A 24-hour M-class flare forecast built from the GOES/XRS event record alone is worth about what recalibration and an eleven-parameter logistic make of it

Calibration and a cheap logistic outrank the hierarchy. Twelve sealed records verify a causal event-record now-cast across calibration, structure, depth, exposure, lead time and flare size — and a second domain, geomagnetic storm counts, returns a negative result: beyond one day, count-only Kp forecasting reduces to 27-day recurrence plus a trailing activity level.

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

Draft — verification paper. It follows the deposited result paper `papers/06-flare-rate-cascade.md` *(concept DOI 10.5281/zenodo.22180472), whose §8 sealed a causal 24-hour M-class now-cast and whose §10 named the re-cuts this paper executes: a recalibration step, a paired bootstrap of the difference rather than two marginal intervals, a per-horizon curve, a real-label X-class run, matched information for the baselines, and a Brier decomposition. Target venue: Space Weather, where the closest precedent (Camporeale & Berger 2025) is resident; the Journal of Space Weather and Space Climate is the in-hand alternative. Zenodo DOI first, then arXiv (physics.space-ph / astro-ph.SR). Every quantitative claim traces to a named, immutable results record under* `out/` *: the flare arc* `v44` *–* `v52` *(Phase 12) and the geomagnetic arc* `v53` *–* `v55` *(Phase 13). GOES/XRS flare event lists: NOAA National Centers for Environmental Information and the Space Weather Prediction Center. Kp: GFZ Helmholtz Centre for Geosciences, Geomagnetic Observatory Niemegk. Dst: World Data Center for Geomagnetism, Kyoto. Figure regenerated by the sealed record that produced it. This paper's own deposit: Zenodo* `10.5281/zenodo.22760517` *(preprint; DOI reserved 2026-09-15; published 2026-09-15 on the owner's sign-off).*

Status note (2026-09-15): the records this paper reads are published on Zenodo and are cited by DOI in §Data availability. The paper itself has no DOI and no priority date yet; deposit, arXiv posting and a domain-collaborator pass are owner actions.

Abstract

Solar flare forecasting is dominated by magnetogram-derived predictors, and the operational systems that use them have repeatedly been found hard to separate from each other or from cheap baselines. That makes a narrower question worth answering precisely: **how much 24-hour M-class forecast skill is in the flare event list by itself, and what part of it is bought by model structure rather than by scaling?** A companion paper fitted an exact hierarchical regime filter — a 2^6-state Markov-switching multifractal on a Poisson counting channel — to fifty years of GOES/XRS daily flare counts and turned it into a causal, filter-only now-cast. This paper verifies that forecast against the field's conventions, on the same sealed trace, with every difference carried by a paired 90-day-block bootstrap on one common panel of resamples rather than by comparing two marginal intervals. The ordering of the answer is the result.

(i) Recalibration is worth more than the hierarchy's entire margin over persistence. A two-parameter Platt map fitted on the training block alone — **a = −0.369440**, **b = +0.924339** — moves the sealed forecast's Brier Skill Score on the 7,305-day verification tail from **+0.3385** to **+0.3645** against that tail's own climatology, a paired **+0.026 [+0.018, +0.035]**, where the filter's whole paired margin over a five-day recency rate is **+0.021 [+0.012, +0.030]**. It buys skill, not a flat reliability diagram: the worst bin still misses by **0.075**.

(ii) An eleven-parameter ridge logistic on the same daily counts beats the calibrated 64-state filter. Seven `log1p` count lags, the trailing 30- and 90-day event-day frequencies and `log1p` of the trailing 7-day sum score **+0.3734**; the paired difference (calibrated cascade − logistic) is **−0.0089 [−0.0165, −0.0019]**. Stacking the 64-state posterior onto the logistic moves the score by **+0.0015 [−0.0022, +0.0051]**.

(iii) Depth buys forecast skill, mostly through calibration, and the hierarchy buys none of the rest. Raw, `k = 6 − k = 2` is **+0.0729 [+0.0527, +0.1069]**; give every entrant its own map and the cheapest depth indistinguishable from `k = 6` is **k = 4** (**+0.0032 [−0.0009, +0.0070]**), while a *free* five-state Poisson hidden Markov model with no hierarchy at all matches the six-layer cascade at **+0.0009 [−0.0090, +0.0097]** — and loses the likelihood race on the same block by ΔBIC **97**.

(iv) The slow rungs were standing in for the solar cycle. Sunspot number and F10.7 in the exposure, on a supply repaired so that no verification bin is beyond issue-time coverage, are worth ΔBSS **+0.0507 [+0.0382, +0.0665]** and repair the quiet-year floor; under the refit the fitted ladder span collapses from **S = 146.8** to **S = 4.16**.

(v) The margin over recency grows with lead time because recency decays faster, not because the filter holds up. Against h-matched persistence it widens from **+0.0210 [+0.0124, +0.0294]** at h = 1 to **+0.0861 [+0.0515, +0.1230]** at h = 14, by which horizon persistence has fallen to **−0.0043**, below the verification climatology. The h-matched logistic still wins at every horizon (**−0.0349 [−0.0499, −0.0230]** at h = 1), and a 365-day moving rate overtakes the raw filter past about three days.

(vi) At the X threshold there is a real signal, and it is a rate effect. Read against real X1.0+ labels the frozen M-ladder's own 64-state belief scores **+0.1022** and beats X persistence by **+0.0718 [+0.0256, +0.1274]** — and a pure size lottery reproduces it, one constant flare-level X fraction **q = 0.0713** applied to the filter's rate, paired difference (state route − rate-only) **+0.0170 [−0.0097, +0.0514]**. The marked channel's primary per-event gain is **+0.0029 [−0.0006, +0.0074]** nats, covering zero, and a power battery recovers a planted ×5.2 swing in the X fraction on only **3/6** seeds: an estimability statement, not a demonstration that flare sizes are regime-independent. What the marks do show is a completeness signature — the M1.0–M1.9 share of M1.0+ events falls with the regime's activity, slope **−0.0789 [−0.1059, −0.0477]**, while flat against the lagged sunspot number (**−0.0559 [−0.1748, +0.0548]**).

(vii) The negative result. Applied to a second domain — the daily count of 3-hour intervals at Kp ≥ 5− over **34,577** calendar days, whose lag-27 storm-day autocorrelation of **0.193** encodes Bartels' solar-rotation recurrence — the same instrument returns **UNDECIDABLE AT THIS N ON P1, S1, S2** on a synthetic two-world atlas, and on the real record **THE CASCADE DOES NOT CARRY SKILL WORTH SERVING**: the calibrated filter beats persistence, recurrence and a seasonal climatology at h = 1, then falls below a flat 27-day recurrence baseline at h = 2 by **−0.0412 [−0.0560, −0.0238]**, with its BIC-selected depth flagged degenerate at **S = 1.00**. The lag-features logistic wins at every horizon and an additive decomposition absorbs its entire long-lead edge: **beyond h = 2, count-only Kp forecasting reduces to 27-day recurrence plus a trailing activity level** — an independent replication, from a different model family, of the lead-time structure Shprits et al. (2019) reported for solar-wind-driven Kp prediction.

We claim a verification, in the conventions of the field, of what the event record supports, and the finding that scaling and the feature set outrank the hierarchy on every axis we could measure. We do not claim an operational forecast, a physical mechanism, or any comparison against magnetogram-based systems, which see information no entrant here sees.

---

1. Introduction

1.1 The question

Operational solar flare forecasting is a mature, comparatively well-instrumented practice with an unflattering verification record. The "all-clear" comparison of Barnes et al. (2016) and the operational-systems comparison of Leka et al. (2019) found numerous methods above the no-skill level but **no single winner**: which method scores best "is decisively a function of flare event definition and the metric used" (Leka et al. 2019). Camporeale & Berger (2025) verified the NOAA Space Weather Prediction Center's own probabilistic M- and X-class forecast over 1998–2024 against zero-cost baselines — persistence, an empirical climatology, naive Bayes and a two-feature logistic regression — and reported that the official forecast does not outperform them on the event-focused metrics, while being poorly calibrated and prone to false alarms, "especially in high-stakes scenarios such as detecting the first flare after extended quiet periods."

Against that background a narrower question is worth answering with care. The GOES/XRS flare event list is public, homogeneous over five decades, and free. **How much 24-hour M-class skill does it carry on its own, and how much of that skill is bought by model structure rather than by scaling?**

The companion paper (`papers/06`) supplied the object this paper verifies: an exact hierarchical regime model — a Markov-switching multifractal cascade (Calvet & Fisher 2001, 2004) on a Poisson counting channel, four free parameters at every depth, an exact 2^k-state forward pass — fitted to the ~50-year daily M1.0+ count series and turned into a causal, filter-only 24-hour now-cast. That paper reported a Brier Skill Score of +0.338 against the verification period's own climatology, an edge over a recency baseline it declined to call significant, and a forecast that over-predicted in all ten reliability bins; and it named, in its §8 and Appendix B, six measurements it had not made — a recalibration step, a *paired* bootstrap of the difference, matched information for the baselines, a per-horizon curve, a real-label X-class run, and a Brier decomposition. This paper makes them, adds three the arc turned up (exposure covariates, the marked size channel, a second domain), and reports the ordering that results.

1.2 What this paper claims, and what it does not

The claim is a verification claim, and it is held narrowly.

We claim that, on this record, at daily cadence, on M-day occurrence: a two-parameter recalibration is worth more than the filter's entire paired margin over a recency baseline; an eleven-parameter ridge logistic on the same counts beats the calibrated filter with an interval excluding zero; the hierarchy's forecast advantage over a free flat hidden Markov model is indistinguishable from zero even where its likelihood advantage is not; and, in a second domain with a known recurrence structure, the same filter carries a one-day edge and nothing beyond it.

We do **not** claim an operational forecast, a comparison against magnetogram-based systems (the information ceiling of `papers/06` §8.4 stands unchanged), or a physical mechanism for the latent states; and we explicitly do not claim, on the evidence of §8, that flare sizes are independent of the activity regime — the marked channel is under-powered at this sample size and says so.

One further scope note belongs here rather than in a footnote. Every in-repo number in this paper is a **sealed** number: it is rendered from a committed record's JSON by the repository's paper-anchor gate and asserted to appear verbatim in this text (§12, Appendix A); literature values are marked as such and are not anchored. Where a metric the field expects is *not* in a sealed record, we do not compute it here. That applies to the True Skill Statistic and Heidke Skill Score at a declared operational threshold, and to reliability diagrams for the flare entrants as figures; §2.3 says exactly what is and is not available, and Appendix B registers the cut that would supply the rest.

