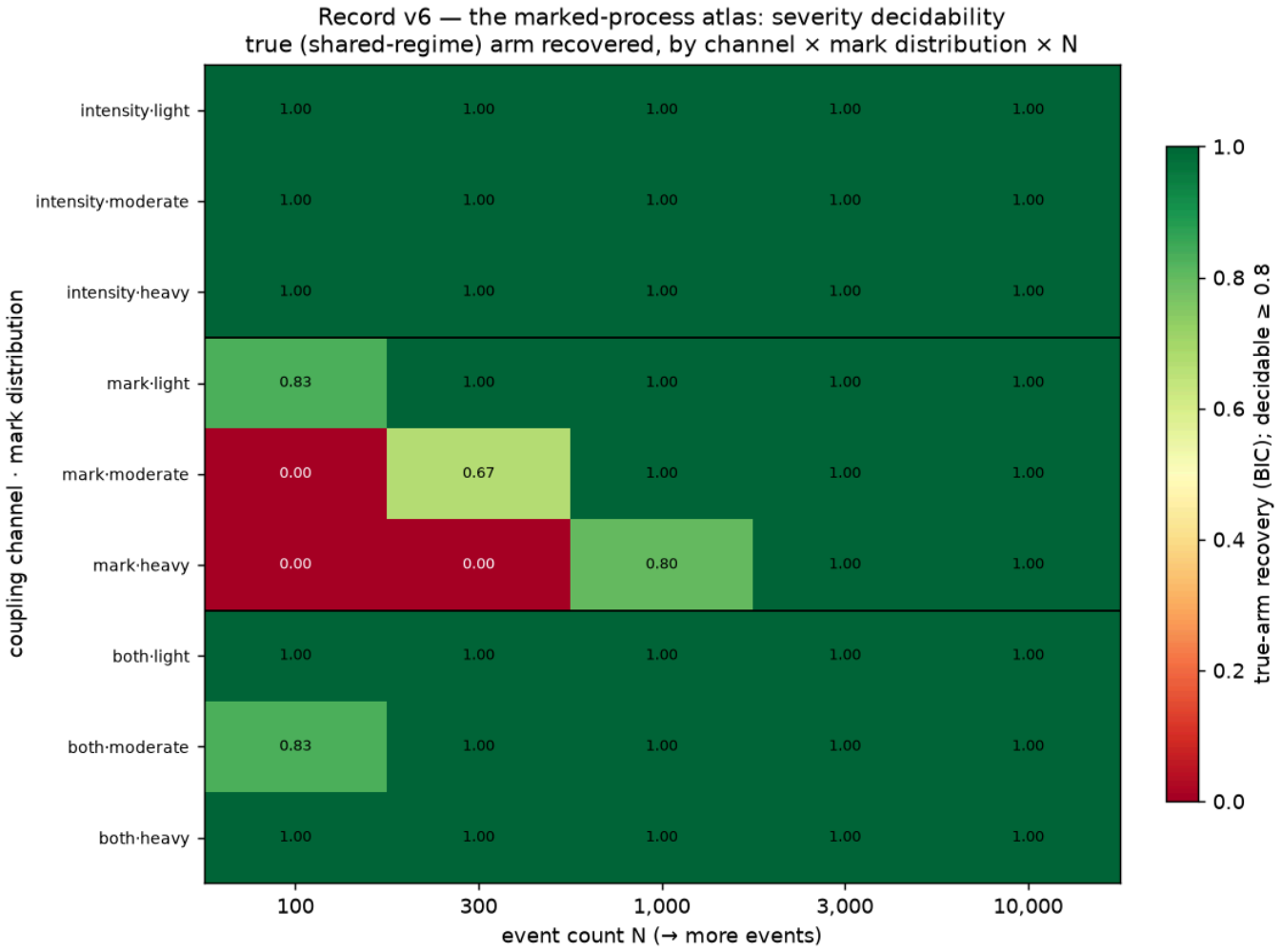

Figure 1. `out/v6/severity_decidability.png` *— the single-stream severity decidability surface. Replication counts per rung run 2–6 and are quoted beside every recovery figure in §5.*

3.2 The cross-domain marked panel atlas (`out/v11`)

`out/v6` maps whether **one** domain's severity clusters. It says nothing about C2, which asks whether a shared latent severity regime drives **many** domains at once — a question that can be answerable from `D` parallel short streams even where no single stream could decide it (the mark-channel analogue of this program's unmarked panel-design gain). Until `out/v11` was built, the cross-domain co-movement read had **no decidability cell to cite** and shipped without the gate every other verdict in this constellation carries. That is recorded in the addendum's own caveats, and `out/v11` is the repair.

Two arms race in the pure severity channel (`κ = 0`, so timing carries nothing): a **panel null** of `D` independent domains (per-domain Poisson counts, per-domain i.i.d. Gaussian marks; 2D+1 parameters) and a **panel regime** in which the `D` domains' mark means share one latent cascade (2D+4 parameters, nesting the null at θ = 0). Grid: `D ∈ {2, 3, 4, 6, 10}`, `N` per domain `∈ {100, 300, 1,000}`, `σ ∈ {0.5, 1.0, 1.8}`, shared-regime strength `θ ∈ {0.3, 0.6, 1.0}`; 65 cells; same 0.8 threshold; anchored against brute-force enumeration in the same F1 battery.

P1 — a shared severity regime is identifiable across domains at all. Yes, and from very few domains: at the reference cell (moderate marks, N/domain = 300, θ = 0.6) the shared-mark panel is recovered from **D = 2**.

P2/P3 — the drivers. The onset `D` is flat at **D = 2** for light and moderate marks at every N/domain in the grid, and moves only on the heavy face, where it falls with sample size: **N/domain = 100 → D = 6; N/domain = 300 → D = 3; N/domain = 1,000 → D = 2**. By shared-regime strength at the reference cell the onset is **D = 2** at θ = 0.30, 0.60 and 1.00 alike — though the θ = 0.30 cell reaches recovery exactly **0.800**, i.e. 4 of 5 replications, sitting precisely on the decidability threshold rather than clearing it. The fitted θ̂ is close to truth throughout (0.293 at θ = 0.30, 0.991 at θ = 1.00), so the arm is measuring what it claims to measure.

P4 — two-front specificity, including the one that matters most. On an independent panel the shared-severity arm falsely wins in **0.000** of replications (5 cells). And on a shared-**timing** panel — counts co-moving across domains, severity i.i.d. — it also fires in **0.000** of replications, with θ̂ ≤ 0.021. A shared timing regime is *not* misread as a shared severity regime. This is the property that licenses reading a cross-domain severity verdict as a statement about severity rather than a re-description of co-moving arrivals, and it is the cross-domain counterpart of the count-channel cancellation of §2.2.

3.3 What the gate refuses

The atlases are used as gates, not as decoration. A verdict ships only from a cell the atlas certifies; where it does not, the committed finding is "undecidable at this N", and this program has published such findings before (the disaster ↔ epidemic cell of `out/v10`, ruled undecidable on exactly this rule after an earlier machinery pass had called it supported). Every real-data read in §5 names the atlas cell it stands on.

---

4. Data, and the tournament as run

The conflict severity mark (open tier). UCDP Georeferenced Event Dataset v24.1 (Sundberg & Melander 2013), `best` fatality estimate per event, global monthly aggregation, **1990–2024**: N = 347,109 events over T = 420 months. The mark law is moderate by the atlas's classification, σ = 0.933; the raw median event is 2 fatalities and the maximum 121,848; 7.9 % of events carry a zero `best` estimate, which the `log(1 + m)` coordinate admits without truncation. This is the one crisis domain with a genuinely open, event-level severity mark, which is why it carries the within-domain result.

The disaster severity mark (registered tier). EM-DAT's impact-threshold subset, Total Deaths per event, 1990–2024: **N = 14,837** events, σ = 1.601 — heavy marks, the harder face of both atlases. The per-event median and maximum are deliberately not quoted here: with an odd N each is one licensed record's Total Deaths field exactly, and the dispersion σ carries the distributional content without disclosing a row. EM-DAT is licensed; its rows never enter a committed file in this repository. Everything reported from it here is a derived aggregate (a monthly mean-severity series, fit margins, a correlation), computed with the owner's file in a git-ignored cache, and the open-tier record remains byte-reproducible from a fresh clone without it. All EM-DAT-derived numbers below are marked **[registered-tier]**.

The roster and the protocol. Three arms — marked null, marked Hawkes, marked MSM-Cox — refit to every stream, BIC computed on a common `n_obs = T` so the comparison is like-for-like, and a held-out log-loss reported beside BIC as an independent check. Each *fitted* arm is fit at **K = 8 seeds** with the best-BIC optimum taken and the multi-seed BIC spread reported as the estimator's own uncertainty; the marked null is closed-form and seed-free, so its single fit carries a spread of zero by construction rather than by measurement. The Hawkes arm is fit by L-BFGS-B on a smooth unconstrained parameterization (§6 explains why this is stated rather than assumed). The mark-shuffle ensemble is **100 surrogates** for the primary and for each cut, fit at the same per-fit budget and with the same converged optimizer as the observed stream. Base seed `20260826`; the numeric record reproduces bit-for-bit.

