



Supplement of

A physically grounded input distance resolves longitude-encoding ambiguity but not specification dependence in ensemble persistent homology of cyclone trajectories

Rongzhen Dai

Correspondence to: Rongzhen Dai (dairongzhen3@gmail.com)

The copyright of individual parts of the supplement might differ from the article licence.

1 **Supplementary Material**

4 Section numbers prefixed S (e.g. Section S1) refer to this document; bare section numbers refer to the main
5 text. Tables in this document are numbered S1, S2, ... in order of appearance. Results of the main text are
6 referred to by the section in which they appear, so that the two numbering schemes cannot be confused.

8 **S1. Values that supersede those of the discussion paper**

9 The pipeline behind the discussion paper was a single-run implementation whose control array was built from
10 an unordered set, so it drew a different control sample in every process even with the random seed fixed, and
11 whose interpreter hash seed was not recorded (§3.3.4). Correcting that, and auditing what the corrected
12 pipeline reproduces, changes several reported values. Table S1 lists every superseded value, what replaces
13 it, and why. The frozen outputs of the superseded runs are in the archive, so each entry can be inspected.

14 **Table S1.** Values reported in the discussion paper that are superseded, the values that replace them, and the
15 reason for each change.

Reported in the discussion paper	Value now reported	Why
D1 pool H ₁ resampling p = 0.130 / 0.009 / 0.214	0.037 / 0.021 / 0.156 (§4.1)	The sphere and cylinder values reflect the original run's process-randomized control order; five independent orderings of the sphere give 0.007–0.018 and five of the cylinder give 0.145–0.212, bracketing the published values. The linear value is not reproducible under any ordering (five give 0.026–0.046), because the linear normalizer of that run was overwritten and cannot be audited.
D1 pool H ₀ resampling p = 0.088 / 0.082 / 0.190	0.052 / 0.059 / 0.067 on the deterministic pipeline; not carried in the manuscript, which reports the H ₀ channel through the placebo calibration of §4.2	The same defect and the same correction. The H ₀ channel was read as calibrated and encoding-robust. The corrected values lie closer together, and the encoding dependence persists in them, but a single-run H ₀ p-value is exactly the quantity §3.5 argues should not be read alone, so §4.1 reports the H ₀ event distance and leaves the channel's verdict to the calibration.

Reported in the discussion paper	Value now reported	Why
Event H_1 distances 0.538 / 0.646 / 0.529	0.353 / 0.656 / 0.546 (§4.1)	The sphere and cylinder normalizers survive in the archived code and reproduce to within a few percent. The linear one does not: 0.538 differs by 52 % from what the §3.3.1 encoding produces, which is how the overwriting was detected.
H_1 placebo false-positive rates 8.2 % / 10.2 % / 14.3 % (4, 5 and 7 of 49), read as an inflation “by a factor 1.6× to 2.9× regardless of encoding choice”	1, 4 and 4 of 49 on the deterministic ordering; pooled 9–12/147 across six auditable orderings, all nominal-compatible (§4.2)	Every cell came from the process-randomized order that also produced the unauditable linear row, so none can be placed among the auditable orderings. Under the reproducible pipeline the inflation does not appear and the factor claim is withdrawn. Separately, the H_0 channel does reach 16 of 147 at one of the six orderings (Section S2); that is a different quantity which happens to share the pooled count of the published H_1 cells.
H_0 placebo false-positive rates 8.2 % / 4.1 % / 6.1 %	2, 1 and 2 of 49 on the deterministic ordering; pooled 4–16/147 across the same six orderings (Section S2)	Same defect. Reporting H_0 as a distribution rather than one ordering also reverses the dimension-asymmetry reading: the H_0 spread is the wider of the two and its upper ordering crosses the nominal rate.
“An intrinsic H_1 inflation that is not curable by lon-encoding choice alone”; “dimension-asymmetric in calibration”	Withdrawn. No general inflation of H_1 over H_0 is claimed (§4.2, §5.2)	Under the linear encoding H_1 (1/49) sits below H_0 (2/49). What survives is narrower: the H_1 verdict sits near α at every ordering while the H_0 verdict is far from it at most.
The diagram distance described as Wasserstein-2	Degree-1 Wasserstein W_1 (§3.2)	<code>persim.wasserstein</code> carries no exponent argument and returns the plain sum of matched costs under a Euclidean ground metric. The description misnamed the implementation while leaving it correct; the true degree-2 distance is now reported separately as a sensitivity (§4.5).

Reported in the discussion paper	Value now reported	Why
Linear-encoding results computed on the archived longitude field	The same results, plus the single-convention value 0.108 and the branch-cut range 0.022–0.108 (§3.3.1, Section S7)	The archived IBTrACS snapshot mixes two longitude conventions, so 2.19 % of fixes sit 1.0 above their counterparts in the normalized coordinate. The sphere, cylinder and great-circle representations are unaffected.
Stratified attribution, cylinder column 0.058 / 0.488 / 0.400 / 0.044 / 0.736	0.048 / 0.590 / 0.338 / 0.306 / 0.788 (§4.3)	The script behind that table carried the same unordered-set defect and depends on an unrecorded hash seed; re-running it under two different seeds gives two different sets. The corrected script gives bit-identical output under both. No cell survived correction before and none does now.
Subsample-size sensitivity, linear column 0.420 / 0.124 / 0.144 / 0.590	0.062 / 0.048 / 0.048 / 0.026 (§4.4)	Same defect and same correction. The regenerated column is monotonic where the earlier one was not; §4.4 reads both as sampling variability of a single seeded realization.
Event statistics quoted without reference to which control diagrams formed the reference set	The reference set is recorded with every result and pinned across resolutions (§3.6)	The fifty reference diagrams are drawn from the pool of null draws, so enlarging the pool changes them: at $B = 10,000$ without pinning, the linear event distance moves from 0.352771 to 0.349897. That is a change in the estimand.
Event windows quoted without stating which date convention defines them	Both conventions stated, with the labelled dates used throughout (§2, Section S5)	The storm-label dates and the Kyoto WDC index-minimum dates differ by one day for two of the three events; recentring on the latter admits 24 storms rather than 25 and gives a basin-matched subsample-averaged separation of 2.01σ against 1.87σ , and a resampling p-value of 0.039 against 0.044. No conclusion changes.