1.3 Contributions

Seven, one per section: the calibration result with its paired interval (§3); the cheap-challenger tournament with a stacking control that returns zero (§4); the first ablation of the cascade's depth in *forecast* rather than likelihood metrics, with the free-HMM comparison at forecast altitude reported beside the ΔBIC that runs the other way (§5); the exposure covariates on a repaired public supply, and what they do to the fitted ladder (§6); the h-matched, leakage-probed per-horizon verification curve (§7); the real-label X-class readout and its rate/size decomposition, with the detection-completeness signature it turns up, reconciled against the sub-day time–energy literature (§8); and a second domain under the same discipline, returning an honest negative (§9).

---

2. Data, records, and the conventions that bind them

2.1 The two series

The flare observation series is the committed per-day trace of record `v44`: daily counts of GOES/XRS M1.0+ flares on an 18,871-bin grid, the five-step era covariate the companion paper fits with, and the sealed causal forecast `p_m24` produced by the frozen six-layer ladder of record `v26`. The event list behind it is the era-stitched NOAA product — NCEI legacy reports 1975–2017, the GOES-R L2 `flsum` science product thereafter — fetched at run time and never committed. Records `v45`, `v46`, `v47`, `v49` and `v52` read that trace and never edit it; `v50` and `v51` additionally read the fetched flare list and public activity indices. **The split is the companion paper's**: fit on the leading 60 % (bins 0–10,956), score the 7,305-day tail `2005.0 ≤ decimal_year < 2025.0`. The 609-day bridge window 2025-01-01 → 2026-09-01, a solar-maximum stretch at 2.6 times the tail's base rate, is scored **separately and never pooled**; every bridge number below is labelled and none is a skill claim.

The second domain's series is the committed anchor file `kp_storm_counts.csv` of record `v53`: the daily count of 3-hour intervals at Kp ≥ 5− (the NOAA G1 storm threshold, in GFZ thirds) over **34,577** calendar days, 1932-01-01 → 2026-08-31, derived from GFZ Potsdam's definitive Kp service (276,616 intervals, no gaps) and released under the feed's CC BY 4.0. WDC Kyoto Dst is validated and pinned in the same record for a later cut, is not modelled here, and its non-commercial clause is named in §**Data availability** so that no consumer inherits it by analogy. The domain was chosen because it carries a **known** recurrent structure — Bartels' 27-day solar-rotation recurrence, visible in the sealed lag-27 storm-day autocorrelation of **0.193** against 0.378 at lag 1 — so any claimed regime skill has a baseline it must beat by construction. Raw index levels are never modelled; the object is a sparse counting channel, the filter's proven regime.

2.2 The three binding conventions

Three conventions bind every record in both arcs.

1. **Everything is fitted on the training block alone** — calibration maps, logistic coefficients, covariate exponents and refit ladders see no data from the scored tail, and rolling-origin variants refit only on data strictly before the day they score.
2. **Every difference is a paired block bootstrap**: ninety-day moving blocks, index sets drawn **once** from a fixed seed and shared by every entrant and both references. No comparison here is a comparison of two marginal intervals.
3. **Every skill score is reported against both references** — the training block's Poisson climatology (0.4186 for flares; the training storm-day frequency 0.2453 for Kp) and the verification tail's own event frequency (0.1441; 0.1695). Unless stated otherwise every BSS below is against the verification tail's own climatology, the reference a space-weather reader assumes and the one Leka et al. (2019) specify for the Brier Skill Score.

Convention 3 is not a formality. Leka et al. (2019) define the BSS reference as “specifically the no-skill climatological forecast of the testing period” and also report a second metric (`MSESS_clim`) referenced instead to a 120-day prior event rate — two defensible choices that move every number. Our two references differ by about 0.25 BSS on the same forecasts. A BSS quoted across papers without naming its reference is not a comparison.

2.3 The metrics that are here, and the ones that are not

Peak True Skill Statistic is sealed for every flare entrant and is quoted where it matters (§3.3, §4). Reliability is sealed as ten-bin tables for the flare entrants and as a committed figure for the geomagnetic entrants (§13); the Murphy decomposition into reliability, resolution and uncertainty is sealed for every entrant in both arcs.

Three things are **not** in the sealed records and are therefore not in this paper: TSS or HSS at a *declared operational threshold* (the TSS values here are maximised over threshold on the test set, an optimistic readout that Leka et al. (2019) warn “can imply performance better than is attainable in an operational setting”); a contingency battery — POD, FAR, CSI, precision — under the community’s `M1.0+/0/24` event-definition convention (Barnes et al. 2016; Leka et al. 2019); and a committed reliability diagram for the flare entrants. We do not compute them here, because computing a number outside a sealed record is the practice this program exists to avoid. They are registered as **P-12.13** today (2026-09-15); Appendix B states the cut.

The cost is real and we name it. TSS is the metric Bloomfield et al. (2012) argued should be the comparison standard for flare forecasting, and a paper that cannot report it at an operational threshold cannot be placed directly beside the operational literature. §4 says what can and cannot be compared without it.

---

3. Calibration: worth more than the hierarchy’s margin over persistence

(Record `v45`.)

3.1 The defect, and the map

On the 7,305-day tail the companion paper’s sealed forecast issues a mean probability of 0.195 where the observed M-day frequency is 0.144, over-issuing in all ten reliability bins. That paper traced the defect to a train/test activity shift — cycles 21–23 in training, the much weaker cycles 24 and 25 in verification — and declined to correct it, on the ground that recalibrating a sealed measurement changes it rather than reports it. Record `v45` reports it instead: it leaves the forecaster untouched and re-scores the same trace under three calibration maps, each fitted on the training block alone.

The primary map is Platt scaling, `p' = σ(a + b·logit p)`, fitted by Newton/IRLS on bins 0–10,956 and never refit: **a = −0.369440**, **b = +0.924339**. The `b < 1` is a shrink toward the base rate and the `a < 0` a downward shift, which is what a filter fitted to an active era needs when applied to a quiet one.

3.2 What it buys

The map moves the tail BSS from **+0.3385** to **+0.3645** against the verification reference — a paired **+0.026 [+0.018, +0.035]**. An isotonic map (49 PAV knots) gains +0.025 and is not distinguishable from it; a rolling-origin Platt refit every 365 bins gains +0.027 over the raw forecast and only +0.0007 over the static map, so the miscalibration is a stable property of the forecaster rather than of this split.

That gain is larger than the hierarchy’s entire margin over recency. On the same resamples the paired difference (sealed cascade − five-day recency EWMA) is **+0.021 [+0.012, +0.030]**. Two things follow. The companion paper’s “near-tie with overlapping 95 % CIs” was an artifact of comparing two marginal intervals:

paired, the difference is cleanly positive. And it is smaller than what a two-parameter rescaling of the same probabilities is worth — a forecast system with a 64-state filter and no calibration step has its priorities inverted.

A control belongs here. The companion paper's §8.6 conceded that the cascade received the era covariate and the EWMA baselines did not. Applied symmetrically, the way the cascade's own emission uses it, the era covariate moves the persistence forecast in 58 of 18,871 bins and the paired ΔBSS is `+0.000000` with a degenerate interval. The asymmetry cost the comparison nothing.

3.3 What it does not buy

The pre-registration asked for every reliability bin within ±0.03 of the diagonal. It is reported as failed: the calibrated forecast's worst bin still misses by **0.075** (raw: 0.170), with an n-weighted mean deviation of 0.024, and the Murphy reliability term falls 0.00398 → 0.00073 with resolution essentially unchanged. A two-parameter map cannot flatten a reliability curve that is not a logit-linear distortion, and the isotonic map, which can represent one, does not do better in Brier and is worse in its worst bin (0.108). On this record **calibration buys skill, not a flat reliability diagram**.

A second negative is a property of the metric rather than the model. Platt is strictly monotone, so calibration cannot move a ranking metric at all: the calibrated cascade keeps the raw peak TSS of 0.621 at a shifted threshold, and the recency baseline keeps the higher 0.633 it already had. A system evaluated only on a ranking metric is blind to the defect this section repairs — the concrete form of the point Leka et al. (2019) make about metric choice, and the reason both are reported.

Scored separately, the 609-day bridge window (base rate 0.368) gives the calibrated cascade +0.277 against the raw +0.228, but the map **over-corrects at the top**: its worst reliability bin moves from 0.252 raw to 0.401, where isotonic (0.194) and rolling-origin (0.193) do not. A map fitted on 1975–2005 is not entitled to a solar-maximum window at 2.6 times the tail's base rate. That is the clearest single caution against shipping a static map into a rising cycle.

---

4. Structure: an eleven-parameter logistic beats the calibrated filter

(Record `v46`.)

Five cheap competitors are raced against the calibrated filter on the identical tail, each fitted on the training block alone, with every feature at bin *t* a function of bins strictly before *t* and no-leakage asserted by truncation.

The winner is a ridge logistic regression on eleven features: seven `log1p` daily count lags, the trailing 30- and 90-day event-day frequencies, and `log1p` of the trailing 7-day count sum, at ridge 10^{-3}. It scores **+0.3734** against the verification reference where the Platt-calibrated 64-state cascade scores +0.3645, and the registered verdict — the paired difference (calibrated cascade − logistic) — is **−0.0089 [−0.0165, −0.0019]**, an interval excluding zero. Against the *uncalibrated* forecast the gap is four times larger (−0.0349 [−0.0494, −0.0235]): most of the published forecaster's distance from a logistic regression was miscalibration, not structure. Either sign was a deliverable for this record; this is the one the data returned, and the margin is about a third of what calibration alone bought in §3.

The stack returns zero. Appending `logit(p_cascade)` as a twelfth feature moves the score by **+0.0015 [−0.0022, +0.0051]**, straddling zero exactly as pre-registered, even though the fit uses the column with a weight of +0.689. The 64-state belief carries no information about tomorrow's M-day that a week of counts and two trailing frequencies do not already carry. This is the cleanest statement in the paper of what the hierarchy is and is not doing at forecast altitude, and it is a null with an interval rather than a failure to reject.

Cheapness has a floor, and the logistic is above it. A two-feature GLM on log-EWMA$_5$ + log-EWMA$_{365}$ scores +0.3315 and a previous-day flag 0.2385; the eleven-feature logistic beats them by +0.042 [+0.028, +0.053] and +0.135 [+0.107, +0.173]. One pre-registered bearing missed and is reported as a construction difference rather than tuned away: the scoping run quoted the two-feature GLM at ≈ +0.348, the literal `log`-EWMA reading gives +0.3315 and the `log1p` twin of the same two features +0.3536. On the bridge window the ordering is unchanged, and the floor entrant is the caution: at a 0.368 base rate two conditional frequencies land within 0.035 BSS of the calibrated 64-state filter.

