Transfer entropy fails its known-answer test on tree rings

A preregistered test on 135 ring-width sites in which summer temperature is known to drive ring width, with planted-direction worlds, time reversal, a shared-trend world and far-site pairings

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F132 in the findings ledger. Honest null (tier D).

Abstract

Transfer entropy is widely used to read which of two time series drives the other. Tree rings give a case where the answer is known: summer warmth shapes the ring laid down that year, and the ring has no meaningful effect on the following summer. We ran a nearest-neighbour transfer entropy estimator on 135 ring-width chronologies from the International Tree-Ring Data Bank, about 78 years each, against June to August temperature, under a plan hash-locked before any estimate on real data. On real data the estimator detected climate to ring at 15.6 percent of sites (8.8 to 26.2) and ring to climate at 11.9 percent (4.9 to 22.5), above the 10 percent reverse limit fixed in advance. Read backwards in time, the same series gave 19.3 and 22.2 percent, so the estimator was reading no arrow of time. With a climate-to-ring coupling planted in synthetic rings, the false direction ranked at or above the true one at 52.6 percent of sites at the real coupling strength and at 40.0 percent with strong coupling, where direction accuracy reached 0.600 against the 0.8 required. The estimator detected coupling and misread its direction. It gets no direction-reading status for this kind of data.

1 What this means for a reader

When a study reports that one yearly record drives another because transfer entropy is larger in one direction, ask whether the method was first run on a case where the direction is known. On short, noisy annual records like tree rings, this standard estimator could not tell the right direction from the wrong one, even when we built the right answer into the data ourselves.

2 Introduction

Schreiber [1] introduced transfer entropy as a measure that separates information exchanged between two systems from information they share through common history, and showed that it can tell a driving element from a responding one. It is now used in climate, ecology, neuroscience and finance to assign a direction to a link between two series. For continuous data it is usually estimated with the nearest-neighbour method of Kraskov, Stoegbauer and Grassberger [2], in the conditional form of Frenzel and Pompe [3].

The method has known failure modes. Smirnov [4] examined spurious causalities with transfer entropy; later papers that cite that work report that the method can fail to recover preset one-way couplings and can give the wrong direction. We read Smirnov's paper only through those citing papers and its title. Runge and colleagues [6] reviewed causal inference from time series in the Earth sciences and set out the conditions under which such methods can be trusted. Sugihara and colleagues [5] proposed convergent cross mapping as a direction test for weakly coupled ecological systems, and we ran it beside transfer entropy as a second probe.

What a user of these methods needs is a known-answer test on data like their own. Annual tree rings and summer temperature give one. Summer warmth shapes ring width at cold sites, and a ring cannot set the weather. Our searches, run on 9 November 2026, found general studies of transfer entropy direction failure and no planted-direction test on ring-width data. That is a search result, not proof that none exists.

3 Preregistration and design

The plan was written as a preregistration and hash-locked on 31 October 2026 at 05:19 UTC, before any transfer entropy, cross mapping or lagged statistic was computed on the real data. Before the lock the estimator was tested on synthetic autoregressive series only, and from the real data only site counts and series lengths were read.

The decision the study informs was written into the same file: may a direction instrument of this kind be used to read the direction of links between other ecological series? The rule set in advance was yes only if the as-written kill condition and the planted-truth anchor both pass, or, for a labelled variant, if the variant's kill condition and its anchor both pass. Otherwise the instrument gets no direction-reading status.

The plan forecast that lag 0 would break the test. A summer and the ring of the same year are simultaneous at annual resolution, so a same-year term carries no time order and can show as a detection in both directions. The design therefore named one variant in advance, the same test with lag 0 removed, and stated that no other variant would be tried.

4 Data and provenance

Rings: ring-width chronologies from the International Tree-Ring Data Bank [7], held by NOAA's National Centers for Environmental Information as public data with a request to cite the contributors. A site entered if its series was gap-free from 1901 with at least 50 years. This gave 135 sites in 57 two-degree blocks: 51 in the Alps, 24 in Fennoscandia and 60 elsewhere, with series of 64 to 91 years and a median of 78.

Climate: June to August mean temperature from the nearest cell of CRU TS 4.10 [8] [9], released under the Open Database License 1.0 with attribution to the Climatic Research Unit, University of East Anglia. June to August precipitation was run as a secondary, descriptive target.