Reported in the discussion paper	Value now reported	Why
Interactive-discussion comment: “the published value, 16, is the single failing tail draw”	Superseded by §4.2 and by the placebo rows above	The attribution to control-order randomization is wrong for the linear row, whose normalizer was overwritten; and the count cannot be placed among the auditable orderings at all.
Interactive-discussion comment: “the corrected, deterministic pipeline at its median implementation is the first documented nominally-calibrated configuration of this class”	Withdrawn. The 49-placebo calibration is nominal in H_1 at every auditable ordering, and no claim of precedence is made (§4.2)	The statement rests on the same seven-implementation campaign whose pooled count cannot be placed among the auditable orderings.
Interactive-discussion comment: a fourth protocol requirement, “report calibration as a distribution over control-resampling implementations”	The protocol carries five requirements and this is not among them; the ordering spread is a diagnostic check on requirement 5 (§3.5, §5.6)	Reordering a uniformly resampled control pool selects a different realisation while the estimand does not move, so the spread diagnoses the resampling resolution rather than forming the quantity to report. What accompanies a p-value is its Monte-Carlo interval.
Interactive-discussion comment: “a small robust false-positive core (two placebo dates under two encodings) persists across implementations and is reported as such”	Not carried	No auditable run reproduces that core. The threshold-straddling population quoted alongside it is carried, and is reported at §4.2 together with the dispersion asymmetry measured on the reproducible cache.
Interactive-discussion comment: “the verdict is single-valued and stable across seven control-draw orderings (per-ordering 2–4/49)”	The verdict is stable across three control-draw orderings, within [0.006, 0.017]; the seven-ordering figure and the 2–4/49 counts are the placebo calibration, a different quantity (§4.5)	Two quantities were combined in one sentence. The verdict was tested on three orderings and the calibration on seven, and the verdict is not stable across the event-subsample axis at all (§4.1).
Interactive-discussion comment: “a full multi-seed distribution with encoding-matched sample indices is left to a larger study”	The multi-seed scan is now reported, with identical sampled point indices across representations (§4.1, Section S6)	The scan was within reach and has been carried out. It shows that the particular combination of verdicts reported in the discussion paper occurs in one of six subsample draws.

16 Two further points about the corrections themselves. The verification script re-derives all 87 registered spec-
17 ifications from the archived data and compares each with the value printed in the manuscript; all 87 agree.
18 What that check does and does not reach is set out in Section S10, and it is narrower than the phrase “all
19 87” suggests. And the pipeline is now deterministic by construction rather than by convention: determinism

is specified as an explicit sort of the stratum and identifier lists, because pinning the interpreter hash seed reproduces one historical run while leaving two logically equivalent implementations free to build a set in different orders.

S2. Auditable-ordering calibration — per-ordering detail

Table S2 gives the pooled placebo false-positive count in each homology channel over the six auditable control-draw orderings, from which the summary of §4.2 is read. The great-circle physical distance at $\lambda = 1$ is nominal across seven orderings, with per-ordering H_1 counts of 2 to 4 of 49. Full per-placebo, per-ordering tables for the encodings and for the physical distance are in the archived `results/placebo_test_*.csv`.

Table S2. Placebo false-positive counts of 147 ($49 \text{ placebos} \times 3 \text{ encodings}$) over the six auditable control-draw orderings, $B = 100$ null draws per placebo. The per-encoding breakdown is recorded for the deterministic ordering; for the five seeded permutations of it the archive records the pooled count, so those are given as the set of values they take rather than assigned to a particular permutation. The three encodings in each count share the same 49 placebo dates and are correlated, so the exact-binomial p-values quoted below understate the sampling uncertainty and are used only to place each count relative to the nominal boundary.

Ordering	H_1 linear	H_1 sphere	H_1 cylinder	H_1 pooled	H_0 pooled
deterministic sort	1	4	4	9	5
five seeded permutations	—	—	—	9, 10, 10, 11, 12	4, 8, 11, 11, 16
range over the six	—	—	—	9–12	4–16

Every H_1 count is nominal-compatible, the largest giving an exact two-sided binomial p of 0.086 against the nominal 5 % rate and the smallest 0.45. The H_0 counts span the wider range in absolute terms and the upper ordering, 16/147, crosses the nominal rate at $p = 0.0035$. What distinguishes the channels is that the H_1 verdict sits near α at every ordering while the H_0 verdict is far from it at most (§5.2). We do not read the H_0 upper ordering as a finding in its own right: the six orderings and the three encodings share one placebo set and one data set, so they are not independent trials.

S3. Pre-registration template and multi-representation agreement rule

The requirements of §5.6 are operationalised below as a form to be completed and deposited **before** the pooled test result is read. Entries in brackets are filled per study.

S3.1 Representation and distance. Primary distance: a physically interpretable metric on the trajectory point cloud (e.g. great-circle geometry plus an explicitly justified attribute weight; here on the IBTrACS best-track archive, Knapp et al., 2010), stated in closed form $d(i, j) = [\dots]$; every free parameter and its sensitivity grid listed $[\dots]$; at least two principled coordinate encodings retained as a robustness check, reported alongside the physical distance and not in place of it.

S3.2 Coordinate conventions are normalised before encoding. Any periodic coordinate is reduced to a single convention across the whole input, and the reduction is stated. Where the source archive mixes

conventions — as IBTrACS does for longitude, §3.3.1 — an encoding that maps the raw field linearly will place two representations of the same location at opposite ends of the normalised range. Where a branch cut is unavoidable, its position is registered in advance and its effect reported (Section S7).

S3.3 The exponent of the matching cost is registered in advance. The exponent is a free parameter of the comparison and not a property of the data, and on this dataset it moves the event-level p-value from 0.003 to 0.087 (§4.6). This is the situation Gelman and Loken (2014) describe: a choice made after seeing the data leaves the nominal p-value meaning something other than what it appears to. Either register the exponent before the pooled result is read, or report a statistic that selects none, in the spirit of the specification-curve construction of Simonsohn et al. (2020) — the specification-averaged construction of §4.7.2 calibrates a family of exponents against the same resampling null at no extra cost. Report the effective number of contributing matching costs for each candidate exponent: it is computable from the diagrams alone and, under the approximations of §4.6.3, envelopes the widening that raising the exponent will produce (Section S8).