Read against the operational literature. This is the shape of result Camporeale & Berger (2025) obtained on a different object. Their Table 2 (M-class, 24 h) reports SWPC Brier 0.11 against climatology 0.13 and persistence 0.18, but TSS 0.44 against persistence 0.46 and a baseline-average 0.47 — the metric reordering again. Two cautions on reading the two sets of numbers side by side: the periods and base rates differ, and a raw Brier score is not comparable across base rates; and our headline is a skill score against a named reference while theirs is a raw Brier plus a contingency battery. Until the registered cut of Appendix B supplies TSS and HSS at a declared threshold under the `M1.0+/0/24` definition, the comparison is qualitative — both papers find cheap count-history baselines hard to beat and calibration a first-order defect of the incumbent. One point cuts the other way: their logistic uses two features (consecutive flare-free days, sunspot number) and has recall 0.15 at 24 h, where ours, on eleven features drawn from the flare history itself, wins. The lesson is not “logistic regression is strong” but “a week of flare history plus two trailing frequencies is a strong feature set” — the same conclusion, from the opposite direction, as Deshmukh et al. (2023), who found added model complexity does not reliably pay on the canonical 24-hour task.

---

5. Depth: what the hierarchy buys as a forecast

(Record `v47`.)

The companion paper chose `k = 6` on likelihood and forecast with it; no record had asked what that depth buys in *forecast* metrics. This record refits the filter on the same training block at `k ∈ {1, 2, 3, 4, 6}` and the free Poisson HMM — a free `J × J` transition with free per-state rates, the flat adversary — at `J ∈ {2, 3, 4, 5, 6}`, walks each forward through the *same* filter code, and scores every one raw and Platt-calibrated. The continuity check is exact: the `k = 6` refit reproduces the sealed ladder to a relative difference of 0.0 in all four parameters and its per-day forecast reproduces the sealed `p_m24` bit for bit, so the refit *is* the sealed entrant.

Depth buys skill; most of it is calibration. Raw, `k = 6 − k = 2` is **+0.0729 [+0.0527, +0.1069]** and every shallower comparison excludes zero, so the pre-registered “depth buys nothing” branch does not fire. Give each entrant its own two-parameter Platt map, fitted on its own training-block forecasts, and the gains collapse by half to two-thirds: `k = 6 − k = 2` falls to +0.0393 [+0.0256, +0.0541], and `k = 6 − k = 4` is **+0.0032 [−0.0009, +0.0070]**, covering zero. The cheapest depth indistinguishable from `k = 6` is **k = 4**. Much of what a deeper ladder contributed was a better-scaled probability, which a two-parameter map supplies directly.

The flat adversary matches the cascade. Against the free Poisson HMM selected on **training** BIC — J = 5, chosen before the tail is looked at — the calibrated paired difference is **+0.0009 [−0.0090, +0.0097]**. A free five-state hidden Markov model with no hierarchy at all matches the geometric six-layer cascade at forecast altitude; calibrated J = 6 edges it (−0.0045 [−0.0135, +0.0025], still covering zero), and on the bridge window the flat J = 5 beats it outright (−0.0219 [−0.0328, −0.0080]).

On the same training block the cascade still wins the *likelihood* race by ΔBIC **97** — 16,358.65 at `k = 6` against 16,456.13 for J = 5, four parameters against twenty-five. The ΔBIC does not convert into a forecast edge on M-day occurrence. That is not a contradiction — a parsimony verdict and a forecast verdict are different questions about the same data, and this record answers the second for the first time — but it is a caution about a common inference, that a model preferred by an information criterion is the model to forecast with.

What this says to the event-statistics lineage. The incumbent description of flare timing is a time-varying Poisson process (Wheatland 2000) with an operational forecaster attached (Wheatland 2004, 2005) that estimates the current rate from recent flares. The cascade is a hierarchical generalisation of it: the same conditionally-Poisson channel, with the single slowly varying rate replaced by a product of switching layers. This section says that the generalisation's *forecast* value on daily M-day occurrence is carried by having several activity states at all, not by their multiplicative, geometric-timescale arrangement — the reading the companion paper reached on likelihood grounds at C1.0, reached here independently in forecast metrics at M1.0. It sits beside two neighbours. Stanislavsky et al. (2020) fitted a two-state hidden Markov model to GOES soft X-ray flux and reported its prediction performance; our free-HMM panel is that family scanned in state count, and finds J = 5 on training BIC with no forecast penalty against a six-layer cascade. Snelling et al. (2020) found excess mutual information in GOES waiting times at ≈ 30 hr against Bayesian-block nonstationary-Poisson surrogates; our finding is compatible and narrower — whatever structure exists beyond the varying rate, a free flat multi-state rate model extracts as much 24-hour forecast value from it as a hierarchy does.

---

6. Exposure: the covariate the slow rungs were standing in for

(Records `v48` and `v51`.)

The sealed now-cast reads only its own count history, and on the tail that shows as a floor. In the nine quiet years (annual M-day frequency below 0.05: 2006–2010, 2016, 2018–2020) it issues 0.0533 a day where the observed frequency is 0.0137 — worse than those years' own climatology, a BSS of −0.1054. A forecaster that cannot tell a deep minimum from a moderate one is not usefully calibrated for the periods when its users most want an all-clear.

Record `v48` put two public, issue-time-available indices into the exposure, `cov_t = era_t × (SN_{t−1}/S̄N)^β₁ × (F_{t−1}/F̄)^β₂`, with WDC-SILSO daily sunspot number and the Penticton/DRAO adjusted F10.7 listing, both lagged one bin, refitting the ladder under every cell of a β grid and selecting β by training-block BIC. It measured a clear gain and then fired its own pre-registered kill rule on **availability**: the NOAA NCEI archival adjusted-F10.7 listing is a frozen product whose last row is 2018-04-30, leaving 2,436 of 7,305 tail bins without a value at issue time.

Record `v51` repaired the supply. A pre-registered source gate scored four live candidates; the NRCan/DRAO `fluxtable.txt` — the Penticton operator's own product — passes, and the splice agrees with the frozen product on 97.40 % of their 4,922 overlap days to within 0.1 sfu. **No tail bin is beyond issue-time coverage**, against 2,436 before. On the repaired supply the pair clears its bar: ΔBSS **+0.0507 [+0.0382, +0.0665]** over the count-only cell, surviving calibration of both (+0.0350 [+0.0249, +0.0482]). The gain lands where the defect was — the quiet-year mean issued falls 0.0533 → 0.0308 against an observed 0.0137, and quiet-year skill against those years' own climatology turns from −0.1054 to +0.0451. Decomposed, the earlier record's own cell is worth +0.0433 [+0.0317, +0.0571] on the spliced series, statistically the same gain it measured on the frozen one, so that gain was real; extending the exponent grid past its boundary adds the rest. Sunspot number alone clears the bar (+0.0269 [+0.0160, +0.0386]) and does **not** repair the quiet years: the floor fix is carried by the radio flux. Two limits travel with this — the selected cell sits at $\beta_2 = 2.5$, the new grid edge, so that exponent is a lower bound and not a measurement; and two strongly collinear covariates load onto the radio flux, leaving a mild sunspot term ($\beta_1 = 0.50$, interior and therefore measured).

The ladder collapses, and that is the finding. Under the refit the fitted span falls from **S = 146.8** to **S = 4.16**. The cascade's slow rungs were tracking the solar cycle; supply the cycle directly and they are not needed. This is the mechanism behind §7's result that a 365-day moving rate overtakes the raw filter past about three days, and it is a caution for any physical reading of fitted rung timescales in a count series spanning several solar cycles: the

slowest rung of such a fit risks being an unlabelled solar cycle. The record's sealed verdict is **THE COVARIATE PAIR IS SERVABLE ON THE SPLICED SUPPLY**; whether anything is served from it is not a question this paper answers.

---

7. Lead time: the per-horizon curve

(Record `v49`.)

Under the frozen ladder — no refit, no fit seed — the filter is read at `h ∈ {1, 2, 3, 5, 7, 14}` days against an h-matched field: persistence and the cycle rate frozen at `t − h`, the count-only logistic with its window shifted by h and fitted pre-2005, and the two constant references. `h = 1` reproduces the sealed trace exactly, and no-leakage is measured rather than assumed: at every h, perturbing a probe bin leaves every forecast for bins before `t₀ + h` bit-identical. This is a verification curve and never a scheduling readout.

Table 1 — BSS(h) on the 7,305-day tail, verification reference 0.1441 (record `v49`; every difference paired on one common panel).

h (days)	cascade (raw)	cascade (Platt)	persistence	cycle rate	logistic	cascade − persistence, paired 95 % CI
1	+0.3385	+0.3645	+0.3175	+0.2052	+0.3734	**+0.0210 [+0.0124, +0.0294]**
2	+0.2527	+0.2968	+0.2231	+0.1999	+0.3065	+0.0297 [+0.0184, +0.0417]
3	+0.2003	+0.2594	+0.1594	+0.1961	+0.2742	+0.0409 [+0.0261, +0.0566]
5	+0.1333	+0.2125	+0.0675	+0.1905	+0.2324	+0.0658 [+0.0446, +0.0901]
7	+0.1123	+0.2013	+0.0403	+0.1872	+0.2259	+0.0721 [+0.0451, +0.1030]
14	+0.0818	+0.1838	**−0.0043**	+0.1816	+0.2125	**+0.0861 [+0.0515, +0.1230]**

Three readings, in order of how much they flatter the filter.

The margin over recency widens with lead, and the reason is recency. The pre-registered deliverable — the first h at which the paired interval against persistence covers zero — does not exist in this grid: it widens from **+0.0210 [+0.0124, +0.0294]** at h = 1 to **+0.0861 [+0.0515, +0.1230]** at h = 14. Table 1 shows why. The cascade decays from +0.3385 to +0.0818 while h-matched persistence decays to **−0.0043**, i.e. by a fortnight the recency rate is worse than the verification climatology. The companion paper's "modest edge over persistence" is a one-day statement, and the general version is mostly a statement about how fast a five-day EWMA forgets.

The count-only logistic still wins at every horizon. Raw, the paired difference is **−0.0349 [−0.0499, −0.0230]** at h = 1, widening to −0.1307 [−0.1900, −0.0819] at h = 14; calibrating the cascade shrinks it but does not close it (−0.0090 [−0.0168, −0.0021] to −0.0291 [−0.0532, −0.0067]).

