Generic early-warning signals are at chance under a matched false-alarm protocol across brain, heart, grid, climate and molecular data

Miroslav Šotek
Anulum Institute
ORCID 0009-0009-3560-0851 protoscience@anulum.li

August 2026

Abstract

Generic early-warning signals (EWS) — the rising variance and lag-one autocorrelation of critical slowing down, and related synchronisation and ordinal-entropy indicators — are widely reported to precede abrupt transitions in the brain, the heart, power systems and the climate. Most of that evidence rests on a retrospective, per-record test: a rising trend measured over one pre-transition window and compared to surrogates of the same record. We evaluate the operational question instead: at a fixed false-alarm budget, does an early-warning detector fire on transitions more often than on no-transition controls? One domain-adaptable detector suite and one matched-false-alarm harness are applied unchanged, through per-domain adapters, to four independent labelled corpora — scalp-EEG seizures, cardiac atrial-fibrillation onsets, power-grid growing oscillations, and palaeoclimate abrupt transitions — and every result is scored with a label-permutation significance test. No detector reaches significance in any domain (every best-member $p \geq 0.05$); the sparse detection observed is what the matched false-alarm rate produces by chance. The canonical literature detector — the Dakos et al. 2008 AR(1)-Kendall-τ trend — run through the same protocol on the same segments fares no better: on its own palaeoclimate records it leads 0 of 6 transitions ($p = 1.00$), and its strongest showing anywhere is 3 of 6 on scalp EEG ($p = 0.067$). Extending the protocol to a molecular fifth modality — dynamical-network-biomarker (DNB) indices — reaches the same conclusion under modality-appropriate nulls: the single-cell transition index of Mojtahedi et al. 2016 clears a matched operating point on 1 of 3 lineages ($p = 0.27$), and the bulk GSE2565 index does not beat a surrogate null granted the same module-selection freedom as the analysis ($p = 0.39$). Because the corpora are small (3–12 transitions per domain), these nulls bound the demonstrated skill rather than proving early warning impossible; what they establish is that the operational bar is materially stricter than the retrospective per-record test on which most published EWS evidence rests. Every alarm and non-detection is sealed in content-addressed, byte-reproducible evidence records; a companion manuscript shows a domain-specific detector clearing the identical bar where a deterministic instability signature exists.

1 Background and question

The theory of critical transitions predicts that a dynamical system approaching a bifurcation recovers more slowly from perturbations, which shows up as a rising variance and a rising lag-one autocorrelation of an observable — critical slowing down (Scheffer et al. 2009). The same idea

motivates rising-synchronisation and ordinal-transition-entropy indicators. These generic early-warning signals have been reported before epileptic seizures, cardiac arrhythmias, power-system instability, ecological collapse and abrupt climate change.

Almost all of that evidence answers a retrospective, per-record question: *within one pre-transition window, is there a statistically significant rising trend of the indicator, relative to surrogate time series of that same record?* (e.g. the Kendall-τ trend test of Dakos et al. 2008). This question is valuable, but it is not the question an operator or clinician faces. The operational question is: *if I set an alarm threshold that fires on at most a fixed fraction of no-transition situations, does the detector then fire on genuine transitions more often than that fraction?* — a matched-false-alarm question, followed by a significance test on the transition hit-rate.

This paper builds the machinery to answer the operational question, applies it across four independent physical domains and one molecular modality with one shared protocol, and runs the canonical literature detector through the identical protocol as a same-segments head-to-head. The result is a null, and the gap between the two questions is the finding.

2 Methods

2.1 The detector suite

One suite of three passive detectors reads a neutral observable bundle and emits per-window scores:

- **Critical slowing down** — the larger of the robust (median/MAD) z-scores of the windowed variance and lag-one autocorrelation of the observable, against a leading baseline.
- **Rising synchronisation** — the robust z-score of the Kuramoto order parameter of a population of phase oscillators.
- **Ordinal-transition entropy** — the robust z-score of the (negated) ordinal-pattern transition entropy of the phase field.
- **Weighted fusion** — the weighted mean of the three members.

The single-series domain (palaeoclimate) carries one scalar observable, so only critical slowing down applies; the multi-node domains (EEG, ECG, grid) carry a population of oscillators and use all four.

2.2 The matched-false-alarm harness

For each domain, a per-domain adapter turns the raw recording into the neutral observable bundle. Each transition is scored on a fixed pre-onset segment ending at the annotated onset, so every window is pre-onset and any alarm is a genuine lead. A no-transition null — a stable, transition-free stretch — is cut into non-overlapping trials of the same length. Each detector's alarm threshold is set continuously to the tightest value holding the trial false-alarm rate at or below a target of 10 % (the quantile of the null alarm scores, with no grid ceiling, so a variance-heavy detector is never silently clipped). A transition is *led* when the detector alarms within its pre-onset segment. The achieved false-alarm rate is recorded alongside every threshold.