S3.4 The subsample realisation is averaged, never selected. Where diagrams are computed from a fixed-size subsample of a larger ensemble, the statistic is averaged over $M \geq [\dots]$ independent subsamples for the event and for every control draw alike. A single realisation is a Monte-Carlo draw and not a result: on this dataset the D1 verdict changes across six event-subsample seeds (§4.1), and the particular combination reported in the discussion paper occurs in one of them.

S3.5 The reference set is registered. The event statistic is a mean distance to a fixed number of control diagrams, and which ones is a choice of the same weight as the resampling resolution. Register the size of the reference set and the rule that draws it, record the drawn indices with the results, and pin them when the number of null draws is changed, since enlarging the null pool otherwise changes the reference set and with it the estimand (§3.6). Report the sensitivity of the headline value to at least $[\dots]$ alternative reference sets drawn under the same rule; on this dataset five alternatives move the linear resampling p-value over 0.036 to 0.047 against 0.037 for the archived set, a spread that drawing more null draws does not reduce.

S3.6 Monte-Carlo reproducibility and calibration. Placebos $n \geq 200$; $B \geq [\dots]$ null draws per placebo, chosen so that the Monte-Carlo interval on a placebo p-value is narrow relative to α ; that interval reported alongside every estimate; pooled H_0 and H_1 false-positive counts reported **separately**, with per-proportion Wilson (1927) intervals.

Varying the control-draw ordering is a **diagnostic sensitivity check, not a reporting requirement**. Reordering a uniformly resampled control pool selects a different Monte-Carlo realisation while the estimand does not move (§3.5), so a spread across orderings measures the finite-B term rather than a specification axis, and reporting the calibration as a distribution over orderings would give that spread a standing it does not have. Run the check once, on enough orderings to confirm that no verdict in the study rests on a single realisation — this study used six for the encodings and seven for the physical distance — and report the deterministic result thereafter, with its Monte-Carlo interval.

Determinism. The control ordering is fixed by an explicit rule — `sorted(...)` on a stated key — and the rule is stated in the methods. Fixing the interpreter’s hash seed is **not** an adequate substitute: it makes one implementation repeat itself, while two implementations that are logically equivalent and differ only in how the control set is built can still produce different orderings under the same seed, and the archived code of this study did (§3.3.4). Report the sorting key, not the seed.

S3.7 Matched null. The control draw preserves the event ensemble’s composition on the axes that could dominate the geometry — at minimum the per-window (seasonal) storm counts, and additionally basin and era — with each stratum’s effect reported separately. Record which axes are *not* matched, and test each at the

unit the draw randomises rather than at the unit the data are stored in: on this dataset the wind-reporting rate of the event set exceeds the control mean per fix but not per storm, and the fix-level gap widens rather than closes under basin matching, so the axis is unmatched by all four designs, and at the storm level the event and control draws are not separable on it under the one matched design tested, nor under the unmatched baseline (Section S8).

S3.8 Pre-registered stratified attribution. Stratification scheme (e.g. basin $\times$ event $\times$ season) fixed before the pooled test is read; multiple-comparison correction stated in advance, here Benjamini and Hochberg (1995) applied within each representation over that representation’s cells; pooled-significant and stratified-null outcomes reported as such, since the pooled and stratified verdicts can disagree in the manner Simpson (1951) describes.

S3.9 Multi-representation agreement — operational decision rule. If the representations disagree on the H_1 verdict, the claim is reported as representation-dependent and **not** as a positive finding, *unless* a physically interpretable distance (Section S3.1) yields a single verdict that is stable across the control-draw ordering, the subsample realisation and the reference set. Where a quantitative threshold is needed, agreement is defined as all retained representations placing the verdict on the same side of α with overlapping Wilson intervals.

S4. Event-window storm membership

The event set is intentionally cross-basin and cross-season: the three benchmark events fall in three different cyclone-season windows (§2). A storm enters the set if any of its fixes falls in any ± 15 -day window (union rule); the ensemble then uses all lifetime fixes of each selected storm. The cross-basin imbalance and the greater recency of the event storms (mean genesis year 2011.6 against 2001.8) are intrinsic to the three calendar choices and are the subject of the matched-null controls of §4.5.

Table S3. Basin $\times$ event distribution of the 25 event-period storms. “Other” captures legacy SH-coded or missing BASIN entries.

Event	Total	WP	NI	EP	SI	SP	Other
Halloween 2003	11	6	1	1	1	0	2
St Patrick 2015	9	2	0	0	4	3	0
Gannon 2024	5	1	1	0	3	0	0
Union (unique)	25	9	2	1	8	3	2

S5. Candidate-event inclusion and exclusion log

Event selection followed an explicit set of inclusion and exclusion criteria, and we list every candidate it was applied to: every geomagnetic storm with minimum Dst ≤ -250 nT since 1981, obtained programmatically from the Kyoto WDC **final** Dst hourly series (World Data Center for Geomagnetism, Kyoto) for 1981–2020 and the **provisional** series for 2021–2024, provisional values being marked with an asterisk. The 1981 start matches the control-pool year range and the consistent-quality satellite era of IBTrACS (Knapp et al., 2010). Values are the monthly-minimum Dst at the storm peak, and **dates in this table are the day of that index minimum**, which for Halloween 2003 and Gannon 2024 is one day later than the storm-label date on which the analysis windows are centred (§2). The analysis uses the §2 dates throughout; recentring on the dates of this table admits 24 storms rather than 25, and that alternative was run through the specification-averaged

analysis and changes no conclusion (Section S1). The two conventions are kept distinct rather than reconciled because this table is built from the Dst series and the event windows are not.

Table S4. All Dst ≤ -250 nT storms, 1981–2024, plus the deepest Cycle-24 storm (†). Kyoto WDC; * = provisional.