Both series were passed through a high-pass filter, each minus a Gaussian smooth with a width of 10 years, and scaled to unit variance. The mean absolute same-year correlation between ring and temperature was 0.218.

5 Methods

The estimator

For a source series X and a target Y, the transfer entropy at lag l is the conditional information

$$T_{X \to Y}(l) = I(Y_t; X_{t-l} \mid Y_{t-1}) \tag{1}$$

estimated in bits with the KSG estimator in conditional form with $k = 4$. It was computed at lags 0, 1, 2 and 3 on one common set of years. The site statistic is the largest of the four:

$$T_{max} = max(T(0), T(1), T(2), T(3)) \tag{2}$$

Forward means climate to ring and reverse means ring to climate.

Surrogate test at each site

Each site statistic was compared with 199 iAAFT surrogates of the source series, each run through the same estimator and maximised over the same lags. The p value is

$$p = \frac{k + 1}{n + 1} \tag{3}$$

with k surrogates at or above the real statistic and $n = 199$, so the smallest possible value is 0.005. A site counts as a detection at p 0.05 or below. The detection rate of an arm is the share of its 135 sites with a detection, and its interval is a block bootstrap over the 57 two-degree blocks.

Direction accuracy

For each arm, direction accuracy is

$$A = \frac{m}{N} \tag{4}$$

where m is the number of sites whose forward statistic exceeds the reverse one and $N = 135$. A coin gives 0.5.

Cross mapping

Convergent cross mapping used an embedding dimension of 3, a delay of 1 year, the full library, leave-one-out prediction and 99 iAAFT surrogates of the predicted series. It was scored with the same detection and accuracy rules.

The variant without lag 0

The preregistered variant is the same test with lag 0 removed from both directions, so eq. 2 takes the largest of lags 1 to 3 for the real data and for the surrogates. It was computed in the same run and is labelled as a variant throughout.

6 Controls and planted checks

Every control and synthetic arm ran in the same run as the real data.

- Far-site pairing: 20 re-pairings in which each ring series meets the climate of a random site at least 1,000 km away, with 99 surrogates each. Its 95th percentile is the bar the real forward rate must beat.
- Time reversal: both series read backwards. A method that reads direction from lag order should weaken or flip.
- No coupling: independent iAAFT copies of each site's ring and climate series.
- Shared trend: independent copies with a shared line and random walk added before the high-pass.
- Common driver: a third iAAFT series drives both ring and climate with a squared coupling of 0.2 and no direct link.

The planted checks build synthetic rings with the true direction known. Each starts from an iAAFT copy S_t of the site's ring series and adds the site's own climate C_t:

$$R_t = S_t + aC_t + 0.4aC_{t-1} \tag{5}$$

At real strength, a is the site's real same-year correlation. In the strong arm the two coefficients are 0.6 and 0.3. Each synthetic arm was drawn once per site.

The planted anchor passes only if, in the strong arm, the forward rate exceeds 0.8 and direction accuracy exceeds 0.8, and if the no-coupling and shared-trend arms keep both rates at or below 0.10.

7 Results

Table 1 gives the detection rates and direction accuracy in every arm, and Figure 1 draws the rates with their intervals.

On real data the estimator detected climate to ring at 15.6 percent of sites (8.8 to 26.2) and ring to climate at 11.9 percent (4.9 to 22.5), above the 10 percent reverse limit fixed in advance. The forward rate beat the far-site pairing, whose mean was 0.056 and 95th percentile 0.082 (p 0.048, the smallest value 20 pairings allow). The reverse rate failed the limit by a narrow margin, and its interval spans 0.10. The kill condition fired. Direction accuracy was 0.519 against a coin's 0.5.

Read backwards in time, the same series gave 19.3 and 22.2 percent, so the estimator was reading no arrow

Table 1. Detection rates over 135 sites, with block-bootstrap 95% intervals, and direction accuracy, as written (lags 0 to 3).