One asymmetry in that comparison is worth naming rather than leaving for a referee to find. The observed stream is fit at K = 8 seeds with the best-BIC optimum taken; each surrogate gets a single converged fit. The record's justification is that the fit is seed-invariant once converged — which the observed multi-seed spread of **5.893e-06** BIC demonstrates directly — and that a one-start and a two-start fit reach the identical optimum on observed and surrogate streams alike, so the multi-start budget does not bias the excess. That is a measured justification rather than an assumption; it is still a justification, and §9 lists re-testing it as a refutation route.

One fit, two figures. The same observed fit feeds the tournament table and the excess block. This sounds like housekeeping and is not: the earlier pass at this tournament shipped two different fits of the same stream into two different figures, and they disagreed by an amount comparable to the effect being reported (§6).

5. Results

5.1 Within conflict: severity is not exchangeable, decisively and seed-robustly

The atlas cell is certified before the read: at N = 347,109 with moderate marks the mark channel is decidable with recovery 1.000 at the cell the gate snaps to, and so is the both-channel cell. Two things must be said about that certification, and neither was said in an earlier draft.

It is 2 of 2 replications. The certifying cell is quoted to three decimal places, but `out/v6` runs `R = 2` replications at `N = 10,000`; "recovery 1.000" there means two out of two. We quote R beside every recovery figure in this paper from here on, as §5.3 already did, and §8 gives the full 2–6 range.

And its `N` axis is not a pure sample-size axis. The atlas's bin count is derived from `N` as `T = min(N/2, 1500)`, so the top rung `N = 10,000` is `T = 1,500` bins at about 6.7 events per bin, while the real conflict stream is `T = 420` bins at about 826. On the axis that governs how many latent regime transitions are actually observed, the real cell sits *below* the certifying cell, not 34× above it, and it lies nowhere on the mapped grid. The gate's snap to the top rung is therefore a certification along a confounded axis.

The confirmatory check we previously offered in its place does not hold either, and we withdraw it. An earlier draft said that what is *not* confounded is the confirmatory recovery bootstrap, run at the actual σ and T with the optimizer this record uses, which recovers the winner in **1.0** and the null in **1.0** of replications (n = 16, events capped at 8,000) — and that this, rather than the snapped atlas cell, is what carries the decidability of this read. It is not. That bootstrap scored each replication by **full-joint BIC arm selection**: whether the generating arm won the roster's BIC argmin. The *count* channel decides that comparison. Record `out/v12` §(c) reruns the identical idiom on marked-Hawkes streams whose mark coefficient is set to **φ = 0** — i.i.d. marks, the severity channel switched off entirely — and the old criterion still returns a winner-recovery of 1.000. A check that certifies streams with no severity structure at all certifies the count channel, not the severity channel. The event cap compounded it: 8,000 events over T = 420 is about 19 per bin against the real stream's 826, i.e. the check sat on essentially the same events-per-bin rung the paragraph above had just called confounding, while the mark driver `d_t = ȳ_t − m₀` has sampling noise σ/√N_t so the whole test's signal-to-noise is set by N_t.

The criterion is **fixed in the code** as of 2026-08-30: the bootstrap now runs the shipped mark-shuffle surrogate ensemble on every replication and asks whether a severity *excess* is recovered on winner truth and not manufactured on null truth — the statistic the verdict is actually read off — at the real events-per-bin, with the mark loop's stationarity `e^{-β}(1+φ) < 1` asserted before simulating and the simulated streams' realised σ recorded. On the φ = 0 counter-cell streams the corrected criterion finds a severity excess in 0.000 of replications. **But it has not been re-run at the real configuration**, because the real UCDP rows are licensed-fetch-at-run-time and are in no clone of this

repository. The honest position of this section is therefore: the snapped atlas cell is a certification along a confounded axis, the confirmatory bootstrap certified the wrong channel, and **decidability of the severity channel at the real configuration is UNCERTIFIED.** What follows is a measurement against a stated null, reported with its rank; it is not an atlas-certified verdict. Re-running the corrected bootstrap on the owner's data is the first item of §9.

The tournament, on BIC and on held-out log-loss, agrees on the ordering:

arm	parameters	BIC	held-out log-loss	converged	multi-seed BIC spread
marked null	3	1,115,749	2223.59	yes	0 (closed-form; one fit)
marked Hawkes	6	950,538	**1770.80**	yes	5.893e-06 (K = 8)
marked MSM-Cox	7	967,534	1826.23	yes	0 (K = 8)

Both coupled arms beat the null by an enormous margin — and essentially all of that margin is arrivals, not severity. The mark-shuffle decomposition:

arm	observed ΔBIC vs null	count channel (surrogate mean)	severity excess	surrogate-rank p	n at or above (of 100)
marked Hawkes	165,211	157,557	**7,654**	0.0099	0
marked MSM-Cox	148,215	147,088	**1,126**	0.0099	0

The verdict is a decisive, seed-robust departure from exchangeability. For the Hawkes arm the excess is 7,654 BIC — **4.6 %** of the raw margin, the other 95 % being the count/onset-timing channel that the surrogate cancels identically. The observed statistic exceeds **all 100** surrogates; the largest surrogate margin is 157,562.8, so the observation clears the ensemble maximum by more than 7,600 BIC against an ensemble whose standard deviation is about 1.0 BIC. The multi-seed interval on the excess is [7,654, 7,654] and the observed BIC moved **5.893e-06** across the eight seed optima, so the separation is not optimizer noise: the estimator's own scatter is about **1.7×10^5** times smaller than the shuffle-to-shuffle scatter the test is measured against. The same tier holds at the least favourable of the K seed optima, which is what "seed-robust" means here and is reported for every cell below.

The rank is the finding. The magnitude is not diagnostic, and we no longer report it as an effect size. Record `out/v12` §(b) measures what it takes to produce an excess of this size with **no severity dynamics of any kind**: streams at this exact configuration whose monthly mean log-severity follows an *exogenous* path, with marks drawn i.i.d. around it, so nothing in the generator depends on the stream's own past. An exogenous monthly-mean drift of about **0.13 log-units** reproduces the full 7,654, and the literal two-population composition cell — two conflict populations of different mean lethality whose

share drifts slowly — reaches 8,872 BIC with a fitted $|\hat{\varphi}|$ of 16.44, i.e. the same order as the real primary's 12.21. A drift of that size is trivial against between-country conflict lethality spread, and it is exactly the compositional alternative §5.6(a) names as the front-runner. So the 7,654 establishes **non-exchangeability of the marks** — which is a real and falsifiable finding — and establishes nothing about the size of any severity mechanism. Every comparison this paper previously built on the magnitude (the 230× disaster ratio, the 220× null-calibration ratio) is withdrawn with it.

And $|\hat{\varphi}|$ is not stable enough to be a scale either. §5.5 withdraws θ as a scale partly because it runs 0.00095, 0.0346, 0.045, 0.0573, 0.1157 across nested subsets of the same stream — two orders of magnitude. The identical instability sits in the coefficient of the *winning* arm, the one this section's mechanism discussion is about. Across the five nested cuts $|\hat{\varphi}|$ runs **12.21, 26.61, 27.45, 1.81, 0.93** — a 30× range on nested subsets of one stream. By the standard this paper already applies to θ, $|\hat{\varphi}|$ is not stable enough to be a scale, and the mechanism reading below should be carried with that attached as well as with the unrecoverable sign.

An external corroboration, from a different design. The reading that conflict severity is generated by within-conflict escalation rather than by an exogenous shared regime is not only ours. Working on within-conflict data for civil and interstate wars 1946–2008, and by an entirely different route, Clauset, Walter, Cederman & Gleditsch find that **escalation dynamics — variation in fighting intensity within a conflict — are the primary mechanism producing large conflicts**, with civil wars tending to de-escalate once very large while interstate wars carry a persistent risk of continued escalation. That is a falsifiable quantitative claim about how conflict *severity* is generated over time, in a domain where such claims are rare, and it points the same way as the self-excitation reading below. It is corroboration by convergence, not replication: different window, different estimator, different level of aggregation, and — importantly — a within-episode mechanism where ours is a stream-level one.

The mechanism — and why we do not name its direction. The Hawkes arm wins the head-to-head by 16,996 BIC and carries the larger excess, so the arm that best describes the data is the one in which the mark mean depends on the recent history of marks, $m_t = m_0 + \varphi \cdot \Sigma\ \text{decay} \cdot (\bar{y} - m_0)$, rather than on a shared latent regime. It is tempting to call that "self-excitation" and say that large magnitudes raise the hazard of the next. **We do not, because the sign of $\hat{\varphi}$ is not recoverable from the sealed record.** The coefficient is *unconstrained in sign* in the estimator — unlike the count-kernel coefficient α, which is softplus-constrained non-negative — and the record stores only its absolute value, $|\varphi| = 12.21$. A negative $\hat{\varphi}$ fits a large-then-small alternation and would raise the excess exactly as persistence does. So what is established is a **dependence** of monthly mean severity on its own recent history, decisively and seed-robustly; whether that dependence is positive feedback or mean reversion is a question the sealed record cannot answer, and recording the signed $\hat{\varphi}$ is a one-line change we name as required next work rather than guess at. This matters downstream: §5.6's discussion of which way a recording artifact would bias the result turns on exactly this sign.

