STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 1–28.
Published online in International Academic Press (www.IAPress.org)

Where does the indeterminacy come from? Deriving the I of neutrosophic statistics from its sources

Maikel Yelandi Leyva-Vázquez[1,2,*], Florentin Smarandache[3]

[1] *Universidad Bernardo O'Higgins, Santiago, Chile*
[2] *Universidad de Guayaquil, Guayaquil, Ecuador*
[3] *University of New Mexico, Gallup, NM, USA*

Abstract In neutrosophic statistics the indeterminate part I of an estimate is usually fixed by hand. We compute I from the records, once the analyst's choices (confidence level, tolerance band, coding of undecided records) are declared, and split it into typed components with provenance: measurement imprecision, which more data do not remove; sampling, which they do; conflict between sources; and hesitancy. We give the exact split for counts and interval data, a test of certified conflict whose false-conflict rate is guaranteed under stated assumptions, the two readings of a hesitant set, a typed arithmetic that keeps the labels through operations, and a table of actions for each component, together with a step-by-step reporting protocol and a checklist. The inferential tools are classical; what is new is the labelled representation and reporting rule built on them, which we map explicitly to its counterparts in interval statistics, imprecise probability, belief functions and hesitant sets. A decision-oriented example shows what the labels can change: among 298 conversations whose majority safety verdict is undecided, the current estimates suggest different follow-up priorities, additional ratings for 229 and examination of the "unsure" answers for 60, a distinction that a single interval does not make; these priorities have not been validated as interventions. With imprecision alone, a fixed I amounts to an interval representation. A single fused interval hull does not retain the identities of disagreeing sources or distinguish the provenance of its width components; the labelled report makes this information explicit. As a conditional illustration, a paired batch-means test on daily rain forecasts from five weather models flags between-model conflict in Guayaquil in 53 % and 76 % of the 90- and 365-day windows and in no Santiago window; its assumptions cannot be verified here, and its level was supported only in simulated designs matched to a persistence diagnostic estimated from the same data, not at 30 days or under strong persistence. In the DICES safety ratings, absence of certified conflict is inconclusive because power is low.

Keywords neutrosophic statistics; typed indeterminacy; provenance of uncertainty; interval data; imprecise probability; certified conflict; Clopper–Pearson bounds; batch means; hesitant sets; affine arithmetic

AMS subject classifications (MSC 2020) 62A01, 62F25, 62G09, 65G30, 03B52

DOI: 10.19139/soic-2310-5070-4832

1. Introduction

Neutrosophic statistics writes an estimate as $N = d + I$, with a determinate part d and an indeterminate part I that is a set rather than a number [1]. The idea has produced a large applied literature, from neutrosophic analysis of variance to acceptance sampling [2, 3]. In most of it I is chosen by the analyst: an interval of plausible values, the range of repeated readings, or a fixed $I \in [I_L, I_U]$ carried through the arithmetic. In an engineering application, the joint roughness coefficient of a rock surface is written as $z = (\mu - \sigma) + 2\sigma I$ with $I \in [0, 1]$, so that the indeterminate part is plus or minus one standard deviation [4]; recent neutrosophic distributions consider both

*Correspondence to: Maikel Yelandi Leyva-Vázquez (Email: maikel.leyvav@ug.edu.ec; ORCID 0000-0001-7911-5879). Universidad de Guayaquil, Guayaquil, Ecuador. F. Smarandache: smarand@unm.edu.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

estimation of interval-valued parameters from interval observations and constructions that add externally specified imprecision [5].

Three critiques make the cost of this choice visible. Woodall et al. [6, 7] argue that neutrosophic statistics, as practised, is interval statistics under a new name. Haq and Woodall [8] show how pairing interval end points can give ranges that are too narrow. Haq and Shabbir [9] show that a neutrosophic random variable with a fixed interval-valued indeterminacy is a deterministic transform of a classical one. Smarandache [10] replied that not every indeterminacy is an interval, and that a hesitant set (a finite set of possible values) is sometimes the better model. We accept the narrow point of the critics (a fixed I adds no inferential content) and the broad point of the reply (indeterminacy is not one thing). The question of this paper is where the I in $N = d + I$ comes from, and how much of it can be computed from the data that generate it.

Our answer is to compute I from the records and to label each part by its source. Four sources are common: (a) *measurement imprecision* (records are sets: an interval reading, an "unsure" answer, a value inside a tolerance band); (b) *sampling* (there are few records); (c) *conflict between sources* (instruments, models or raters issue opposite verdicts); and (d) *hesitancy* (someone states a finite set of possible values). The computation does not remove the analyst: the confidence level, the tolerance band, the coding of undecided answers, the choice of sources and the reading of a hesitant set are decisions, and the results depend on them. What changes is that they are declared and that the results are reported as functions of them.

Why the source matters. The four sources call for different actions, and a single width hides which one applies. In the DICES safety ratings of Section 6.3, 298 conversations have an undecided majority verdict. For 229 of them the estimated identified interval lies on one side of the majority threshold, so additional ratings may settle the verdict; for 60 the "unsure" answers straddle the threshold, so reducing sampling uncertainty alone may not settle it and the meaning of "unsure" deserves examination. One interval per conversation reports all 298 as "undecided"; the labelled I suggests different follow-up priorities. These priorities come from estimates and have not been validated as interventions. Similarly, an average of five weather models hides that one model systematically forecasts more rain days than the others; a conflict statement keeps that model identified.

What is new and what is classical. The inferential machinery is classical: Clopper–Pearson bounds, outer confidence intervals for interval data and partially identified parameters, the stationary block bootstrap, batch-means t tests, Holm's procedure and affine arithmetic. We do not claim a new statistical inference. The contribution is a representation and a reporting rule built on these tools. Its main elements are the following.

(C1) An explicit, labelled split of the width of N into a measurement and a sampling part, with provenance, exact for counts and for interval data (Proposition 1).
(C2) Conflict reported outside I, as a test of opposite verdicts with a stated false-conflict rate and power (Propositions 4 and 5), including a paired version for sources observed on the same units, with its level proved under explicit conditions and a counterexample where they fail. The guarantee is that of the theorem under its assumptions; in our rain application those assumptions cannot be verified, so the rain results are illustrative and conditional.
(C3) The disjunctive and conjunctive readings of a hesitant set, with conflict under the conjunctive reading evaluated for the chosen proposition (Proposition 6).
(C4) An arithmetic that keeps the labels through operations and recomputes conflict source by source (Section 4), continuing our earlier article [17].
(C5) Practical guidance: what to do with each component (Table 14), a step-by-step reporting protocol with a checklist (Section 7), and the computational cost of each step.

These are supported by simulations, partly preregistered, and by three applications, whose provenance and inferential scope are stated in Table 13. Immediate identities, the scope of the rule $I^2 = I$, indexed hesitant arithmetic and the full arithmetic examples are in the Supplement.

Related work. Typed uncertainty is familiar outside neutrosophy: aleatoric and epistemic uncertainty in machine learning [18], conflict and non-specificity in credal sets [19], ontic and epistemic readings of set-valued data [20, 21], confidence intervals with imprecise data [22] and partial identification [23, 15, 16]. When imprecision is the only indeterminacy, a credal (lower-probability) representation suffices: normalised neutrosophic probabilities are exactly conjugate pairs of lower and upper probabilities, with indeterminacy equal to imprecision, and neutrosophy adds content only with conflict or typed indeterminacy [40]. Fuzzy statistics [24, 25] and symbolic data analysis [26, 27] are the established frameworks for imprecise and interval-valued records; Section 6.5 compares our implementations of their standard analyses with the labelled I. Between-source differences are the object of heterogeneity statistics in meta-analysis [13, 14]; opposite verdicts with support and refutation summing to more than one are the mass of the empty set in the transferable belief model [28], Dempster's conflict [29] and Belnap's value "both" [30]. Single-valued neutrosophic hesitant fuzzy sets [31] take T, I and F as elicited inputs; Torra [32] showed that the envelope of a hesitant element is an intuitionistic pair. Affine arithmetic goes back to Hansen [33] and to Comba and Stolfi [34]; see [35, 36]. Our earlier article [17] proved that $a + bI$ numbers and affine forms give the same ranges under $I = (1 + \varepsilon)/2$ and introduced multi-source and hesitant $a + bI$ numbers. The present paper continues the layered view of neutrosophic statistics [37, 38]; the conflict component is taken from [39], and the library `rnel` [41] already labels the width of count-based estimates by type.

Where each source sits in the existing literature. Each of the four sources has an established treatment outside neutrosophy, and the labelled I uses those treatments rather than replacing them. Table 1 maps each source to its counterpart, to the tool we use for it, and to what the label adds. In short, the contribution is to organise established outputs in one report: a source-labelled width decomposition, a separate source-specific conflict statement, and the assumptions and actions associated with each (Table 14). This comparison concerns the summaries and implementations used here; it does not imply that interval, belief-function or other uncertainty frameworks cannot retain source information or report conflict separately (Dempster's rule, for example, computes a conflict coefficient that can be reported, and the transferable belief model keeps the unnormalised mass). Interval-valued and hyper-interval-valued neutrosophic sets [42] generalise the *representation* of indeterminacy; our question is instead how much of it can be computed from records.

Table 1. The four sources of indeterminacy: established counterparts, the tool used here, and what the label adds.

Source	Established counterpart	Tool used here	What the label adds
Measurement imprecision	interval-valued data [27]; partial identification [23]; credal sets [19] (measurement-error models address a related but different problem)	undecided share u/n; identified set; W_{imp}	the part of the width that more records of the same kind do not remove, reported separately from sampling
Sampling	confidence intervals, survey sampling, bootstrap [50]	Clopper–Pearson bounds; outer intervals; block bootstrap	the part that more records remove, with its dependence model and whether its coverage was checked
Conflict	Dempster's conflict and the empty-set mass [29, 28]; heterogeneity in meta-analysis [13, 14]; interval non-overlap [12]; robust aggregation, which can downweight discordant sources	certified separation; paired batch-means test with Holm	a statement outside the width that keeps the disagreeing sources identified, with an error rate, its conditions and its power
Hesitancy	hesitant fuzzy sets [32, 31]; epistemic and ontic set-valued data [20]	disjunctive and conjunctive readings (Prop. 6)	distinguishes one holder's uncertainty from separate holders' assertions, with conflict evaluated for the chosen proposition

Roadmap and running example. Section 2 gives a short primer and the definitions; Section 3 the split, the conflict tests and the hesitant readings; Section 4 the arithmetic; Section 5 the simulations; Section 6 the applications and a decision-impact example; Section 7 the reporting protocol; Section 8 the discussion,

computational cost and limitations. A running example accompanies the definitions: two sources report $[4, 6]$ and $[5, 7]$ (imprecise, compatible), or outer confidence intervals $[4.0, 4.1]$ and $[6.9, 7.0]$ around crisp estimates (precise, contradictory). Both have the same hull, $[4, 7]$; the labels tell them apart (end of Section 3).

2. Setting and definitions

A short primer. A *neutrosophic number* $N = d + I$ has a determinate part d and an indeterminate part I, a set of values (here an interval $[0, W]$, so $N \in [d, d + W]$) [1]. A *neutrosophic triple* (T, I, F) for a proposition gives degrees of truth (support), indeterminacy and falsity (refutation); unlike a probability, $T + I + F$ need not equal one. In the normalised one-source construction of Definition 1, $T + I + F = 1$, so $[T, 1 - F]$ is an ordered interval of width I: in the Clopper–Pearson version an outer confidence interval for the identified set of $P(A)$, in the IDM version a pair of lower and upper probabilities. This interval reading does not extend unchanged to fused triples. Under existential fusion (Definition 2) the *glut* is $G = T + I + F - 1$, the probability that at least one source supports and at least one refutes the proposition (Belnap's value "both" [30]); it can be positive even when $T + F \leq 1$, and when $I = 0$ it is positive exactly when $T + F > 1$. *Typed* (or labelled) indeterminacy means that each part of I carries the name of its source. Table 2 is a glossary of the terms used.

Table 2. Glossary.

Term	Meaning in this paper
Source	an instrument, model or pool of raters that produces records about the same units
Record; support, refute, undecided	a point, interval or finite set; it supports A if it lies in A, refutes A if it lies outside, and is undecided otherwise
I_{meas}, I_{samp}	measurement and sampling components of I (Definition 1)
W_{imp}, W_{samp}, W_{gap}	imprecision, sampling and plug-in spread of the width of N for interval data (Definition 4)
Glut G	in Definition 2, $G = T + I + F - 1$, the probability of at least one supporting and one refuting verdict; when $I = 0$, $G > 0$ is equivalent to $T + F > 1$
Certified conflict	two sources' confidence sets are disjoint; its false-conflict rate is bounded under stated assumptions (Proposition 4)
Flagged conflict	conflict found by a test whose assumptions are not verified for the data (used in the rain application)
Hesitant set; disjunctive, conjunctive reading	a finite set of possible values; held by one holder, or one value per holder (Definition 5)
Labelled (typed) arithmetic	operations that keep the type of each indeterminacy symbol (Section 4)
Refined range; affine envelope	the reported range of a computed result; the enclosure that carries the typed decomposition

A *source* $k = 1, \ldots, K$ (an instrument, a forecast model, a pool of raters) produces *records* about a unit. A record is a set: a point, an interval or a finite set of values. For a proposition A ("rain on day t", "$\theta \leq c$"), a record *supports* A if it lies inside A, *refutes* A if it lies inside not-A, and is *undecided* otherwise. Source k yields counts a_k (support), r_k (refute) and u_k (undecided), $n_k = a_k + r_k + u_k$.

Definition 1 (One-source triple)
With $\text{CP}_\gamma(x, n)$ the one-sided lower Clopper–Pearson bound at confidence $1 - \gamma$ [43], the confidence version is

$$T = \text{CP}_{\alpha/2}(a, n), \quad F = \text{CP}_{\alpha/2}(r, n), \quad I = 1 - T - F, \qquad I_{\text{meas}} = \frac{u}{n}, \quad I_{\text{samp}} = \left(\frac{a}{n} - T\right) + \left(\frac{r}{n} - F\right).$$

The IDM version with prior strength $s > 0$ [11] is $T = a/(n + s)$, $F = r/(n + s)$, $I_{\text{meas}} = u/(n + s)$ and $I_{\text{samp}} = s/(n + s)$.

Definition 2 (Unit level: existential fusion)
Each source issues one categorical verdict, support with probability T_k, refute with probability F_k and neither otherwise, independently of the others. The fused triple is $T = 1 - \prod_k (1 - T_k)$, $F = 1 - \prod_k (1 - F_k)$, $I = \prod_k (1 - T_k - F_k)$ and $G = T + I + F - 1$.

G is the glut, the mass of Belnap's value "both". The construction is the unnormalised conjunctive rule of the transferable belief model [28]; $G = P(\text{some source supports and some source refutes})$, it is non-decreasing in every T_k and F_k, and it equals Dempster's conflict [29] (Supplement, Proposition S1). Its reading as a neutrosophic triple is from [39]. If two sources each report a 70 % forecast and their verdicts are drawn independently, $G = 0.42$ although the sources agree [44]; we therefore apply Definition 2 only to verdicts actually issued, and make claims that the *sources* contradict each other at the parameter level.