Arm	Climate to ring	Ring to climate	Accuracy
Real data	0.156 (0.088 to 0.262)	0.119 (0.049 to 0.225)	0.519
Real data, time reversed	0.193 (0.128 to 0.280)	0.222 (0.154 to 0.327)	0.511
No coupling	0.074 (0.034 to 0.115)	0.037 (0.010 to 0.065)	0.578
Shared trend	0.126 (0.072 to 0.191)	0.170 (0.109 to 0.250)	0.489
Common driver	0.059 (0.023 to 0.112)	0.096 (0.049 to 0.154)	0.541
Planted, real strength	0.193 (0.104 to 0.315)	0.185 (0.108 to 0.292)	0.474
Planted, strong	0.978 (0.954 to 1.000)	0.919 (0.857 to 0.963)	0.600
Real precipitation	0.067 (0.028 to 0.115)	0.096 (0.053 to 0.147)	0.578

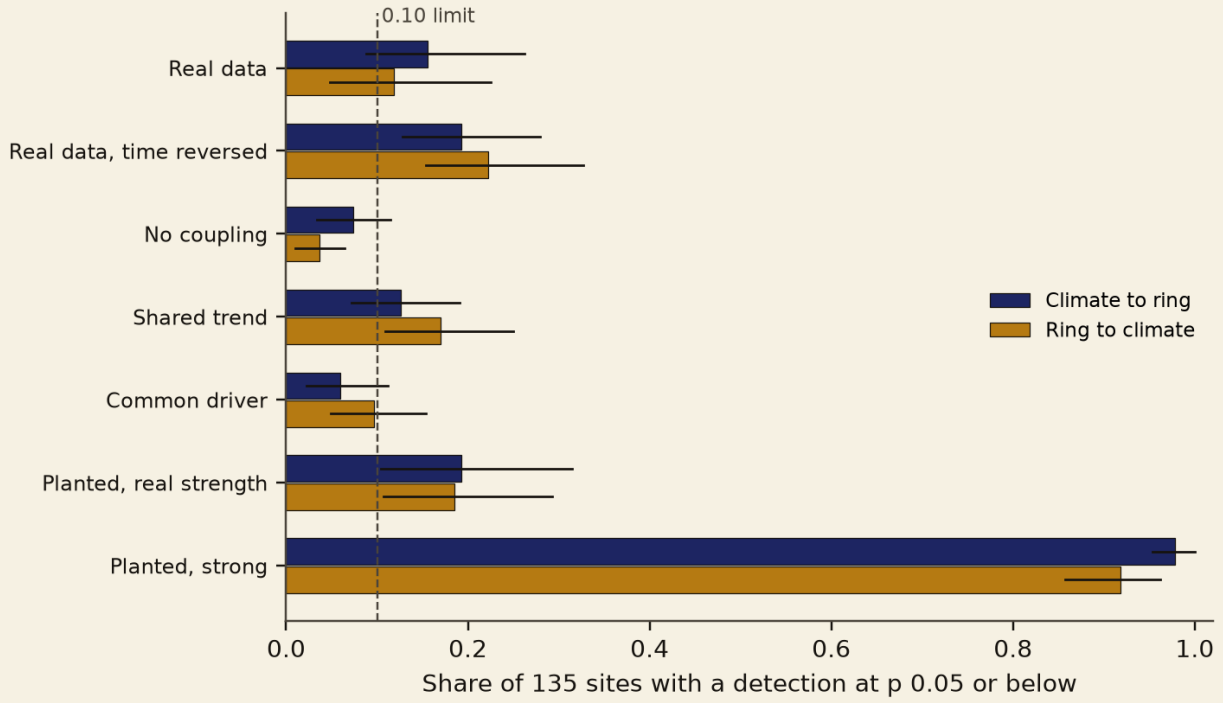


Figure 1. Detection rates in each arm, climate to ring (navy) and ring to climate (ochre), with block-bootstrap 95% intervals. The dashed line is the 0.10 limit. Drawn from the run's result file.

of time. The mean transfer entropy at lag 0 was 0.044 bits in both directions, which fits the forecast that the same-year term carries no order.

The planted anchor failed. With a climate-to-ring coupling planted in synthetic rings, the false direction ranked at or above the true one at 52.6 percent of sites at the real coupling strength and at 40.0 percent with strong coupling, where direction accuracy reached 0.600 against the 0.8 required. With strong coupling, detection rose to 0.978 forward and 0.919 reverse: the estimator saw the coupling and could not place its direction. The no-coupling arm stayed inside its limit (0.074 and 0.037). The shared-trend world produced 12.6 and 17.0 percent, above the 0.10 limit even after the high-pass. The common-driver arm stayed under 0.10 (0.059 and 0.096). At a coupling near 0.2 this says more about low power than about any ability to separate a shared driver from a direct link.