What the atlas offers in support of even the arm-selection reading is narrower than it may look, and we state the gap. M4 shows the trichotomy separating above its floor, but it measures that separation at the atlas's *own* mark-coupling strength (θ = 0.6), not at whatever strength operates here; the gate as coded certifies MSM-Cox cells and snaps N = 347,109 to the grid's top rung of 10,000, so it never certified the marked-Hawkes optimizer the winner is read off — a limitation this program's own audit of the first pass recorded. The confirmatory recovery bootstrap that an earlier draft offered here as the thing that does apply directly is withdrawn above: it scored full-joint arm selection, which the count channel decides. Against that, this program's flagship result is that Hawkes and Cox descriptions of the same clustering are separable only in part of the parameter space, both marked arms carry an excess here, and the sealed record's own caveat says the mechanism "can be undecidable at strength". **That monthly mean severity is not exchangeable is the robust finding; which model describes the departure is weaker; and the direction of the dependence is not established at all.**

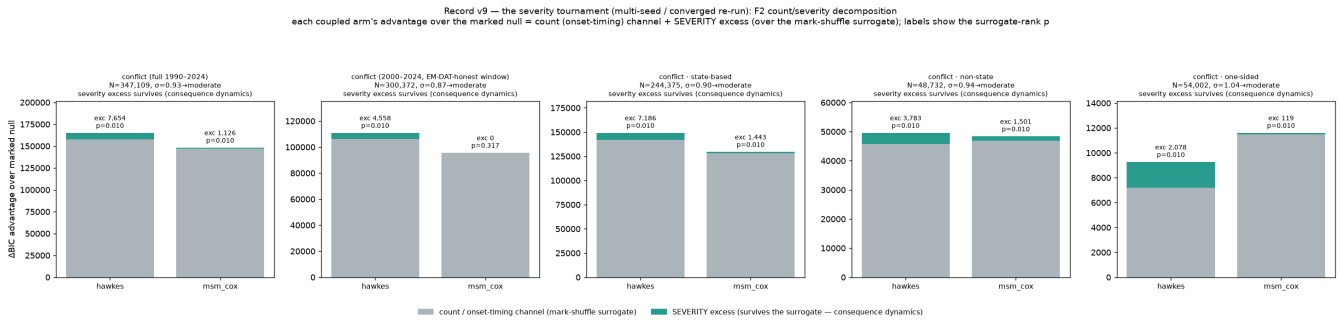

Figure 2. `out/v9/severity_decomposition.png` *— the F2 decomposition for the conflict primary. The tall grey bar is the count channel, present identically in the observed stream and in every surrogate; the excess on top of it is the statistic this section reports. §5.1 and* `out/v12` *state what that excess does and does not establish.*

5.2 The five conflict cuts: multiplicity, and what the p-values can and cannot say

The primary is accompanied by four pre-declared cuts of the same stream. All five read the same way:

cut	N	σ	severity excess (Hawkes)	as % of raw margin	rank p	worst-seed p	seed-robust
conflict, full 1990–2024 (primary)	347,109	0.933	**7,654**	4.6 %	0.0099	0.0099	yes
conflict, 2000–2024 (EM-DAT-honest window)	300,372	0.872	**4,558**	4.1 %	0.0099	0.0099	yes
conflict · state-based	244,375	0.903	**7,186**	4.8 %	0.0099	0.0099	yes
conflict · non-state	48,732	0.942	**3,783**	7.6 %	0.0099	0.0099	yes
conflict · one-sided	54,002	1.044	**2,078**	22.4 %	0.0099	0.0099	yes

Three statements about multiplicity are owed to a reader here, and we make them rather than leave the table to imply more than it contains.

First, these are not five independent tests. The four cuts are nested or overlapping subsets of the primary stream — a shorter window and three violence-type partitions — so they are a *robustness* surface, not replication. Their agreement shows that the effect is not carried by one window or one violence type; it does not multiply the evidence. The most informative cell in the table is the last, though not for the reason a glance at N would suggest — the **non-state** cut is the smallest by event count (N = 48,732) and the one-sided cut is second. What distinguishes the one-sided cut is that it is the **heaviest-marked** (σ = 1.044) and has **by far the smallest count-channel margin** (7,190 BIC of surrogate mean, against 157,557 in the primary), and it nonetheless carries the *largest* fraction of its raw margin in the severity channel, 22.4 %. That is the pattern one would expect if the severity dynamics are real and the count channel is merely large where the stream is large.

Second, `p = 0.0099` is a resolution floor, not an estimate. It is exactly 1/101, the smallest value an add-one rank test against 100 surrogates can return, and every cell in the table sits on it because in every cell the observation exceeded all 100 surrogates (`n_at_or_above = 0`). The tier "decisive" is defined in this program as p < 0.01, which a 100-surrogate ensemble clears only just; with 99 surrogates the floor would be 1/100 = 0.010 and the same observations would read "marginal". **The tier is sensitive to the ensemble size, and no reader should treat these five p-values as five independent small numbers.** What is *not* floor-limited is the separation itself: the observed statistic clears the surrogate maximum by thousands of BIC in every cut, which is the quantity to argue with.

Third, we do not convert that separation into a σ or a tail probability. The earlier pass at this tournament did, and the resulting "13σ" was withdrawn as uncalibrated (§6). The rank is what the test rests on; the BIC separation is reported as a descriptive distance; no normal tail is asserted.

5.3 Within disasters: decidable, and near-negligible [registered-tier]

Run with the owner's EM-DAT file, on the same protocol and behind the same atlas gate — though with much less headroom than the conflict cell had. N = 14,837 is about **1.5×** the atlas cap of 10,000, not 34× as in §5.1; the snapped cell (mark channel, heavy marks, N = 10,000) reads recovery 1.000 on two replications (`R = 2`, as the conflict cell's certifying cell also is — §5.1), and the heavy mark-only floor itself sits at N = 1,000 on recovery exactly 0.800 (4 of 5 replications):

arm	observed ΔBIC vs null	count channel	severity excess	rank p	n at or above (of 100)
marked Hawkes	9,403	9,369	**33.6**	0.010	0
marked MSM-Cox	9,944	9,944	**0.1**	0.079	7

The rank p's are quoted at the addendum's own precision. They are the same add-one ranks as §5.2's, on the same 100-surrogate ensemble: the Hawkes cell's 0.010 is the identical 1/101 = 0.0099 floor, and the MSM-Cox cell's 0.079 is 8/101. §5.2's caution about the resolution floor applies here unchanged. We no longer quote the ratio of this excess to conflict's as an effect-size comparison: per §5.1 and `out/v12`*, neither magnitude is diagnostic of a severity mechanism, so their ratio is not either.*

Disaster severity departs from exchangeability only **faintly**: the Hawkes excess is decidable by rank but amounts to 33.6 BIC — under 0.4 % of its own raw count-channel margin, and under 0.5 % of conflict's 7,654 on the identical protocol. The MSM-Cox arm finds essentially nothing (0.1 BIC, p = 0.079, seven surrogates at or above the observation). This is on heavy σ = 1.601 marks that both atlases identify as the hardest face of the grid.

And this cell should be read against a null calibration — which now exists, and is not the one an earlier draft used. That draft measured 33.6 BIC against a known-null calibration "whose false excesses topped out near 23.5 BIC", making the disaster excess "about 1.4×" a ceiling. **That ruler is withdrawn.** It had no generating script, no pinned seed, no committed JSON and no sealed record anywhere in this repository; every occurrence of it was prose, and the occurrence in the tournament's own sealed record was a string constant in the runner, i.e. a number printed without being computed.[1] It was also measured under the superseded truncated optimizer and applied under the converged one, at conflict scale and applied at disaster scale.

The calibration is now run and sealed as `out/v12` §(a): true-null streams — the mark channel exchangeable by construction — at **both** real configurations, under the converged optimizer, from documented seeds. The measured false-excess ceiling is **+7.19 BIC** at the conflict configuration (N = 347,109, σ = 0.933, T = 420) and **+4.02 BIC** at the disaster configuration (N = 14,837, σ = 1.601, T = 420) — a single-digit BIC scale. The withdrawn figure was about 3.3× too large at the configuration it was nominally measured at, and about 5.9× too large at the configuration it was actually applied to, in the direction that *understated* how far this cell sits above a true-null baseline: against the disaster configuration's own measured ceiling the 33.6 BIC excess is about **8.4×** the ceiling, not the 1.4× the withdrawn ruler implied. That makes it a real departure from exchangeability rather than a marginal one. What it is *not* is a mechanism or a magnitude, for the reasons §5.1 gives.

No mechanism is read from this cell, and none should be. A 33.6 BIC excess against a competing arm's 0.1 is not a mechanism separation; `out/v6` simulates marked-Hawkes truth **only at moderate marks**, so the heavy face has no Hawkes-truth mirror cell to certify a recovery there; and unlike the conflict primary, no confirmatory recovery bootstrap was run for the disaster cell. The sealed record's standing caveat — that the mechanism can be undecidable at strength — binds here with far more force than at the conflict cell, where §5.1 does venture a reading. What this cell licenses is the *presence* of a faint excess, at a rank the ensemble can just resolve, and nothing about its cause.

This is a **corrected** reading and supersedes an earlier one. The pre-correction pass, fitting the Hawkes arm with a truncated optimizer, reported disaster severity as *exchangeable in time* — flat, no excess at all. The converged fit finds a small but rank-decidable excess instead. The two readings differ in kind, not just in magnitude ("no clustering" versus "faint clustering"), and the change is a consequence of the estimator fix of §6, not of any change in data or protocol. Both the superseded and the corrected readings are in the repository's record ledger; the corrected one is the number to cite.