Definition 3 (Parameter level: certified conflict)
For a parameter θ, source k reports a confidence set $[L_k, U_k]$ for its target θ_k (for a chance, $[T_k, 1 - F_k]$). The sources are in *certified conflict* when two of the sets are disjoint; the certified gap is $\max(0, \max_k L_k - \min_k U_k)$.

Definition 4 (Interval data and the neutrosophic number)
Source k has interval records $[\ell_{ki}, \upsilon_{ki}]$, $i = 1, \dots, n_k$, identified interval for the mean $[\mathrm{E}\,\ell, \mathrm{E}\,\upsilon]$, estimate $[\bar{\ell}_k, \bar{\upsilon}_k]$ and outer confidence interval $L_k = \bar{\ell}_k - t\, s_\ell/\sqrt{n_k}$, $U_k = \bar{\upsilon}_k + t\, s_\upsilon/\sqrt{n_k}$, t the Student quantile at $1 - \alpha/(2K)$ (the Bonferroni version of [15, 16]). The reported number is $N = d + I$ with $d = \min_k L_k$, $I \in [0, W_H]$, $W_H = \max_k U_k - \min_k L_k$, and $W_H = W_{\text{samp}} + W_{\text{gap}} + W_{\text{imp}}$, where, with $H_M = [\min_k \bar{\ell}_k, \max_k \bar{\upsilon}_k]$, $W_{\text{samp}} = W_H - |H_M|$, $W_{\text{gap}} = \max(0, \max_k \bar{\ell}_k - \min_k \bar{\upsilon}_k)$ (plug-in gap: descriptive, it absorbs sampling noise) and $W_{\text{imp}} = |H_M| - W_{\text{gap}}$.

Definition 5 (Hesitant set)
A hesitant set $h = \{v_1, \dots, v_m\}$ of possible values of θ is read *disjunctively* when one holder states that θ is one of them, and *conjunctively* when holder j asserts $\theta = v_j$ [20]. Disjunctively, $N_h(A) = 1[h \subseteq A]$, $\Pi_h(A) = 1[h \cap A \neq \emptyset]$ and the triple is $(N_h, \Pi_h - N_h, 1 - \Pi_h)$. Conjunctively, the holders are dogmatic sources and Definition 2 applies.

What is computed and what is chosen. The decomposition is computed from the records once the analyst has fixed the proposition and its threshold, the tolerance band that maps a record to support, refutation or undecided (δ in Section 6.1), the coding of undecided answers, the sources, the confidence level α (or the IDM strength s), the dependence model and, for hesitant sets, the reading. Two objects are reported: the width of N, split into measurement imprecision, sampling and (with several sources) a plug-in spread; and, separately, a conflict statement. Conflict is not a component of the width; it is a relation between sources. We call a conflict *certified* only when the conditions of Proposition 4 or 5 are argued to hold for the data; when they cannot be verified, as in the rain application, we say that the test *flags* a conflict.

3. Results

Proposition 1 (Labelled split)
(a) *Counts.* In Definition 1, $T + I + F = 1$, $I = I_{\text{meas}} + I_{\text{samp}}$ and both parts are non-negative. For the IDM, $I_{\text{samp}} = s/(n + s) \to 0$. For Clopper–Pearson, $0 \le I_{\text{samp}} \le 2/(n + 1) + \sqrt{2(1 - \alpha/2)/\alpha}\,/\sqrt{n + 2}$ for every outcome, so $I_{\text{samp}} \to 0$ deterministically. For i.i.d. records with probabilities (p_a, p_r, p_u), $I_{\text{meas}} \to p_u$ almost surely in both versions. (b) *Interval data.* In Definition 4 the three components are non-negative and add up to W_H; with one source $W_{\text{gap}} = 0$ and $W_{\text{imp}} = \bar{\upsilon} - \bar{\ell}$. The certified gap never exceeds W_{gap}.

Proof
(a) The sums hold by construction, and $\mathrm{CP}_\gamma(x, n) \le x/n$ gives $I_{\text{samp}} \ge 0$. For $x \ge 1$, $\mathrm{CP}_\gamma(x, n)$ is the γ quantile of $\mathrm{Beta}(x, n - x + 1)$, with mean $x/(n + 1)$ and standard deviation at most $1/(2\sqrt{n + 2})$; by Cantelli's inequality the γ quantile is at least mean minus $\sigma\sqrt{(1 - \gamma)/\gamma}$, so $0 \le x/n - \mathrm{CP}_\gamma(x, n) \le 1/(n + 1) + \sqrt{(1 - \gamma)/\gamma}\,/(2\sqrt{n + 2})$; for $x = 0$ the difference is 0. The strong law gives $u/n \to p_u$. (b) $[\bar{\ell}_k, \bar{\upsilon}_k] \subseteq [L_k, U_k]$ gives $W_H \ge |H_M|$; $\max \bar{\ell} - \min \bar{\upsilon} \le \max \bar{\upsilon} - \min \bar{\ell}$; and $\max L - \min U \le \max \bar{\ell} - \min \bar{\upsilon}$. □

Proposition 2 (Coverage)
If the records are i.i.d. with probabilities (p_a, p_r, p_u), the Clopper–Pearson triple satisfies $P(T \le p_a,\ F \le p_r) \ge$

$1-\alpha$, so $[T, 1-F]$ contains the identified set $[p_a, p_a+p_u]$ of $P(A)$ with probability at least $1-\alpha$. For the IDM triple with $0<p_a<1$, $P(T \le p_a) \to 1/2$; if $p_a \in \{0,1\}$, $P(T \le p_a) = 1$ for every n.

Proof
$a \sim \mathrm{Bin}(n, p_a)$, each Clopper–Pearson margin is exact at $\alpha/2$, and the union bound gives $1-\alpha$. For the IDM, $a/(n+s) - p_a = (a/n - p_a) - s\,a/(n(n+s))$, the second term is $O(1/n)$ and $\sqrt{n}(a/n - p_a)$ is asymptotically normal with mean zero. If $p_a = 0$ then $T = 0$; if $p_a = 1$ then $T = n/(n+s) < 1$. □

The IDM is a model of prior ignorance, not a confidence procedure; Proposition 2 says only that its bounds should not be read as confidence bounds.

Proposition 3 (Masking)
(a) *Imprecision.* For dogmatic interval sources, "$\theta \le c$" is a glut exactly for $c \in [\min_k U_k, \max_k L_k)$. Replacing any $[L_k, U_k]$ by a set that contains it can only shrink this set; in Definition 2, moving mass from T_k or F_k to the undecided part cannot increase G. (b) *Sampling.* Fewer records do not give nested intervals, so (a) does not apply. If the targets share a point, the probability that a sub-sample shows a certified conflict is at most the false-conflict probability of Proposition 4.

Proof
(a) Source k supports "$\theta \le c$" iff $U_k \le c$ and refutes it iff $L_k > c$; decreasing an L_k or increasing a U_k shrinks $[\min U, \max L)$; the last statement is the monotonicity of G. (b) The sub-sample is itself a sample to which Proposition 4 applies. □

With $K = 2$ sources whose records support or refute with probability $1/2$ each, the first 40 of 400 records showed a certified conflict that the full sample did not have in 13 of 20 000 replicates, all accounted for by the exact false-conflict probability at $n = 40$ (6.5×10^{-4}, expected 13.1).

Proposition 4 (False-conflict control, conditional on simultaneous coverage)
Suppose the targets of all sources share a common point θ^* and the confidence sets cover their targets simultaneously with probability at least $1-\alpha$. Then the probability of a certified conflict is at most α. If two targets are disjoint and every set shrinks to its target, certified conflict occurs with probability tending to one.

Proof
With probability at least $1-\alpha$ every set contains its target and hence θ^*, and then the certified gap is zero. In the second part the gap between shrinking sets converges to the positive gap between the targets. □

Simultaneous coverage holds exactly for Clopper–Pearson sets at level α/K with independent records (Bonferroni) and asymptotically for Definition 4. When it fails, Proposition 4 gives no guarantee: the false-conflict rate may still be small (if the limits of different sources fail together) or may not. Section 5 shows that the seasonally adjusted block-bootstrap limits used in earlier versions of this work do not reach simultaneous coverage in finite windows with persistence; for that reason the conflict claims of Section 6.1 use the paired test below, whose level does not rest on the coverage of all limits (for those data it rests instead on assumptions that cannot be verified; Section 6.1).

What certified conflict is. Definition 3 tests the hypothesis that the sources' targets share a point (for partially identified targets, that their identified sets overlap) against the alternative that two of them are separated, which is what "the sources issue opposite verdicts" means. It is the multi-source version of confidence-interval non-overlap, which is conservative as a test of difference [12]. A test of equality (analysis of variance, Cochran's Q and I^2 [13, 14]) answers whether the sources differ; certified conflict answers the narrower question whether their sets of plausible values are disjoint, which with identified sets of positive width is not the same question. Its price is power: in the preregistered simulations the false-conflict rate never exceeded 0.002 in the i.i.d. designs, while the power at $n = 100$ per source and a gap of 0.2 was 0.32–0.40 and below 0.02 at $n = 20$; for two pools of 36 raters with majority proportions 0.40 vs 0.60, 0.35 vs 0.65 and 0.30 vs 0.70 it is 0.04, 0.19 and 0.50 (exact, Section 6.3). Absence of certified conflict is weak evidence of agreement unless the power is reported.

Paired sources. When the sources are observed on the same units (the five rain models forecast the same days), conflict between two of them can be tested on paired daily data. For an ordered pair (j, k) let S_{jt} and R_{kt} indicate that source j supports and source k refutes on day t, and

$$D_{jk} = p_{a,j} - (1 - p_{r,k}) = \frac{1}{n} \sum_t \mathrm{E}\,[S_{jt} + R_{kt} - 1],$$

the distance by which the lower end of j's identified set exceeds the upper end of k's, over the n days used. Its estimate is the mean of the single daily series $d_t = S_{jt} + R_{kt} - 1$, whose variance uses the covariance between the two sources.

Proposition 5 (Paired certified conflict)
Let $m \geq 0$ be a margin and $\gamma \in (0, 1/2)$. (a) *What conflict requires.* There is a threshold c with $p_{a,j} > c + m/2$ and $1 - p_{r,k} < c - m/2$ if and only if $D_{jk} > m$. Moreover $D_{jk} + D_{kj} = -(p_{u,j} + p_{u,k}) \leq 0$, so at most one direction of a pair can be in conflict. (b) *Fixed threshold.* If c is fixed in advance and π_j^+, π_k^- are valid p-values for "$p_{a,j} \leq c + m/2$" and "$1 - p_{r,k} \geq c - m/2$", then $\max(\pi_j^+, \pi_k^-)$ is a valid p-value for "not (j above and k below)" (intersection–union test). (c) *Paired versus non-overlap.* If $(\hat{p}_{a,j}, \hat{p}_{r,k})$ is normal with standard deviations σ_a, σ_r and any correlation ρ, the non-overlap rule and the paired rule "$\hat{D}_{jk} > z_{1-\gamma}\,\mathrm{sd}(\hat{D}_{jk})$" both have level γ at $D_{jk} \leq 0$, and the paired rule rejects whenever the non-overlap rule does. (d) *Batch means.* Drop the first $n - b\lfloor n/b \rfloor$ days and split the remaining ones into b consecutive batches of equal length; let $\bar{d}_1, \ldots, \bar{d}_b$ be the batch means, $\hat{D}_{jk}$ their average (the mean of d_t over the kept days), s_b their standard deviation and $t_{jk} = (\hat{D}_{jk} - m)\sqrt{b}/s_b$; the target is D_{jk} over the kept days. Under $D_{jk} \leq m$, $P(t_{jk} > t_{b-1,1-\gamma}) \leq \gamma$ in each of the two cases: (d1) the $\bar{d}_i$ are independent normal with a common variance and arbitrary means whose average is D_{jk}; (d2) they are independent normal with a common mean $\mu \leq m$ and arbitrary variances, and $2\gamma \leq 0.083$ [76, 73]. When the means and the variances both differ, no such bound holds (Remark 1). (e) *Holm across pairs.* If every ordered pair has a p-value valid under its own null at levels up to α/m_0, where m_0 is the number of true nulls, Holm's procedure [64] over the $K(K-1)$ ordered pairs declares a false conflict with probability at most α, whatever the dependence between the p-values. By (a) at most one direction of each pair is in conflict at $m = 0$, so $m_0 \geq K(K-1)/2$ and validity is needed only up to $2\alpha/(K(K-1))$ (0.005 for $K = 5$), within the range of (d2).

Proof
(a) If $D_{jk} > m$, take $c = (p_{a,j} + 1 - p_{r,k})/2$; then $p_{a,j} - c = c - (1 - p_{r,k}) = D_{jk}/2 > m/2$. The converse adds the two inequalities; the identity follows from $p_a + p_r + p_u = 1$. (b) The null is the union of the two component nulls; under either one the corresponding component, hence the maximum, is valid [69]. (c) $\mathrm{sd}(\hat{D}_{jk}) = (\sigma_a^2 + \sigma_r^2 + 2\rho\sigma_a\sigma_r)^{1/2} \leq \sigma_a + \sigma_r$. (d1) With equal batch lengths $\hat{D}_{jk} = \bar{\bar{d}}$. Write $\bar{d}_i = \mu_i + \sigma\varepsilon_i$; $\bar{\bar{d}}$ and the residuals $\bar{d}_i - \bar{\bar{d}}$ are independent, $(b-1)s_b^2/\sigma^2$ is noncentral chi-square with noncentrality $\sum_i(\mu_i - \bar{\mu})^2/\sigma^2$, stochastically no smaller than the central one, and $P(\bar{\bar{d}} - m > c\,s_b/\sqrt{b} \mid s_b)$ decreases in s_b for $c > 0$ and increases in $\bar{\mu} \leq m$; at $\bar{\mu} = m$ and equal μ_i it is exactly γ. (d2) At $\mu = m$ the t statistic is symmetric about zero, so the one-sided level is half the two-sided one, which the cited theorem bounds for two-sided levels up to 0.083; a common shift of all batch means leaves s_b unchanged and lowers t_{jk} pointwise, so $\mu < m$ gives a smaller rejection probability. (e) A false rejection requires some true null to have $p \leq \alpha/m_0$ at Holm's first step involving a true null, so the union bound over the m_0 true nulls gives α with marginal validity at level α/m_0. □

Remark 1 (Both heterogeneities at once). Take $b = 10$ independent normal batch means with means $(-0.9, 0.1, \ldots, 0.1)$ (average 0, the null value), standard deviation 1 for the first and $\tau \to 0$ for the others, and the one-sided level $\gamma = 0.0025 = 0.05/20$ of the first Holm step. Then $t = Z/|Z - 1|$ with Z standard normal, and the test rejects iff $Z \in (c/(1+c), c/(c-1))$, $c = t_{9,0.9975} = 3.69$: probability 0.131 (Monte Carlo, 1 000 000 draws: 0.131), more than fifty times the nominal level. With a common variance and the same means the rate is 0.0019; with a common mean and the same variances it is 6.0e-06; for standard-deviation ratios 1, 2, 4, 10 and 100 between the first batch and the rest it is 0.0017, 0.0033, 0.0083, 0.0451, 0.1297 (`code/counterexample_v06.py`). Seasonal data can have both kinds of heterogeneity, so the level of the batch-means rule in the application is not

guaranteed by (d); simulation in designs with seasonality in both the mean and the variance of d_t gives it only limited support (Section 5).