Peak date	Min Dst (nT)	Cycle	Decision	Reason
1981 Apr 13	−311	21	excluded	Cycle 21, outside the three recent cycles; NH spring, low TC activity
1982 Jul 14	−325	21	excluded	Cycle 21
1982 Sep 06	−289	21	excluded	Cycle 21
1986 Feb 09	−307	22	excluded	Cycle 22; NH winter, low TC activity
1989 Mar 14	−589	22	excluded	Cycle 22; NH winter, minimal TC activity
1989 Sep 19	−255	22	excluded	Cycle 22
1989 Oct 21	−268	22	excluded	Cycle 22
1989 Nov 17	−266	22	excluded	Cycle 22
1990 Apr 10	−281	22	excluded	Cycle 22; NH spring
1991 Mar 25	−298	22	excluded	Cycle 22; NH spring
1991 Oct 29	−254	22	excluded	Cycle 22
1991 Nov 09	−354	22	excluded	Cycle 22
1992 May 10	−288	22	excluded	Cycle 22
2000 Apr 07	−292	23	excluded	NH spring, low cross-basin TC activity
2000 Jul 16	−300	23	excluded	single-basin (NWP only) in window
2001 Mar 31	−387	23	excluded	extreme magnitude, but a late-March window has inadequate cross-basin TC activity; Halloween 2003 has comparable magnitude with full seasonal coverage
2001 Apr 11	−271	23	excluded	NH spring

Peak date	Min Dst (nT)	Cycle	Decision	Reason
2001 Nov 06	−292	23	excluded	season-end, marginal cross-basin count
2003 Oct 30 (Halloween)	−383	23	included	extreme magnitude, cross-basin storms in window
2003 Nov 20	−422	23	excluded	within the Halloween 2003 ± 2 yr buffer; not independent
2004 Nov 08	−374	23	excluded	within the Halloween 2003 ± 2 yr buffer
2015 Mar 17 (St Patrick)	−234[†]	24	included	strongest Cycle 24 storm; below −250 nT, and Cycle 24 had no ≤ -250 nT event
2024 May 11 (Gannon)	−406[*]	25	included	strongest Cycle 25 event to date; full modern coverage
2024 Oct 11	−333 [*]	25	excluded	within the Gannon 2024 ± 2 yr buffer

Twenty-four candidates; three included, one per recent solar cycle. The selection balances magnitude, sufficient cross-basin TC activity to support an ensemble test, one benchmark per recent cycle, and independence under the ± 2 -year control-pool buffer. The Cycle 21–22 events predate the three cycles benchmarked here, and restricting to Cycles 23–25 also keeps the event windows within the consistent-quality modern intensity record, which matters because era is the leading confounder identified by the matched-null analysis (§4.5). We do not claim this is the only defensible selection; a reanalysis can adopt a different rule — for example all $\text{Dst} \leq -300$ nT regardless of TC activity, accepting reduced ensemble support — and check stability. **This criterion set is also the structural limit of §5.5:** because sufficient cyclone activity is one of the inclusion conditions, event eligibility depends in part on the quantity the test examines.

[†] *St Patrick 2015 (Dst −234 nT)* is below the −250 nT threshold and is the deepest Cycle 24 storm; excluding it would drop all Cycle 24 representation. Its cell in the stratified attribution of §4.3 survives correction in no representation, so it behaves like the other four cells.

A note on index versions, since these values are checkable. The widely cited −223 nT for St Patrick 2015 and −412 nT for Gannon 2024 are the provisional real-time values; the final (2015) and provisional (2024) Kyoto indices used here give −234 and −406, verified against the hourly Dst series. The 2005 May 15 storm reaches −263 nT in provisional Dst and −247 nT in the final index, below the threshold, and so does not appear in Table S4.

S6. ENSO and solar state of the windows, and the subsample seed scan

Climate state. A natural question is whether the event windows differ systematically from the placebo windows in climate or solar state, which could make the placebo false-positive rate reflect climatic rather than methodological variability. ENSO phases follow the NOAA CPC Oceanic Niño Index ($\text{ONI} \geq +0.5$ El Niño, ≤ -0.5 La Niña), computed on ERSSTv5 (Huang et al., 2017), and solar phase the SILSO sunspot record in its revised calibration (Clette et al., 2014). All three events fall in neutral-to-weak-El-Niño conditions, none in a strong El Niño or a La Niña, and by construction all three sit near solar-active phases, because extreme geomagnetic storms themselves cluster near solar maximum. The 49 placebo years span the full ENSO range, containing strong El Niño years (1997, 1998), multiple moderate-to-strong La Niña years (1999, 2000, 2007, 2008, 2010, 2011, 2020) and neutral years, and likewise span the solar cycle from years near minimum (1986, 1996, 2008, 2019, 2020) to years near maximum (1989, 1990, 1991, 2000).

In ENSO terms the three events sit in the middle of the placebo distribution rather than at an extreme, so the event set is not ENSO-selected relative to the placebos. In solar terms the events are, as expected, more active than a random placebo, but that is the exposure the study probes and not a confound of the calibration. That the same placebo dates — with climate and solar state held fixed — change false-positive status purely when the control-draw order is permuted (§4.2) shows a procedural contribution to the observed instability, since a climatic or solar driver cannot produce a verdict change at fixed climate and fixed solar state. It does not follow that between-placebo variability is free of climatic or solar components, and §4.2 reports the placebo rejection rate as folding methodological behaviour together with climate, era and basin variability.

Subsample seed scan. Each persistence diagram is computed from 600 of the event ensemble’s 1,880 fixes, and which 600 is a Monte-Carlo draw. Section 4.1 reports the range of the $D1\ H_1$ resampling p-value under six draws; the detail is as follows. Seed 2026 was frozen before the analyses of §§4.1–4.5 and is not a post-hoc selection. The five further seeds were added afterwards, with the control-draw ordering held at the deterministic sort, every other setting held fixed, and **identical sampled point indices across representations**, so that the rows differ only in the subsample. A positive control reproduces the archived cylinder row bit for bit (0.153, 0.283, 0.617, 0.607, 0.389) and the archived great-circle value at seed 2026 (0.012).