5.4 Across domains: no shared severity regime [registered-tier]

This is the polycrisis headline in the consequence channel, and it is a negative.

Over **288 common-support months**, the detrended contemporaneous correlation of conflict and disaster monthly mean log-severity is **−0.1147** (precision-weighted −0.0994) against a **5,000-shift** circular-shift null with mean +0.0081 and standard deviation 0.0791 — one-sided positive-coupling **p = 0.9492**, two-sided p = 0.1358, −1.55 standard deviations on the wrong side of zero. The ensemble's *resolution* is lower than 5,000 suggests and we say so: the shifts are drawn with replacement from the at most `T − 1 = 419` distinct circular shifts available, so 5,000 draws are a heavily duplicated sample of 419 distinct surrogates. For a negative this is harmless — a duplicated ensemble cannot fabricate a non-significant result — but the quoted p has 419-fold granularity, not 5,000-fold, and enumerating all 419 exhaustively would be both cheaper and exact. The ±6-month lead–lag scan finds no lag at which the correlation turns positive; it ranges from −0.142 (conflict leading by 5 months) to −0.050 (conflict lagging by 5 months), with the contemporaneous cell at −0.115.

There is no shared severity regime across these two domains. Severity does not co-move: not contemporaneously, not at any lag in the scan, and not in the direction the claim requires. And the lag scan's own multiplicity works in the negative's favour — thirteen chances to find a positive correlation produced none, so the reading is safe against exactly the multiple-comparison objection that would undermine a positive.

The atlas certification, and its four limits. The operative cell is D = 2 domains with per-domain event counts of 347,109 and 14,837 — both above the panel atlas's top rung of 1,000 per domain — at moderate (σ = 0.933) and heavy (σ = 1.601) marks. `out/v11` certifies decidability there: on its *heaviest* face, D = 2 at N/domain = 1,000 recovers the shared regime at 1.000 — on **three replications** (`R = 3` at that rung; §8), which is the whole of the evidence behind the three decimal places. So this is a **decidable** negative rather than an undecidable one — for the instrument the atlas maps, at that replication count.

Five conditions attach to that certification and we state them as prominently as the certification itself.

- **P4's timing-leak result is structural, not a measurement about this statistic.** The panel-regime arm hard-fixes `κ = 0`, so it *cannot express* count coupling at all; its 0.000 leak rate on a shared-timing panel follows from the parameterisation rather than from anything measured, and it says nothing about a correlation of `Σy/N` ratios, which

the atlas never computes. The real-data statistic has exposures the panel likelihood does not: the sampling error of `ȳ_t` scales as `σ/√N_t`, so co-moving counts co-move the two series' precisions, and in real data a month's event composition — hence its mean log-severity — is not independent of its count. **The specificity of the correlation statistic against co-moving counts is unmapped.** The cheap missing check is the correlation of the two domains' detrended monthly *counts*, which neither the record nor this paper reports.

- **P4 is also one-sided.** It measures false-*positive* rates only: 0.000 on an independent panel and 0.000 on a shared-timing panel. Whether co-moving counts could *mask* a real shared severity regime is not mapped — `out/v11` contains no cell crossing a shared timing regime with a true shared severity one. Both specificity families are mapped only at moderate marks and N/domain = 300 (five cells each), not on the heavy face the operative cell occupies.
- **It was applied after the fact.** The cross-domain read was run before `out/v11` existed — its own caveats say so, in the record — and the certification here is obtained by querying the sealed panel atlas at the operative cell, not by re-running the co-movement read with the gate attached in-line. Attaching the gate to the *existing* correlation statistic would not move the correlation reported here; the higher-powered re-run named in §9 runs a different arm (the panel likelihood) and would produce different numbers. Either way the provenance is retrospective, and the gated re-run is the honest next increment, registered in this program's prompt ledger rather than claimed here.
- **The atlas certifies a likelihood arm; the real-data read uses a correlation.** `out/v11` maps the recoverability of a shared severity cascade by the *panel likelihood*; the statistic actually computed across domains is a detrended correlation against a circular-shift null, which is a lower-powered instrument. The certification therefore bounds what the panel likelihood could have seen, and the correlation can only have seen less. This makes the negative *conservative in mechanism* but *weaker in power* than the atlas cell implies, and it is the second reason the gated re-run is the named next step.

5.5 The effect-size floor, stated [registered-tier]

A negative is only as strong as the smallest effect it could have detected, and this one has a number — with its coordinates, which matter as much as the number.

The panel atlas's weakest gridded shared-regime strength is **θ = 0.30**, and at D = 2 that cell reaches recovery exactly **0.800** — the threshold, on four of five replications, not a comfortable margin. Nothing below θ = 0.30 is mapped at all. Two caveats attach to reading it as *the* floor for this negative, and both are extrapolations rather than measurements:

- **The θ axis is mapped at one (σ, N/domain) point only** — moderate marks, N/domain = 300. There is no θ = 0.30 cell on the heavy face anywhere in the atlas, and the operative panel is *mixed*, with one heavy-marked domain (EM-DAT, σ = 1.601). On the heavy face even the *stronger* θ = 0.6 fails at D = 2 below N/domain = 1,000

(recovery 0.000 at N/domain = 100 and 0.600 at N/domain = 300), which is why the certification of §5.4 leans on the real panel's very large per-domain counts. Reading $\theta \gtrsim 0.3$ as the floor at N/domain $\gg$ 1,000 on a mixed-σ panel extrapolates along both axes.

- **The atlas certifies the panel likelihood; the statistic run is a correlation.** As §5.4 says, the instrument actually applied is lower-powered than the one the grid maps, so $\theta \gtrsim 0.3$ is the floor for a test that has not yet been performed. The operative floor for the test that *was* performed is at least that, and is not itself mapped.

An earlier draft offered a scale for that floor — the conflict fit's own mark-coupling coefficient, $\theta = 0.045$, about 7× below the weakest mapped cell. **We withdraw it as a scale**, because three things are wrong with the comparison and each is enough on its own.

- **It comes from the arm this fit rejects.** θ is the shared-regime (MSM-Cox) coefficient; Hawkes beats that arm by 16,996 BIC.
- **It is not the same parameter as the atlas's θ, despite sharing a symbol.** The single-stream MSM-Cox arm fits `κ` freely and here returns $\hat{\kappa} = 0.676$, so the latent path its θ loads on is driven overwhelmingly by the count channel; the panel-regime arm defines θ with `κ ≡ 0`, a purely mark-driven cascade. θ answers "how much do conflict marks load on a conflict-*count*-driven regime", which is a different question from "how strong is a shared severity cascade".
- **It is not stable enough to be a scale.** Across the paper's own five cuts of the *same* stream, θ runs 0.00095, 0.0346, 0.045, 0.0573, 0.1157 — two orders of magnitude on nested subsets. And `out/v6`'s *zero*-mark-coupling cells (marked-null and intensity-only truth) return $|\theta|$ up to **0.039** even at the grid's largest N, so a fitted 0.045 is not established as distinguishable from zero at all.

What survives is only the direction of the comparison, stated without a multiplier: whatever mark coupling a shared-regime description attributes to these streams is small in absolute terms, and the region the cross-domain grid maps begins well above it. Producing a real scale would mean fitting the panel arm itself on the real marks — the same gated re-run §5.4 and §9 name.

So the claim this paper licenses is precise, and narrower than "polycrisis severity coupling does not exist":

> Between global monthly conflict severity and global monthly disaster severity, 1990–2024, there is no shared severity regime of the strength the panel atlas maps — $\theta \gtrsim 0.3$ at its reference cell, by the panel-likelihood instrument the atlas certifies. Weaker regimes, and regimes only a lower-powered correlation statistic could have been expected to catch, are outside the mapped region and this result is silent about them.

That sentence, with its qualifications intact, is what this paper offers in place of an unfalsifiable one. It can be refuted by measurement — with more domains, longer series, or the higher-powered instrument — and §9 says how.

5.6 The two leading alternative explanations, neither excluded

The mark-shuffle null is exchangeability of severity across the whole stream. Rejecting it establishes that the observed monthly allocation of severity departs from a single global law. It does **not** establish, on its own, that anything propagates. Two explanations compete with consequence dynamics for that departure, and we could not exclude either. We set them out at the length the positive result's interpretation deserves.

(a) Compositional heterogeneity — the one this design is weakest against. The surrogate permutes 347,109 marks uniformly, which imposes one severity law on every month of a 35-year, ~190-country stream. But the global conflict stream is a slowly-changing mixture of persistent, heterogeneous sub-streams: which wars are running, in which countries, between which actors, each with its own lethality distribution. Membership turns over on a scale of years. That alone produces months whose mean log-severity departs from the global mean in a temporally correlated way — **with no severity dynamics of any kind**. And the test is exquisitely sensitive to it: at ~826 events per month and σ = 0.933, the shuffle null's standard deviation for a monthly mean is about 0.03 log-units, so composition-driven drift far smaller than the between-conflict spread in lethality is enough to reject exchangeability decisively.

We regard this as **co-leading with recording, and more likely than recording to be the whole story**, for two reasons. It requires no artifact — only that wars differ in lethality and do not all run at once, which is not in question. And it predicts precisely what §5.2 observes: an excess that appears in every cut of the stream, since every cut is still a mixture.