Three further points fix the implementation. The reference distribution is always Student's t with $b-1$ degrees of freedom; no normal approximation is used in the test, but the normality of the batch means is itself an approximation that improves with the batch length (3 days at $W=30$, 9 at $W=90$, 36 at $W=365$ with $b=10$), which is one reason we draw no conclusion from 30-day windows. A pair whose b batch means are all equal ($s_b=0$) gets $p=1$ and is never certified. For dependent series the batch means are only asymptotically independent and normal (batch length growing, b fixed [71, 72]). Holm and Bonferroni reject *some* pair in exactly the same windows; the gain over separate intervals comes from pairing, not from Holm. A max-type step-down [70] would use the dependence between pairs, but our resampling-based versions did not keep their level (Section 5).

Proposition 6 (The two readings of a hesitant set)
(a) For every event A, the disjunctive triple of h and the triple of its hull $[\min h, \max h]$ satisfy $T_h \geq T_{\text{hull}}$, $F_h \geq F_{\text{hull}}$ and $I_h \leq I_{\text{hull}}$, with equality for all half-lines; the inequality is strict exactly when A meets the hull but not h, or when $h \subseteq A$ but the hull is not contained in A. (b) The conjunctive glut equals the disjunctive indeterminacy: $G_{\text{conj}}(A) = I_{\text{disj}}(A) = \Pi_h(A) - N_h(A)$, and the conjunctive triple has $T+F = 1+G$.

Proof
(a) $h \subseteq$ hull; for a half-line $(-\infty, c]$, $h \subseteq A$ iff $\max h \leq c$ iff the hull is inside, and h meets A iff $\min h \leq c$. (b) Conjunctively $T = \Pi_h(A)$, $F = 1 - N_h(A)$ and $I = 0$. □

Smarandache's temperature example [10, §2.12] illustrates (a): with $t = 10 + I$, $I \in \{5, 30, 35\}$, the event "$18 \leq t \leq 38$" gets $(T, I, F) = (0, 0, 1)$ from the hesitant set and $(0, 1, 0)$ from the hull $[15, 45]$. Part (b) states the exchange of indeterminacy for glut between the epistemic and ontic readings of set-valued data [20, 21]: when the elements come from different holders, the indeterminacy of the set reappears as conflict.

Worked example (same hull, different content). Two sources report $[4, 6]$ and $[5, 7]$ (imprecise, compatible), or outer confidence intervals $[4.0, 4.1]$ and $[6.9, 7.0]$ around crisp estimates (precise, contradictory). Both hulls give $N = 4 + I$, $I \in [0, 3]$. The labelled width is $W_{\text{imp}} = 3$ in the first case and $W_{\text{gap}} = 2.9$, $W_{\text{samp}} = 0.1$ in the second (certified gap 2.8); for "$\theta \leq 5.5$" the first case gives $(T, I, F) = (0, 1, 0)$ and the second $(1, 0, 1)$ with $G = 1$. A forest plot would also show the difference; what the labelled number adds is the explicit split, the label and a rule for the conflict claim.

4. Arithmetic of labelled neutrosophic numbers

If X and Y carry labelled indeterminacy, what is the labelled indeterminacy of $X + Y$, XY or X/Y? We ask that the result enclose every value the inputs admit, that the same quantity used twice take the same value while independent quantities do not cancel, and that the labels keep their reading as far as possible. Interval arithmetic already propagates uncertain coefficients through applied models, for example input–output models solved by interval Gauss–Seidel iteration [45]; what we add is that each part of the width keeps its type.

Definition 6 (Labelled neutrosophic number of one source)
$N = d + \sum_k w_k I_k$, $I_k \in [0, 1]$, where each symbol k has a type $\tau(k) \in \{\text{meas}, \text{samp}, \text{mix}\}$ and a name (the quantity it belongs to). Symbols with different names vary independently; the same symbol in two numbers is the same unknown. The typed widths are $w_\tau(N) = \sum_{\tau(k)=\tau} |w_k|$.

Without the types, Definition 6 is the multi-source number $a + \sum_i b_i I_i$ of [17, Definition 1], which writes Smarandache's refined indeterminacies [46, pp. 98–99] with one symbol per source; we attach a type and a name to every symbol. Through $I_k = (1 + \varepsilon_k)/2$, N becomes an affine form $c + \sum_k r_k \varepsilon_k$, $r_k = w_k/2$ [34, 35]; sums and scalar multiples are exact on the range and the typed widths obey the triangle inequality (Supplement, Lemma S2). A source with outer interval $[L, U]$ around $[\bar{\ell}, \bar{v}]$ enters as $c = (\bar{\ell} + \bar{v})/2$, $r_{\text{meas}} = (\bar{v} - \bar{\ell})/2$ and

$r_{\text{samp}} = \max(\bar{\ell} - L, U - \bar{v})$. The type mix is never an input; it is created by nonlinear operations. A confidence interval is propagated as if its end points were hard bounds: if each input interval covers with probability $1 - \alpha_i$, the computed range covers with probability at least $1 - \sum_i \alpha_i$, whatever the operation.

Write $X = x_0 + \sum_i x_i \varepsilon_i$ and $Y = y_0 + \sum_j y_j \varepsilon_j$. Then $XY = x_0 y_0 + \sum_k (y_0 x_k + x_0 y_k)\, \varepsilon_k + Q$ with $Q = \sum_{i,j} x_i y_j \varepsilon_i \varepsilon_j$.

Definition 7 (Typed remainder)
Group the terms of Q by the pair of types of their symbols: (meas, meas) gives type meas, (samp, samp) gives samp, every other pair gives mix. For a group g, with diagonal terms ($\varepsilon_i^2 \in [0, 1]$) and off-diagonal pairs,

$$\ell_g = \sum_{i \in g} \min(0, x_i y_i) - \sum_{i<j \in g} |x_i y_j + x_j y_i|, \qquad u_g = \sum_{i \in g} \max(0, x_i y_i) + \sum_{i<j \in g} |x_i y_j + x_j y_i|,$$

and Q_g is replaced by $(\ell_g + u_g)/2 + \varepsilon_g (u_g - \ell_g)/2$ with a fresh symbol ε_g of type g. For a square, $Q \in [0, R^2]$ with $R = R_{\text{meas}} + R_{\text{samp}} + R_{\text{mix}}$, split as R^2_{meas} (meas), R^2_{samp} (samp) and the rest (mix). For f convex or concave on the range $[a, b]$ of X (the reciprocal when $0 \notin [a, b]$; powers), $f(X)$ is replaced by its Chebyshev line plus an error symbol of radius δ [35], split into the error on the measurement-only range (meas), on the sampling-only range (samp, capped so the parts add up to δ) and the rest (mix). Division is $X \cdot (1/Y)$.

Proposition 7 (Enclosure and vanishing of the sampling-dependent remainders)
(a) For every admissible value of the input symbols, the true value of XY, X^2, X^n and X/Y lies in the range of the computed number. (b) A mix group of a product contains only terms in which one factor is attached to a sampling (or mix) symbol. (c) For a finite sequence of these operations on inputs whose ranges stay in a fixed bounded set, with every reciprocal applied to a number whose range stays in $(-\infty, -m] \cup [m, \infty)$ for a fixed $m > 0$, if all input sampling widths tend to zero with the other coefficients fixed, then $w_{\text{samp}} + w_{\text{mix}}$ of the result tends to zero.

Proof
(a) $Q = \sum_g Q_g$ and each $Q_g \in [\ell_g, u_g]$ because $\varepsilon_i^2 \in [0, 1]$ and $|\varepsilon_i \varepsilon_j| \le 1$; for a square, $0 \le (\sum_i x_i \varepsilon_i)^2 \le R^2$; for the Chebyshev step $|f(x) - \alpha x - \zeta| \le \delta$ on $[a, b]$; composition preserves enclosure. (b) By Definition 7. (c) By induction: linear sampling and mix coefficients are bounded combinations of input sampling coefficients; for a reciprocal on $[a, b] \subseteq [m, M]$ the minimax error is Lipschitz in the end points with a constant depending on m and M, so its sampling and mix parts are at most a constant times $R_{\text{samp}} + R_{\text{mix}}$. □

The conditions in (c) are needed: for Y_n with range $[1/(2n), 3/(2n)]$, a pure sampling width $1/n \to 0$, the range of $1/Y_n$ has width $4n/3 \to \infty$ (Supplement, Section S3, which also shows that the measurement label of a nonlinear result depends on the linearisation point and is therefore a convention).

Remark 2 (Which labels are identified and which are conventions). Conditional on fully specified input affine forms, including the identities and types of shared symbols, sums, differences and scalar multiples have algebraically determined typed coefficients: each coefficient of the result is a fixed linear combination of input coefficients of the same type. This is conditional algebraic determination, not statistical identification of the input labels from the data. Products, powers and reciprocals additionally require declared rules for allocating cross terms $x_i y_j \varepsilon_i \varepsilon_j$ of different types and the parts of a Chebyshev error, which depend on the linearisation point; these allocations are *conventions*. The excess of the affine envelope over the refined range is left unattributed, as explained below. Proposition 7 establishes enclosure, and vanishing sampling and mix widths, for the allocation of Definition 7 under its boundedness and separation-from-zero conditions; it does not establish invariance under arbitrary reallocations (representing the range of X^2, $X = \varepsilon_{\text{meas}}$, by $1/2 + \varepsilon_{\text{mix}}/2$ encloses it but leaves a positive mix width without any sampling). Reports should give the refined range as the result and the typed decomposition as that of the chosen affine envelope.

Two objects, not one. The labelled affine form and interval arithmetic are both enclosures, so their intersection is also one. We report two objects and keep them apart: (1) the *refined range*, the intersection of the two enclosures, which is the reported range of the result; and (2) the *typed decomposition of the labelled affine envelope*,

$w_{\mathrm{meas}} + w_{\mathrm{samp}} + w_{\mathrm{mix}}$, which adds up to the width of the affine envelope and not to the width of the refined range. For XY in Table 3 the refined range is $[2, 3.64]$ (width 1.64) while the decomposition $1.23 + 0.48 + 0.05$ adds up to the envelope width 1.76. We do not rescale the labels to the refined width, because no attribution rule for the excess of the envelope over the intersection follows from the construction. The interval comparator uses dedicated powers: the interval square of $x \in [-1, 1]$ is $[0, 1]$, while the product $x \cdot x$ of two independent occurrences is $[-1, 1]$; the dependency problem of interval arithmetic is shown by $z(1 - z)$, where each occurrence of z is treated as independent.

Table 3. Worked examples (selection; the full table with $X + Y$, $X - Y$, X^2 and X^3 is Table S1). $X = 2 + 0.4I^X_{\mathrm{meas}} + 0.2I^X_{\mathrm{samp}}$, X' an independent replicate, $Y = 1 + 0.3I^Y_{\mathrm{meas}} + 0.1I^Y_{\mathrm{samp}}$, $z = I_{\mathrm{meas}}$, $x = -1 + 2I_{\mathrm{meas}}$. Interval = interval arithmetic with every occurrence independent and dedicated powers (*: the product $x \cdot x$ gives $[-1, 1]$); one I = every quantity written with the same symbol I and reduced with $I^2 = I$ and Smarandache's division rule [1, pp. 35–38]. (1) refined range = intersection of the interval enclosure and the labelled affine envelope (reported range). (2) the labelled affine envelope and its typed decomposition; "sum" = $w_{\mathrm{meas}} + w_{\mathrm{samp}} + w_{\mathrm{mix}}$ = width of the envelope, not of (1). True ranges on a grid of 41 values per symbol.

				(1) refined	**(2) labelled affine envelope**				
Expression	True range	Interval	One I	range	range	w_{meas}	w_{samp}	w_{mix}	sum
$X - X$	$[0, 0]$	$[-0.6, 0.6]$	$[0, 0]$	$[0, 0]$	$[0, 0]$	0	0	0	0
$X - X'$	$[-0.6, 0.6]$	$[-0.6, 0.6]$	$[0, 0]$	$[-0.6, 0.6]$	$[-0.6, 0.6]$	0.8	0.4	0	1.2
XY	$[2, 3.64]$	$[2, 3.64]$	$[2, 3.64]$	$[2, 3.64]$	$[1.88, 3.64]$	1.23	0.48	0.05	1.76
X/Y	$[1.4286, 2.6]$	$[1.4286, 2.6]$	$[1.8571, 2]$	$[1.4286, 2.6]$	$[1.2877, 2.6]$	0.907	0.3439	0.0614	1.3123
$z(1 - z)$	$[0, 0.25]$	$[0, 1]$	$[0, 0]$	$[0, 0.25]$	$[0, 0.25]$	0.25	0	0	0.25
x^2	$[0, 1]$	$[0, 1]^*$	$[1, 1]$	$[0, 1]$	$[0, 1]$	1	0	0	1

Where a quantity appears twice ($X - X$, $z(1 - z)$), the affine form is exact and interval arithmetic is not; where each input appears once and the function is monotone (XY, X/Y), interval arithmetic is exact and the affine envelope is wider, the known trade-off [35]; the refined range is exact in every row of Tables 3 and S1. The one-I column cancels $X - X$ correctly, but a single shared symbol makes independent quantities move together ($X - X' = 0$) and the reduction misses interior extremes ($z(1 - z)$, x^2): $I^2 = I$ is the exact arithmetic of a two-valued indeterminacy $I \in \{0, 1\}$, not of an interval (Supplement, Proposition S4). In a random check of 20 000 expressions no computed range missed a sampled value (Supplement, Section S3). Hesitant sets are handled by indexing their values, so that the same index used twice is the same value (Supplement, Section S4).

4.1. Conflict: operate per source, then decompose

Conflict is a relation between sources, so it is not carried as a width: the operation is applied to each source and Definitions 3–4 are applied to the results. Let $S_k \subseteq \mathbb{R}^p$ be the set of input vectors that source k admits, g the operation and $E_k \supseteq g(S_k)$ the computed enclosure.

Proposition 8 (Conflict under operations)
(a) If the S_k have a common point, the outputs are not in certified conflict. (b) If $p = 1$ and g is continuous and injective on the union of the pairwise disjoint intervals S_k, the exact outputs are pairwise disjoint; an affine map $a\theta + b$, $a \neq 0$, multiplies the certified gap by $|a|$; the gap of the enclosures never exceeds that of the exact outputs, so a certified conflict after an operation is not an artefact of the arithmetic, although loose enclosures can hide one. (c) A non-injective map can remove a certified conflict. (d) If the S_k meet coordinate by coordinate but not jointly, an operation can create one.

Proof
(a) If $\theta^* \in \bigcap_k S_k$ then $g(\theta^*) \in \bigcap_k E_k$. (b) Continuous images of intervals are intervals, injectivity keeps them disjoint, and $E_k \supseteq g(S_k)$ moves lower ends down and upper ends up. (c), (d) Sources $[-2.0, -1.9]$ and $[1.9, 2.0]$ are in certified conflict but θ^2 gives the same interval for both; if source 1 has $X \in [0, 1]$, $Y = X$ and source 2 has $X \in [0, 0.5]$, $Y = X + 0.5$, the X and Y intervals overlap but $X - Y$ is 0 for one and -0.5 for the other (Supplement, Section S6). □

If the input sets cover their target vectors simultaneously with probability at least $1-\alpha$ and the targets share a point, the probability of a certified conflict after any operation is at most α.