2.3 The permutation significance test

A detection count leaves one question open: does the detector lead transitions more often than the matched false-alarm rate explains? We answer it with a label-permutation (exchangeability) test. Each segment — transition or null — alarms or not at the calibrated threshold. Under the null hypothesis that transition segments are indistinguishable from nulls, the assignment of the "transition" label over the pooled segments is exchangeable. We draw 10 000 random transition-sized subsets of the pooled alarm outcomes (fixed seed, so the p-value is byte-reproducible), build the null distribution of the lead count, and report the one-sided p-value with an add-one correction.

2.4 The competitor head-to-head

The canonical literature detector is Dakos et al. 2008: the Kendall rank correlation τ of the windowed lag-one autocorrelation against time — a rising-AR(1) trend. We implement it as a per-segment score, calibrate a matched-false-alarm τ threshold on the null segments exactly as above, and run the same permutation test on the resulting alarms. Because the two detectors read the same one-dimensional signal (the detrended proxy residual for palaeoclimate, the cross-node Kuramoto order parameter for the multi-node domains), the head-to-head is a same-segments, same-budget, same-test comparison in which only the detector differs.

2.5 Corpora (citation-only)

Domain	Corpus	Transitions	Null
Brain (scalp EEG)	CHB-MIT, chb01 (Shoeb 2009, PhysioNet)	6 seizures	interictal records
Heart (ECG)	MIT-BIH AFDB (Moody & Mark 1983)	6 AF onsets	sinus stretches
Grid (PMU)	PSML 23-bus (Zheng et al. 2021)	12 growing osc.	damped disturbances
Palaeoclimate	Dakos et al. 2008 records	6 of 8 evaluated	stable pre-approach

Table 1: The four physical corpora.

All raw data are public and cited; none is redistributed. Only derived, hash-sealed, byte-reproducible evidence records are committed. Each detector's alarm — or non-detection — is sealed into a content-addressed (canonical-JSON SHA-256) record, so the result, including every sealed non-detection, is auditable and reproduces bit-for-bit from a fresh run.

2.6 The molecular fifth modality (DNB)

Dynamical-network-biomarker early warning (Chen et al. 2012) is a different modality: few timepoints, many genes, and a cross-sample statistic computed over the sample population at each timepoint, so the sliding-window suite does not apply. Its index rises to a peak at the tipping point. Two published index forms are implemented, each with a modality-appropriate honest null:

- **Single-cell (Mojtahedi et al. 2016).** The transition index is the ratio of mean signed gene–gene to mean signed cell–cell correlation over the curated panel (which *is* the critical module, so no selection step exists). We take the per-lineage index trajectory and its bootstrap standard error from the published supplement (not re-derived from the confounded raw qPCR); the pre-transition rising limb (Days 0, 1, 3) is scored by its least-squares slope; the matched-false-alarm null is a temporally-shuffled surrogate resampled from the published

mean and SE; and the three lineages' alarms are tested by the same label-permutation core as the physical domains.

- **Bulk (GSE2565 phosgene lung injury).** Here the module must be *selected*, which introduces selection freedom: the composite index $\mathrm{SD}_{\mathrm{in}} \cdot \mathrm{PCC}_{\mathrm{in}}/\mathrm{PCC}_{\mathrm{out}}$ is evaluated on a module chosen to peak at the transition, so a rise is partly guaranteed by construction. The null therefore re-selects the entire module on each surrogate: it shuffles the timepoint labels across the arm's samples and reruns module selection from scratch, so the surrogate has the same selection freedom as the real analysis. The observed rising slope is ranked against these selection-controlled surrogates for a one-sided p-value.

Both DNB paths seal hash-addressed, byte-reproducible records exactly as the physical domains do.

3 Results

3.1 The suite across four physical domains

At a matched false-alarm rate (target 10 %), detection is sparse everywhere, and no detector reaches significance in any domain (Table 2).

Domain	Best member	Led / N	Permutation p
Brain (EEG)	critical slowing down	2 / 6	0.215
Heart (ECG)	synchronisation	2 / 6	0.219
Grid (PMU)	critical slowing down	3 / 12	0.195
Palaeoclimate	critical slowing down	1 / 6	0.514

Table 2: Best suite member per domain at a matched 10 % false alarm.

The fusion never leads more transitions than its best single member, so there is no robust fusion advantage. The observed lead counts are within what the matched false-alarm rate produces by chance (expected counts 0.6–1.7 across the domains).

3.2 Head-to-head: suite vs Dakos AR(1)-Kendall-τ

Running the canonical Dakos detector through the identical protocol on the identical segments gives Table 3.