And it is now quantified rather than merely named. `out/v12` §(b) measures the drift-equivalence curve at exactly this configuration: streams whose monthly mean log-severity follows an *exogenous* path, with marks drawn i.i.d. around it, so nothing in the generator depends on the stream's own past. Exogenous drifts with `sd(ȳ)` of 0.03, 0.06, 0.10 and 0.14 log-units produce excesses of 191, 1,354, 4,025 and 9,272 BIC respectively, and a literal two-population composition model — two conflict populations of different mean lethality whose share drifts slowly — reaches 8,872 BIC with a fitted `|φ|` of 16.44, the same order as the primary's 12.21. Interpolating, a drift of about **0.13 log-units** reproduces the full 7,654. Against the between-country spread in conflict lethality that is a small number. This does not show that composition *is* the explanation; it shows that composition is *sufficient*, and that no part of the observed magnitude discriminates against it.

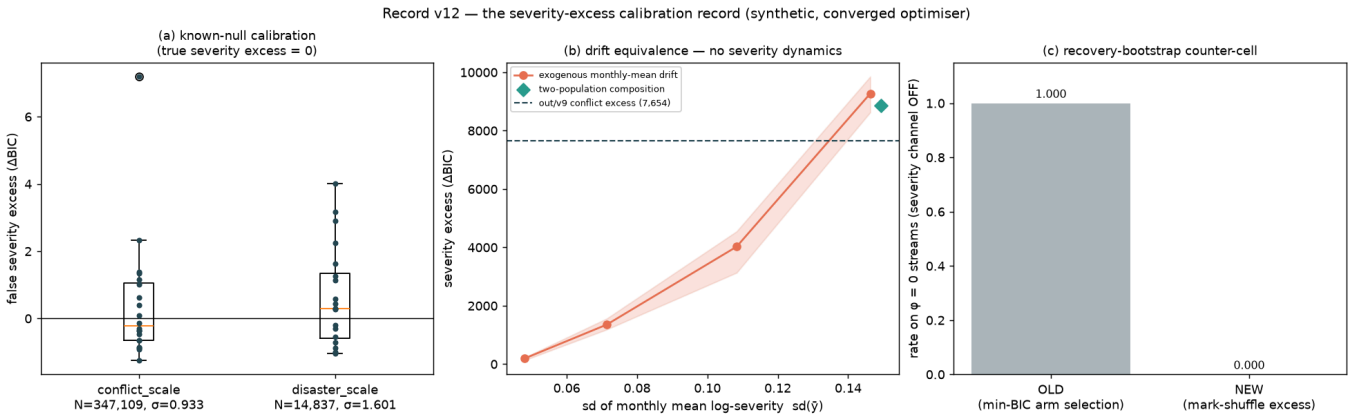

Figure 3. `out/v12/calibration.png` *— left, the false-excess distribution on true-null streams at both real configurations; centre, the drift-equivalence curve, with the sealed conflict excess drawn as a dashed line and the two-population composition cell as a diamond; right, the counter-cell, in which the withdrawn criterion returns 1.000 and the corrected one 0.000 on streams whose severity channel is switched off.*

The compensating fact is that, unlike the recording objection, **this one is checkable now, on open data, and cheaply.** Refit the mark mean with country or dyad fixed effects and ask whether the excess survives; or run the test within-conflict and pool. `out/v4` already established for the unmarked panel that heterogeneous loadings change decidability, and the marked panel atlas's own recorded next cell is the heterogeneous-domain extension. Until one of those is run, **"conflict severity clusters in time" should be read as "the monthly allocation of conflict severity is not exchangeable", with composition as the front-running explanation.** That is a weaker claim than the one this paper set out to test, and it is the claim the evidence supports.

(b) Severity-selective recording. Seismology litigated this objection first, and it deserves to be met head-on. Davidsen & Green (2011) argued that reported magnitude correlations are largely an artifact of **catalogue incompleteness**: after a large earthquake the detection threshold rises, small events are missed, and the *recorded* magnitudes in that window are biased upward — which random labelling then scores as magnitude clustering. The crisis analogue is immediate: after a major crisis event, attention rises, coding effort changes, and the set of events reaching the catalogue may shift systematically in magnitude.

Three things we previously offered as defences do not survive scrutiny, and we withdraw them rather than leave them for a referee.

- **The count channel's cancellation excludes less than it appears to.** Holding `N_t` fixed removes the artifact's contribution to the *model comparison* against a homogeneous-count null. It does not remove the artifact's imprint on the *marks*, which is what the excess measures — and coverage variation is not separable into a count part and a mark part in the first place. A month in which coverage improves gets more events, and the extra events are the marginal, smaller ones, so `ȳ_t` falls exactly when `N_t` rises. A rate-varying reporting regime is therefore **not** excluded; it is one of the ways the artifact would act.

- **We cannot say which way the artifact would bias the result.** The earlier version of this section argued that elevated attention pulls more small events into the catalogue, producing large-then-small structure and biasing the excess *downward*. That argument is wrong twice over. The excess is a two-sided contrast against exchangeability, and — per §5.1 — the mark coefficient φ is unconstrained in sign, so large-then-small structure raises the excess exactly as persistence does. The direction of the bias is therefore undetermined, and would remain so even if the artifact's own direction were known.
- **The violence-type partitions are not independent source ecologies.** The three sum exactly to the primary (244,375 + 48,732 + 54,002 = 347,109): they are one catalogue, coded by one project under one `best`/`low`/`high` convention from one sourcing pipeline, split on a type-of-violence field. A catalogue-wide recording artifact operates in all three *by construction*, so their agreement is not evidence against one. What the agreement does show is the narrower thing §5.2 already claims — that the effect is not carried by one violence type.

What genuinely bounds the recording objection is thinner. The **2000–2024 cut** removes the steepest part of the crisis-reporting ramp and the excess survives it at 4,558 BIC, 4.1 % of that cut's raw margin — a stationarity check, not a completeness control. And Xiong et al. (2023) answer the incompleteness objection with *laboratory* catalogues, where completeness is controlled by construction; crisis data has no laboratory, and that defence cannot be borrowed.

What would settle either explanation. For composition: a fixed-effects or within-conflict refit, and the heterogeneous-domain panel the atlas already names as its next cell. For recording: a refit under an explicitly time-varying magnitude-detection threshold; a dual-coded severity comparison against an independently coded conflict catalogue; and this program's wired-but-unrun vintage control on UCDP's un-revised monthly release, which would show directly whether retrospective revision creates the structure. None is run here. The positive result of §5.1 should be carried with all of them attached.

The cross-domain negative's exposure is different, and it is two-sided. We previously wrote that a shared attention or reporting regime could only inflate a *positive* co-movement, so a negative was robust to it. That is not right, and the observed data are the reason to say so. Attention is rivalrous in the newswire both catalogues are coded from: a month dominated by a large conflict can crowd out disaster coverage, depressing recorded disaster completeness and hence recorded mean severity in that month — a **negative** cross-domain correlation that would *mask* a real positive coupling. The observed correlation is negative at the contemporaneous cell and at all thirteen lags, which is consistent with a rivalrous-attention signature as well as with an absence of coupling. The honest statement is that a shared ramp would inflate a positive, a rivalrous regime would induce a negative, and this design distinguishes neither from the absence of a shared severity regime.

6. The v7 → v9 correction

The result in §5.1 is the second pass at this tournament. The first pass (`out/v7`) reported the same direction with a different magnitude, a σ-tier that does not exist, and two of five cuts whose verdicts flipped on an arbitrary seed. The episode is reported in full because the failure mode is one that marked-point-process surrogate designs are generically exposed to.

What went wrong. The marked-Hawkes arm was fit by a Nelder–Mead simplex over six parameters at an iteration cap of 400. It never converged — `converged: false` on the headline arm in the primary and in all four cuts, while the null and MSM-Cox arms converged everywhere. Nothing in the record's caveat list, the findings ledger, or the repository README mentioned convergence at all. In a program whose entire discipline is stated caveats, an unstated one on the headline arm is the breach; the audit that found it is the repair.

What that cost, quantitatively. An adversarial re-check recomputed every headline quantity from the sealed record at full precision and instrumented the shipped call graph. Three findings:

1. **The record contained two fits of the same stream, at the same budget, 2,426 BIC apart** — differing only in the multi-start seed. That gap was 46 % of the reported excess and 6.1× the surrogate standard deviation the excess was being divided by. The two published figures inherited it: they disagreed by exactly that amount, one printing each fit.
2. **The σ-tier was uncalibrated and was withdrawn.** The denominator was the surrogate-to-surrogate reproducibility of a *truncated* fit, not the sampling error of the observed statistic. Measured directly on synthetic streams matched to the primary cell and run through the shipped code verbatim, roughly **96 % of the surrogate spread was optimizer noise**: holding the fit seed fixed collapsed the ensemble standard deviation from ≈ 1.31 to ≈ 0.06. A "13σ" whose σ is mostly estimator scatter asserts a normal tail no quantity in this program could support, and the companion arm's z-scores of **502 to 1,114** were worse. The tier was withdrawn and the surrogate **rank** — which is what the tail test had actually been resting on — was promoted to the verdict statistic.
3. **Two of five cuts flipped on the seed** and were quarantined pending a re-run: the one-sided cut and the 2000–2024 window.
4. **And the withdrawal took two further cuts' tiers with it.** State-based and non-state had each read `decisive` — but at 40 surrogates apiece, where the add-one rank floor is 1/41 = 0.024 and the rank alone cannot reach a $p < 0.01$ tier. Their escalation had run *entirely* through the withdrawn z, so their tiers fell with the σ. Four of the five cuts' published tiers were affected, not two, and a section claiming to tell the correction "in full" has to say so.