Figure 2 sets direction accuracy in each arm against the coin and the required bar.

The variant without lag 0 passed on real data, 8.9 percent forward and 3.7 percent reverse, and missed the planted effect: with strong coupling it detected the forward link at 2.2 percent of sites, so its pass cannot be

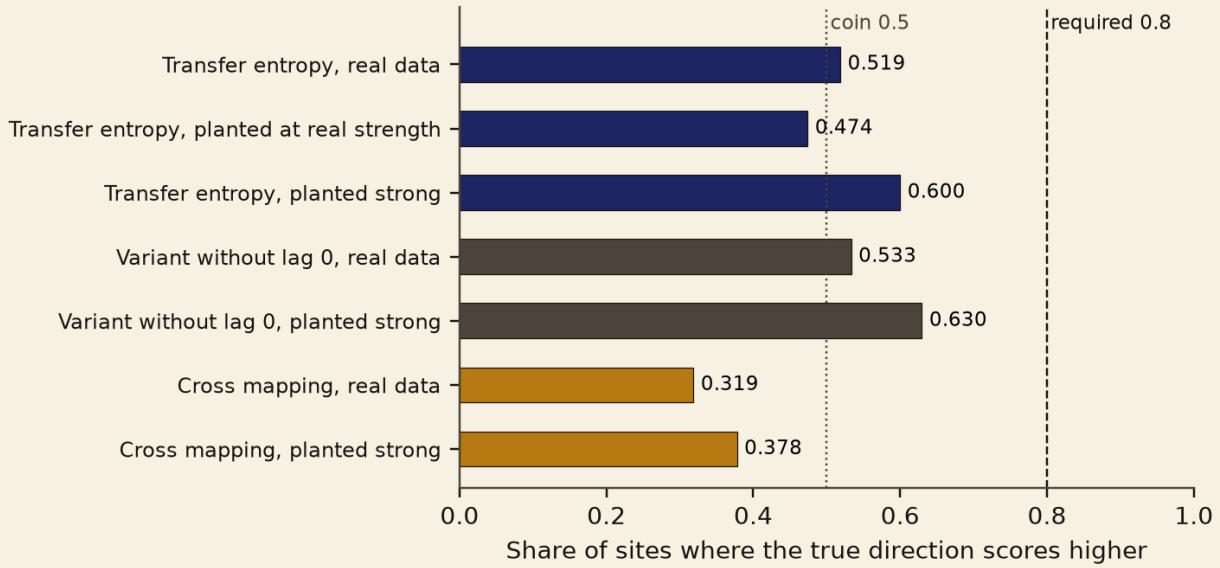


Figure 2. Direction accuracy for transfer entropy as written (navy), the variant without lag 0 (grey) and convergent cross mapping (ochre). The dotted line is a coin and the dashed line the preregistered bar.

credited. Its real forward rate beat its pairing 95th percentile of 0.060 and its accuracy was 0.533. In the planted strong arm its accuracy was 0.630, and its shared-trend rates were 0.104 both ways, just over the limit.

Convergent cross mapping pointed the wrong way on real data: 18.5 percent forward against 28.9 percent reverse, accuracy 0.319. With strong planted coupling its accuracy was 0.378.

Correcting across sites with the Benjamini-Hochberg rule left no detections on real data in any direction. With 199 surrogates the smallest p is 0.005, so the correction over 135 sites can reject only when a large share of sites sit near that floor. It did so in the strong planted arm (0.978 forward) and nowhere in the real data. Over the six preregistered family tests, the reverse rate against its nominal level (q 0.002) and both cross-mapping rates (q below 1e-7) were rejected; the forward and variant-forward rates against the pairing gave q 0.057, limited by the 20-pairing floor.

By region, Fennoscandia, where the mean same-year correlation was 0.318, gave 0.375 forward and 0.292 reverse; the Alps 0.137 both ways; other sites 0.083 and 0.033.

Table 2 sets the forecasts written before the run against the outcomes.