The cycle rate overtakes the raw filter past about three days. Cascade − cycle is +0.1332 [+0.1004, +0.1734] at h = 1, covers zero at h = 3 (+0.0042 [−0.0271, +0.0356]) and is significantly negative from h = 5 (−0.0572 [−0.0920, −0.0237]) through h = 14 (−0.0998 [−0.1468, −0.0634]). Past three days the raw filter carries nothing a 365-day moving rate does not already have — §6's ladder collapse, seen from the forecast side. Calibration is also worth more the further out one goes (+0.026 at h = 1, +0.102 at h = 14), as the belief relaxes toward its stationary distribution.

---

8. Rare events and marks: the X tile is a rate effect

8.1 The real-label X-class readout

(Record `v50`.)

The companion paper carried an X-class BSS of +0.082 and withdrew it as a synthetic-label diagnostic: the harness never read an X label, it thinned the M-class counts binomially at 6 % and scored the M-forecast against those draws, while 568 real letter-X rows sat unused in the same fetched product. Record `v50` is the measurement that withdrawal promised. The thinning is absent from its path; the occurrence series is the real ≥ X1.0 daily count on the same grid, and the tail carries 122 X-days at a verification climatology of 0.016701 — an order of magnitude below the M-class base rate.

Reading the frozen M-ladder's own predictive 64-state belief at the X threshold — a per-latent-state X-day rate fitted on the training block and applied causally — scores **+0.1022** after a training-block Platt map and beats h = 1 X persistence by **+0.0718 [+0.0256, +0.1274]**, with no replicate at or below zero. A separately refit X ladder also works (k = 6: +0.0796, +0.0492 [+0.0322, +0.0786] over persistence), and at this threshold depth does buy skill (k = 6 − k = 2 is +0.0296 [+0.0101, +0.0516]) where the M-class ablation's did not. One structural reading comes with it: squashing the sealed scalar `p_m24` through a logistic on its logit scores only +0.0194, indistinguishable from persistence, while the same fit's 64-state belief supports +0.0884 raw — the latent state carries information the one-dimensional output has thrown away. With 122 positive days the intervals are wide and the two winning routes are not separable; nothing here restores the withdrawn +0.082, scored against different labels at a different base rate.

8.2 The size channel: a rate effect, stated as an estimability result

(Record `v52`.)

An X-*day* rate is a rate × size product, so §8.1 leaves the question open: does the latent regime set flare **size**, or only flare **rate**? The standard self-organized-criticality reading says it need not. Record `v52` reads the marks: the NOAA-reported peak flux of every M1.0+ flare, as an interval (the reporting quantum, or the whole class decade for the 646 decade-censored 1975–2001 rows), conditioned on the frozen ladder's predictive 64-state belief at the event's bin, with no refit. Two regime-independent laws (a threshold-truncated lognormal; a tapered Pareto, which wins the training block) are raced against per-state families on held-out per-event log-probability, with three mark-shuffle nulls, a detection-completeness control, and a semi-synthetic power battery run **first**.

The firmer half first: §8.1's X-class skill is a rate effect. On the identical tail and panel, the frozen ladder's X-day probability under a *pure size lottery* — its state-mixed rate thinned by one constant flare-level X fraction **q = 0.0713**, the training block's — scores +0.0852 calibrated where the per-state route scored +0.1022. The calibrated paired difference (state route − rate-only) is **+0.0170 [−0.0097, +0.0514]**, covering zero, and the rate-only forecast beats X persistence by +0.0548 [+0.0089, +0.1074] — the state route's own margin. The X tile needs no size physics: the state carries X information through the rate.

The second half is a null with a stated power. The primary per-event gain of the shrunk 64-state family over the pooled law reads **+0.0029 [−0.0006, +0.0074]** nats per event at M1.0. It beats all 300 surrogates of the three shuffle nulls (rank floor p = 0.0099 on each) and still fails its own interval by 0.0006, and it does not climb the threshold ladder: +0.0035 [+0.0006, +0.0065] at M2.0, then −0.0029 [−0.0076, +0.0033] at M5.0. The registered rule required both the interval and the null bars at M1.0 *and* the interval at M5.0; neither holds, so the sealed sentence is the null one. The battery is what makes that sentence honest rather than empty. On a synthetic null world the primary instrument fires 0/6 at M1.0, so specificity holds; on a planted world whose X fraction moves by ×5.2 across the latent activity range it recovers the plant on only **3/6** seeds at M1.0, while the *parameter* — a linear index slope — is recovered almost exactly (mean −0.0518 against a planted −0.05). The held-out mark

likelihood is a low-power instrument at N = 8,393 flares, because a few very large events dominate every log-likelihood difference. So: **this record does not show that flare sizes are regime-free; it shows that at this N the mark likelihood cannot see a regime dependence smaller than a ×5 X-fraction swing, that the record carries at most a hair of one, and that §8.1 needed none.**

8.3 The completeness signature, and the sub-day literature reconciled

The marked channel does produce one positive that is not about the Sun. The M1.0–M1.9 share of M1.0+ flares falls from 0.632 to 0.505 across the belief's activity deciles — a logistic slope of **−0.0789 [−0.1059, −0.0477]** per unit of latent activity — while being flat against the lagged sunspot number (**−0.0559 [−0.1748, +0.0548]**). The deficit of small flares tracks the *flaring regime*, not the 11-year background. A genuine shift in the size distribution with activity and self-inflicted threshold censoring on active days look alike in this probe; but the slope in the tapered-Pareto index dies above threshold (−0.0452 per unit at M1.0, −0.0015 at M5.0) where the battery says a real link of that size would still register at M5.0. A size effect that vanishes above threshold while the small-flare share falls with the flaring regime is the detection-completeness signature, and it is the economical reading.

That mechanism is documented. Wheatland (2001) measured **obscuration** — the soft X-ray flux of a large flare suppressing detection of events in its decline — and estimated that "in the absence of obscuration the number of flares above C1 class would be higher by (75 ± 23) %". Aschwanden & Freeland (2012) built an automated detector sensitive to about five times more events than the NOAA catalog and reported the same direction, modelling it as a flare pile-up bias that steepens the rise-time distribution at cycle maximum. Our probe reads that effect through a latent activity state instead of through cycle phase, and it is stronger against the state than against the sunspot number — what one expects if the censoring is driven by the local flaring rate rather than the slow background.

The reconciliation with the sub-day result belongs here explicitly, because it looks like a conflict and is not. Lippiello, de Arcangelis & Godano (2010) raced the same GOES catalog against an energy-decorrelated reshuffled catalog and found genuine time–energy correlations: a deficit of events at $\Delta t < 2$ h after energetic flares, an excess for $\Delta t \in [2\text{ h}, 4\text{ h}]$ decaying to background at longer separations, and a tendency for flares close in time to have the *second* event large. Our record finds no resolvable dependence of flare size on the latent regime. The two are about different timescales and different aggregation levels. Theirs is an event-pair statement at separations of hours, conditioned on the immediately preceding flare; ours is a daily statement, conditioned on a latent state whose fastest rung is of order a day. A 2–4 h aftershock-like rate excess is aliased into a daily count as overdispersion and absorbed by the counting channel as *rate* — at our resolution it is not a size statement at all. Their $\Delta t < 2$ h deficit is obscuration, the Wheatland (2001) effect our completeness probe sees from the other end. And their "second event large" correlation lives at sub-day separation, which daily binning cannot represent: our null is silent about it rather than in conflict with it. A sub-daily cut is the measurement that would engage them directly, and Appendix B names it.

---

9. A second domain: geomagnetic storm counts, and the negative result

9.1 The atlas: the instrument does not decide at this N

(Records `v53`, `v54`.)

Record `v53` seals no model: it fetches, pins, validates and licenses the two geomagnetic feeds and seals the statistics the atlas cites as landmarks. Its 60 % split falls at 1988-10-19, with training P(storm day) 0.2453 against the tail's 0.1695 — the two references differ by more than in the flare line, and every entrant inherits the shift. Record `v54` then asks whether the instrument can tell a planted hierarchy from the domain's known structure at the real N, building two worlds on the real calendar grid, both riding one deterministic-plus-smooth modulation (a

lagged solar-cycle envelope, an equinoctial semiannual term, a 27.27-day recurrence with a drifting complex amplitude) and differing only in whether a hierarchy is present: a single free two-state switch, or a six-layer cascade.

The sealed verdict is **UNDECIDABLE AT THIS N ON P1, S1, S2**. The power arm that matters holds — on the cascade world the calibrated filter beats persistence and recurrence at h = 1 on every seed, and the planted ladder is recovered (k* = 6 on 3/3 seeds, $\hat{S}$ 144.1 / 129.2 / 112.5 against a planted 120). The three failing arms are the diagnostic ones. The depth gain from one layer to k* is only 1.45–1.78× larger in the cascade world than in the switch world, against a pre-registered 3× (the flare line's null read ≈ 7×): a null carrying a season, a recurrence and a two-day switch *is* multi-timescale structure, and a low-order cascade absorbs it. The switch world selects k* = 4 with a compact, non-degenerate ladder on 3/3 seeds — a four-layer ladder manufactured where no hierarchy exists. And the cascade never beats the free HMM on the null world, on two seeds losing by an interval excluding zero where the rule asked for one covering zero.

The atlas is also honest about an **under-matched null**, and this scopes everything below: the switch world's lag-27 storm-day autocorrelation is 0.021 / 0.025 / 0.019 against the real record's **0.193**, and its decadal cycle contrast 1.17 / 1.11 / 1.11 against the real 3.06. The null is *weaker* in the very structure the specificity arms test, which makes those arms easier to pass — and they still fail.

9.2 The real tournament: a one-day edge, then recurrence

(Record `v55`.)

The same machinery, unchanged, is pointed at the real series: every entrant fitted on the training block and walked forward at h ∈ {1, 2, 3, 5, 7, 27} against one paired panel, with the mandatory recurrence baseline (four lagged counts at 27-day offsets plus two harmonics) and a seasonal climatology in the field.

Table 2 — BSS(h) on the 13,831-day Kp tail, verification reference 0.1695 (record `v55`, primary threshold Kp ≥ 5−).

h (days)	cascade (Platt)	persistence (Platt)	recurrence	recurrence + level	logistic	free HMM (Platt)	seasonal clim.
1	**+0.1499**	+0.1037	**+0.0575**	+0.1024	**+0.2043**	+0.1557	−0.0319
2	+0.0163	+0.0303	+0.0575	+0.0939	+0.1045	+0.0165	−0.0319
3	−0.0076	+0.0176	+0.0575	+0.0916	+0.0946	−0.0090	−0.0319
5	−0.0203	+0.0111	+0.0575	+0.0901	+0.0907	−0.0232	−0.0319
7	−0.0249	+0.0069	+0.0575	+0.0890	+0.0892	−0.0277	−0.0319
27	**−0.0407**	+0.0425	**+0.0511**	+0.0848	**+0.0862**	−0.0407	−0.0319

The first pre-registered sentence is **THE CASCADE DOES NOT CARRY SKILL WORTH SERVING**. At h = 1 the calibrated filter beats every rival with a paired interval clear of zero — persistence by +0.2225 [+0.1962, +0.2485], recurrence by +0.0924 [+0.0771, +0.1119], the seasonal climatology by +0.1818 [+0.1627, +0.2055]. At h = 2 it falls **below** the flat 27-day recurrence baseline by **−0.0412 [−0.0560, −0.0238]**, the first failing pair, and stays below it at every longer horizon. Calibrated against calibrated, its edge over recency is likewise a one-day edge: +0.0462 [+0.0345, +0.0563] at h = 1, −0.0140 [−0.0267, −0.0041] at h = 2.