Domain	Suite best (led/N, p)	Dakos AR(1)-Kendall-τ (led/N, p)
Palaeoclimate	CSD 1/6, $p = 0.514$	0/6, $p = 1.000$
Grid	CSD 3/12, $p = 0.195$	1/12, $p = 0.713$
Heart	sync 2/6, $p = 0.219$	0/6, $p = 1.000$
Brain (EEG)	CSD 2/6, $p = 0.215$	3/6, $p = 0.067$

Table 3: Same-segments, same-budget, same-test head-to-head.

On its own palaeoclimate records — the data Dakos et al. analysed — the AR(1)-Kendall-τ detector, at a matched operating point, leads zero of six transitions ($p = 1.00$): it does not beat

chance, and it does no better than the suite. The same holds for grid and heart. The one exception is scalp EEG, where the rising-AR(1) trend is the strongest signal in the study (3 of 6, $p = 0.067$) — marginally short of significance, and better than the suite on that domain.

3.3 Reading the null

At a matched operating point with a permutation significance test, no detector reaches significance at these corpus sizes — neither the suite nor the canonical literature detector, in any of the four domains. Two readings follow. First, the corpora are small: a modestly skilled detector could remain non-significant at $n = 6$–12, so the null bounds the demonstrated skill rather than establishing absence of skill (Section 5). Second, independently of power, the same detectors on the same data pass the retrospective per-record test used in the literature and fail the operational one used here; the choice of test changes the conclusion. The closest approach to significance — the AR(1) trend on scalp EEG (3 of 6, $p = 0.067$) — is consistent with the seizure-prediction literature's use of autocorrelation features and warrants a larger EEG corpus.

3.4 The molecular fifth modality

The same conclusion holds when the protocol is carried to a molecular modality (Table 4).

Case	Test	Result	p
Single-cell (Mojtahedi 2016)	matched-FA + permutation, 3 lineages	1 / 3 lineages clear 10 %	0.266
Bulk (GSE2565 phosgene)	selection-controlled surrogate rank	rise does not beat surrogates	0.385

Table 4: The molecular modality under modality-appropriate nulls.

The single-cell transition index rises 2.8-fold toward the erythroid bifurcation — a clear effect size — yet at its published four-timepoint resolution only the strongest of three lineages clears a matched false-alarm operating point, and the corpus-level permutation test is not significant ($p = 0.27$); the binding constraint is the temporal resolution, not the effect size. In the bulk GSE2565 case, the phosgene-exposed arm's apparent DNB rise (slope 1.15) sits below the 90th percentile of surrogates that were allowed to re-select the biomarker module on shuffled time (surrogate-rank $p = 0.385$), and it barely exceeds the air control's own selected-module rise ($p = 0.414$). On this corpus, granting the null the same selection freedom as the analysis absorbs most of the apparent signal — a result consistent with selection freedom explaining much of the rise, subject to the caveats of Section 5. Both molecular records are hash-sealed and byte-reproducible.

4 Discussion

The evaluation gap is the finding. The difference between "there is a detectable rising trend in this record's approach" (retrospective, per-record, vs surrogates) and "the detector fires on transitions more than on controls at a fixed false-alarm budget" (operational, matched, significance-tested) is exactly the gap between the literature's positive results and this study's null. That gap — not any single detector — is what the four-domain replication and the Dakos head-to-head expose. A positive early-warning claim intended for operational use should be required to clear the operational bar; most published EWS results have cleared only the retrospective one. To be explicit about scope: at these corpus sizes the nulls bound the *demonstrated* skill of the detectors tested; they are not evidence that operational early warning is impossible in these domains.

Selection freedom is a hidden cost of self-selecting detectors. The bulk-DNB case adds a specific methodological lesson: when a detector chooses its own module (or feature set) to peak at the transition, an honest null must be granted the same choice. A surrogate that re-selects from scratch on shuffled time absorbs most of the apparent signal on this corpus — so a fair null for a self-selecting detector is not an afterthought but the core of the test.

What survives the null. What this protocol produces even when detection fails is an auditable evaluation object: a matched false-alarm operating point, a permutation p-value, and a content-addressed evidence record for every transition, including every non-detection. Negative operational results of this kind are currently rare in the EWS literature, and they are the necessary baseline against which any claimed operational skill has to be measured. A companion manuscript (same repository) shows the complementary positive case: a domain-specific detector that reads a deterministic physical instability signature — the growth rate of a power-grid electromechanical mode — clears the identical bar decisively, while the same detector idea on a domain without such a signature does not.