8 Discussion

The test has a known answer, and the estimator did not find it. On real data the two directions were detected at similar rates, reversing time made no difference, and when we planted the right direction the estimator ranked the wrong one at or above it at about half the sites. This is an honest null: the method's failure on this kind of data, measured against a truth we built in.

Two mechanisms fit the pattern. The same-year term is symmetric, as forecast, and it dominates the estimate at these lengths. The lagged terms are small, and conditioning on the ring's own previous year can absorb a planted lag-one effect, because the previous ring already holds the previous summer. That second reading is an inference from a look after the run: real data showed mean transfer entropy at lag 1 of 0.035 bits forward and 0.002 reverse, an asymmetry in the known direction, while the planted lag-one arms showed about 0.01 or less. We did not test it.

The result agrees in kind with the general failures reported for transfer entropy [4] and with the caution of Runge and colleagues [6]. Its contribution is narrower: a measured failure on public ring-width data, at the length and

Table 2. Forecasts fixed in the preregistration against what the run gave.

Quantity	Forecast	Outcome
Forward rate, as written	0.35 (0.20 to 0.55)	0.156
Reverse rate, as written	0.30 (0.15 to 0.50)	0.119
Kill condition fires	probability 0.75	fired
Forward rate above 0.5 fails	probability 0.8	failed
Variant forward rate	0.15 (0.07 to 0.30)	0.089
Variant reverse rate	0.07 (0.03 to 0.12)	0.037
Variant accuracy	0.58 (0.50 to 0.68)	0.533
Variant kill condition passes	probability 0.45	passed
Cross mapping direction test passes	probability 0.25	failed
Planted strong forward rate	above 0.9	0.978
Planted real-strength forward rate	0.35	0.193
Shared trend under 0.10	expected	no, 0.126 and 0.170
Common driver above 0.10, failing	probability 0.85	passed

coupling strength that ring studies have.

Two later rounds asked the same question with a different instrument: lagged Granger tests read across the whole panel of sites at once. On 161 held-out high-latitude sites (finding F675), the panel index pointed from summer to ring at +1.206 (95% interval +0.871 to +1.591) against a far-pairing 95th percentile of +0.213, and flipped under time reversal. Two of its locked first-difference checks missed by small margins (0.584 against 0.6, and 0.117 against 0.10), so its own kill condition fired. On a third panel of 145 held-out sites (finding F676), with the planted prior-summer term fitted to the measured sign, the panel index was +1.204 against a pairing 95th percentile of +0.461, every binding check but one passed, and the kill condition did not fire. That panel method earned a limited direction status at that coupling strength. Neither round changes the result for transfer entropy read site by site.

9 Limitations

- Power is low at about 78 years and a ring-temperature correlation near 0.2. This is a failure on short annual series. A longer or monthly design might behave differently.
- Rings here are width only, which carries less of the summer warmth signal than wood density.
- With 199 surrogates the smallest p is 0.005, so correcting across 135 sites could reject nothing at the detection rates the real data showed. That is a design error we should have caught before the lock.
- The far-site pairing ran 20 times, so its p value cannot go below 0.048.
- The planted prior-summer term in eq. 5 was positive. The later round on 161 sites found the real prior-summer coefficient negative at most sites (median -0.127, positive at 21.7 percent), so the planted world did not match the real lag structure at those sites.
- CRU temperature is a gridded product with thin station coverage in early decades. Ring-to-climate influence is taken as absent at site scale on these time scales; if it were not, the reverse rule would mislabel.
- Scoring was done by our own code and was not blind.

10 Verified

Verified: on 9 November 2026 the site, block and length counts, every rate, interval and accuracy in Table 1, the pairing mean, percentile and p value, the lag-0 and lag-1 mean transfer entropy, the false-ranking shares, the variant's rates and accuracy, the cross-mapping rates and accuracy, the family q values, the per-site correction result and the regional rates were re-read from the run's result file and match the findings ledger. The forecasts in Table 2 were re-read from the preregistration. The SHA-256 of the preregistration matches the hash recorded

in the result file, and the SHA-256 of the result file matches the abbreviated hash in the ledger. Both figures were drawn by a script that reads only that result file. The later-round values in the discussion and limitations were read from the findings ledger and their published write-ups; their result files were not opened for this paper. The study was not re-run for this paper. The literature was found by searches run on 9 November 2026 and read at the level of abstracts, publisher pages and search summaries; Smirnov's paper was read only through papers that cite it.