The flare line's findings reappear intact. The lag-features logistic — the §4 feature idiom, h-shifted — beats the cascade at every horizon (−0.0544 [−0.0635, −0.0453] at h = 1, −0.1269 [−0.1552, −0.1019] at h = 27), and the free three-state Poisson HMM beats it at h = 1 too (−0.0058 [−0.0090, −0.0025]). The fit itself declines to be read as a hierarchy: the BIC-selected depth is k* = 3 with a ladder flagged degenerate at **S = 1.00** — all three rungs at one ≈ 4.6-day rate — with k = 1 and k = 6 flagged as well and a non-monotone BIC curve, while the free three-

state HMM describes the training block better than any cascade depth by 445 BIC. The same fit protocol recovered clean six-layer ladders where a hierarchy was planted (§9.1). On the real record there is no multiplicative hierarchy resolvable beyond one storm timescale — less structure than the atlas's own null world.

9.3 The registered sentence and the measured sentence, both kept

The second pre-registered sentence reads, in its registered form, `COUNT-ONLY KP FORECASTING DOES NOT REDUCE TO RECURRENCE PLUS CLIMATOLOGY AT ANY LEAD IN THE GRID`, because the logistic beats the recurrence baseline at every horizon (+0.1468 [+0.1327, +0.1621] at h = 1, +0.0351 [+0.0254, +0.0473] at h = 27), so the registered crossing horizon does not exist. We keep that sentence, because the registration is what was pushed before the run.

We also keep what the record measured, and say why the letter reads as it does. The registered baseline carries a **fixed** training-block climatology while the tail's storm-day frequency is 31 % lower than the training block's. An additive decomposition entrant — the recurrence baseline's features plus the h-shifted trailing 30- and 90-day storm-day frequencies, an *adaptive* climatology, and no daily lag, added after the first pass and never gating — absorbs the logistic's entire long-lead edge: logistic − (recurrence + level) is +0.003 [+0.000, +0.006] at h = 3, +0.001 [−0.000, +0.002] at h = 5, +0.000 [−0.000, +0.001] at h = 7 and +0.001 [−0.000, +0.003] at h = 27, while (recurrence + level) − recurrence is +0.034 [+0.024, +0.046] at h = 27. So: **beyond h = 2, count-only Kp forecasting reduces to the 27-day recurrence plus a trailing activity level.** The cascade's skill is the storm's own persistence, read for one day. The consumer gate the registration attached to a positive Sentence 1 is therefore closed, and the stage it gated stays registered and unexecuted. Both sentences read the same at the sensitivity threshold Kp ≥ 6−.

This replicates, from a different model family and a different predictor set, a result already in the Kp literature. Shprits, Vasile & Zhelavskaya (2019) fed L1 solar-wind measurements and Kp history to neural networks against persistence and recurrence baselines and concluded that "for a longer-term forecast of more than 2 days, predictions can be made using recurrence only, while solar wind measurements provide very little improvement for a forecast with long horizon times." Our entrants see no solar wind at all — only the index's own count history — and the crossing lands in the same place: a one-day edge for the driven model, and from h = 2 a flat recurrence baseline (in the spirit of the recurrence index of Sargent 2021) that nothing looking only at the recent past beats. That two quite different model families, given quite different information, degrade to the same recurrence floor at the same lead time is a stronger statement about the domain than either result alone.

---

10. Discussion

The ordering is the message. Across twelve sealed records and two domains, every axis we could measure put scaling and features above structure. A two-parameter calibration map was worth more than the hierarchy's entire margin over a recency baseline (§3). An eleven-feature logistic on the same counts beat the calibrated 64-state filter, and stacking the filter's posterior onto it returned zero (§4). Half to two-thirds of what depth appeared to buy was calibration, and what remained was matched by a free five-state hidden Markov model with no hierarchy at all, while the cascade won the likelihood race on the same block by ΔBIC 97 (§5). Supplying the solar cycle as an exposure covariate bought more than depth did and collapsed the fitted ladder by a factor of 35 (§6). In a second domain the fitted ladder collapsed to degeneracy outright (§9.2).

We resist the conclusion that the hierarchy is worthless, because that is not what was measured: the companion paper's likelihood result stands, and this paper measures that it does not convert. The useful statement is narrower and firmer. **If the goal is a 24-hour occurrence forecast from an event list, spend the effort on the probability scale and the feature set, not on the latent architecture.** If the goal is a description of the rate process, a different verdict is available and the companion paper reports it.

What it says to the forecasting literature. Camporeale & Berger (2025) found the operational SWPC forecast unable to beat persistence, climatology, naive Bayes and a two-feature logistic on event metrics. We find the same shape of result inside a research forecaster considerably more structured than SWPC's procedure, and can add the decomposition: the incumbent's gap to a cheap model was four times larger before recalibration than after. Barnes et al. (2016) and Leka et al. (2019) established that ranking depends on the event definition and the metric; §3.3 is the sharpest instance, where a monotone recalibration changes the Brier Skill Score by more than the model comparison it is embedded in and changes a ranking metric by exactly zero. Deshmukh et al. (2023) found simpler models competitive on the canonical 24-hour task; §§4–5 are two further instances, one in feature space and one in latent-state space. Three baselines earned their place and should be standard in event-record forecasting: a *paired* bootstrap of the difference rather than two marginal intervals, which reversed one published reading (§3.2); a calibrated version of *every* entrant, since an uncalibrated comparison mostly measures scaling; and, in any domain with known periodicity, a recurrence baseline carrying an *adaptive* climatology (§9.3).

On the McIntosh-rate lineage and the GOES rescaling. The operational climatology this kind of forecast is implicitly measured against is a McIntosh-class Poisson rate table (McCloskey et al. 2018, on the Bloomfield et al. 2012 construction). Janssens et al. (2025) revisited those rates under the GOES/XRS rescaling SWPC adopted in December 2019, recomputing 1976–2019 flare intensities and finding 55 % and 52 % more M- and X-class events, which moves every class-conditional rate in the table. Our records handle the same discontinuity with a coarser instrument: a five-step per-era covariate on the latent rate, carried unchanged through both arcs, plus a unit-harmonisation cell in the marked channel that multiplies GOES-R-era fluxes by the SWPC long-channel factor 0.70. That harmonisation lifted the marked channel's primary gain to an interval excluding zero (+0.0035 [+0.0010, +0.0069]) — so part of what looked like a size signal is a units signal a rate covariate never touched. A re-cut of both arcs under the uniform Hudson et al. (2024) recalibration factor is registered in Appendix B as the robustness run this paper owes.

Limitations

Scope. This is a verification paper, not a forecasting system: nothing here is operational and no product decision follows from it. Every entrant is blind to magnetograms, so the information ceiling of the companion paper's §8.4 governs every number above; the flare tables read as "the skill in the event record alone", a floor for magnetogram-informed work rather than a competitor to it. §8.2's null is an estimability statement and is stated as one — the battery recovers a planted ×5.2 X-fraction swing on 3/6 seeds — and §9.1's atlas did not decide, so every positive sentence in §9.2 is scoped as an estimability statement and the null world is under-matched in exactly the structure the specificity arms test.

Metrics. There is no TSS or HSS at a declared operational threshold and no contingency battery; the TSS values quoted are maximised over threshold on the test set. No reliability diagram is committed for the flare entrants, only ten-bin tables and the Murphy decomposition. The BSS reference choice moves every number by about 0.25 and both are reported throughout. Until P-12.13 (Appendix B) no like-for-like comparison with the operational literature is available.

Apparatus. Every flare number rests on a single 60/40 split of a single record at daily resolution; there is no rolling-origin verification with activity-matched training and test blocks, and §3.2's rolling-origin map is a calibration control, not a verification protocol. The bridge window is 609 days at 2.6× the tail's base rate, flatters every forecaster, is never pooled and is never a claim. The β grid of §6 is still on a boundary in the F10.7 exponent, so that exponent is a lower bound, and two collinear covariates mean the split between them is not identified. The depth ablation fixes k* by BIC and does not swap it post hoc: in the geomagnetic arc the non-degenerate k = 4 and k = 5 fits forecast better than the selected k* = 3 at every horizon (+0.104 [+0.082, +0.128] at h = 27) and are still below recurrence; both readings are sealed and neither gates. At h = 27 the k* = 3 filter no longer responds to a 27-day-old count — a mixing-time fact, not leakage; the no-leakage probes hold on all 36 cells.

Data and instrument. The era covariate is a coarse instrument for the GOES rescaling, five steps against a uniform factor, and the paragraph above names the robustness run that would price it. Detection completeness is a live systematic, not a controlled one (§8.3): the probe cannot separate a real activity-dependent shift in the size distribution from self-inflicted censoring, and we report the economical reading rather than a measurement. Daily binning aliases all sub-day structure into overdispersion, so this analysis cannot engage the hours-scale time–energy literature on its own ground. The Kp count channel has bounded support (at most eight storm intervals per day) while every entrant here is a Poisson-channel model; a binomial channel is a named next cut.

And one that is not a limitation of the analysis. No domain-collaborator pass yet: a solar-physics and a geomagnetism referee should both be resisted into §8.3 and §9.3 respectively.