5. Simulations

Preregistered i.i.d. part. The design, success criteria and seeds of studies S1–S4 (Table 4) were frozen in the project repository before any run (git commit c424482; the time stamp is that of a private repository, not of a third-party registry).

Table 4. Preregistered simulation, i.i.d. records ($\alpha = 0.05$). CP = Clopper–Pearson.

Study	Design	Criterion	Result
S1.1	one source, 3 scenarios × 6 sizes, 2000 reps	CP joint coverage ≥ 0.935	min 0.9515
S1.2	same	I_{samp} decreasing, < 0.05 at $n = 3000$	pass
S1.3	same	$\lvert\overline{I_{\text{meas}}} - p_u\rvert < 0.01$ for $n \geq 1000$	max 0.0006
S1.4	IDM, $s = 2$, $n = 3000$	report only	coverage 0.166, 0.213, 0.034
S2.1	$K \in \{2, 3, 5\}$, no true conflict (45 cells)	false-conflict rate ≤ 0.065	max 0.002
S2.2	true gap 0.2	power ≥ 0.80 at $n = 500$	0.999, 1.000, 1.000
S2.3	gap masked by imprecision	rate ≤ 0.065	max 0.0005
S2.4	unit level, categorical verdicts	MC G = closed form	within 3 MC SE
S3.1	interval data, no bias	false certified gap ≤ 0.07	max 0.005
S3.2	same	W_{samp} decreasing in n	pass
S3.4	bias 1.5, rounding w	power ≥ 0.80 ($w \leq 0.5$); ≤ 0.07 ($w = 2$)	1.000; 0.000
S4	10 000 hesitant sets and events	0 violations of Prop. 6	0; strict in 19.9 %

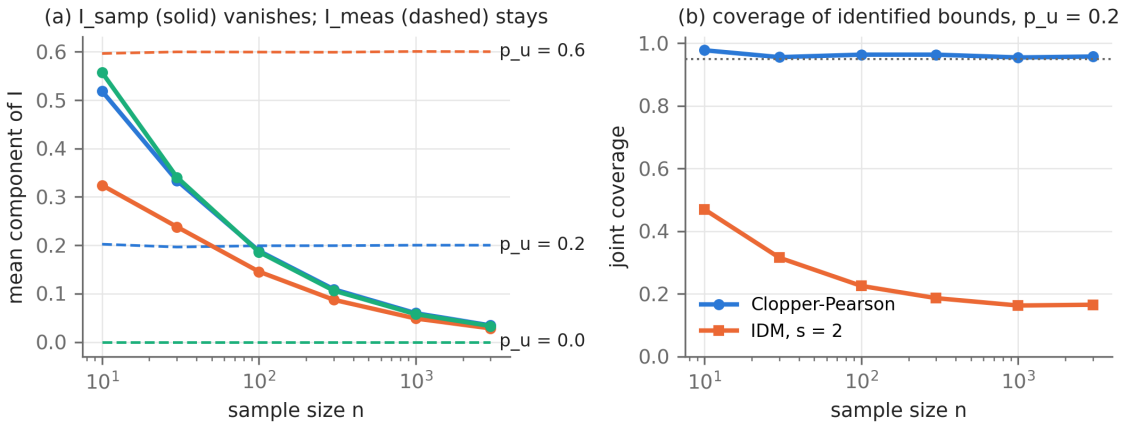


Figure 1. Simulation S1. (a) Mean sampling component (solid) and measurement component (dashed) for three scenarios; (b) joint coverage of the identified bounds, Clopper–Pearson versus IDM ($p_u = 0.2$).

Dependent sources: the generator. To check the rules of Section 6.1 we simulate $K = 5$ sources $X_{kt} = z_t + (A + c_k)\sin(2\pi t/365) + b_k + s_k(t)\, e_{kt}$, with a common AR(1) weather state z_t (coefficient ϕ_z), errors e_{kt} that are AR(1) with their own coefficient ϕ_e and independent across sources, and $s_k(t) = \mathrm{sd}_e(1 + v_k \cos(2\pi t/365))$; a source supports "rain" when $X_{kt} \geq \delta$ and refutes it when $X_{kt} < -\delta$. Each replicate generates three years, fits the seasonal mean as in the application and analyses one window of $W \in \{30, 90, 365\}$ days; the true D_{jk} of every window is computed in closed form, so each rejected pair is classified as a true or a false conflict. Four studies use this family: a dependence study with $\phi_e = \phi_z/2$ (v0.4, not preregistered; Table S4), a preregistered

study of the paired tests (v0.5, `PREREG_v05.md`), a study of simultaneous coverage (v0.5, not preregistered) and a preregistered study with own persistence and seasonal heterogeneity (v0.6, `PREREG_v06.md`, frozen before any run, with one amendment after a 100-replicate smoke test and before the full run).

Paired tests (preregistered, v0.5). With the error persistence tied to the common state ($\phi_e = \phi_z/2$), identical, highly correlated, compatible-but-different, non-seasonal, Markov and fixed-threshold null designs, the preregistered block-bootstrap paired test failed its size criterion (familywise false-conflict rate up to 0.155), while the batch-means version BM-10 (amendment, before any real-data run) kept it (largest rate 0.051 against a limit of 0.064; Table S5). With the exact v0.6 batch construction the largest rates over the same null designs were 0.031, 0.048 and 0.047 at $W = 30, 90, 365$ (2000 new replicates per design, not preregistered; Table S6). Its mean power over 30 alternative designs was 0.434, against 0.178 for the seasonally adjusted bootstrap rule.

Simultaneous coverage. In the same family, the empirical simultaneous coverage of all $2K = 10$ seasonally adjusted bootstrap limits (nominal 0.95) was 0.231, 0.425, 0.643 at $W = 30, 90, 365$ with $\phi_z = 0.95$, and at least 0.667, 0.843 and 0.881 for $\phi_z \leq 0.5$ (Table S7). Under persistence one window sees few independent weather episodes, and no resampling of the window's own days can show how much its frequencies would change in another window. The sampling components computed with these limits are therefore potentially underestimated, and their nominal coverage is not validated in these scenarios. In those designs the conflict rules still kept their false-conflict rates because the limits of different sources failed together, in the same direction, when the shared state moved them.

Own persistence and seasonal heterogeneity (preregistered, v0.6). That last observation depends on the errors of different sources moving together. We therefore preregistered 81 designs (1000 replicates each; bootstrap $B = 999$) in which each source's errors have their own persistence $\phi_e \in \{0.3, 0.5, 0.8, 0.95\}$, with $\phi_z \in \{0, 0.5\}$ and $\mathrm{sd}_e \in \{0.6, 1.5\}$, so that the limits of different sources need not fail together; and designs with seasonality in both the mean and the variance of the paired series ($c = (0.6, -0.6, 0, 0, 0)$, $v = (v, -v, 0, 0, 0)$, $v \in \{0, 0.7\}$; over a full 365-day window every D_{jk} is exactly zero, and over the 360 days kept by BM-10 it is close to zero, while the batch means differ through the year in mean and variance). To place the data among the designs we use one persistence diagnostic, computed identically on simulated and real series: the lag-1 autocorrelation of each source's support indicator minus the mean indicator of all sources, averaged over sources. It is 0.18 in Guayaquil (stationary-bootstrap 95 % interval 0.13–0.20, 30-day blocks) and 0.04 in Santiago (-0.03–0.08) at $\delta = 0.5$, 0.10–0.29 in the designs with $\phi_e \leq 0.5$ and 0.26–0.61 with $\phi_e \geq 0.8$. Santiago's value lies below every simulated design. Table 5 gives the results.

Four results matter for the application. First, in the 4 null designs per window length whose diagnostic lies inside the 95 % interval of the Guayaquil value, BM-10's largest false-conflict rates were 0.069, 0.053 and 0.053 at $W = 30, 90, 365$, against the limit $0.05 + 2 \times 0.0069 = 0.0638$ (Monte Carlo SE about 0.007): it kept its level at 90 and 365 days and not at 30 days. Over all designs with $\phi_e \leq 0.5$ (diagnostic up to 0.29) the largest rates were 0.064 at 90 days, a borderline failure in a design with diagnostic 0.29, and 0.058 at 365 days; the smallest diagnostic of a design in which BM-10 failed was 0.26 at 90 days and 0.30 at 365 days, but the diagnostic does not order failure exactly. Second, the reading rule as first frozen allowed a conditional claim only if BM-10 failed solely in designs with $\phi_e = 0.95$. It also failed with $\phi_e = 0.8$ (up to 0.164 at 90 days and 0.066 at 365 days), so under that rule no window length would admit a conflict claim. Amendment A1, made after the smoke test had shown failures at $\phi_e \geq 0.8$ (simulated output only, before any v0.6 real-data run, but after the v0.5 real-data results were known), replaced it with the rule that the designs whose diagnostic brackets the data decide. The rain results below rest on that amended rule and are conditional and partly post hoc; the support they receive from simulation is limited to the few designs per window length whose diagnostic falls in an interval estimated from the same data. Third, with seasonality in both the mean and the variance of the paired series, BM-10 stayed far below the level (largest rate 0.021). These designs did not reproduce the data's heterogeneity, however: in Guayaquil 40 % of the (pair, 365-day window) combinations have a batch in which d_t is constant (a dry spell), against 0–6 % in these designs. A further check, not preregistered, with a dry season (identical models, or models 1 and 2 with opposite seasonal offsets in

Table 5. Own-persistence and seasonal-heterogeneity designs (preregistered, v0.6): largest familywise false-conflict rate over the designs in each row (limit 0.0638), $K = 5$, $\alpha = 0.05$, 1000 replicates per design. Lag-1 diag. = range of the persistence diagnostic (data: 0.18 Guayaquil, 0.04 Santiago). V04 = seasonally adjusted bootstrap rule; CP = nominal Clopper–Pearson; BM-b = paired batch-means test with b equal batches and Holm; BM-10s = seasonally centred variant; cov. V04 = smallest simultaneous coverage of the V04 limits. All designs per row and per window length are in Table S8.

Design		W	lag-1 diag.	V04	CP	BM-10	BM-5	BM-20	BM-10s	cov. V04
own persistence	$\phi_e \le 0.5$	30	0.10–0.29	0.088	0.022	0.112	0.087	0.277	0.115	0.59
		90	0.10–0.29	0.027	0.031	0.064	0.062	0.106	0.061	0.74
		365	0.10–0.29	0.012	0.025	0.058	0.055	0.053	0.077	0.86
own persistence	$\phi_e \ge 0.8$	30	0.26–0.61	0.869	0.783	0.820	0.524	0.923	0.826	0.02
		90	0.26–0.61	0.672	0.873	0.646	0.245	0.884	0.665	0.10
		365	0.26–0.61	0.325	0.915	0.218	0.078	0.437	0.296	0.35
own persistence, $\delta = 0.5$	$\phi_e = 0.95$	30	0.61–0.61	0.500	0.453	0.464	0.201	0.664	0.479	0.08
		90	0.61–0.61	0.210	0.411	0.186	0.058	0.419	0.204	0.21
		365	0.61–0.61	0.007	0.126	0.005	0.002	0.019	0.008	0.46
seasonal mean and variance	$\phi_e = 0.5$	30	0.24–0.29	0.006	0.002	0.014	0.020	0.046	0.013	0.45
		90	0.23–0.29	0.001	0.000	0.021	0.024	0.028	0.020	0.65
		365	0.23–0.29	0.000	0.001	0.016	0.019	0.016	0.048	0.82

mean and variance; diagnostic 0.10–0.20), produced such batches in at least 98 % of the combinations, and BM-10's largest false-conflict rate was 0.015 (`code/sim_v06_extra.py`). Fourth, under strong own persistence ($\phi_e \ge 0.8$) every rule failed: the largest rates were 0.820, 0.646 and 0.218 for BM-10 and 0.869, 0.672 and 0.325 for the bootstrap rule at $W = 30, 90, 365$. When each source's errors persist for weeks, a window's frequencies of identical sources differ by more than any within-window variance estimate shows, and no rule built on the window alone can tell this from conflict. The bootstrap rule's false-conflict rates were small with weak own persistence (at most 0.027 and 0.012 at 90 and 365 days) even though its simultaneous coverage was only 0.739–0.888 and 0.860–0.919; its control in these designs is empirical and not explained by Proposition 4. The seasonally centred variant BM-10s exceeded the limit at 365 days (0.077) and is not used. BM-5 was the most robust version (largest rate 0.078 at 365 days under strong own persistence) but flags about half as many Guayaquil windows as BM-10 on the same days (26 % against 53 % at 90 days, 32 % against 76 % at 365 days; Section 6.1). With one biased source ($b_1 = 0.5$, $\phi_e = 0.5$) BM-10 had power 0.11–0.46 and 0.41–0.98 at 90 and 365 days, against 0.07–0.12 and 0.25–0.73 for the bootstrap rule.

Number of batches on common days (not preregistered, v0.6.1). In the v0.6 construction the versions with $b = 5$, 10 and 20 dropped the earliest $n \bmod b$ days of a window, so they analysed slightly different periods (at 365 days $b = 5$ used all 365 days and $b = 10, 20$ the last 360; at 90 days $b = 20$ used the last 80). We therefore re-ran the batch-means rules on a common set of days: the days kept by the primary rule ($b = 10$: 30, 90 and 360 days for $W = 30, 90, 365$), with every $b \in \{5, 10, 15, 20\}$ that divides that length ($b = 20$ only at 365 days; `paired_conflict.common_days`). In the 63 null designs of Table 5 (1000 new replicates each, seed 20261023, `code/sim_v061_common.py`; Table S9), among the 4 designs per window length whose diagnostic lies in the Guayaquil interval the largest false-conflict rates were 0.053, 0.059 and 0.060 for $b = 5, 10, 15$ at 90 days and 0.058, 0.054, 0.048 and 0.051 for $b = 5, 10, 15, 20$ at 365 days (limit 0.0638), and 0.080, 0.068 and 0.079 at 30 days, above the limit. Under strong own persistence every version failed, more so with more batches (at 365 days, from 0.096 with $b = 5$ to 0.460 with $b = 20$). This support rests on four designs per window length.

6. Applications

The three applications are *methodological demonstrations*, not empirical validations of the method: they show what the labelled I reports on real data and what it changes, and the provenance and inferential scope of each result are

summarised in Table 13. In each application the components are *illustrated* on data whose generating process is unknown; the exact derivation is that of Proposition 1.