11 Corrections

- 9 November 2026: first published version. No correction to this finding is recorded in the ledger.

12 Reproduction

- Data: ITRDB ring-width chronologies gap-free from 1901 with at least 50 years; June to August temperature from the nearest CRU TS 4.10 cell.
- High-pass both series with a 10-year Gaussian and scale them.
- At each site compute eq. 1 both ways at lags 0 to 3 with the conditional KSG estimator, take eq. 2, and compare with 199 iAAFT surrogates of the source by eq. 3.
- Repeat on the time-reversed series, the 20 far-site pairings and the synthetic arms, including the planted rings of eq. 5, and score eq. 4.
- Repeat without lag 0 for the variant, and run convergent cross mapping with embedding dimension 3.
- Preregistration SHA-256: 54fc1c57bcd3fe569455c2bf4203e4a4c59054b3a81099266db058289a7089da. Plan SHA-256: 7a1ef29b4cc14024600c542bbe2fcfb706e7086845a4014fc501becb18884d4c. Run 20261031T051913Z. Result file SHA-256: 420dc9e020273a5680c67fd7186e08b5de6e7862da2d7859 ed5adf3de5843ef7.
- Cost: the analysis took about five minutes (315 seconds) on one desktop CPU.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the run. A verdict passes only when every rule holds. Values are read from the run's result file.

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **transfer entropy**. The information, in bits, that the past of one series adds about the next value of a second series beyond what the second series' own past already gives; a larger value one way than the other is commonly read as the direction of influence. Field: information theory (Schreiber 2000). Excludes plain correlation or mutual information, which are the same in both directions, and any proof of a physical cause.
- **convergent cross mapping**. A test that rebuilds the state of one series from its own lagged values and asks how well that rebuilt state predicts the other series; the driven series is expected to predict the driver. Field: nonlinear dynamics and ecology (Sugihara and colleagues 2012). Excludes regression of one series on the lagged values of the other.
- **known-answer test**. A test of a method on a case where the true answer is fixed by physics or by construction, so that any wrong reading is a fault of the method. Field: measurement and method validation. Excludes agreement with another method, which can share the same fault.
- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the ques-

Table 3. Preregistered verdicts and the rules that decided them.

Verdict	Deciding rule	Value	Holds
Kill condition survived, as written	forward rate above pairing 95th percentile (0.082); reverse rate at most 0.10	0.156; 0.119	no
Planted anchor	strong forward rate above 0.8; strong accuracy above 0.8; no-coupling rates at most 0.10; shared-trend rates at most 0.10	0.978; 0.600; 0.074 and 0.037; 0.126 and 0.170	no
Variant kill condition survived	variant forward above its pairing 95th percentile (0.060); variant reverse at most 0.10; variant accuracy above 0.5	0.089; 0.037; 0.533	yes
Variant planted anchor	the planted anchor's rules on the variant	0.022; 0.630; 0.067 and 0.074; 0.104 and 0.104	no
Common driver	common-driver rates at most 0.10	0.059; 0.096	yes
Variant common driver	variant common-driver rates at most 0.10	0.037; 0.052	yes
P1	forward rate above 0.5	0.156	no
P2	reverse rate at most 0.05	0.119	no
P3 cross mapping	forward rate above reverse; reverse at most 0.10	0.185 against 0.289	no
Decision	as written: kill condition and anchor both pass; or variant: both variant rules pass	no; no	no status

tion, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (meta-science). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.