5 Limitations

- **Small corpora, low power.** With 6–12 transitions per physical domain — and 3 lineages (single-cell) or 1 exposed arm (bulk) in the molecular modality — the significance tests have limited power; "at chance" bounds the demonstrated skill and does not prove early warning impossible.
- **One competitor.** Only the Dakos AR(1)-Kendall-τ detector was run head-to-head; modern approaches (e.g. learning-based tipping-point predictors) were not tested and could behave differently.
- **Single-subject EEG; within-record climate null.** The EEG corpus is one subject (chb01); the palaeoclimate null is the stable pre-approach interval of each record, which carries that record's own variability and so is conservative.
- **Parameterisation.** One reasonable choice of window, step, baseline and target false alarm was used per domain; a sweep was not performed.
- **DNB caveats.** The single-cell index is taken from the published supplement (its bootstrap SE drives the surrogate), not re-derived from the raw qPCR; the bulk case has one exposed arm, so it is a single-transition surrogate test, not a corpus, and its module search is a greedy reimplementation of the DNB selection rather than the authors' exact procedure. The molecular conclusion is bounded accordingly: no significance at this resolution, and no surrogate-beating rise under a selection-controlled null — not a refutation of the DNB method.

6 Reproduction

Every result regenerates deterministically (no randomness in the pipelines; the permutation test uses a fixed seed). The pipelines are pure Python (NumPy/SciPy stack; pinned lock files ship in the repository); a full re-run of the four physical domains and both DNB cases completes in minutes on a workstation once the cited corpora are downloaded. With each raw corpus in a directory:

```
python bench/early_warning_leadtime_eeg.py     DATA OUT   # brain
python bench/early_warning_leadtime_cardiac.py DATA OUT   # heart
python bench/early_warning_leadtime_grid.py    DATA OUT   # grid
python bench/early_warning_leadtime_climate.py DATA OUT   # palaeoclimate
python bench/head_to_head_ar1_kendall.py OUT \
    --climate-dir C --grid-dir G --cardiac-dir H --eeg-dir E
python -m bench.early_warning_dnb          OUT   # single-cell DNB
python -m bench.early_warning_dnb_bulk     DATA OUT   # bulk GSE2565 DNB
```

The single-cell DNB case reads only the embedded published summary, so it needs no external data. The sealed evidence and aggregate results are committed under `examples/real_data/` in the source repository (https://github.com/anulum/scpn-phase-orchestrator). A fresh run reproduces every `content_hash` bit-for-bit; committed integrity tests recompute each sealed hash from the payload alone.

7 Data availability

Raw corpora are public and cited, not redistributed:

- CHB-MIT Scalp EEG Database — https://physionet.org/content/chbmit/
- MIT-BIH Atrial Fibrillation Database — https://physionet.org/content/afdb/
- PSML power-system dataset (Zheng et al. 2021) — the 23-bus millisecond-level PMU measurements.
- Dakos et al. 2008 palaeoclimate records — https://github.com/earlywarningtoolbox/datasets
- Mojtahedi et al. 2016 single-cell transition index — Table S2 of the paper (index + bootstrap SE per lineage per day).
- GSE2565 phosgene lung-injury expression — https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2565

References

[1] Scheffer M, Bascompte J, Brock WA, et al. Early-warning signals for critical transitions. *Nature* 461:53 (2009).

[2] Dakos V, Scheffer M, van Nes EH, Brovkin V, Petoukhov V, Held H. Slowing down as an early warning signal for abrupt climate change. *PNAS* 105(38):14308 (2008).

[3] Dakos V, Carpenter SR, Brock WA, et al. Methods for detecting early warning signals of critical transitions in time series. *PLoS ONE* 7(7):e41010 (2012).

[4] Shoeb A. Application of machine learning to epileptic seizure onset detection and treatment. PhD thesis, MIT (2009); CHB-MIT via PhysioNet.

[5] Moody GB, Mark RG. A new method for detecting atrial fibrillation using R-R intervals. *Computers in Cardiology* 10:227 (1983).

[6] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* 101(23):e215 (2000).

[7] Zheng X, et al. A multi-scale time-series dataset with benchmark for machine learning in electricity (PSML) (2021).

[8] Chen L, Liu R, Liu Z-P, Li M, Aihara K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Scientific Reports* 2:342 (2012).

[9] Mojtahedi M, Skupin A, Zhou J, et al. Cell fate decision as high-dimensional critical state transition. *PLoS Biology* 14(12):e2000640 (2016).

[10] Sciuto AM, Phillips CS, Orzolek LD, Hensley JL, Chang L-Y, Nadadur SS. Genomic analysis of murine pulmonary tissue following carbonyl chloride inhalation. *Chemical Research in Toxicology* 18(11):1654 (2005); GSE2565.

[11] Mormann F, Andrzejak RG, Elger CE, Lehnertz K. Seizure prediction: the long and winding road. *Brain* 130(2):314 (2007).