6.1. Rain forecasts from five models

Data and provenance. Daily precipitation for 1 January 2023 to 31 December 2025 was downloaded on 10 October 2026 from the Open-Meteo Historical Forecast API [48, 77] for Guayaquil (requested point -2.19, -79.89) and Santiago de Chile (-33.45, -70.66), variable `precipitation_sum` (mm), `timezone=auto` (America/Guayaquil; America/Santiago, with daylight saving time), so that a day is a local calendar day as aggregated by Open-Meteo. The model identifiers are `gem_seamless`, `gfs_seamless`, `icon_seamless`, `jma_seamless` and `ukmo_seamless`, the five that met the preregistered 90 % completeness rule (four others were dropped); each response is answered at the model's nearest grid point, whose coordinates differ by model (Table S2). The Historical Forecast API builds a continuous series by concatenating the first hours of successive model runs [77]. The series are therefore short-range forecasts concatenated across runs, not forecasts at a fixed lead time. The Previous Runs API provides fixed-lead forecasts, but a common record for all five selected models does not cover the full 2023–2025 period. The series also concatenate the operational model versions in use at each date; we did not adjust for version changes during 2023–2025, which Open-Meteo itself warns about. ERA5 [49] came from the Open-Meteo archive API with the same settings. We use the 1094 days on which all five models and ERA5 are present. The stored JSON responses, their SHA-256 hashes and the reconstructed request URLs are in the code archive (`code/openmeteo_metadata_v06.py`).

Estimand. The proposition is "precipitation ≥ 1 mm on day t". A model supports it when its forecast is at least $1+\delta$ mm, refutes it below $1-\delta$ mm and is undecided in between; we report $\delta \in \{0, 0.1, 0.25, 0.5, 0.75\}$. For a window of W days ($W = 30, 90, 365$, ends every 30 days) the target of model k is the expected window frequency of each verdict; its identified set for "rain" is $[p_{a,k}, 1-p_{r,k}]$. Windows overlap, so the shares below are descriptive summaries over windows, not independent replicates.

Test used for the conflict claims. Conflict claims are made with the paired test of Proposition 5: for each of the 20 ordered pairs, the one-sided batch-means t test of $D_{jk} \leq 0$ with $b = 10$ equal batches of the raw daily series d_t (the five earliest days of a 365-day window are dropped, so the target is the frequency over its last 360 days; 30- and 90-day windows are used whole), $p = 1$ when the batch means are all equal, and Holm across the pairs. This rule was preregistered as an amendment in v0.5, where the preregistered decision rule had kept the bootstrap rule as primary because the paired rule did not dominate it in power; we use it here because the bootstrap limits fail simultaneous coverage, so Proposition 4 does not cover them. This change of primary rule is a deviation from the v0.5 decision rule, made after seeing the coverage results, and is declared as such.

What the theorem guarantees and what these data allow. Proposition 5(d)–(e) guarantees the false-conflict rate only when the batch means are independent and normal with a common variance or with a common mean. For these data neither assumption can be verified: the batch means come from serially dependent daily series, and the batches differ in both mean and variance, the case of Remark 1 in which no bound holds. The level of the test on these data therefore rests on simulation alone, and that support is limited: the test kept its level in the 4 simulated designs per window length whose persistence diagnostic lies in the 95 % interval estimated from the same data (Guayaquil 0.18, interval 0.13–0.20), not at 30 days, and not under strong own persistence ($\phi_e \geq 0.8$), where every rule failed (Section 5). Whether the models' own errors are in fact as weakly persistent as the diagnostic indicates cannot be checked with three years of data. We therefore do not call the rain results certified conflict with a guaranteed false-conflict rate. We report the windows *flagged* by the paired batch-means test, at 90 and 365 days only (amended reading rule of `PREREG_v06.md`), as illustrative evidence conditional on the persistence diagnostic.

Descriptive rule. The seasonally adjusted stationary block bootstrap of the earlier versions is reported as descriptive. In a window of n days let $Y_t \in \{0,1\}^{2K}$ collect the support and refutation indicators, $\hat{\theta}$ their window means, $\hat{C}_t$ the two-harmonic seasonal mean of Y_t fitted on the three years and clipped to $[0,1]$, and $Z_t = Y_t - \hat{C}_t$; the stationary bootstrap [50, 51, 52] resamples the same days for all models [53], $\theta_b^* = \hat{\theta} + n^{-1}\sum_t (Z_{i_b(t)} - \bar{Z})$, the block length is the largest Politis–White value [54, 55] over the $2K$ residual series, $\hat{c}$ is the $1-\alpha$ quantile of $\max_{j:s_j>0}(\theta_{b,j}^* - \hat{\theta}_j)/s_j$, and the limits are $T_k = \max(0, \hat{p}_{a,k} - \hat{c}\, s_{a,k})$, $F_k = \max(0, \hat{p}_{r,k} - \hat{c}\, s_{r,k})$ ($B = 2000$; a constant coordinate gets the limit $\hat{\theta}_j$). Its sampling components I_{samp} and its shares of disjoint limits are given for comparison; as shown in Section 5, its limits undercover and its sampling components may be underestimated. We did not use a seasonal block bootstrap that resamples corresponding calendar periods across years [74]; with only three annual cycles available, the coverage of that alternative was not evaluated or established in this study.

Results. Table 6 and Figure 2 give the results at $\delta = 0.5$. In Guayaquil the measurement component is about 0.095 at every W. The paired test flags between-model conflict in 53 % of the 90-day and 76 % of the 365-day windows (30 days: 22 %, no claim). The share falls with the tolerance, from 29 % at $\delta = 0.75$ to 85 % at $\delta = 0$ at $W = 90$ and from 64 % at $\delta = 0.75$ to 96 % at $\delta = 0$ at $W = 365$ (Table 7), and it depends strongly on the number of batches. On the same days (Table 8), $b = 5$ and $b = 15$ give 26 % and 71 % at 90 days, and $b = 5$, 15 and 20 give 32 %, 92 % and 100 % at 365 days, against 53 % and 76 % with $b = 10$: the share rises with b, over 32 %–100 % at 365 days (28 %–100 % in v0.6, where $b = 5$ also used the five earliest days). All four versions kept their level in the simulated designs matched to the diagnostic (Section 5), so the simulation does not say which b to prefer; $b = 10$ was fixed in advance. In Santiago, $b = 5$ flags 2 of the 25 365-day windows (8 %; 4 % in v0.6, on 365 days) and the other b none. With a margin $m = 0.05$ (the identified sets must be separated by five percentage points) the shares are 26 % and 48 %. The pairs flagged at 365 days are mostly JMA against the others: JMA > GFS (16); JMA > ICON (13); JMA > GEM (5) of 25 windows. JMA forecasts the most rainy days in every 365-day window. The data are in the both-heterogeneities case of Remark 1: in 40 % of the (pair, 365-day window) combinations some batch has a constant d_t, and where all batches vary, the largest within-batch standard deviation is a median 3.6 times the smallest; the dry-season designs of Section 5, which reproduce this feature, kept BM-10's level. No pair had all batch means equal with a positive estimate (zero-variance pairs occurred in 300 of 950 window–tolerance cells, always with $\hat{D}_{jk} \le 0$). The descriptive bootstrap rule gives similar shares (59 % and 72 % at 90 and 365 days) and the v0.5 batch construction gives 80 % at 365 days.

Table 6. Rain-day frequency, five models, $\delta = 0.5$ mm: labelled components (means over windows) and share of windows with between-model conflict. Flagged = paired batch-means test ($b = 10$, Holm over 20 ordered pairs), conditional and illustrative: its level for these data is supported only in simulated designs matched to a persistence diagnostic estimated from the same data (Section 6.1); ‡ 30-day windows: the test did not keep its level in simulation, no claim. Descriptive: SB-seas = seasonally adjusted block bootstrap, whose limits undercover (its I_{samp} may be underestimated); CP = nominal Clopper–Pearson (assumes independent days). Plug-in disjoint = estimated identified sets disjoint (no error control). 36, 34 and 25 windows for $W = 30, 90, 365$.

			I_{samp} (descriptive)		flagged	descriptive		plug-in
Location	W	I_{meas}	CP	SB-seas	paired BM-10, cond.	SB-seas	CP	disjoint
Guayaquil	30	0.095	0.306	0.251	0.22‡	0.39	0.08	1.00
Guayaquil	90	0.094	0.194	0.210	0.53	0.59	0.59	1.00
Guayaquil	365	0.097	0.130	0.177	0.76	0.72	0.88	1.00
Santiago	30	0.030	0.266	0.178	0.00‡	0.00	0.00	0.67
Santiago	90	0.031	0.152	0.143	0.00	0.00	0.00	0.82
Santiago	365	0.029	0.079	0.088	0.00	0.00	0.00	1.00

Santiago. At $\delta = 0.5$ the paired test ($b = 10$) flags no Santiago window at 90 or 365 days, and few at small tolerances (6 % and 20 % of the 90- and 365-day windows at $\delta = 0$). The models do not agree exactly: the estimated identified sets are disjoint in 67 %, 82 %, 100 % of the 30-, 90- and 365-day windows, but the largest estimated

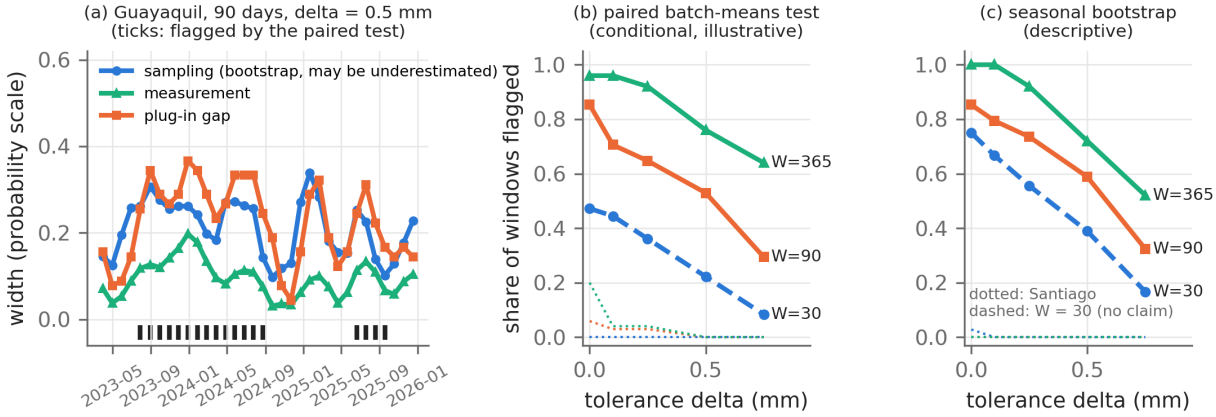


Figure 2. (a) Guayaquil, 90-day windows, $\delta = 0.5$ mm: sampling component (seasonally adjusted bootstrap; may be underestimated), measurement component and plug-in gap; ticks mark windows flagged by the paired test. (b) Share of windows flagged by the paired batch-means test (conditional, illustrative) as a function of the tolerance δ; (c) the same for the descriptive bootstrap rule. Solid and dashed lines Guayaquil (dashed: 30-day windows, no claim), dotted lines Santiago.

Table 7. Share of windows flagged by the paired batch-means test ($b = 10$; conditional, Section 6.1) as a function of the tolerance δ; in parentheses, the descriptive bootstrap rule.

Location	W	$\delta = 0$	0.1	0.25	0.5	0.75
Guayaquil	90	0.85 (0.85)	0.71 (0.79)	0.65 (0.74)	0.53 (0.59)	0.29 (0.32)
Guayaquil	365	0.96 (1.00)	0.96 (1.00)	0.92 (0.92)	0.76 (0.72)	0.64 (0.52)
Santiago	90	0.06 (0.00)	0.03 (0.00)	0.03 (0.00)	0.00 (0.00)	0.00 (0.00)
Santiago	365	0.20 (0.00)	0.04 (0.00)	0.04 (0.00)	0.00 (0.00)	0.00 (0.00)

Table 8. Sensitivity to the number of batches b, all on the same days (those kept by the primary rule $b = 10$; b only among the divisors of that length), $\delta = 0.5$: share of windows flagged by the paired batch-means test (Holm over 20 ordered pairs; conditional, Section 6.1). ‡ 30-day windows: no claim. All tolerances: Table S10; level of each version in simulation: Table S9.

Location	W	days kept	$b = 5$	$b = 10$ (primary)	$b = 15$	$b = 20$
Guayaquil	30‡	30	0.19	0.22	0.25	–
	90	90	0.26	0.53	0.71	–
	365	360	0.32	0.76	0.92	1.00
Santiago	30‡	30	0.00	0.00	0.00	–
	90	90	0.00	0.00	0.00	–
	365	360	0.08	0.00	0.00	0.00

separation $\max_{jk} \hat{D}_{jk}$ averages 0.024 and 0.014 at 90 and 365 days (at 365 days, about five rain days a year), and the batch-means test that two models' midpoints differ, whatever their side, succeeds in 0 % and 12 % of the windows. The Santiago models differ by small amounts that neither rule resolves as opposite verdicts. This absence of flagged conflict is inconclusive as evidence of agreement: power against substantively relevant separations was not estimated for Santiago, and its persistence diagnostic lies below every simulated design; the small observed separations are descriptive and do not substitute for a power assessment.

Comparison with ERA5 (descriptive). A fixed-threshold test (Proposition 5(b)) needs c fixed in advance. The ERA5 rain-day frequency of the three years ($c = 0.635$ in Guayaquil, 0.111 in Santiago) is estimated from the same days and ERA5 is itself a model, so we use it only descriptively and draw no test conclusion from it; an externally

fixed threshold, for example a climatology from an earlier period, would be needed for a confirmatory version. Descriptively, most models lie below the ERA5 frequency in both cities, and in Guayaquil the flagged conflict is between models that lie below it by different amounts. On single days, glut days (some model forecasts rain and another a dry day, $\delta = 0.5$) were 37.7 % of the days in Guayaquil and 7.4 % in Santiago, and the majority verdict disagreed with ERA5 on 32.6 % of decided glut days against 22.3 % of unanimous days in Guayaquil (difference 0.103, stationary-bootstrap 95 % interval [0.036, 0.162]) and on 30.1 % against 2.2 % in Santiago (interval [0.180, 0.376]). This is consistent with the spread–skill relation of ensemble forecasting [56], but it concerns concatenated short-range series checked against a reanalysis, not forecast skill at a stated lead time against observed rain.

6.2. Rock joint roughness: dispersion and heterogeneity

Chen, Ye and Du [4] report, for 24 orientations and 10 sample lengths of a rock joint, the mean μ and standard deviation σ of the joint roughness coefficient (JRC) over profiles, and write each cell as $z = (\mu - \sigma) + 2\sigma I$. We transcribed their Table 1 and reproduced their neutrosophic average for $L = 10$ cm ($6.9427 + 4.3055\, I$). Their 2σ describes the variability of the JRC across profiles, the engineering quantity they study. Read as uncertainty about the cell's mean JRC, the computed sampling component is $2t\sigma/\sqrt{n}$; the number of profiles n per cell is not reported, so all results below are conditional on an assumed n. The computed width is 1.24 times the hand-chosen one at $n = 5$, 0.72 at $n = 10$ and 0.28 at $n = 50$.

Treat the 24 orientations as sources for the JRC at length L. They are different populations by design (anisotropy is the object of [4]), so a certified gap between them is heterogeneity beyond dispersion, not contradictory evidence about one quantity. The hand intervals $[\mu - \sigma, \mu + \sigma]$ all share a point at $L = 10$ cm; the orientation means would be separated by the Bonferroni rule of Definition 3 once n reaches 15 profiles per cell at 10 cm, 9 to 12 at 20–90 cm and 7 at 100 cm, if those were the sample sizes (Figure 3; conditional on the assumed n and on independent profiles). A one-way analysis of variance from the published summaries detects that the orientations differ with 2 to 5 profiles per cell; the difference between the two thresholds reflects the conservativeness of the Bonferroni separation rule over 24 groups. A direction-free neutrosophic number for a given length thus mixes anisotropy with dispersion.