---

11. Conclusions

1. **Recalibration outranks the hierarchy's margin over recency.** A training-block Platt map (**a = −0.369440**, **b = +0.924339**) moves the sealed 24-hour forecast from **+0.3385** to **+0.3645**, a paired **+0.026 [+0.018, +0.035]**, where the filter's whole paired margin over a five-day recency rate is **+0.021 [+0.012, +0.030]**. It buys skill and not reliability: the worst bin still misses by **0.075**, and a monotone map cannot move peak TSS at all.
2. **An eleven-parameter logistic on the same counts wins**, **+0.3734** against the calibrated cascade's +0.3645, paired **−0.0089 [−0.0165, −0.0019]**; stacking the 64-state posterior onto it is worth **+0.0015 [−0.0022, +0.0051]**.
3. **Depth buys forecast skill mostly through calibration, and the hierarchy buys none of the rest.** Raw `k = 6 − k = 2` is **+0.0729 [+0.0527, +0.1069]**; calibrated, **k = 4** is indistinguishable from k = 6 (**+0.0032 [−0.0009, +0.0070]**) and a free five-state Poisson HMM matches the cascade (**+0.0009 [−0.0090, +0.0097]**) while losing the likelihood race by ΔBIC **97**.
4. **The slow rungs were the solar cycle.** Sunspot number and F10.7 in the exposure are worth **+0.0507 [+0.0382, +0.0665]** on a repaired supply, repair the quiet-year floor, and collapse the fitted ladder span from **S = 146.8** to **S = 4.16**.
5. **The lead-time margin over recency is a statement about recency.** It widens from **+0.0210 [+0.0124, +0.0294]** at h = 1 to **+0.0861 [+0.0515, +0.1230]** at h = 14 because persistence falls to **−0.0043**; the logistic wins at every horizon (**−0.0349 [−0.0499, −0.0230]** at h = 1) and a 365-day rate overtakes the raw filter past about three days.
6. **The X-class tile is real and is a rate effect.** Real-label BSS **+0.1022**, beating X persistence by **+0.0718 [+0.0256, +0.1274]**, and reproduced by one constant flare-level X fraction **q = 0.0713** (state route − rate-only **+0.0170 [−0.0097, +0.0514]**). The marked channel's primary gain is **+0.0029 [−0.0006, +0.0074]** nats and the battery recovers a planted ×5.2 swing on **3/6** seeds, so "the cascade sets when, not how big" is an estimability statement. The one clear mark result is a completeness signature: slope **−0.0789 [−0.1059, −0.0477]** against latent activity, **−0.0559 [−0.1748, +0.0548]** against lagged sunspot number.
7. **The second domain returns a negative.** On **34,577** days of Kp storm counts (lag-27 autocorrelation **0.193**) the atlas reads **UNDECIDABLE AT THIS N ON P1, S1, S2** and the real tournament **THE CASCADE DOES NOT CARRY SKILL WORTH SERVING** — a one-day edge, then a fall below flat 27-day recurrence at h = 2 by **−0.0412 [−0.0560, −0.0238]**, with the selected ladder flagged degenerate at **S = 1.00**. The calibrated cascade runs **+0.1499** at h = 1 to **−0.0407** at h = 27 while the recurrence baseline holds **+0.0575 → +0.0511** and the lag-features logistic **+0.2043 → +0.0862**. Beyond h = 2, count-only Kp forecasting reduces to recurrence plus a trailing activity level, replicating Shprits et al. (2019) from a different model family.

The claim is the narrow one: a verification, in the field's conventions, of what a 24-hour event-record flare forecast is worth, and an honest negative in a second domain. The result is ready for owner review and a collaborator pass, and is reproducible from the named records.

12. Reproducibility

Everything in this paper regenerates from the named records behind the repository's apparatus gate, which is data-free.

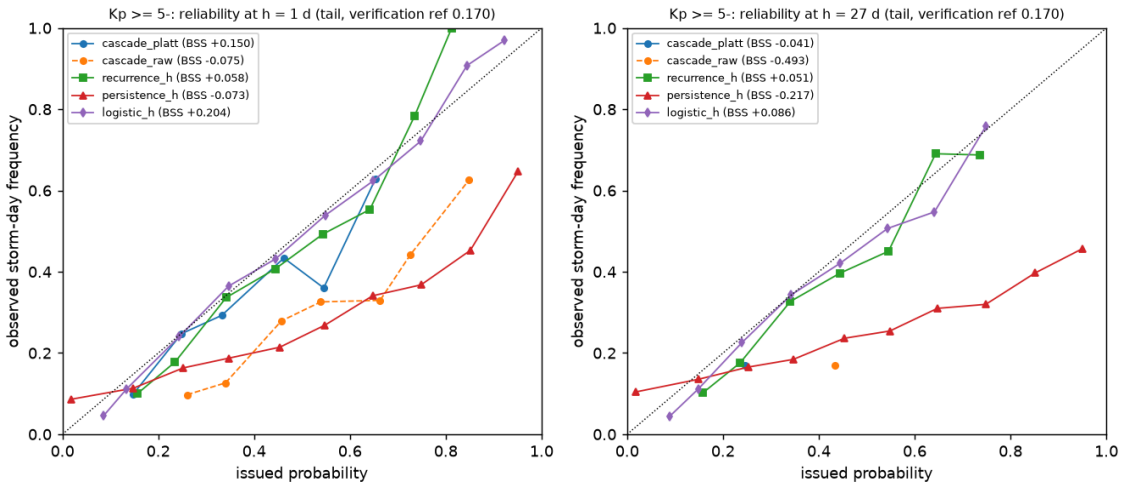
```
python3 -m venv venv && venv/bin/pip install -r requirements-dev.txt
./verify.sh                                          # apparatus gate, data-free
venv/bin/python -m fenocosm.p103_nowcast_calibration   # -> out/v45/  (offline)
venv/bin/python -m fenocosm.p104_nowcast_challengers   # -> out/v46/  (offline)
venv/bin/python -m fenocosm.p105_nowcast_ablation      # -> out/v47/  (offline)
venv/bin/python -m fenocosm.p107_nowcast_horizon       # -> out/v49/  (offline)
venv/bin/python -m fenocosm.p113_geomag_tournament     # -> out/v55/  (offline)
venv/bin/python -m fenocosm.goes_fetch                 # -> data/goes/ (network)
venv/bin/python -m fenocosm.p108_nowcast_xclass        # -> out/v50/
venv/bin/python -m fenocosm.p109_nowcast_f107_supply   # -> out/v51/  (+ two covariate feeds)
venv/bin/python -m fenocosm.p110_flare_marks           # -> out/v52/  (+ the SILSO feed)
```

Records `v45`, `v46`, `v47`, `v49` and `v55` need no network and no fetched data: the committed `out/v44/` trace **is** the flare observation series and `out/v53/kp_storm_counts.csv` the geomagnetic one. `v50`, `v51` and `v52` need the NOAA flare fetch, and `v51` and `v52` additionally need the public activity feeds. Each record carries an offline `verify_seal()` or reproduction test that re-derives its own table from its committed inputs.

Every **bolded in-repo** number in this paper is additionally bound to a sealed record's JSON by the repository's paper-anchor gate, which renders the value from the record and asserts it appears verbatim here; a re-cut that moves a number fails the gate until the paper is swept in the same commit. Appendix A is the map.

13. Figure

- `../out/v55/p113_reliability.png`



../out/v55/p113_reliability.png

— the h = 1 and h = 27 reliability diagrams for the geomagnetic tournament at Kp ≥ 5− (record `v55`, §9.2),

drawn by the sealed record's own module `fenocosm.p113_geomag_tournament` from the committed anchor series. At h = 1 the calibrated filter's bulk bin over-issues by 0.05 — the base-rate shift a training-block map cannot see — and by h = 27 the filter has relaxed onto its own climatology.

No figure is drawn for the flare arc: its reliability tables are sealed as numbers and a committed diagram is part of the registered cut in Appendix B.

---

14. References

Every entry below was checked against a primary source — the Crossref registry record, the publisher page, or the author's own deposited preprint — in the pass that produced this draft.

- **Aschwanden, M. J., & Freeland, S. L. 2012**, "Automated Solar Flare Statistics in Soft X-Rays over 37 Years of GOES Observations: The Invariance of Self-organized Criticality during Three Solar Cycles," *ApJ*, 754, 112. doi:10.1088/0004-637X/754/2/112 (arXiv:1205.6712) *(§8.3: ≈ 5× the NOAA catalog's sensitivity, and the solar-cycle-dependent flare pile-up bias).*
- **Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016**, "A Comparison of Flare Forecasting Methods. I. Results from the 'All-Clear' Workshop," *ApJ*, 829, 89. doi:10.3847/0004-637X/829/2/89 *(Paper I; the event-definition notation and the benchmark design).*
- **Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012**, "Toward Reliable Benchmarking of Solar Flare Forecasting Methods," *ApJL*, 747, L41. doi:10.1088/2041-8205/747/2/L41 *(the case for TSS as the comparison standard, and the McIntosh-class Poisson method: maximum TSS 0.53 for M-class over 24 hr).*
- **Brier, G. W. 1950**, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
- **Calvet, L. E., & Fisher, A. J. 2001**, "Forecasting multifractal volatility," *Journal of Econometrics*, 105, 27–58. doi:10.1016/S0304-4076(01)00069-0
- **Calvet, L. E., & Fisher, A. J. 2004**, "How to Forecast Long-Run Volatility: Regime Switching and the Estimation of Multifractal Processes," *Journal of Financial Econometrics*, 2, 49–83. doi:10.1093/jjfinec/nbh003
- **Camporeale, E., & Berger, T. E. 2025**, "Verification of the NOAA Space Weather Prediction Center Solar Flare Forecast (1998–2024)," *Space Weather*, 23, e2025SW004546. doi:10.1029/2025SW004546 (arXiv:2508.01114) *(the closest precedent and this paper's template; Table 2 supplies the 24-hour M-class baseline comparison quoted in §4).*
- **Deshmukh, V., Baskar, S., Berger, T. E., Bradley, E., & Meiss, J. D. 2023**, "Comparing feature sets and machine-learning models for prediction of solar flares," *A&A*, 674, A159. doi:10.1051/0004-6361/202245742
- **Hudson, H. S., Cliver, E. W., White, S. M., et al. 2024**, "The Greatest GOES Soft X-ray Flares: Saturation and Recalibration over Two Hale Cycles," *Solar Physics*, 299, 39. doi:10.1007/s11207-024-02287-x *(the 1.43 recalibration factor; §10 and Appendix B).*
- **Janssens, J., Delouille, V., Clette, F., & Andries, J. 2025**, "Solar flare rates and probabilities based on the McIntosh classification: Impacts of GOES/XRS rescaling and revisited sunspot classifications," *JSWSC*, 15, 14. doi:10.1051/swsc/2025007 *(55 % and 52 % more M- and X-class events over 1976–2019 under the rescaling; §10).*
- **Jolliffe, I. T., & Stephenson, D. B. (eds.) 2012**, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed. (Chichester: Wiley). doi:10.1002/9781119960003
- **Leka, K. D., Park, S.-H., Kusano, K., et al. 2019**, "A Comparison of Flare Forecasting Methods. II. Benchmarks, Metrics, and Performance Results for Operational Solar Flare Forecasting Systems," *ApJS*, 243, 36. doi:10.3847/1538-4365/ab2e12 (arXiv:1907.02905) *(the BSS reference convention quoted in §2.2, the `MSESS_clim` alternative, and the "no single winner" finding).*