Two features of the scan matter for how §4.1 should be read. The verdicts are not stable across this axis: the great-circle $\lambda = 1$ verdict crosses α in one of six draws, the linear encoding in one of six, the unit-sphere encoding in three of six, and the cylinder encoding in none. Expressed as separations, seed 2026 gives 2.91 standard deviations for the great-circle representation against 1.16 to 1.75 for the other five, so the frozen seed is the outlier of the six. The relative structure, by contrast, is stable: under every seed the sphere p-value is the smallest of the three encodings and the cylinder p-value the largest. What is not reproducible is the position of any of them relative to 0.05, and the particular combination reported in the discussion paper occurs in one of the six draws. Section 4.7.1 replaces the single subsample with an average over ten, which is the treatment this axis requires; Section S3.4 states it as a pre-registration requirement.

S7. Longitude convention and the branch cut

Section 3.3.1 records that the archived IBTrACS snapshot does not use a single longitude convention: LON runs from -179.8° to $+266.9^\circ$, with 8,911 of 406,059 fixes (2.19 %) recorded east of $+180^\circ$ and 67,257 carrying the negative representation of the same meridians. The linear encoding maps that field with $\text{lon}_n = (\text{lon} + 180)/360$ and therefore returns 0.0006 to 1.2414 rather than $[0, 1]$, placing two representations of one meridian 1.0 apart. The sphere, cylinder and great-circle representations pass longitude through a periodic function before any distance is taken and are unaffected. The measured consequence, with everything else

held fixed, is in Table S5.

Table S5. Linear-encoding D1 test under four longitude treatments, 1,000 control draws, deterministic ordering, seed 2026. The first row is a positive control: it reproduces the archived event distance and resampling p-value bit for bit, so the differences in the remaining rows come from the longitude treatment alone. Run logs: `results/independent_checks/lon_arch.log` (as archived), `lon_wrap.log` (single convention), `rot0.log`, `rot90.log` and `rot180.log` (branch cut rotated). Five logs give four rows because `rot0.log` repeats the single-convention treatment at zero rotation and reproduces the value of `lon_wrap.log`; the positive control inside `rot0.log` compares that value against the as-archived one instead, and therefore prints a failure that is a mis-set reference rather than a failed reproduction (Section S10). These runs are not among the 87 registered specifications (§3.7).

Longitude handling	Event distance	σ	resampling p (H_1)	Event H_1 features
As archived	0.352771	2.13	0.037	83
Single convention, cut at $\pm 180^\circ$	0.328413	1.29	0.108	81
Single convention, cut rotated $+90^\circ$	0.322854	1.52	0.081	80
Single convention, cut rotated $+180^\circ$	0.369943	2.53	0.022	83

The linear entry of §4.1 is therefore one point in a range that a free and previously unregistered choice spans, and the linear columns of §§4.2–4.4 were all computed under the archived convention. We read this as a further instance of the dependence the paper reports and not as a separate defect of it; Section S3.2 states the corresponding pre-registration requirement.

S8. The exponent diagnostic — decomposition, scatter, and what is not controlled

Where the exponent effect comes from. Table S6 decomposes the matching cost that carries the event H_1 excess, alongside the persistence profile of §4.6.1. It is measured from the optimal assignments themselves, averaged over the fifty reference diagrams.

Table S6. Decomposition of the matching cost by persistence bin, under the great-circle distance at $\lambda = 1$, computed from the single archived event sub-sample. Columns: the fraction of event features in the bin matched to the diagonal rather than to a control feature; the bin’s share of Σc_i and of Σc_i^2 ; and the fraction of its features whose matched partner differs between $p = 1$ and $p = 2$. The bins are those of the persistence profile in §4.6.1. Positive control: the $p = 1$ costs sum to the archived event distance, 0.545580, bit for bit. Across ten sub-samples the direction is invariant — the most persistent bin’s share of Σc^2 exceeds its share of Σc in every one — while the shares themselves move, that bin’s share of Σc^2 spanning 48.7 % to 63.3 % and the four bins below 0.04 spanning 26.7 % to 34.1 %, so the entries locate the excess rather than measure it to the tenth of a percent.

Persistence bin	To diagonal	Share of Σc	Share of Σc^2	Rematched
[0, 0.005)	57 %	12.3 %	2.5 %	17 %
[0.005, 0.010)	43 %	10.9 %	4.3 %	21 %

Persistence bin	To diagonal	Share of Σc	Share of Σc^2	Rematched
[0.010, 0.020)	35 %	14.8 %	6.4 %	37 %
[0.020, 0.040)	15 %	26.5 %	17.9 %	43 %
[0.040, 0.080)	8 %	13.3 %	13.2 %	34 %
[0.080, 1.000)	52 %	14.2 %	48.7 %	14 %
Control features to diagonal	—	8.1 %	6.9 %	46 %

The diagonal-matching column shows why reasoning from persistence lands in the right place even though a feature’s cost equals its persistence only when it is matched to the diagonal: the fraction of event features matched to the diagonal falls from 57 % in the lowest bin to 8–15 % in the bins just above 0.02, so for the small features the cost is proportional to the persistence, and it is those features that squaring demotes.

Scatter of the counts. The effective counts of §4.6.3 are computed against fifty reference diagrams, and the point estimates quoted there are means over that set, which has substantial spread: for the great-circle $\lambda = 1$ representation the count under the summed statistic ranges from 16.2 to 34.5 across the fifty references, and under the squared statistic from 1.8 to 13.4. Propagated through $\sqrt{(N_1/N_2)}$, that scatter puts the envelope for a single representation between about 1.3 and 3.7 — wider than the 1.91 to 2.60 spread of the envelope across the six specifications. Single-representation values should be read with that in mind.