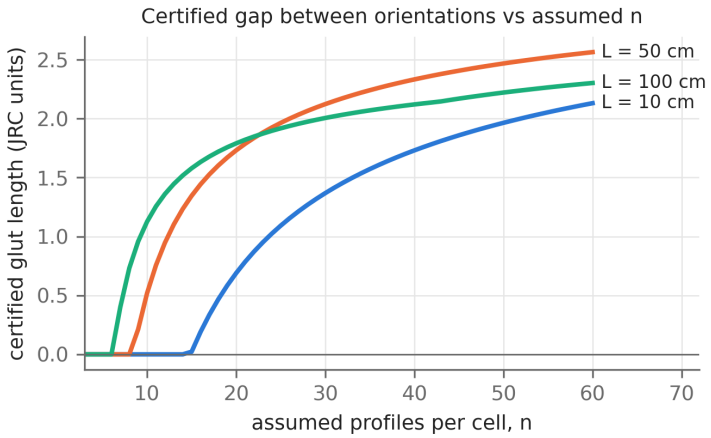


Figure 3. Rock joint roughness [4]: nominal separation gap between the 24 orientations as a function of the assumed number of profiles per cell, for three sample lengths, under independent normal profile measurements; a conditional sensitivity calculation.

6.3. Raters with undecided answers: DICES-990

The DICES-990 dataset [57] has 72 103 ratings of 990 conversations by 172 raters from India and the United States (29 to 45 raters per item and locale, median 36), each answering whether a conversation is unsafe with Yes, No

or Unsure. The proposition is "most raters of a pool would rate this conversation unsafe"; an Unsure answer is an undecided record and the two locales are the sources.

Definition 8 (Majority-verdict conflict)
Two pools are in majority-verdict conflict when one pool's confidence set for the proportion of Yes lies above $1/2$ and the other pool's set lies below $1/2$ (each at level α/K).

Pooling the raters, we used the Clopper–Pearson construction under a working i.i.d. categorical model for the item's ratings. The majority verdict was decided at the stated nominal level for 692 conversations (56 unsafe and 636 safe) and undecided for 298. Among the undecided conversations, the estimated identified interval lay strictly on one side of $1/2$ for 229, straddled $1/2$ for 60, and touched the boundary for 9. These are descriptive classifications of estimated intervals; their implications for follow-up are discussed in Section 6.4. The mean sampling component per pool was 0.305 with all raters and 0.40 and 0.55 with 20 and 10 raters per pool, while the Unsure share was 0.052.

No conversation showed majority-verdict conflict between the locales, nor between men and women, nor for the harmful-content question. Under the weaker Definition 3 the confidence sets of the two locales are separated for 4 conversations (largest gap 0.027); with 990 items and no multiplicity correction this does not by itself establish locale disagreements. The absence of majority-verdict conflict is inconclusive: with 36 raters per pool and no Unsure answers (the best case), the exact power of interval separation is 0.04, 0.19 and 0.50 for majority proportions 0.40 vs 0.60, 0.35 vs 0.65 and 0.30 vs 0.70, and that of Definition 8 is 0.01, 0.055 and 0.22; Definition 8 reaches power 0.8 against 0.30 vs 0.70 only with 72 raters per pool, and against 0.35 vs 0.65 with 133. The design has low power even for differences of 0.4 in proportion. Here majority-verdict conflict denotes opposite majority judgments in locale-specific rater populations; it does not imply contradictory estimates of a single common population parameter.

6.4. Decision impact: what the labels change

The labelled split changes an applied conclusion when the same width calls for different actions depending on its source. The three applications give one case each; all three use only results already reported above.

DICES: more raters or a closer look at "unsure". With one interval per conversation (the hull of its identified set and sampling error), all 298 conversations without a decided majority verdict receive the same report, "undecided". Among them, the estimated identified interval lies strictly on one side of $1/2$ for 229, straddles $1/2$ for 60, and touches the boundary for 9. These descriptive categories suggest different follow-up priorities: additional ratings for the first group, and examination of the meaning and coding of "unsure" for the second. Measurement imprecision need not be zero in the first group. Because the classification uses estimated rather than known population identified sets, it does not establish that additional ratings will resolve the first group or cannot help the second. The 60 cases are 20.1 % of the undecided items; this is a descriptive share, not an estimate of wasted annotation effort or achievable savings.

Rain: averaging or investigating. A forecaster who averages the five models obtains one rain-day frequency per window. In Guayaquil the flagged conflicts at 365 days are mostly JMA against the other models, and JMA forecasts the most rain days in every 365-day window. The labelled report therefore says not to average without first examining why JMA forecasts more rain days than the others; an average would place the combined forecast between incompatible verdicts. Because these flags are conditional (Section 6.1), the action is "investigate before combining", not "discard JMA".

Rock roughness: dispersion or uncertainty. The hand-chosen width 2σ of [4] is the variability of the roughness across profiles, a property of the rock. Read as uncertainty about the mean roughness of a cell, the computed sampling width is 1.24, 0.72 and 0.28 times the hand-chosen one at 5, 10 and 50 profiles per cell. A design decision that needs the mean (for example, a shear-strength estimate from the mean JRC) should use the computed

sampling width, which shrinks with more profiles; one that needs the spread of the surface should use 2σ, which does not. Pooling the orientations into one direction-free number mixes the anisotropy with both.

6.5. Comparison with implementations of fuzzy and symbolic statistics

This comparison was not preregistered. We implemented ourselves, from the cited formulas, the standard analysis of each framework, and the codings of undecided records are our choices; the results describe these implementations, not every analysis possible within each framework. We chose, for each framework, a standard illustrative analysis relevant to the question at hand (the fuzzy confidence interval and fuzzy mean of [25], the functional test of equal fuzzy means of [60, 61], the descriptive statistics and centre-and-range regression of symbolic data analysis), because these are common first choices in each framework; they are illustrative implementations, not definitive benchmarks, and stronger variants exist.

Implementations. (F) *Fuzzy statistics with fuzzy data* [25]: each record is a fuzzy number, the estimate is the fuzzy mean, and the cut of the fuzzy confidence interval at level λ is the union of the classical intervals over the crisp samples in the λ-cut of the fuzzy sample; a rain forecast in the tolerance band becomes the triangular number $(0, m, 1)$, $m = (x - 0.5)/1.0$, a DICES Unsure becomes $(0, 0.5, 1)$; at the 0-cut the interval is an exact Clopper–Pearson interval, and at $\lambda > 0$ the fractional counts make it a numerical extension without a coverage guarantee. Disagreement is read as $c = 1 - \mathrm{hgt}(\bigcap_k \tilde{C}_k)$ [58]. (G) *Fuzzy random variables* [24] as functional data: equality of fuzzy means tested with the D_θ metric ($\theta = 1/3$) [59] by bootstrap [60, 61]; for rain, days resampled jointly, independently or in stationary blocks; for DICES, a permutation test between the two pools of each item with 99 999 random permutations and $p = (1 + b)/(B + 1)$ [62]. (S) *Symbolic data analysis* [26, 27]: a day is the interval of the five forecasts, a JRC cell is $[\mu - \sigma, \mu + \sigma]$; descriptive statistics follow [65] and regression the centre-and-range method [66, 67]. (IDM) Walley's model with $s = 2$, with disjoint intervals read as conflict (not a test).

Table 9. What our implementations of the rival analyses report on the same data, and what the labelled I reports.

Analysis	Observed on our data
Labelled I (this paper)	measurement and sampling parts of the width, additive and labelled; a plug-in spread; a conflict statement outside I with a stated error rate and power. Does not test whether sources *differ*, model interval data, or grade disagreement
Fuzzy CI [25]	support of the fuzzy estimate has width I_{meas}; the 0-cut of the fuzzy CI at level α/K has width $I_{\mathrm{meas}} + I_{\mathrm{samp}}$ (every window); the parts are one fuzzy shape, not labelled; $c = 1$ at the 0-cut coincides with nominal certified conflict by construction; graded c has no error control
Functional fuzzy ANOVA [60, 61]	tests whether sources differ; much more power than certified conflict; empirical size 0.042–0.060 in our null designs
Symbolic data analysis [27]	describes and regresses interval data; no sampling control in its descriptive core
IDM [11]	additive split $u/(n + s)$, $s/(n + s)$; disjointness is not a test

Rain. Two identities hold in every window up to rounding (largest difference 1e-16): the support of Viertl's fuzzy estimate has width I_{meas}, and the 0-cut of the fuzzy confidence interval at level α/K has width $I_{\mathrm{meas}} + I_{\mathrm{samp}}$ of the nominal Clopper–Pearson triple. The graded $c > 0$ flags 79 % of the Guayaquil 90-day windows without error control. The functional analysis of variance rejects equal means in 91 % (Guayaquil) and 35 % (Santiago) of the 90-day windows with block resampling, and in 100 % of the Santiago 365-day windows. In a known-truth simulation (2000 replicates per design) with identical sources the largest rates over 12 null designs were 0.000 for certified conflict, 0.003 for $c > 0$ and 0.547 for IDM disjointness; with sources that differ but have overlapping identified sets, the ANOVA rejected in 29.8 % of replicates (correctly, for its question) and certified conflict in 0.0 %; with contradictory sources (gap 0.2), the ANOVA rejected in 99.9 % and certified conflict in 31.2 %.

Rock roughness and DICES. Symbolic data analysis decomposes the dispersion across orientations that the hand-chosen 2σ hides: the within-interval share of the symbolic variance falls from 0.50 at 10 cm to 0.26 at 100 cm, and its centre-and-range regression over the 240 cells gives $\mu = 11.26 - 0.94 \ln L$ ($R^2 = 0.28$) and $2\sigma = 6.38 - 0.97 \ln L$ ($R^2 = 0.70$), a model of the scale effect that the labelled analysis does not attempt. In DICES, the fuzzy permutation test gives $p < 0.05$ for 219 conversations against 49.5 expected if no conversation differed. The main multiplicity correction is Holm's [64], valid under any dependence between the per-item p-values (each of which is a valid permutation p-value for its own item): 1 conversation survives it (adjusted $p = 0.0297$). This rests on 2 exceedances in 99 999 random permutations: the upper 95 % Clopper–Pearson limit for that item's exact permutation p-value is 7.2e-05, whose Holm adjustment would be 0.072, so the single family-wise rejection is fragile to Monte Carlo error. Benjamini–Hochberg [63] retains 4, but its false-discovery-rate guarantee needs independence or positive regression dependence, which is doubtful here because the same raters rate many items; we report it as exploratory. Benjamini–Yekutieli [75], valid under any dependence, retains 0 (smallest adjusted p 0.167). Table 10 lists the four conversations selected by the exploratory BH procedure. In all four the US pool rates the conversation mostly safe while the Indian pool is split or leans unsafe; none has opposite certified majorities. The fuzzy test thus detects at most one difference between locales with family-wise control, and that one is fragile; it shows no contradiction between majority verdicts.

Table 10. DICES-990, overall safety rating: the four conversations with BH-adjusted $p \leq 0.05$ in the functional fuzzy permutation test between rater locales. n = raters per pool; raw $p = (1 + \#\{\text{permutation statistics} \geq \text{observed}\})/(1 + 99\,999)$; Holm (main conclusion), Benjamini–Hochberg (exploratory) and Benjamini–Yekutieli adjusted over 990 items (`results/dices_multiplicity_v06.csv`).

Item ID	n_{IN}	n_{US}	raw p	Holm (main)	BH (exploratory)	BY
332	32	40	3×10^{-5}	0.030	0.022	0.167
970	38	37	6×10^{-5}	0.059	0.022	0.167
602	39	34	8×10^{-5}	0.079	0.022	0.167
624	33	41	9×10^{-5}	0.089	0.022	0.167

Where the rivals do better. The functional ANOVA is the right tool, with much higher power, when the question is whether sources differ. Symbolic regression models interval data: adding the range of the five log-forecasts to their centre raises R^2 for ERA5 log-precipitation from 0.471 to 0.484 in Guayaquil ($p = 1 \times 10^{-7}$; serial dependence makes this p optimistic) but not in Santiago. Viertl's fuzzy confidence interval holds the same information as our two count components plus a graded shape, and comes with fuzzy tests [68]. What the labelled I adds is the explicit, additive split with labels and provenance, a separate conflict statement with a stated error rate, and a rule for what to do with each part (Table 14).

Trade-offs: when to choose which. *Data.* The labelled I needs records that can be classified as support, refutation or undecided, or interval records, and several sources for a conflict statement; fuzzy statistics needs a membership function for each record, which is an additional modelling choice (our triangular codings), and symbolic analysis needs interval or distributional records. *Sample size.* A conflict statement needs more data than a test of difference: on the same simulated data the functional ANOVA detected differences that certified conflict did not, and for a gap of 0.2 certified conflict had power 0.32–0.40 with 100 records per source and above 0.99 with 500 (Sections 3 and 5). *Computation.* The count-based summaries and paired-window calculations had low runtimes in the reported benchmarks (Section 8); the estimated DICES permutation-test total was about four minutes. Complete simulation runs were not benchmarked. *Expertise.* The labelled report uses only standard confidence bounds and tests, but it requires the analyst to declare the coding, the tolerance and the sources. We would choose the functional ANOVA when the question is whether sources differ, symbolic regression when the object is to model interval-valued data, Viertl's fuzzy interval when a graded summary is wanted, and the labelled I when the question is which components of uncertainty additional data may reduce and whether sources issue opposite verdicts.

7. A reporting protocol

The steps below turn the results into a workflow; Table 11 is the corresponding checklist of what a report must declare.

1. *State the proposition and the sources.* Fix the proposition A (with its threshold) and decide what a source is. Sources should be alternative measurements of the same quantity; if they are different populations by design (orientations of a rock joint, locales of raters), a separation between them is heterogeneity, not contradictory evidence (Section 6.2).
2. *Code the records.* Fix the tolerance band that maps a record to support, refutation or undecided, and the coding of "unsure" answers. If the choice is not dictated by the measurement process, report the results over a grid of tolerances (as Table 7 does).
3. *Choose the confidence level and the dependence model.* Use Clopper–Pearson bounds (Definition 1) under an i.i.d. categorical sampling model, or outer intervals (Definition 4), at level α/K for K sources. For serially dependent records, use a block method and report it as descriptive unless its coverage has been checked for data of that kind. Choose α for the decision at stake, not by convention; when many items are analysed, state whether the error rate is controlled per item, family-wise (Bonferroni or Holm over items) or as a false discovery rate, as in the DICES analyses (Sections 6.3 and 6.5). The tolerance band should come from the measurement process (for example, the resolution of the instrument or a minimal relevant difference); if it does not, report a grid.
4. *Compute and report the split.* Report I_{meas} and I_{samp} (or W_{imp}, W_{samp} and the plug-in spread) next to N. Choose the follow-up action by comparing the identified set and its uncertainty with the decision threshold, rather than by selecting the largest width component alone. If the estimated identified set lies on one side of the threshold but the outer confidence interval crosses it, additional records may resolve the decision; if the population identified set straddles the threshold, reducing sampling uncertainty alone cannot do so. Because that population set is unknown, report the uncertainty of this classification and the assumptions used for sample-size planning.
5. *Test for conflict, and report its conditions and power.* Use interval separation (Proposition 4) when valid simultaneous coverage is available; independence between sources is not required. For sources observed on the same units, the paired batch-means test with Holm (Proposition 5) may improve power by using their covariance; apply it as a guaranteed test only when its batch assumptions are justified. Check its assumptions; if they cannot be verified, say "flagged", not "certified", and give the simulation support. Report the power against a relevant alternative before reading the absence of conflict as agreement.
6. *Read hesitant sets.* Determine whether the alternatives describe one holder's uncertainty or separate holders' assertions; use the disjunctive reading in the first case and the conjunctive reading in the second. Under the conjunctive reading, report conflict about the chosen proposition only when some holders support it and others refute it (Proposition 6).
7. *Propagate.* When the result enters further computation, use the labelled arithmetic (Section 4), report the refined range as the result, treat the typed widths of nonlinear results as conventions (Remark 2), and recompute conflict source by source.
8. *Act.* Use Table 14 to turn each component into an action.