What survived the audit unchanged. The *direction* of the primary. Both known optima of the observed stream exceeded all 100 surrogates, and the published number had used

the *worse* of the two optima, so the error direction was conservative. It was a calibration failure, not an inflated finding.

One reassurance offered here at the time is withdrawn. That paragraph also cited "a 24-stream known-null calibration at the real scale" producing "0 of 24 false decisives, with false excesses topping out near 23 BIC against a reported effect roughly 220× larger". No such calibration exists in this repository: no script, no seed, no JSON, no sealed record — prose only (§5.3, footnote). It is replaced by `out/v12` §(a), which measures the false-excess ceiling under the converged optimizer at both configurations and finds **+7.19 BIC** at the conflict configuration — a single-digit scale, about a third of the withdrawn figure. The measured calibration supports the same *direction* (no true-null stream manufactures anything of the observed size), and the "220×" ratio is not restated, because §5.1 withdraws effect-size ratios built on the excess magnitude altogether.

What the re-run changed. Three fixes, all of which ride in the code rather than in the prose: the Hawkes arm moved to L-BFGS-B on a smooth unconstrained parameterization (opt-in, so the sealed atlas still regenerates bit-for-bit); every arm fit at K = 8 seeds with best-BIC taken and the multi-seed spread reported as the uncertainty; and one shared observed fit feeding both the tournament table and the excess block, so the two figures can no longer disagree. On the real conflict cell the Hawkes arm now converges and reaches the same optimum from every seed — the multi-seed BIC spread collapses from 2,426 to **5.893e-06**.

What the re-run found. Both quarantined cuts resolve to the same decisive, seed-robust positive as the rest. The primary's converged magnitude is 7,654 — *larger* than the truncated fit's 5,262, precisely as the audit predicted, since a better optimum on the observed stream can only widen the gap to a surrogate ensemble that was already being fit to its own optimum. And the disaster cell's reading changed in kind (§5.3): what the truncated fit reported as exchangeable severity is, converged, a faint but rank-decidable excess.

The generalisable lesson. A surrogate test compares an observed fit to an ensemble of surrogate fits. If the optimizer's scatter is comparable to the between-surrogate scatter, the test is measuring the optimizer. The diagnostic is cheap and we recommend it as routine for any marked-point-process surrogate design: **fit the observed stream at K seeds and compare the multi-seed spread to the surrogate ensemble's standard deviation.** Here that ratio is now about 6×10^{-6} to 1. In the first pass it was 6 to 1, in the wrong direction, and nothing in the output would have said so.

The superseded record remains sealed on disk beside the corrected one; neither was edited, which is what allows the two to be compared quantity by quantity.

---

7. What this does *not* say

- **It does not reopen the onset-time question.** The count channel's very large advantage over the null is *within-domain* arrival clustering in conflict; it is held

identical by the mark-shuffle surrogate — the count data are the same on both sides of the difference — and it is not the cross-domain onset quantity `out/v1`, `out/v3` and `out/v5` tested. Those verdicts stand exactly as published. A severity result within conflict does not resurrect an onset-time coupling, and the cross-domain severity negative does not add to the onset-time negative — different channels, designed to be.

- **It does not say polycrisis is false.** It says that one specific, stated, measurable form of the claim [registered-tier: the disaster half of the comparison is EM-DAT-derived, so this bound is not reproducible from a fresh open clone] — a shared severity regime of the strength the panel atlas maps ($\theta \gtrsim 0.3$ at its reference cell, by the panel-likelihood instrument, with the two extrapolations §5.5 names) between global monthly conflict and disaster severity, 1990–2024 — is not present in these data. Systemic-risk interdependence, longer-horizon chains, domain pairs not tested here, coupling in second moments rather than means, and coupling below or beside the mapped region are all untouched.
- **It does not establish that severity propagates within conflict either.** The robust finding is narrower: the monthly allocation of conflict severity is not exchangeable given the counts. Two explanations for that departure are not excluded (§5.6) — a slowly-changing mixture of heterogeneous conflicts, which requires no dynamics at all, and severity-selective recording — and the first is the front-runner.
- **It does not establish a mechanism, a magnitude, or even a direction.** The best-fitting arm is the mark-history one, but the atlas does not certify that arm at this cell, the confirmatory recovery bootstrap that was offered in its place certified the count channel and is withdrawn (§5.1), both coupled arms carry an excess, the fitted $|\hat{\varphi}|$ runs $12.21 \to 27.45 \to 0.93$ across nested cuts of one stream, and the *sign* of the mark coefficient is not recoverable from the sealed record — so "self-excitation" is a word this paper declines to use for its own result. `out/v12` adds that an excess of this size is reproduced by streams with no severity dynamics at all, so the magnitude carries no mechanism information either.
- **It is not a forecast, a warning, or a ranking.** Every number is a retrospective model-comparison on public, aggregate data. No country, no period, and no domain is scored, and no posterior regime path is published as a rankable quantity (F7).
- **It does not claim a new statistical instrument.** The mark-shuffle surrogate is random labelling; the arms are standard marked point processes. What is new is the application, the gate, and the stated floor.

---

8. Limitations

The mean channel only. Both atlases and the tournament modulate the **mean** log-severity: a hot regime raises the typical magnitude. That is the cleanest and most anchorable reading of "severity level", but the canonical object of the multifractal machinery used here is the *volatility* of magnitude, and a scale/volatility modulation — hot months being more *dispersed* in severity rather than uniformly deadlier — is a sibling

channel that is not tested. It is registered as the next measurement in this program. A shared cross-domain severity regime acting on second moments would be invisible to every number in this paper.

Composition is not conditioned out anywhere. The single largest limitation of the within-domain result is §5.6a's: the surrogate tests exchangeability against one global severity law, and a heterogeneous, slowly-turning mixture of conflicts rejects that with no dynamics at all. Nothing in this paper conditions on country, dyad, or conflict episode. This is the first thing a reader should discount the positive by, and the first thing a follow-up should run.

One open severity mark. Conflict is the only crisis domain with an open, event-level severity mark. The within-domain positive rests on it alone; the cross-domain negative rests on exactly two domains, one of them licensed. `D = 2` is the minimum panel the atlas certifies, not a comfortable one, and the panel-design gain the atlas measures is precisely the argument for running this test again with more domains.

Gaussian log-marks against a heavier real tail. The marks are modeled as lognormal magnitudes; real crisis severity is heavier-tailed than lognormal. What the atlases measure about this is narrow and we do not stretch it: `out/v6` shows that a heavier mark law **lowers recovery at matched N** (mark-only recovery 0.667 → 0.000 at N = 300, 1.000 → 0.800 at N = 1,000) and therefore **raises the decidability floor**. That is a statement about *detectability*, not about the magnitude a measured BIC excess would take under a different mark law, and we make no prediction of the second kind in either direction. The consequence that does follow is asymmetric between the two verdicts: a raised floor makes the **cross-domain negative less conservative than it looks**, since a Pareto severity law would need more domains or more events to reach the same verdict, while for the within-domain positive a heavy-tailed refit is simply an open question — §9 lists it as a refutation route precisely because the answer is not implied by anything measured here. The disaster stream, at $\sigma = 1.601$, is already on the heavy face of both grids.

Grid resolution and Monte-Carlo noise in the gates. Both atlases report recovery fractions estimated at 2–6 replications per cell, and both onsets are resolution-limited: the true onset lies between the last undecidable and the first decidable rung. Two cells that carry weight in the argument sit exactly on the 0.8 threshold — the heavy-mark single-stream cell at N = 1,000 (0.800) and the $\theta = 0.30$ panel cell at D = 2 (0.800, i.e. 4 of 5 replications) — where a single replication would move them. They are quoted as thresholds met, not as margins.

Vintage. UCDP GED is annually revised, so the conflict series is a retrospective corpus edited with post-hoc knowledge, not a real-time vintage. The one genuinely out-of-sample control available to this program — UCDP's monthly Candidate Events release — is wired but not yet run against the severity channel, and is registered as a separate measurement. Nothing here is described as held-out in the vintage sense.

The reporting channel — and the limit of what the surrogate buys. Detection and definition are crisis data's photon channel, and this program's standing rule is that no

claim ships without a declared artifact channel. The mark-shuffle surrogate holds the counts and the *global* marginal severity distribution exactly fixed, so a reporting regime that changes *how many* events are recorded is differenced out. It does **not** hold the *local* marginal fixed, so a reporting regime that changes *which magnitudes* are recorded, or how they are estimated, over time is not controlled — and that is precisely the mechanism whose seismological version has been argued to explain away magnitude correlations entirely. §5.6 treats it as the leading alternative explanation for the within-conflict positive, states the partial checks that survive it, and names the three controls that would settle it. For the cross-domain read the exposure is different, and it is two-sided rather than smaller. Linear detrending blocks a shared *linear* ramp only; and while a shared attention ramp would inflate a *positive* co-movement, a **rivalrous** attention regime — one crisis crowding another out of the same newswire — induces a *negative* one and could mask a real coupling. §5.4 and §5.6 carry that correction; an earlier draft of this paper claimed the confound ran in one direction only, and it does not.