Direction and size of the overshoot. The envelope assumes uncorrelated contributions of comparable relative variance and the same matched costs under both exponents; the third condition fails measurably, the assignment changing for 21 % to 35 % of matched pairs between $p = 1$ and $p = 2$. Across the six archived specifications the envelope exceeds the measured widening in all six, by 5 % to 54 % and by 23 % on average, and across the sixty subsample-by-representation combinations it exceeds it in every one, with a median overshoot of 24 % on the ten-subsample median reading. We attach no inferential p-value to that count: resampling the null draws places the envelope above the measurement in every one of 20,000 bootstrap replicates for each representation tested, so sixty-for-sixty is a structural property of these data and not a set of independent trials. Its root-mean-square error is 0.44 against a standard deviation of 0.30 in the measured widenings themselves, so as a point predictor of magnitude it does not improve on reporting their mean. What the six specifications support is the direction of the widening together with an empirical envelope on its size; the rank correlation is 0.71 on the archived subsample (exact permutation $p = 0.068$ over all 720 orderings) and 0.83 on the median of ten subsamples (exact $p = 0.029$), and the six points come from one dataset and one set of 1,000 control draws.

Two checks that the association is not a proxy for something simpler. Neither ingredient predicts the widening on its own: feature count alone gives a rank correlation of -0.20 and the concentration of the persistence distribution alone -0.37 , both far from significance, while the composite gives 0.83. Holding feature count fixed by partial rank correlation raises the association to 0.87 rather than lowering it.

Two caveats on what the quantity is a diagnostic of. First, it envelopes the widening of the null and not the shape of the p-value curve: the $\lambda = 0$ great-circle representation has the smallest ratio of the six, and its p-value still rises steadily with the exponent, because its event-to-control ratio falls over the same range. Second, the counts are computed from event-to-reference matching costs, while the quantity they are compared against is a property of the control null. An alternative built only from control-to-reference costs was tested: it lowers the rank correlation from 0.714 to 0.600 and reduces the systematic overshoot from six of six to four of six, while improving the root-mean-square error from 0.444 to 0.310, which only matches the constant baseline of 0.303. At $n = 6$ the two readings are not separable, so we report the mismatch as a limitation and retain

260 the estimator as defined.

261 **An axis the matched nulls do not control.** The fraction of fixes carrying an observed WMO_WIND value is
262 41.3 % for the event set against a control-draw mean of 38.0 %, placing the event at the 81st percentile of
263 1,000 draws. The intensity channel is 66.1 % imputed (§3.3) and the residual difference of \$4.5 is confined
264 to that channel, so reporting rate is a candidate covariate, and none of the four matched designs controls it.
265 Aggregated to the storm, however, the event and control distributions are indistinguishable (Mann–Whitney
266 $p = 0.993$ for reporting rate, $p = 0.394$ for storm length), so the difference is one of fix counts. Because
267 the draw randomises storms and not fixes, we tested the axis at that unit under the unmatched baseline draw
268 and one of the four matched designs, the basin-matched draw. Over 1,000 draws the event sits at the 54th
269 percentile of the control means under the uniform draw and at the 78th under the basin-matched draw, with
270 a median Mann–Whitney p of 0.61 in both and $p < 0.05$ reached in 0.4 % and 0.7 % of draws against the 5
271 % expected by chance; the fix-level percentile meanwhile rises from 81 to 93, which is the same imbalance
272 re-weighted by storm length and not a second one. At the unit the test randomises, then, the event and the
273 control draws are not separable on this covariate under both draws tested, and we report no fifth matched null.
274 Two qualifications: the per-window, era-matched and joint designs were not tested at the storm level, and the
275 one matched design that was tested moved the event from the 54th to the 78th percentile rather than towards
276 the middle, so the reading should not be extrapolated to the three untested designs. The check remains
277 the right one for a study that read an intensity-channel residual as physical, where the fix-level weighting
278 would enter the statistic directly. Source: v7/code/redteam_checks/rt5_windrate_matched.py, log
279 v7/results/_log_rt5_windrate_matched.txt; two positive controls and a fixed draw order.

280

281 S9. Analysis specifications and the pre-registration record

282 This section documents the analysis specifications that were varied, the pre-registered decision rules that
283 were fixed before the runs they govern, and the positive control attached to each. It is the evidence that the
284 reported specification dependence was mapped systematically rather than assembled from whichever runs
285 proved convenient.

286 S9.1 Two pre-registered decision rules

287 Two pre-registration documents were deposited before the runs they govern, and both are reproduced in the
288 archive.

289 The **first** fixed, on 2026-07-18 and deposited on 2026-07-19, the criteria by which two candidate remedies for
290 the H_1 calibration — a z-score normalisation of the event-to-null distance, and an L^2 -persistence-landscape
291 summary in place of the diagram distance — together with a nuisance-dimension control would be judged to
292 restore or fail to restore nominal calibration. The primary decision was a pooled exact two-sided binomial
293 test of the 147-placebo H_1 false-positive count against the nominal 5 % rate, with compatibility declared at p
294 ≥ 0.05 . Both remedies were run under that rule and neither restored calibration relative to the baseline (§4.2).

295 The **second**, fixed and deposited on 2026-07-19 before the runs it governs, addressed a sharper question:
296 whether the calibration verdict itself depends on the implementation. Five further reproducible control-draw
297 orderings were generated as seeded permutations of the deterministic sort, applied through a dedicated order-
298 ing generator so that the main resampling stream was untouched, and each was run through the full 49-placebo
299 calibration under the three encodings. The decision rule was fixed before those five were known: with M
300 the median of the seven pooled H_1 false-positive counts, $M \geq 13$ would mean that the typical implementation
301 exceeds the binomial boundary and the over-rejection is real though implementation-variable, while $M \leq 12$

would mean that the typical implementation is compatible with nominal and the earlier inflated reading is one tail of a distribution straddling the threshold. The boundary of 13 is the largest count with exact two-sided binomial $p \geq 0.05$ at $n = 147$. The seven counts came out as $\{9, 9, 10, 10, 11, 12, 16\}$, with median 10 and binomial $p = 0.339$ at the median, so the second branch holds: the calibration verdict is implementation-dependent, and the pipeline is reported as nominal with a residual ordering spread (§4.2). A parallel pre-registered run isolated the library-version factor by repeating the D1 test and the calibration under numpy 1.26 against 2.4 with the ordering held fixed; the resampling p-values were bit-identical and the pooled placebo counts equal, so the version factor is negligible and the difference from the discussion paper is attributed to the control ordering and the missing-wind fallback.