In decision-tree form: if a conflict is certified (or flagged), do not combine the sources before explaining the disagreement; otherwise, if the power to detect a relevant conflict is low, report the absence as inconclusive; then, if the estimated identified set lies on one side of the decision threshold while the confidence set crosses it, collect more records, and if the identified set straddles the threshold, examine the measurement or the question.

8. Discussion

Does a computed I answer the critique? Partly. The measurement and sampling components are interval and imprecise-probability statistics; computing them answers the charge that I is fixed by hand, once the

Table 11. Checklist for a labelled report.

Item	What to report
Proposition and threshold	A, its threshold and why
Sources	what a source is; whether the sources measure the same quantity
Coding	tolerance band; coding of undecided answers; sensitivity grid if not fixed by the measurement
Confidence level	α and the multiplicity adjustment across sources
Dependence model	sampling model (e.g. i.i.d. categorical records), or the block method and whether its coverage was checked
Components	I_{meas} and I_{samp} (or W_{imp}, W_{samp}, plug-in spread) with N
Conflict	test used; its assumptions and whether they were verified ("certified" or "flagged"); disagreeing sources; power against a stated alternative
Hesitant sets	reading (one holder or several)
Propagated results	refined range; typed widths as conventions for nonlinear operations
Status	provenance (frozen, amended, exploratory) and inferential scope (proved, conditional, descriptive) (Table 13)
Code	where the code and the data are

analyst's choices are declared, and it answers [9] in the sense that a computed I shrinks with data in a way that carries information. It does not give neutrosophic statistics inferential content beyond interval statistics for these components; with imprecision alone a credal set suffices [40]. Certified conflict is also classical in its inferential content: a test of separation, here a paired batch-means test with Holm's procedure. The contribution is a representation and a reporting rule: the source-labelled report retains the identities of disagreeing sources alongside the triple and accompanies each inferential conflict statement with its test, assumptions, error control and power. A fused triple alone does not identify the disagreeing sources. Dempster–Shafer theory represents the same situation through the mass of the empty set [28]; neutrosophic triples represent it directly. For hesitant sets, the disjunctive reading represents one holder's uncertainty, which never carries more indeterminacy than its hull; under the conjunctive reading, conflict about a chosen proposition occurs only when some holders support it and others refute it.

Which neutrosophic form separates imprecision from contradiction. The limitation pointed out by the critics belongs to the form $N = d + I$ with a fixed I: combining sources into one such number takes the hull of their sets, and the hull cannot tell imprecise, compatible sources from precise, contradictory ones (the running example gives $4 + I$, $I \in [0, 3]$, in both cases). The neutrosophic triple can: because $T + F$ may exceed one, the contradictory case is the glut $(1, 0, 1)$ for "$\theta \leq 5.5$", while the imprecise case is $(0, 1, 0)$; a probability or an interval cannot represent the first. The triple alone, however, says that some sources disagree, not which ones, with what error rate or with what power. Table 12 summarises the three forms; its second case uses point estimates 4.05 and 6.95 with the stated symmetric outer intervals, and is schematic: error control requires simultaneous coverage at the declared level, and power is not determined by the interval end points alone.

Table 12. Which form separates imprecision from contradiction (running example of Section 3).

Form	Imprecise vs contradictory sources	What is reported
$N = d + I$, fixed or hull I	not separated	one interval, $4 + I$, $I \in [0, 3]$, in both cases; equivalent to retaining a single interval hull
Triple (T, I, F) with glut	separated	$(0, 1, 0)$ against $(1, 0, 1)$; the disagreeing sources, the error rate and the power are not reported
Labelled I and conflict statement (this paper)	separated	$W_{\mathrm{imp}} = 3$ against $W_{\mathrm{gap}} = 2.9$, $W_{\mathrm{samp}} = 0.1$; a conflict statement that names the two sources, with the certified gap (2.8), the test, its conditions, its false-conflict rate and its power

Which procedure supports which conclusion. Two things must be kept apart. Proposition 5 guarantees the false-conflict rate of the paired batch-means test under its stated assumptions, and that guarantee is general. Its

application to the rain data is not covered by it: the assumptions cannot be verified there, and the form the data take is the one of Remark 1, where no bound holds. The rain conflict results are therefore illustrative and conditional: the test's level for data like these was supported only in a few simulated designs whose persistence matches a diagnostic estimated from the same data, at 90 and 365 days, not at 30 days and not under strong own persistence, under a reading rule amended after a simulation smoke test; and the shares change markedly with the number of batches. Santiago's absence of flagged conflict rests on the same test and on the direct diagnosis of small, same-side differences (Santiago's persistence diagnostic lies below every simulated design, so the test's power there is not bracketed by the simulation). The bootstrap limits, their sampling components and their shares are descriptive: their simultaneous coverage was below nominal in every dependent design we ran, so their sampling components are potentially underestimated and Proposition 4 does not apply to them. The ERA5 comparison is descriptive. The DICES conclusion rests on Holm's procedure, and its one rejection is fragile to Monte Carlo error. Table 13 gives the status of every result.

Table 13. Provenance, inferential scope and outcome of each result, reported separately. Provenance: frozen before the relevant run, amended after pilot results, or exploratory (not preregistered). Scope: proved under stated assumptions, conditionally supported, or descriptive (no error control claimed). "Inconclusive" describes an outcome, not provenance. Preregistering a simulation does not establish a theorem's assumptions for an empirical application.

Result	Provenance	Inferential scope	Outcome
Propositions 1–6, 7, 8	–	proved under stated conditions	–
Simulations S1–S4 (Table 4)	frozen (commit c424482)	simulation evidence	all criteria met
Paired-test designs v0.5 and v0.6	frozen; BM-10 rule added as an amendment before real-data runs	simulation evidence	level kept at 90/365 days in matched designs; failed at 30 days and under strong own persistence
Reading rule for the rain claims	amended after a smoke test (A1)	–	under the first frozen rule no claim would have been made
Dependence (v0.4), coverage, extra and common-day studies	exploratory	simulation evidence	reported in Section 5
Rain: flagged conflict, 90 and 365 days	applied with the amended rule	conditional (assumptions not verifiable; diagnostic from the same data)	53 % and 76 % of Guayaquil windows; none in Santiago
Rain: 30-day windows	–	no claim	–
Rain: bootstrap components; ERA5 comparison	descriptive analyses	descriptive	limits undercover; threshold from the same data
Rock roughness	exploratory	conditional (assumed profiles per cell)	separation from 7–15 profiles
DICES: split of the 298 undecided verdicts	exploratory	descriptive	229 / 60 / 9
DICES: majority-verdict conflict	exploratory	valid under the binomial model	none found; inconclusive (low power)
DICES: fuzzy test, Holm	exploratory	FWER control conditional on valid item-level permutation p-values; BH exploratory	one fragile rejection
Comparison with rival implementations	exploratory	descriptive	Section 6.5
Computational times	exploratory	descriptive (one machine)	Section 8

When the labels help and when they add little. The labels add little when there is one source and no undecided records: then $I_{\mathrm{meas}} = 0$, I_{samp} is the width of an ordinary confidence interval, and the labelled number is that interval written differently. They also add little when the decision does not depend on why the uncertainty is there, for example when only the total width is compared with a tolerance. They help when the width calls for different actions depending on its source (more records against a better instrument, as in DICES), when several sources may

issue opposite verdicts that an average or a hull would hide (as in the rain data), and when a computed quantity is propagated and the analyst needs to know which part of its width more data would remove.

Computational cost and computational choices. All times are medians on one laptop (Intel Core, Python 3.14, NumPy 2.4), of five runs for the count and paired calculations and three runs for the bootstrap (`code/timing_v07.py`; the permutation test is timed by `code/timing_v07_dices.py`). For synthetic count arrays with 1980 cells, the number of DICES item–pool cells, the Clopper–Pearson calculation took 3 ms, and for 198 000 synthetic cells 0.32 s. The paired batch-means test with Holm over 20 ordered pairs took about 0.1–0.2 ms per window; multiplying the 365-day benchmark by the 95 windows of one city and tolerance gives a rough estimate of 0.01 s, which was not timed directly. The descriptive block bootstrap with $B = 2000$ takes 0.01 s, 0.05 s and 0.15 s for windows of 90, 365 and 1094 days, linear in B (0.011 s with $B = 500$ to 0.097 s with $B = 4000$ at 365 days) and roughly linear in the window length. Complete simulation runs were not benchmarked. For the functional fuzzy permutation test with 99 999 permutations, the median runtime across the first five DICES items was 0.23 s, corresponding to an extrapolated total of about four minutes for 990 items. The method is therefore feasible on a laptop for data of the size used here; for much larger numbers of windows or designs the simulations should be run in parallel, which is straightforward because replicates are independent. As for computational choices, in the last 365-day Guayaquil window at $\delta = 0.5$ the mean sampling component, averaged over models and ten seeds, was 0.1486–0.1488 for B from 500 to 4000 (standard deviation over seeds 0.0016–0.0039). This is a Monte Carlo stability check for that summary and window, not a validation of coverage or of all interval endpoints. The summary was more sensitive to the block length: 0.149 with the data-driven Politis–White length (31 days) and 0.126, 0.141, 0.148 and 0.144 with fixed mean block lengths of 5, 10, 20 and 40 days. The 5- and 10-day blocks produced smaller estimated sampling components than the data-driven length in this window; the relationship was not monotone across all tested block lengths, and the estimates remain descriptive.

Table 14. What each component means for the analyst; actions are prompts to be read together with the position of the identified set relative to the decision threshold (Section 7, step 4).

Component	Reduced by	Action
Measurement imprecision	better instruments, finer records, narrower tolerance	do not collect more of the same records; report I_{meas} as the part that more data will not remove
Sampling	more records	compute the sample size at which the decision would be made; report I_{samp} with its dependence model and whether its coverage was checked
Conflict (certified)	nothing within one source	do not hide the discrepancy in a single average; investigate its origin and justify any combination; report the disagreeing sources and the conditions of the test
Conflict (not certified)	more records, if power is low	report the power before reading the absence as agreement
Hesitancy, one holder	elicitation	treat as imprecision; check whether the event is non-convex (Proposition 6(a))
Hesitancy, several holders	nothing within the panel	evaluate conflict for the chosen proposition (Proposition 6(b))

Limitations. Certified conflict is conservative and has low power at moderate sample sizes. The paired rule's level is proved for independent normal batch means with a common variance, or with a common mean; with both heterogeneities no bound holds (Remark 1). In the rain application these assumptions cannot be verified, so its level there rests on simulation alone, in a few designs whose persistence matches a diagnostic estimated from the same data, which include seasonal heterogeneity in mean and variance and, in a non-preregistered check, dry spells; we therefore present the rain results as illustrative and conditional, not as conflict with a guaranteed false-conflict rate. The reading rule that allows the conditional claim was amended after a simulation smoke test; under the rule as

first frozen no conflict would have been claimed. The rule failed at 30 days and under stronger own persistence, where every rule we tried failed, and our persistence diagnostic is estimated from three years and does not order failure exactly; if the models' own errors are more persistent than it shows, the flagged shares may contain false conflicts. The shares depend on the tolerance and strongly on the number of batches, both fixed in advance (at 365 days, on the same days, 32 %–100 % with 5 to 20 batches; all versions kept their level in the designs matched to the diagnostic, so the simulation does not single out a b). The sampling components of the rain application come from bootstrap limits that undercover and may be underestimated; we have no validated alternative for them. The rain series are concatenated short-range forecasts over model version changes, compared with a reanalysis. The number of JRC profiles is unknown, and the DICES Unsure category is derived from several sub-questions. The arithmetic uses floating point without directed rounding, and the labels of a nonlinear result are a convention.

Failure modes: when the decomposition can mislead. (i) *Wrong sources.* If the sources are different populations by design, a separation between them is heterogeneity, not conflict (rock roughness). (ii) *Dependence.* Under unmodelled serial or cluster dependence the sampling component can be underestimated, particularly when dependence is positive, and nominal conflict tests may exceed their stated error rate; under strong own persistence every rule we tried failed. Raters who rate many items induce dependence across items; a hierarchical or mixed-effects model of items and raters would be the natural next step for rater data, and we have not fitted one. (iii) *Coding.* The split moves with the tolerance and the coding of undecided answers; when these are not fixed by the measurement process, a single choice can make imprecision look like sampling or conflict. (iv) *Low power.* With few records per source the absence of conflict is not agreement. (v) *Nonlinear propagation.* The typed widths of nonlinear results are conventions, and reciprocals near zero make every width explode. (vi) *Unverifiable time-series assumptions.* We did not perform independent-period calibration, fixed-lead reanalysis, model-version adjustment or seasonal robustness analysis in this revision. These analyses would require additional provenance information, a different forecast sample (a common fixed-lead record for all five selected models does not cover the full 2023–2025 period) or smaller subsets of the available record, and a subset already examined would not become a held-out validation. Robustness across these dimensions therefore remains unassessed, and the rain results remain conditional and descriptive. The approach is not suited to questions about whether sources differ at all (use a test of equality), to modelling interval-valued data (use symbolic regression), or to data whose records cannot be coded as support, refutation or undecided without arbitrary thresholds.

9. Conclusion

The I of a neutrosophic number can be computed from the records once the analyst's choices are declared, and split, with labels, into what more data will remove (sampling) and what they will not (measurement imprecision and one-holder hesitancy). Conflict between sources is reported separately, with its conditions and power, and it is recomputed after every operation; whether those conditions hold has to be argued in each application. The inferential tools are classical; the contribution is a reporting rule that keeps provenance, maps each source to its established counterpart, and comes with a protocol, a checklist and an action for each component, with low computational cost for the count-based summaries. The labels matter when the same width calls for different actions: in the DICES ratings they suggest different follow-up priorities for the undecided verdicts, additional ratings for 229 and examination of "unsure" for 60 (priorities not validated as interventions), and the absence of conflict between locales cannot be read as agreement because power is low; in the rain data they keep a disagreeing model identified instead of averaging it away, as illustrative evidence conditional on the models' own-error persistence being as weak as a diagnostic estimated from the same data indicates, while in Santiago the models differ slightly on the same side; in the rock-roughness data they distinguish dispersion from uncertainty about the mean. They add little with one source and no undecided records. The applied results are demonstrations, not validations; their provenance and scope are stated in Table 13.