- **Lippiello, E., de Arcangelis, L., & Godano, C. 2010**, "Time-energy correlations in solar flare occurrence," *A&A*, 511, L2. doi:10.1051/0004-6361/200913784 (arXiv:1003.0308) *(the $\Delta t < 2$ h deficit, the 2–4 h excess and the $\approx 1/t$ decay after main flares; reconciled in §8.3).*
- **Matzka, J., Stolle, C., Yamazaki, Y., Bronkalla, O., & Morschhauser, A. 2021**, "The Geomagnetic Kp Index and Derived Indices of Geomagnetic Activity," *Space Weather*, 19, e2020SW002641. doi:10.1029/2020SW002641
- **McCloskey, A. E., Gallagher, P. T., & Bloomfield, D. S. 2018**, "Flare forecasting using the evolution of McIntosh sunspot classifications," *JSWSC*, 8, A34. doi:10.1051/swsc/2018022
- **Sargent, H. H. 2021**, "A Revised 27-day Recurrence Index," arXiv:2101.02155. doi:10.48550/arXiv.2101.02155 *(the construction of the recurrence index §9.3 gestures at; the baseline used here is a lagged-count-plus-harmonics regression, not this index).*
- **Shprits, Y. Y., Vasile, R., & Zhelavskaya, I. S. 2019**, "Nowcasting and Predicting the Kp Index Using Historical Values and Real-Time Observations," *Space Weather*, 17, 1219–1229. doi:10.1029/2018SW002141 *(the lead-time structure §9.3 replicates).*
- **Snelling, J. M., Johnson, J. R., Willard, J., Nurhan, Y., Homan, J., & Wing, S. 2020**, "Information Theoretical Approach to Understanding Flare Waiting Times," *ApJ*, 899, 148. doi:10.3847/1538-4357/aba7b9
- **Stanislavsky, A. A., Nitka, W., Małek, M., Burnecki, K., & Janczura, J. 2020**, "Prediction performance of Hidden Markov modelling for solar flares," *JASTP*, 208, 105407. doi:10.1016/j.jastp.2020.105407
- **Wheatland, M. S. 2000**, "The Origin of the Solar Flare Waiting-Time Distribution," *ApJ*, 536, L109–L112. doi:10.1086/312739
- **Wheatland, M. S. 2001**, "Rates of Flaring in Individual Active Regions," *Solar Physics*, 203, 87–106. doi:10.1023/A:1012749706764 *(obscuration; "in the absence of obscuration the number of flares above C1 class would be higher by (75 ± 23) %", §8.3).*
- **Wheatland, M. S. 2004**, "A Bayesian Approach to Solar Flare Prediction," *ApJ*, 609, 1134–1139. doi:10.1086/421261
- **Wheatland, M. S. 2005**, "A statistical solar flare forecast method," *Space Weather*, 3, S07003. doi:10.1029/2004SW000131 *(the operational event-statistics forecaster; a recency rate estimator, and therefore closer to this paper's persistence baseline than to its 365-day one).*
- **Woodcock, F. 1976**, "The Evaluation of Yes/No Forecasts for Scientific and Administrative Purposes," *Monthly Weather Review*, 104, 1209–1214. doi:10.1175/1520-0493(1976)104<1209:TEOYFF>2.0.CO;2

Instrument and data. GOES/XRS flare event lists: NOAA National Centers for Environmental Information (NCEI) legacy reports, 1975–2017, and the GOES-R L2 XRS flare summary product, 2017–present. Kp: GFZ Helmholtz Centre for Geosciences, Geomagnetic Observatory Niemegk, definitive values, data doi:10.5880/Kp.0001 (CC BY 4.0). Dst: World Data Center for Geomagnetism, Kyoto, final and provisional values, data doi:10.17593/14515-74000 (registered in the Handle System and resolving to the NICT/ISDS data-DOI landing page, which returned 503 at the time of writing; Nose, Iyemori, Sugiura & Kamei 2015, after Sugiura 1964). Sunspot number: WDC-SILSO, Royal Observatory of Belgium. F10.7: Dominion Radio Astrophysical Observatory, Penticton, via NRCan and NOAA NCEI.

The MSM-Cox model definition, its closed forms and its metrology are in `papers/01-correspondence-metrology.md`; the tournament that produced the forecaster verified here is in `papers/06-flare-rate-cascade.md`.

Data availability & acknowledgments

This paper is itself deposited as a standalone Zenodo preprint at 10.5281/zenodo.22760517 (Markdown, PDF and figure files; DOI reserved 2026-09-15; published 2026-09-15 on the owner's sign-off), citing every record it reads by DOI in its metadata. The source repository is private and remains private.

The flare event lists are the public NOAA products named above, fetched at run time and **not committed**; the frozen per-day trace, its SHA-256 and its fetch provenance are sealed in `out/v44/`. The geomagnetic feeds are fetched at run time and not committed; the **derived** daily storm-interval count series `kp_storm_counts.csv` is committed in `out/v53/` under the Kp feed's CC BY 4.0.

The records, by DOI. The source repository is private and remains private, so records are cited by DOI, never by repository path.

record	what it carries here	DOI
`v44`	the sealed per-day forecast trace and the bridge window (§2.1)	10.5281/zenodo.22759669
`v45`	the calibration maps, the paired re-scoring, the era-symmetry control (§3)	10.5281/zenodo.22759673
`v46`	the cheap-challenger tournament and the stack control (§4)	10.5281/zenodo.22759675
`v47`	the depth and flat-HMM ablation in forecast metrics (§5)	10.5281/zenodo.22759677
`v48`	the first covariate run and the quiet-year floor (§6)	10.5281/zenodo.22759679
`v49`	the per-horizon verification curve (§7)	10.5281/zenodo.22759683
`v50`	the real-label X-class re-run (§8.1)	10.5281/zenodo.22759685
`v51`	the repaired F10.7 supply and the extended β grid (§6)	10.5281/zenodo.22759687
`v52`	the marked flare channel, its battery and the completeness probe (§8.2–8.3)	10.5281/zenodo.22759689
`v53`	the geomagnetic data layer and the pinned streams (§2.1, §9.1)	10.5281/zenodo.22759693
`v54`	the synthetic two-world atlas (§9.2)	10.5281/zenodo.22759695
`v55`	the real Kp tournament and both sentences (§§9.2–9.3)	10.5281/zenodo.22759697

The companion paper is `papers/06-flare-rate-cascade.md`, concept DOI 10.5281/zenodo.22180472.

Acknowledgments and terms. GOES/XRS soft-X-ray flare event lists courtesy of NOAA's National Centers for Environmental Information and Space Weather Prediction Center. *Kp and ap indices: GFZ Helmholtz Centre for Geosciences, Geomagnetic Observatory Niemegk (Matzka et al. 2021, doi:10.1029/2020SW002641; doi:10.5880/Kp.0001), CC BY 4.0.* Dst index: World Data Center for Geomagnetism, Kyoto — the WDC's terms permit scientific use without restriction but **prohibit commercial application of the indices** and require acknowledgment and citation of the data DOI; nothing Dst-derived is committed in this program, and the clause is named here so that no downstream consumer inherits it by analogy from the Kp series, which is CC BY 4.0. Sunspot number courtesy of WDC-SILSO, Royal Observatory of Belgium, Brussels (daily total sunspot number, Version 2.0), distributed under **CC BY-NC 4.0**; like the Dst clause that non-commercial term is named here so that no consumer inherits a licence by analogy from the CC BY 4.0 Kp series, and nothing SILSO-derived is committed in this program. F10.7 solar radio flux courtesy of the Dominion Radio Astrophysical Observatory, Penticton, operated by the National Research Council of Canada with Natural Resources Canada, under Government of Canada terms (non-commercial reproduction with attribution; NRC requests a reference where the data are used).

All analysis code — the calibration maps, the challengers, the ablation, the covariate exposure, the h-step filter, the marked channel and the geomagnetic tournament — is new Fenocosm work built on the vendored CC BY 4.0 MSM-Cox generator and exact forward algorithm (see PROVENANCE.md). All original content is CC BY 4.0, © 2026 Evan Tabak Atlas.

Appendix A — Provenance table (every headline number → its sealed record)

Every **bolded in-repo** number in this paper is bound to a sealed record's JSON by the repository's paper-anchor gate (`fenocosm.verify.check_paper_anchors`), which renders the value from the record and asserts it appears verbatim here; a re-cut that moves a number fails the gate until the paper is swept in the same commit. External numbers — literature values, and anything not produced by a sealed record — are **not** bolded and **not** anchored, and are marked *external* below.

number (as used)	record	file → JSON key
Platt map **a = −0.369440**, **b = +0.924339**	v45	`p103_nowcast_calibration.json` → `calibration_maps.platt.a`, `.b`
tail BSS **+0.3385** (raw) → **+0.3645** (calibrated)	v45	`verdict.raw_bss`, `.calibrated_bss`
calibration paired ΔBSS **+0.026 [+0.018, +0.035]**	v45	`verdict.calibrated_delta_bss`, `.calibrated_delta_bss_ci95`
cascade − persistence paired **+0.021 [+0.012, +0.030]**	v45	`verdict.raw_minus_persistence`, `.raw_minus_persistence_ci95`
worst calibrated reliability-bin deviation **0.075**	v45	`verdict.reliability_max_abs_dev_calibrated`
mean issued 0.195 → 0.168 vs observed 0.144; raw worst bin 0.170; peak TSS 0.621 / 0.633; isotonic and rolling-origin columns; the 58-of-18,871 era-symmetry control; the bridge window's 0.252 → 0.401	v45	`verdict.*`, `windows.tail.entrants.*`, `windows.bridge.entrants.*`, `era_covariate.bins_where_era_moves_persistence`
logistic tail BSS **+0.3734**	v46	`p104_nowcast_challengers.json` → `verdict.bss.logistic`
verdict: calibrated cascade − logistic **−0.0089 [−0.0165, −0.0019]**	v46	`verdict.delta_bss`, `.ci95`
stack − logistic **+0.0015 [−0.0022, +0.0051]**	v46	`verdict.stack_minus_logistic`, `.stack_minus_logistic_ci95`
raw cascade − logistic −0.0349 [−0.0494, −0.0235]; glm2 +0.3315 and its `log1p` twin +0.3536; the previous-day flag +0.2385; the eleven sealed logistic weights; the bridge ordering	v46	`verdict.raw_minus_logistic*`, `verdict.bss`, `glm2_log1p_sensitivity`, `coefficients.logistic`, `windows.bridge`
raw depth gain `k = 6 − k = 2` **+0.0729**	v47	`p105_nowcast_ablation.json` → `verdict.depth_deltas.k6_minus_k2.raw`