S9.2 The specification matrix

Table S7 lists the eleven axes along which the analysis was varied, what each does to the verdict, and the positive control that gates it. A positive control is a known archived value that the new code must reproduce before any new number from it is reported; four of the runs behind this table were stopped by their own positive control and were re-derived before any conclusion was drawn from them.

Table S7. Specification axes, their effect on the verdict, and the positive control for each. Values in the middle column are H_1 resampling p-values unless stated otherwise.

Axis	What is varied	What it does to the verdict	Positive control
1	Ground metric of the matching cost (Euclidean, sup) $\times$ exponent (1, 2)	Changing the ground metric moves nothing (0.037 $\rightarrow$ 0.037 linear, 0.012 $\rightarrow$ 0.015 great-circle); changing the exponent moves everything (0.012 $\rightarrow$ 0.151 great-circle). The difference lies in how the costs are aggregated	The twelve $p = 1$ Euclidean cells reproduce the archived values bit for bit
2	Event-subsample seed (six draws)	The combination of verdicts reported in the discussion paper occurs in one draw of six; the relative order of the three encodings holds in all six (§4.1)	Cylinder row reproduces the archived five-seed values bit for bit
3	Intensity weight $\lambda \in \{0, 0.5, 1\}$	0.157 / 0.062 / 0.012 at $M = 1$ across three orderings; the gradient compresses to a factor of 1.75 after subsample averaging (§4.7.2)	Seed 2026 reproduces the archived value 0.012

Axis	What is varied	What it does to the verdict	Positive control
4	Matched-null design (four) $\times$ metric (W_1 , W_2 , W_∞)	All twelve W_1 cells below α ; none of the twenty-four W_2 and W_∞ cells below α , with the event between the 68th and 87th percentile in every W_∞ cell (§4.5)	Twelve $p = 1$ cells reproduce archived values; a purpose-written exact bottleneck agrees with <code>persim.bottleneck</code> on 300 archived diagram pairs with zero deviation
5	Determinism rule for the matched draw (pinned hash seed, explicit sort)	All four designs stay below α under both rules, so the defect is a reproducibility failure and not a verdict change; two independent explicit-sort implementations still differ (0.009 against 0.021 on basin $\times$ era) unless the position of the missing-basin category is also fixed	Pinned-hash-seed run reproduces the archived values
6	Missing-wind imputation rule (per-draw median, global constant)	0.012 against 0.034 — a factor of three, with the direction preserved	Recount reproduces the 66.1 % missing rate from the raw archive
7	Calibration metric (W_1 , W_2 , exact bottleneck)	All three are well calibrated: pooled placebo H_1 20/343 for W_1 (binomial $p = 0.46$) and 25/343 for W_2 ($p = 0.062$), with the bottleneck at comparable magnitude. The exponent's effect on the event verdict is therefore not a calibration failure	The 7×49 W_1 placebo results reproduce the archived values bit for bit

Axis	What is varied	What it does to the verdict	Positive control
8	Homology channel reported (H_1 alone, H_0 and H_1)	H_1 {9,9,10,10,11,12}/147 and H_0 {4,5,8,11,11,16}/147 — the H_0 spread is the wider, which withdraws the earlier reading that only the H_1 swing crosses the threshold (Section S2)	Same runs as axis 7
9	Persistence scale examined	The excess sits at intermediate persistence (99.6th percentile) while the largest single feature is at the 82nd; three binnings agree qualitatively (§4.6.1)	The reconstructed sub-sample reproduces the archived diagram: 90 features, total persistence 0.9076
10	The two unregistered choices together: exponent $\times$ subsample	Family p-values 0.032 / 0.052 / 0.054 / 0.081 at $M = 1$ become 0.082 / 0.140 / 0.122 / 0.134 at $M = 10$; all four exponent-bearing families move in the same direction (§4.7.2)	$M = 1$ reproduces the archived event distances exactly; restricted to a single cell the minimum-based combination returns the archived p-value exactly
11	Reading below the threshold	The event is above the control median in all 42 cells, at a median 86th percentile, with an event-to-control ratio in [0.964, 1.338] and the linear representation alone rising with the exponent (§4.7.2)	Effect-size and percentile tables recomputed from the archived null distributions

Two constraints on how Table S7 may be read. The axes are not independent: every cell comes from the same 25 event storms, the same 1,020-storm control pool and, at $M = 1$, the same event subsample, so counts across cells describe consistency and are not tests. And the table records what was varied, which is a lower bound on what could have been: the window-cropping rule, the subsample size at the level of the physical distance, and the reporting-rate covariate of Section S8 were left fixed.

S10. Sidedness, two-sided companion values, and what the verification covers

Sidedness. The resampling p-value of §3.5 is the fraction of null draws whose distance is at least the event distance, which is one-sided: the pre-specified alternative is that the event ensemble lies farther from the reference set than the controls do. That convention was fixed in the discussion-paper version and is not changed here. Because sidedness is a convention rather than a result, Table S8 gives the two-sided companion of every value on which a threshold statement in this paper rests, under both of the conventions in use, and the two do not agree.

Doubling is the conservative conversion and assumes a symmetric null. The empirical two-sided count makes no symmetry assumption, though it does take the null median as the centre: it asks how many null draws lie at least as far from that median as the event does, in either direction. These nulls are right-skewed and the event sits well into the upper tail, so in most cells no control draw falls as far below the median as the event lies above it, and the empirical count returns the one-sided value. Readers who prefer the symmetric convention should read the doubled column; readers who prefer to make no distributional assumption should read the empirical one.

Table S8. One-sided resampling p-values as reported in the main text, with their two-sided companions under both conventions. Every row is generated by code/twosided.py from the archived null draws and is not typed: each cell first reproduces the archived one-sided value to within 5×10^{-4} and is dropped, with the failure printed, if it does not. All nine rows pass that check. The first five rows are at $B = 10,000$, the twelve matched-null cells at $B = 1000$.