The cross-domain read's provenance and instrument, as stated in §5.4: a retrospective atlas certification, and a correlation statistic where the atlas maps a likelihood.

Single aggregation. Both cross-domain series are global monthly aggregates. The country × domain panel — the cross-section the aggregation discards — is where this program's onset-time work found its power, and the severity analogue of that design has not been run.

---

9. What would refute this, and the agenda it opens

The point of writing a claim down is that someone can take it away. We therefore state the refutation conditions for each of this paper's three results, in the form a disagreeing analyst would need.

To confirm or dissolve the within-conflict positive (C1). The finding as stated — the monthly allocation of conflict severity is not exchangeable — is robust. What is open is *why*, and one thing is open before that: whether the cell is decidable at all.

- *the decidability certificate itself, first.* §5.1 withdraws the confirmatory recovery bootstrap that was carrying this read: it scored full-joint BIC arm selection, which the count channel decides, and `out/v12` §(c) shows it returns a winner-recovery of 1.000 with the mark coefficient set to zero. The criterion is fixed in the repository's code — it now bootstraps the mark-shuffle severity excess at the real events-per-bin, asserting the mark loop's stationarity before simulating — but **it has not been run at the real configuration**, because the real UCDP rows are fetched at run time and are in no clone. Running it is a single command on the owner's data and is the first item on this agenda. Until it is run, the positive is a measurement against a stated null and not an atlas-certified verdict.

Then the two routes that would settle *why*, in rough order of cheapness:

- *composition* — the front-running alternative (§5.6a). Refit the mark mean with country or dyad fixed effects, or run the test within-conflict and pool, and ask whether the excess survives once the mixture is conditioned out. This needs only open data and the existing code path. If the excess vanishes, the paper's positive becomes a statement about conflict heterogeneity rather than about severity dynamics, and we would report that;
- *the direction of the dependence* — record the **signed** $\hat{\varphi}$ (the estimator stores only `|φ|`) for the primary and every cut. Until that is done, nothing in this program distinguishes persistence from alternation in the mark channel, and §5.6's discussion of which way a recording artifact would bias the result cannot be closed;

and then the channels the design was built to exclude:

- *the count channel* — show that the mark-shuffle surrogate fails to hold the counts fixed in some respect that matters, or that the excess is a fixed leak from the count margin. It is not: across the five cuts the count channel spans a factor of **twenty-two** (157,557 BIC in the primary down to 7,190 in the one-sided cut) while the excess fraction moves the other way, 4.1 % to 22.4 %. That is not a clean dissociation — the excesses themselves are ordered roughly as the cuts' sample sizes are — but it rules out a constant proportional leak;
- *the estimator* — show that the observed fit's advantage over the surrogate fits comes from an asymmetry in fitting budget or convergence. This is exactly the failure that produced the first pass's withdrawn tier, so the diagnostic is published with the result: the observed multi-seed BIC spread is 5.893e-06 against a surrogate ensemble standard deviation near 1.0 BIC, and observed and surrogate are fit at the same per-fit budget with the same converged optimizer. The one asymmetry that remains is §4's — K = 8 observed seeds against one converged fit per surrogate — and the direct test is to refit the whole ensemble at K seeds and see whether the excess moves;
- *the mark law* — show that the Gaussian log-mark assumption creates the excess. The direct test is to refit with a Pareto or generalized-Pareto mark law and ask whether the excess survives. Nothing measured here predicts the outcome: `out/v6` establishes only that a heavier tail lowers recovery at matched N and raises the decidability floor, which is about detectability rather than about the size of a measured excess. That makes this a genuinely open route rather than a formality;
- *the recording process* — the strongest route, and the one seismology used (§5.6). Show that the catalogue's magnitude completeness varies with recent activity in a way that reproduces the excess: refit under a time-varying magnitude-detection threshold, compare against an independently coded conflict catalogue, or run the un-revised monthly vintage. Any of the three could eliminate the result, and we would report it.

To refute the cross-domain negative (C2), the routes are cleaner still, because a negative bounded by a stated floor is refuted by any positive above that floor:

- *more domains* — the panel atlas's own gain result says the shared-regime onset falls with `D`. Adding epidemic, economic, or food-price severity marks to the panel raises power directly, and the atlas already maps what each additional domain buys;

- *the panel likelihood instead of the correlation* — run the `out/v11` panel regime arm on the real marks rather than the reduced correlation statistic. This is the higher-powered instrument the atlas actually certifies, and it is the named next increment in this program's ledger;
- *the country × domain panel* — the aggregation used here discards the cross-section, and a shared severity regime that acts on a subset of countries would be diluted to invisibility by global aggregation;
- *the volatility channel* — a shared regime acting on severity *dispersion* rather than severity *mean* would be invisible to this design, and the sibling measurement is registered;
- *a weaker regime, honestly bounded* — any demonstration of a shared severity regime with $\theta < 0.3$ would fall in the region this design does not map, and would refute nothing here while still being a real finding. That asymmetry is the reason the floor is in the abstract.

The agenda. What this paper contributes to the polycrisis field is less the verdicts than the shape of the object: a claim with a null, a gate, a floor, and a refutation condition. Three consequences follow for the field's own programme.

First, severity marks belong in the corpora. The field's harmonized cross-domain corpus deliberately excludes event magnitudes. On the evidence here, magnitudes are where the only decisive departure from a null in this program's entire severity and onset-time programme has been found — even if §5.6 leaves open whether that departure is dynamics or composition — and a corpus without marks cannot ask the question at all. A marked corpus is also what would *settle* that ambiguity, since composition is only conditionable when the sub-stream labels travel with the marks. A harmonized, marked, cross-domain crisis corpus is a tractable and high-value data contribution.

Second, decidability should be reported before verdicts, not after. The most consequential single number in this paper is not the excess; it is the D-onset of the panel atlas, because it says how many domains an analyst needs before a cross-domain severity verdict means anything. A field whose central claim is about cross-domain interaction and which has never published a power curve for that interaction is a field in which no negative can be believed and no positive can be checked. The two atlases here are released for reuse and are queryable programmatically.

Third, the plausible form of the claim should be tested before the strong one is defended. Onset-time entanglement is the version of the claim that headlines carry and it is the one this program has found fragile three times. Consequence propagation is the version the field's own mechanism vocabulary supports, and it now has a first measurement: real, and domain-local.

An open invitation to adversarial collaboration on this program's discrimination apparatus is published separately in this repository; it extends to every refutation route above.

---

10. Data and code availability

Open tier — reproducible from a fresh clone. UCDP GED v24.1 is fetched at run time by the repository's fetch script and never committed. The three records this paper reads are committed in full: `out/v6/` (the single-stream marked atlas, with `atlas_marked.json`, its `RECORD.md`, seeds, and three figures), `out/v11/` (the marked panel atlas, `atlas_marked_panel.json`, `RECORD.md`, three figures), and `out/v9/` (the severity tournament, `severity.json`, `RECORD.md`, two figures). The calibration record `out/v12/` (`calibration.json`, `RECORD.md`, one figure) is entirely synthetic and reads no data at all; it regenerates from its own pinned seed with `python x10_severity_calibration.py`. Each `RECORD.md` carries the exact regeneration commands and the pinned base seed (`20260825` for `out/v6`, `20260826` for `out/v9` and `out/v11`); each numeric record reproduces bit-for-bit from its seed on the stated environment. The superseded first pass at the tournament remains sealed at `out/v7/` and is not edited.

Registered tier — reproducible only with a licensed file. The EM-DAT disaster-severity results of §5.3 and §5.4 require the EM-DAT public export, which is licensed and is never committed to this repository. With that file placed in the git-ignored data cache, the addendum regenerates from its own pinned seed. Only license-clean derived statistics appear here: the event count, the mark dispersion σ, the monthly mean-severity series, the fit margins and their surrogate decompositions, the cross-domain correlations and their null moments, and the lag scan. Per-event order statistics — a median or a maximum, each of which discloses one record's field exactly — are not repeated in this paper, and, as of 2026-08-30, no longer appear in `out/v9/emdat_addendum.md` either: that file published two of them, in breach of its own header and of this program's Standing Rule 7, and both have been redacted in place under a dated intervention block, with the report writer fixed so no future run re-emits them. That is the only in-place edit any sealed record in this repository has received; the repair, its authority, and the record-immutability manifest that now pins the corrected bytes are documented in `FINDINGS.md`.

Verification. The repository gate runs a vendored-code checksum pin, an anchor battery in which every likelihood used here is verified against brute-force path enumeration and known-truth recovery, a compile sweep, and a citation-consistency check. That check includes a number-anchor registry which binds a chosen set of open-tier headline numbers in this paper to the sealed record JSON each is read from, and fails if any drifts. It is a **drift detector on a registered subset**, not a coverage proof: a number that is not registered is not checked, which is how the withdrawn calibration ruler of §5.3 survived a green gate for as long as it did. The registered-tier numbers cannot be bound this way at all, because no EM-DAT-derived value is committed; they are marked in place and traced to the addendum.

Deposit DOIs (published 2026-08-31). This paper and every record it reads are deposited separately.