number (as used)	record	file → JSON key
[+0.0527, +0.1069]		
calibrated `k = 6 − k = 4` **+0.0032 [−0.0009, +0.0070]**; cheapest equivalent depth **k = 4**	v47	`verdict.depth_deltas.k6_minus_k4.platt`; `verdict.cheapest_equivalent_k`
calibrated `k = 6 −` best `J` **+0.0009 [−0.0090, +0.0097]**	v47	`verdict.k6_minus_bestJ.platt`
likelihood margin ΔBIC **97**	v47 *(derived)*	`hmm_fits.5.bic − cox_fits.6.bic`
the k- and J-panels' BICs, held-out values and calibrated scores; the bridge J = 5 reversal; the exact continuity certificate against the sealed ladder	v47	`cox_fits.*`, `hmm_fits.*`, `windows.*`, `continuity_vs_v26`
covariate ΔBSS **+0.0507 [+0.0382, +0.0665]** and the sealed verdict **THE COVARIATE PAIR IS SERVABLE ON THE SPLICED SUPPLY**	v51	`p109_nowcast_f107_supply.json` → `kill_rule.selected_dbss_verification`, `.selected_ci95`, `.verd`
refit ladder span **S = 146.8 → S = 4.16**	v51	`grid[β=0].ladder.S`, `grid[β₁=0.5, β₂=2.5].ladder.S`
the quiet-year floor 0.0533 → 0.0308 and −0.1054 → +0.0451; the SN-only arm +0.0269 [+0.0160, +0.0386]; the earlier cell +0.0433 [+0.0317, +0.0571]; 0 of 7,305 tail bins beyond coverage against 2,436; the 97.40 % splice agreement; (β_1, β_2) = (0.50, 2.50)	v51	`quiet_year_floor`, `entrants.*.{raw,platt_calibrated}.quiet_years`, `kill_rule.*`, `covariates.f107_spliced`, `selection`
the first covariate run and its	v48	`p106_nowcast_covariates.json` → `kill_rule`, `f107_subtail`, `quiet_year_floor`

number (as used)	record	file → JSON key
availability kill		
cascade − persistence **+0.0210 [+0.0124, +0.0294]** (h = 1) and **+0.0861 [+0.0515, +0.1230]** (h = 14)	v49	`p107_nowcast_horizon.json` → `crossing.per_h.{1,14}.raw`
h-matched persistence BSS **−0.0043** at h = 14	v49	`curve.14.tail.persistence_h.bss_vs_verification`
cascade − logistic **−0.0349 [−0.0499, −0.0230]** at h = 1	v49	`curve.1.paired.cascade_raw_minus_logistic_h.verification`
Table 1's window (**7,305** days at reference **0.1441**) and all of its rows; the calibrated columns; the cascade − cycle crossing at h = 3 and its negative arm from h = 5; the per-h Platt maps and Murphy terms; the leakage probes; the bridge curve	v49	`sealed_anchors.n_tail`, `.tail_base_rate`, `curve.*.tail`, `curve.*.paired`, `curve.*.platt_maps`, `leakage_probes`, `curve.*.bridge`
real-label X BSS **+0.1022**, beating X persistence by **+0.0718 [+0.0256, +0.1274]**	v50	`p108_nowcast_xclass.json` → `kill_rule.best_bss_vs_verification`, `.best_vs_persistence`
568 X flares, 122 tail X-days, climatology 0.016701; the k = 6 X ladder +0.0796 and +0.0492 [+0.0322, +0.0786]; k = 6 − k = 2 +0.0296 [+0.0101, +0.0516]; the scalar route's +0.0194	v50	`x_series`, `entrants.*`, `paired_comparisons.*`
constant flare-level X fraction **q = 0.0713**	v52	`p110_flare_marks.json` → `v50_decomposition.q_train_event`
state route − rate-only **+0.0170 [−0.0097, +0.0514]**	v52	`v50_decomposition.paired_comparisons.v50_state_fraction_platt_minus_rate_only_platt.verifica`

number (as used)	record	file → JSON key
primary per-event gain **+0.0029 [−0.0006, +0.0074]** at M1.0	v52	`thresholds.M1.0.primary.gain`
battery recovery on the stronger planted world, **3/6** seeds at M1.0	v52	`battery.summary["planted_theta0.05@M1.0"].n_recovered / .n_seeds`
low-band share slope **−0.0789 [−0.1059, −0.0477]** per net unit	v52	`detection_probe.vs_belief_net.slope_per_net`
low-band share slope vs lagged sunspot number **−0.0559 [−0.1748, +0.0548]**	v52	`detection_probe.vs_sunspot_number.slope_per_log1p_SN`
the M2.0 and M5.0 arms; the three null ranks; the unit-harmonised +0.0035 [+0.0010, +0.0069]; the ×5.2 planted swing; the index-slope readout; the 0.632 → 0.505 decile shares; the sealed sentences	v52	`thresholds.*`, `sensitivity_M1.unit_harmonised`, `battery.generator.x_fraction_span_over_net`, `detection_probe.low_band`, `verdict`
34,577 calendar days; lag-27 storm-day autocorrelation **0.193**	v53	`p111_geomag_data.json` → `grid.n_days`; `kp_streams["5-"].acf_storm_day["27"]`
19,157 storm intervals (6.93 %), mean 0.554/day, P(storm day) 0.215, lag-1 0.378 and lag-7 0.068; the split at 1988-10-19 and its two references; the CSV's licence, SHA-256 and attribution	v53	`kp_streams["5-"]`, `grid`, `anchor_csv`
atlas verdict **UNDECIDABLE AT THIS N ON P1, S1, S2**	v54	`p112_geomag_atlas.json` → `verdict.word`
the depth-gain contrast 1.45–1.78×; the switch world's k* = 4	v54	`summary.*`, `verdict.per_seed`, `matching_table.worlds`, `matching_table.real`

number (as used)	record	file → JSON key
and S ≈ 2; the recovered cascade ladders; the under-matched null's lag-27 0.021 / 0.025 / 0.019 against the real 0.193 and its cycle contrast 1.17 / 1.11 / 1.11 against 3.06		
Sentence 1 **THE CASCADE DOES NOT CARRY SKILL WORTH SERVING**; the h = 2 failing pair **−0.0412 [−0.0560, −0.0238]**	v55	`p113_geomag_tournament.json` → `sentence_1.word`; `sentence_1.checks[h=2, recurrence_h]`
cascade (Platt) **+0.1499** at h = 1 and **−0.0407** at h = 27; recurrence **+0.0575** → **+0.0511**; logistic **+0.2043** → **+0.0862**	v55	`bss_curve_verification["5-"].{1,27}`
k* = 3 ladder flagged degenerate at **S = 1.00**	v55	`results["5-"].kscan.panel["3"].ladder.S`, `.degenerate`
Table 2's window (**13,831** days at reference **0.1695**) and all of its rows; the h = 1 rival margins; the calibrated-persistence pairs; the registered Sentence 2 and the decomposition entrant's intervals; the reference gap +0.1831 / +0.1499; the HMM panel and the 445-BIC margin; the k = 4 / 5 readout; the Kp ≥ 6− sensitivity; the closed Stage-3 gate	v55	`results["5-"].split.n_tail`, `reference_gap.verification_ref`, `bss_curve_verification`, `sentence_` `sentence_2`, `results["5-"].{kscan,hmm_scan,per_h,supplementary_pairs}`, `stage3_gate`
every literature value in §§1, 2.3,	*external*	see §14; each verified against a primary source

number (as used)	record	file → JSON key
4, 5, 8.3, 9.3 and 10		

Appendix B — The registered next cut, and companion work

P-12.13 — the event-metric battery (registered 2026-09-15). The measurement this paper owes the operational literature, and the one §2.3 declines to make outside a record: for every entrant in records `v45` – `v51` and `v55`, on the same tails and the same paired panel, a full contingency battery under the community's event-definition convention (`M1.0+/0/24` in the Barnes et al. 2016 notation) at a **declared** probability threshold fixed on the training block rather than maximised on the test set — POD, FAR, CSI, precision, accuracy, frequency bias, TSS and HSS — reported beside the threshold-maximised TSS the current records seal, so that the optimism of the latter is priced; committed reliability diagrams for the flare entrants beside the sealed bin tables; and the Brier decomposition into reliability, resolution and uncertainty already sealed, redrawn per entrant. Bloomfield et al. (2012) is the standard the TSS column answers to and Leka et al. (2019) the discipline the threshold choice answers to. Nothing in §§3–9 changes if it comes back; what changes is whether this paper's entrants can be placed beside an operational table.

The GOES rescaling robustness run. §10 records that these arcs carry the satellite-era discontinuity as a five-step per-era multiplier on the latent rate, not as Hudson et al.'s (2024) uniform 1.43 factor, and that Janssens et al. (2025) find 55 % and 52 % more M- and X-class events over 1976–2019 under the SWPC rescaling. A re-cut of the calibration, challenger and ablation records on a uniformly rescaled event list — a new record beside each, never an edit — is what would price the covariate's coarseness. Record `v52`'s unit-harmonisation cell is a partial, marks-only version of it and already moves a headline interval.

A sub-daily cut. §8.3 reconciles this paper's daily null against Lippiello et al. (2010) by timescale and aggregation level, and cannot do better than that at daily resolution. A 6-hour or 12-hour cut of the same tournament, with an overdispersed (negative-binomial) emission in the roster, is what would let the analysis engage the waiting-time and time–energy literature on its own ground and separate "multi-state rate structure" from "daily counts are not conditionally Poisson."

A completeness-safe subsample. §8.3's probe identifies a detection-completeness signature and cannot separate it from a physical shift. Repeating the depth, challenger and marked-channel measurements on a subsample restricted to low-background intervals — the measurement a solar referee will ask for first — is the control that would settle it.

The geomagnetic arc's named next cuts (from record `v55`): a recurrence baseline carrying an *adaptive* climatology as the registered domain baseline; depth selection by training-block forecast score rather than BIC; a binomial counting channel for the bounded count; solar-wind and F10.7 covariates; and the Dst storm-onset stream, which record `v53` pins and validates and which carries the WDC non-commercial clause with every number derived from it.

Companion work, named and out of scope. The model and its metrology are in `papers/01-correspondence-metrology.md`; the likelihood tournament that produced the forecaster verified here, with its own limitations, is in `papers/06-flare-rate-cascade.md`. Neither is restated here, and neither number in them is changed by this paper: this paper measures what the forecast is worth, not what the record prefers.