Result	one-sided (reported)	doubled	empirical two-sided
Linear modular, archived convention	0.0343	0.0686	0.0343
Unit-sphere	0.0120	0.0240	0.0120
Cylinder	0.1516	0.3032	0.1726
Great-circle $\lambda = 1$, basin-matched, $M = 10$	0.0427	0.0854	0.0427
The same under index-minimum dates	0.0329	0.0658	0.0329
Per-window count, three orderings	0.006 / 0.010 / 0.004	0.012 / 0.020 / 0.008	as one-sided
Genesis basin, three orderings	0.003 / 0.004 / 0.003	0.006 / 0.008 / 0.006	as one-sided
Genesis era, three orderings	0.043 / 0.034 / 0.009	0.086 / 0.068 / 0.018	as one-sided
Joint basin $\times$ era, three orderings	0.012 / 0.013 / 0.011	0.024 / 0.026 / 0.022	as one-sided

The convention changes two of the twelve matched-null verdicts and no others. Under doubling the event-control difference survives the per-window, basin and joint designs in all three orderings and fails the era-matched design in two of its three; under the empirical count it survives all twelve. Either reading places the failure in the design that §4.5 already reports as the weakest of the four, so the sidedness convention sharpens a caveat the paper states rather than adding a new one. For the specification-averaged values the question does not arise in the same way: those are already above α one-sided, and doubling takes 0.082, 0.122 and 0.134 to 0.164, 0.244 and 0.268, so the resolution statement of §4.7.3 only strengthens.

What the 87-specification verification covers. The verification script re-derives every registered specification from the archived data and compares it with the value printed in the manuscript, and all 87 agree. What it compares is the **resampling p-value** of every registered specification and, for the specifications computed at $p = 1$ under the Euclidean ground metric, also the **archived event distance**, to a tolerance of 10^{-9} . Several classes of number in this paper sit outside that comparison, and each is anchored elsewhere (Table S9).

Table S9. Classes of number in this paper that lie outside the 87-specification verification, and where each is anchored.

Outside the 87	Where it is anchored
Null means and null standard deviations wherever printed, and event distances for specifications other than the $p = 1$ Euclidean ones (§4.1, §4.5)	The same per-specification result files, which carry all four quantities alongside the p-value
The five values recomputed at $B = 10,000$ (§3.5), on which every threshold-crossing statement in this paper rests	Their own result files, one per specification, which also record the pinned reference-set indices
Eighteen of the thirty-six cells of the matched-null table in §4.5: the three stratified designs under the true W_2 and the exact W_∞ . The basin-matched design is registered under all three metrics; the stratified designs are registered only at $p = 1$	The run logs of those stratified re-runs
Quantities derived from two specifications — the measured widening of §4.6.3 is a ratio of two null distributions — and the two-sided companions of Table S8	The archived null draws, from which <code>code/twosided.py</code> recomputes them under a per-cell positive control
The 49-placebo calibration of §4.2 and the stratified and subsample tables of §§4.3–4.4	Their own run logs and regenerated tables; the §§4.3–4.4 tables additionally reproduce bit for bit under two different interpreter hash seeds
The longitude-convention and branch-cut runs of §3.3.1 and Section S7	Their run logs, with the caveat below

Those longitude logs need one caveat, because a reader who opens them will meet it. Only `lon_arch.log` carries a positive control, and it passes, reproducing the archived event distance and p-value bit for bit. `rot0.log` carries a check that prints a failure: it compares its own event distance, 0.328413, against the archived 0.352771, which are the single-convention and as-archived quantities respectively and are not meant to agree. The reference in that check is set wrongly; the run itself is sound, and its value is the one reported in Table S5. `lon_wrap.log`, `rot90.log` and `rot180.log` carry no positive control at all. `lon_wrap.log` and `rot0.log` do agree bit for bit on the event distance and the p-value, but the encodings they apply are the same expression differing by an added zero, so that agreement is not a check of the encoding. What it does cross-check is the control-draw code: the two scripts were written separately and subset the control storms by different means, one by a boolean mask on the storm identifier and one by a pre-built positional index. That is a cross-check of the resampling implementation, not a check against an archived value. We report this rather than leave a reader to find it, and a study repeating these runs should give each of them an archived value to reproduce first.

The 87-specification statement should therefore be read as covering the registered p-values and the $p = 1$ Euclidean event distances, and not as covering every number in the paper.

373 **References cited in this supplement**

- 374 Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach
375 to multiple testing, *J. Roy. Stat. Soc. B*, 57, 289–300, <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>,
376 1995.
- 377 Clette, F., Svalgaard, L., Vaquero, J. M., and Cliver, E. W.: Revisiting the sunspot number, *Space Sci. Rev.*,
378 186, 35–103, <https://doi.org/10.1007/s11214-014-0074-2>, 2014.
- 379 Gelman, A. and Loken, E.: The statistical crisis in science, *Am. Sci.*, 102, 460, <https://doi.org/10.1511/2014.111.460>,
380 2014.
- 381 Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T.
382 M., Vose, R. S., and Zhang, H.-M.: Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5),
383 *J. Climate*, 30, 8179–8205, <https://doi.org/10.1175/JCLI-D-16-0836.1>, 2017.
- 384 Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., and Neumann, C. J.: The Interna-
385 tional Best Track Archive for Climate Stewardship (IBTrACS), *B. Am. Meteorol. Soc.*, 91, 363–376,
386 <https://doi.org/10.1175/2009BAMS2755.1>, 2010.
- 387 Simonsohn, U., Simmons, J. P., and Nelson, L. D.: Specification curve analysis, *Nat. Hum. Behav.*, 4,
388 1208–1214, <https://doi.org/10.1038/s41562-020-0912-z>, 2020.
- 389 Simpson, E. H.: The interpretation of interaction in contingency tables, *J. Roy. Stat. Soc. B*, 13, 238–241,
390 <https://doi.org/10.1111/j.2517-6161.1951.tb00088.x>, 1951.
- 391 Wilson, E. B.: Probable inference, the law of succession, and statistical inference, *J. Am. Stat. Assoc.*, 22,
392 209–212, <https://doi.org/10.1080/01621459.1927.10502953>, 1927.
- 393 World Data Center for Geomagnetism, Kyoto: Geomagnetic Dst index, available at: [https://wdc.kugi.kyoto-](https://wdc.kugi.kyoto-u.ac.jp/dstdir/)
394 [u.ac.jp/dstdir/](https://wdc.kugi.kyoto-u.ac.jp/dstdir/), last access: 30 April 2026.