- **kill condition**. A rule written in the preregistration that, when met, decides against the claim with no room for judgement after the run. Field: research method (our usage, after Popper's falsification). Excludes reasons found after the run for setting a result aside.
- **high-pass filter**. A filter that keeps fast variation and removes slow trends; here each series minus a Gaussian smooth of width 10 years. Field: signal processing. Excludes a fixed polynomial detrend.
- **KSG estimator, conditional form**. The k-nearest-neighbour estimate of information from samples given by Kraskov, Stoegbauer and Grassberger and extended to conditional information by Frenzel and Pompe, here with 4 neighbours and the maximum norm. Field: statistics and information theory. Excludes estimates that first bin the data into histograms.
- **iAAFT surrogate**. A copy of a series built by repeated rounds of Fourier phase randomisation and rank reordering, so that it keeps both the original spectrum, and with it the autocorrelation, and the original set of values, while its timing against any other series is lost. Field: time-series statistics, surrogate data. Excludes a shuffle of single years, which destroys the autocorrelation.
- **detection rate**. The share of sites at which the score in one direction beats its own surrogates at p 0.05 or below. Field: statistics, our usage. Excludes a rate corrected for testing 135 sites, which is reported separately.
- **block bootstrap**. An interval made by resampling whole 2 degree blocks of sites with replacement, so that neighbouring sites that share climate move together and the interval is not too narrow. Field: statistics. Excludes a bootstrap over single sites, and a binomial interval that treats sites as independent.
- **direction accuracy**. The share of sites at which the climate-to-ring score is larger than the ring-to-climate score; 0.5 is what a coin would give. Field: method validation, our usage. Excludes whether either score is significant.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental design. Excludes a comparison with a different dataset or with a published number.
- **far-site pairing**. A control in which each ring series is paired with the climate of a random site at least 1,000

km away, so that real coupling is removed while both series keep their own form. Field: statistics, our usage. Excludes pairings with nearby sites, which share regional climate.

- **time reversal**. The same test run on both series read backwards in time; a method that reads direction from lag order should weaken or flip. Field: time-series analysis. Excludes a control on coupling strength.
- **planted check**. Data built so that the right answer is known, run through the same analysis before the real data, to show that the analysis can see the effect when it is there. Field: research method; called a positive control in experimental science. Excludes any check made on the real data, which cannot show what the analysis misses.
- **Benjamini-Hochberg correction**. A rule that adjusts a set of p values so that, among the tests called significant, the expected share of false calls stays at or below a chosen rate, here 0.05; the adjusted value is called q. Field: statistics, multiple testing. Excludes the stricter Bonferroni rule, which bounds the chance of any false call.
- **honest null**. A working test that found no effect beyond a stated bound, a forecast that failed, or an error of our own found and fixed, reported with the same weight as a positive result. Field: research method (our usage). Excludes a void run, and a result left unpublished because it was negative.
- **Granger test**. A regression test of whether the past values of one series improve the prediction of another beyond that series' own past; for Gaussian data it is equivalent to transfer entropy. Field: econometrics and time-series statistics. Excludes a test of same-year effects, and proof of a physical cause.
- **verdict**. The pass or fail of one preregistered rule set against the run's numbers, read off by code. Field: research method (our usage). Excludes an interpretation written after reading the numbers.

15 References

[1] Schreiber, T. (2000). Measuring information transfer. Physical Review Letters 85, 461. doi:10.1103/PhysRevLett.85.461

[2] Kraskov, A., Stögbauer, H. and Grassberger, P. (2004). Estimating mutual information. Physical Review E 69, 066138. doi:10.1103/PhysRevE.69.066138

[3] Frenzel, S. and Pompe, B. (2007). Partial mutual information for coupling analysis of multivariate time series. Physical Review Letters 99, 204101. doi:10.1103/PhysRevLett.99.204101

[4] Smirnov, D. A. (2013). Spurious causalities with transfer entropy. Physical Review E 87, 042917.

[5] Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M. and Munch, S. (2012). Detecting causality in complex ecosystems. Science 338, 496-500. doi:10.1126/science.1227079

[6] Runge, J., Bathiany, S., Bollt, E. and others (2019). Inferring causation from time series in Earth system sciences. Nature Communications 10, 2553. doi:10.1038/s41467-019-10105-3

[7] International Tree-Ring Data Bank. NOAA National Centers for Environmental Information, World Data Service for Paleoclimatology. Ring-width chronologies. https://www.ncei.noaa.gov/pub/data/paleo/treering/

[8] Harris, I., Osborn, T. J., Jones, P. and Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Scientific Data 7, 109. doi:10.1038/s41597-020-0453-3

[9] Climatic Research Unit, University of East Anglia. CRU TS 4.10 monthly temperature and precipitation, 0.5 degree. Open Database License 1.0. https://crudata.uea.ac.uk/cru/data/hrg/