- This paper — `10.5281/zenodo.22180511`.

- Record `out/v6` (the single-stream marked atlas) — `10.5281/zenodo.22180519`.
- Record `out/v9` (the severity tournament, converged multi-seed re-run) — `10.5281/zenodo.22180521`.
- Record `out/v10` (the cross-domain coupling matrix, machinery-fix re-run) — `10.5281/zenodo.22180523`.
- Record `out/v11` (the marked panel atlas) — `10.5281/zenodo.22180525`.
- Record `out/v12` (the severity-excess calibration record) — DOI reservation pending; it is deposited in the same batch as the records above and this paper's citation of it is by record id until the identifier is assigned.

Each assigned DOI above is **published**: the deposits went live on Zenodo on **2026-08-31** and every assigned identifier resolves. The `out/v12` bullet is the one exception — no identifier for it is recorded in this repository, and its citation stays by record id until one is. The source repository is private and will remain so, so each deposit is standalone and carries its own reproduction materials; deposit metadata deliberately omits repository links. *(Corrected 2026-08-31 — publication day: until today this paragraph read "reserved on Zenodo and its deposit is a private draft … the record is not yet published", with " (Zenodo DOI reserved; publication pending)" wherever a DOI appeared. That staging state ended when the owner pressed Publish on 2026-08-31.)*

Pre-registration. Both atlases' expectations (M1–M4 for `out/v6`, P1–P4 for `out/v11`) are committed verbatim in their run scripts and were printed to the run log before any cell was fit. The severity tournament's expectation is committed in the same way and is quoted in `out/v9/RECORD.md`.

References

(Verified against the actual sources before release — Standing Rule 11. Bibliographic metadata for every DOI below was confirmed against Crossref, with OpenAlex and Semantic Scholar as cross-checks; the two grey-literature bylines and the Xiong et al. surrogate construction were confirmed by reading the sources directly. Where two candidate papers are commonly conflated, the one not cited here is named so a reference manager cannot silently substitute it.)

- Brosig, M. 2025. How do crises spread? The polycrisis and crisis transmission. *Global Sustainability* 8, e11. doi:10.1017/sus.2025.14. (Self-described as "a hypothesis-proposing study" that "does not offer empirical testing which is left for future research".)
- Clauset, A., Walter, B. F., Cederman, L.-E., & Gleditsch, K. S. 2026. Escalation dynamics and the severity of wars. *Science Advances* 12(32), eaeb1705. doi:10.1126/sciadv.aeb1705. (Preprint: arXiv:2503.03945, March 2025. The corroborating conflict-severity anchor of §5.1. **Not** to be confused with Clauset, A. 2018, Trends and fluctuations in the severity of interstate wars, *Science Advances* 4(2), eaao3580 — the "long peace" paper, which addresses a different question.)

- Clinet, S. 2021. Quasi-likelihood analysis for marked point processes and application to marked Hawkes processes. *Statistical Inference for Stochastic Processes* 25(2), 189–225. doi:10.1007/s11203-021-09251-7 (online first 2021; print issue 2022; preprint arXiv:2001.11624). (Single-authored; the asymptotic theory under which a marked-Hawkes quasi-likelihood estimator is well behaved.)
- Corral, Á. 2006. Dependence of earthquake recurrence times and independence of magnitudes on seismicity history. *Tectonophysics* 424(3–4), 177–193. doi:10.1016/j.tecto.2006.03.035. (The null-side anchor: recurrence times depend on history, magnitudes do not. **Not** Corral, Á. 2006, Universal earthquake-occurrence jumps, correlations with time, and anomalous diffusion, *Physical Review Letters* 97, 178501, which concerns waiting times and spatial jumps, not magnitude correlation.)
- Davidsen, J., & Green, A. 2011. Are earthquake magnitudes clustered? *Physical Review Letters* 106(10), 108502. doi:10.1103/PhysRevLett.106.108502. (The incompleteness-artifact objection adopted in §5.6 as this paper's leading alternative explanation.)
- Delannoy, L., Verzier, A., Bastien-Olvera, B. A., Benra, F., Nyström, M., & Søgaard Jørgensen, P. 2025. Dynamics of the polycrisis: temporal trends, spatial distribution, and co-occurrences of national shocks (1970–2019). *Global Sustainability* 8, e24. doi:10.1017/sus.2025.10008. (The field's most quantitative study; magnitudes excluded by the authors as "inconsistent across datasets", no null model reported.)
- Diggle, P. J. 2013. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3rd ed. Chapman & Hall/CRC. doi:10.1201/b15326. (Random labelling as a formal null.)
- Filimonov, V., & Sornette, D. 2015. Apparent criticality and calibration issues in the Hawkes self-excited point process model. *Quantitative Finance* 15(8), 1293–1314.
- Grabarnik, P., Myllymäki, M., & Stoyan, D. 2011. Correct testing of mark independence for marked point patterns. *Ecological Modelling* 222(23–24), 3888–3894. doi:10.1016/j.ecolmodel.2011.10.005. (Mark-permutation envelopes can be miscalibrated under an inhomogeneous or clustered ground process — the calibration caveat of §2.2.)
- Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58(1), 83–90.
- Homer-Dixon, T., Renn, O., Rockström, J., Donges, J. F., & Janzwood, S. 2022. A call for an international research program on the risk of a global polycrisis. (Cascade Institute, Technical Paper 2022-3.)
- Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. 2008. *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley. doi:10.1002/9780470725160. (§5 sets out the random-labelling versus independent-marks distinction the count-channel cancellation of §2.2 relies on.)
- Lawrence, M. 2022. Polycrisis: why we must turn this meme into a big idea. Cascade Institute, 12 December 2022. (Single-authored; a shorter version appeared in *The Conversation Canada*, 11 December 2022. Frequently misattributed — the byline is Lawrence's, not Homer-Dixon's.)

- Lawrence, M., Homer-Dixon, T., Janzwood, S., Rockström, J., Renn, O., & Donges, J. F. 2024. Global polycrisis: the causal mechanisms of crisis entanglement. *Global Sustainability* 7, e6. doi:10.1017/sus.2024.1. (The definition and the three-pathway taxonomy — common stresses, domino effects, inter-systemic feedbacks — this paper operationalizes; it contains no null, no estimator, and no data. With Lawrence et al. 2024, *Polycrisis Research and Action Roadmap*, Cascade Institute, the source of the call for "formal models of crisis interaction".)
- Lippiello, E., de Arcangelis, L., & Godano, C. 2008. Influence of time and space correlations on earthquake magnitude. *Physical Review Letters* 100(3), 038501. doi:10.1103/PhysRevLett.100.038501. (**Not** Lippiello et al. 2007, *Physical Review Letters* 98, 098501, "Dynamical scaling in branching models for seismicity", which is a different result.)
- Lippiello, E., Godano, C., & de Arcangelis, L. 2012. The earthquake magnitude is influenced by previous seismicity. *Geophysical Research Letters* 39(5), L05309. doi:10.1029/2012GL051083. (Note the author order differs from the 2008 paper.)
- Ogata, Y. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83(401), 9–27.
- Petrillo, G., & Zhuang, J. 2022. The debate on the earthquake magnitude correlations: a meta-analysis. *Scientific Reports* 12, 20683. doi:10.1038/s41598-022-25276-1. (Twenty-nine papers; concludes the question is not yet answerable at the required standard. See also Petrillo & Zhuang 2023, Verifying the magnitude dependence in earthquake occurrence, *Physical Review Letters* 131(15), 154101, doi:10.1103/PhysRevLett.131.154101.)
- Rakowski, J. J., et al. 2025. Characterizing the global polycrisis: a systematic review of recent literature. *Annual Review of Environment and Resources* 50, 159–183. (One dynamic model among 59 analyzed publications.)
- Sundberg, R., & Melander, E. 2013. Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research* 50(4), 523–532. (Release used here: UCDP GED v24.1.)
- Taroni, M. 2024. Are the magnitudes of earthquakes in Southern California, with incompleteness removed, correlated? *Geophysical Journal International* 236(3), 1596–1600. doi:10.1093/gji/ggae007. (Single-authored; a deflationary result.)
- Xiong, Q., Brudzinski, M. R., Gossett, D., Lin, Q., & Hampton, J. C. 2023. Seismic magnitude clustering is prevalent in field and laboratory catalogs. *Nature Communications* 14, 2056. doi:10.1038/s41467-023-37782-5. (The closest sibling of the surrogate used here: the magnitude-difference distribution compared against the same quantity computed after randomly shuffling the order of the catalogued events. The prior art for the mark-shuffle design; the laboratory arm of its incompleteness defence has no crisis analogue.)
- Zscheischler, J., & Seneviratne, S. I. 2017. Dependence of drivers affects risks associated with compound events. *Science Advances* 3, e1700263. doi:10.1126/sciadv.1700263. (The compound-extremes template for testing cross-

driver dependence against an explicit independence null — cross-variable within one domain, and the acknowledged precedent for the null-hypothesis form used here.)

Full program reference list: `SPECS.md` *§10.*

---

1. The withdrawn ruler is recorded here rather than silently dropped because a reader of the previous draft would otherwise be unable to tell what happened to it. It was unregenerable by anyone, including the author, and it is a direct counterexample to this paper's own provenance sentence — which is why §10's statement of that discipline is also narrowed below. ↩