Declarations

Data and code availability. Code, tests, the preregistrations (`PREREGISTRATION.md`, `PREREG_v05.md`, `PREREG_v06.md`), the list of deviations and all outputs are in a release archive (version 0.7, which also contains the Supplement and a script that regenerates the numbers of this article from the code) deposited at Zenodo (DOI: 10.5281/zenodo.23789946); its README maps every table and figure to the script that produces it; the computational benchmarks are generated by running `code/timing_v07.py` followed by `code/timing_v07_dices.py`, which adds the permutation-test measurements and the extrapolated total to `results/timing_v07.json`. Data: Open-Meteo forecasts (CC BY 4.0); ERA5 through Open-Meteo (contains modified Copernicus Climate Change Service information); Chen et al. [4] Table 1 (CC BY 4.0); DICES-990 (CC BY 4.0, Google LLC). The Supplement contains Propositions S1–S4, Tables S1–S10 and the provenance of the rain data.

Funding. None. **Competing interests.** F. Smarandache founded neutrosophic theory and edits neutrosophic journals; M. Y. Leyva-Vázquez is an editor of neutrosophic journals. **Use of AI tools.** Claude (Anthropic) assisted with literature search, code, simulation and drafting under the authors' direction; the authors checked every reported number against the code outputs and are responsible for the content. **Author contributions.** M.Y.L.-V.: conceptualisation, methodology, software, analysis, writing. F.S.: conceptualisation (neutrosophic statistics and hesitant indeterminacy), writing (review).

REFERENCES

1. F. Smarandache, *Introduction to neutrosophic statistics*, Sitech & Education Publishing, Craiova, 2014. arXiv:1406.2000, doi:10.48550/arXiv.1406.2000.
2. M. Aslam, *Neutrosophic analysis of variance: Application to university students*, Complex & Intelligent Systems, vol. 5, no. 4, pp. 403–407, 2019. doi:10.1007/s40747-019-0107-2.
3. M. Aslam, *Product acceptance determination with measurement error using the neutrosophic statistics*, Advances in Fuzzy Systems, vol. 2019, Article 8953051, 2019. doi:10.1155/2019/8953051.
4. J. Chen, J. Ye, and S. Du, *Scale effect and anisotropy analyzed for neutrosophic numbers of rock joint roughness coefficient based on neutrosophic statistics*, Symmetry, vol. 9, no. 10, Article 208, 2017. doi:10.3390/sym9100208.
5. N. Odat, *Neutrosophic Kies distribution with basic properties and engineering applications*, Statistics, Optimization & Information Computing, vol. 16, no. 5, pp. 5148–5170, 2026. doi:10.19139/soic-2310-5070-4115.
6. W. H. Woodall, A. R. Driscoll, and D. C. Montgomery, *A review and perspective on neutrosophic statistical process monitoring methods*, IEEE Access, vol. 10, pp. 100456–100462, 2022. doi:10.1109/ACCESS.2022.3207188.
7. W. H. Woodall, C. King, A. R. Driscoll, and D. C. Montgomery, *A critical assessment of neutrosophic statistical methods*, Quality Engineering, vol. 37, no. 4, pp. 631–644, 2025. doi:10.1080/08982112.2025.2482198.
8. A. Haq and W. H. Woodall, *A critique of neutrosophic statistical analysis illustrated with interval data from designed experiments*, Journal of Quality Technology, vol. 57, no. 3, pp. 257–264, 2025. doi:10.1080/00224065.2025.2509618.
9. A. Haq and J. Shabbir, *Revisiting the neutrosophic random variable*, Open Journal of Statistics, vol. 16, no. 5, pp. 370–381, 2026. doi:10.4236/ojs.2026.165018.
10. F. Smarandache, *Neutrosophic statistics is an extension of interval statistics, while plithogenic statistics is the most general form of statistics (second version)*, International Journal of Neutrosophic Science, vol. 19, no. 1, pp. 148–165, 2022. doi:10.54216/IJNS.190111.
11. P. Walley, *Inferences from multinomial data: Learning about a bag of marbles*, Journal of the Royal Statistical Society: Series B, vol. 58, no. 1, pp. 3–34, 1996. doi:10.1111/j.2517-6161.1996.tb02065.x.
12. N. Schenker and J. F. Gentleman, *On judging the significance of differences by examining the overlap between confidence intervals*, The American Statistician, vol. 55, no. 3, pp. 182–186, 2001. doi:10.1198/000313001317097960.
13. W. G. Cochran, *The combination of estimates from different experiments*, Biometrics, vol. 10, no. 1, pp. 101–129, 1954. doi:10.2307/3001666.
14. J. P. T. Higgins and S. G. Thompson, *Quantifying heterogeneity in a meta-analysis*, Statistics in Medicine, vol. 21, no. 11, pp. 1539–1558, 2002. doi:10.1002/sim.1186.
15. G. W. Imbens and C. F. Manski, *Confidence intervals for partially identified parameters*, Econometrica, vol. 72, no. 6, pp. 1845–1857, 2004. doi:10.1111/j.1468-0262.2004.00555.x.
16. J. Stoye, *More on confidence intervals for partially identified parameters*, Econometrica, vol. 77, no. 4, pp. 1299–1315, 2009. doi:10.3982/ECTA7347.
17. F. Smarandache and M. Y. Leyva Vázquez, *Números neutrosóficos de indeterminación intervalar de la forma a + bI: Un análisis comparativo con aritmética intervalar y afín, con extensiones multifuente y vacilantes*, Serie Científica de la Universidad de las Ciencias Informáticas, vol. 19, no. 3, 2026. Preprint: SciELO Preprints, doi:10.1590/SciELOPreprints.16004.
18. E. Hüllermeier and W. Waegeman, *Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods*, Machine Learning, vol. 110, no. 3, pp. 457–506, 2021. doi:10.1007/s10994-021-05946-3.

19. J. Abellán, *Uncertainty measures on probability intervals from the imprecise Dirichlet model*, International Journal of General Systems, vol. 35, no. 5, pp. 509–528, 2006. doi:10.1080/03081070600687643.
20. I. Couso and D. Dubois, *Statistical reasoning with set-valued information: Ontic vs. epistemic views*, International Journal of Approximate Reasoning, vol. 55, no. 7, pp. 1502–1518, 2014. doi:10.1016/j.ijar.2013.07.002.
21. D. Dubois and H. Prade, *Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets*, Fuzzy Sets and Systems, vol. 192, pp. 3–24, 2012. doi:10.1016/j.fss.2010.11.007.
22. D. Tagarro, R. Pérez-Fernández, and E. Miranda, *Confidence intervals with imprecise data*, Statistical Papers, vol. 67, no. 2, Article 39, 2026. doi:10.1007/s00362-026-01818-x.
23. C. F. Manski, *Partial identification of probability distributions*, Springer, New York, 2003. doi:10.1007/b97478.
24. M. L. Puri and D. A. Ralescu, *Fuzzy random variables*, Journal of Mathematical Analysis and Applications, vol. 114, no. 2, pp. 409–422, 1986. doi:10.1016/0022-247X(86)90093-4.
25. R. Viertl, *Statistical methods for fuzzy data*, Wiley, Chichester, 2011. doi:10.1002/9780470974414.
26. L. Billard and E. Diday, *From the statistics of data to the statistics of knowledge: Symbolic data analysis*, Journal of the American Statistical Association, vol. 98, no. 462, pp. 470–487, 2003. doi:10.1198/016214503000242.
27. L. Billard and E. Diday, *Symbolic data analysis: Conceptual statistics and data mining*, Wiley, Chichester, 2006. doi:10.1002/9780470090183.
28. P. Smets, *The combination of evidence in the transferable belief model*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 5, pp. 447–458, 1990. doi:10.1109/34.55104.
29. A. P. Dempster, *Upper and lower probabilities induced by a multivalued mapping*, The Annals of Mathematical Statistics, vol. 38, no. 2, pp. 325–339, 1967. doi:10.1214/aoms/1177698950.
30. N. D. Belnap, *A useful four-valued logic*, in Modern Uses of Multiple-Valued Logic, edited by J. M. Dunn and G. Epstein, Reidel, Dordrecht, pp. 5–37, 1977. doi:10.1007/978-94-010-1161-7_2.
31. J. Ye, *Multiple-attribute decision-making method under a single-valued neutrosophic hesitant fuzzy environment*, Journal of Intelligent Systems, vol. 24, no. 1, pp. 23–36, 2015. doi:10.1515/jisys-2014-0001.
32. V. Torra, *Hesitant fuzzy sets*, International Journal of Intelligent Systems, vol. 25, no. 6, pp. 529–539, 2010. doi:10.1002/int.20418.
33. E. R. Hansen, *A generalized interval arithmetic*, in Interval Mathematics, edited by K. Nickel, Lecture Notes in Computer Science, vol. 29, Springer, Berlin, pp. 7–18, 1975. doi:10.1007/3-540-07170-9_2.
34. J. L. D. Comba and J. Stolfi, *Affine arithmetic and its applications to computer graphics*, in Anais do VI Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens (SIBGRAPI'93), Recife, Brazil, pp. 9–18, 1993.
35. J. Stolfi and L. H. de Figueiredo, *An introduction to affine arithmetic*, TEMA – Tendências em Matemática Aplicada e Computacional, vol. 4, no. 3, pp. 297–312, 2003. doi:10.5540/tema.2003.04.03.0297; https://tema.sbmac.org.br/tema/article/view/352.
36. L. H. de Figueiredo and J. Stolfi, *Affine arithmetic: Concepts and applications*, Numerical Algorithms, vol. 37, no. 1–4, pp. 147–158, 2004. doi:10.1023/B:NUMA.0000049462.70970.b6.
37. M. Y. Leyva Vázquez and F. Smarandache, *A layered framework for neutrosophic statistics: Foundational distinctions, empirical validation, and operational implementation*, Hacettepe Journal of Mathematics and Statistics, advance online publication, 2026. doi:10.15672/hujms.1942616.
38. M. Y. Leyva Vázquez and F. Smarandache, *Clarifying the layers of neutrosophic statistics: A taxonomy, corrected simulation, and publication protocol*, Preprint, Zenodo, 2026. doi:10.5281/zenodo.19778333.
39. M. Y. Leyva Vázquez and F. Smarandache, *Measuring neutrosophic triples from multi-source evidence: The glut as Dempster's conflict and the conformal guarantee*, Zenodo, 2026. doi:10.5281/zenodo.23217578.
40. M. Y. Leyva-Vázquez and F. Smarandache, *In what sense does neutrosophic probability generalise lower probability? Exact reduction, coherence tests and a paraconsistent lifting*, Submitted to Statistics, Optimization & Information Computing, 2026.
41. M. Y. Leyva-Vázquez and F. Smarandache, *rnel: Refined Neutrosophic Evidential Logic* (version 0.3.0) [computer software], Zenodo, 2026. doi:10.5281/zenodo.23178344.
42. T. Fujita, A. Heilat, and A. K. Das, *HyperInterval-valued and SuperHyperInterval-valued fuzzy/neutrosophic set*, Statistics, Optimization & Information Computing, vol. 16, no. 2, pp. 1003–1032, 2026. doi:10.19139/soic-2310-5070-3750.
43. C. J. Clopper and E. S. Pearson, *The use of confidence or fiducial limits illustrated in the case of the binomial*, Biometrika, vol. 26, no. 4, pp. 404–413, 1934. doi:10.1093/biomet/26.4.404.
44. C. Osswald and A. Martin, *Understanding the large family of Dempster–Shafer theory's fusion operators—A decision-based measure*, in 2006 9th International Conference on Information Fusion, IEEE, pp. 1–7, 2006. doi:10.1109/ICIF.2006.301631.
45. M. A. Benhari and M. Kaicer, *Analysis of uncertainty in the Leontief model by interval arithmetic*, Statistics, Optimization & Information Computing, vol. 13, no. 5, pp. 2011–2026, 2025. doi:10.19139/soic-2310-5070-2279.
46. F. Smarandache, *Neutrosophic precalculus and neutrosophic calculus*, EuropaNova, Brussels, 2015. doi:10.5281/zenodo.31821.
47. R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to interval analysis*, SIAM, Philadelphia, 2009. doi:10.1137/1.9780898717716.
48. P. Zippenfenig, *Open-Meteo.com weather API* [computer software], Zenodo, 2024. doi:10.5281/zenodo.7970649.
49. H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, et al., *The ERA5 global reanalysis*, Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999–2049, 2020. doi:10.1002/qj.3803.
50. D. N. Politis and J. P. Romano, *The stationary bootstrap*, Journal of the American Statistical Association, vol. 89, no. 428, pp. 1303–1313, 1994. doi:10.1080/01621459.1994.10476870.
51. H. R. Künsch, *The jackknife and the bootstrap for general stationary observations*, The Annals of Statistics, vol. 17, no. 3, pp. 1217–1241, 1989. doi:10.1214/aos/1176347265.
52. S. N. Lahiri, *Resampling methods for dependent data*, Springer, New York, 2003. doi:10.1007/978-1-4757-3803-2.
53. D. S. Wilks, *Resampling hypothesis tests for autocorrelated fields*, Journal of Climate, vol. 10, no. 1, pp. 65–82, 1997. https://doi.org/10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2.

54. D. N. Politis and H. White, *Automatic block-length selection for the dependent bootstrap*, Econometric Reviews, vol. 23, no. 1, pp. 53–70, 2004. doi:10.1081/ETC-120028836.
55. A. Patton, D. N. Politis, and H. White, *Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White*, Econometric Reviews, vol. 28, no. 4, pp. 372–375, 2009. doi:10.1080/07474930802459016.
56. J. S. Whitaker and A. F. Loughe, *The relationship between ensemble spread and ensemble mean skill*, Monthly Weather Review, vol. 126, no. 12, pp. 3292–3302, 1998. https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2.
57. L. Aroyo, A. Taylor, M. Díaz, C. Homan, A. Parrish, G. Serapio-García, V. Prabhakaran, and D. Wang, *DICES dataset: Diversity in conversational AI evaluation for safety*, in Advances in Neural Information Processing Systems 36, pp. 53330–53342, 2023. doi:10.52202/075280-2321.
58. D. Dubois and H. Prade, *Representation and combination of uncertainty with belief functions and possibility measures*, Computational Intelligence, vol. 4, no. 3, pp. 244–264, 1988. doi:10.1111/j.1467-8640.1988.tb00279.x.
59. W. Trutschnig, G. González-Rodríguez, A. Colubi, and M. Á. Gil, *A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread*, Information Sciences, vol. 179, no. 23, pp. 3964–3972, 2009. doi:10.1016/j.ins.2009.06.023.
60. M. Á. Gil, M. Montenegro, G. González-Rodríguez, A. Colubi, and M. R. Casals, *Bootstrap approach to the multi-sample test of means with imprecise data*, Computational Statistics & Data Analysis, vol. 51, no. 1, pp. 148–162, 2006. doi:10.1016/j.csda.2006.04.018.
61. G. González-Rodríguez, A. Colubi, and M. Á. Gil, *Fuzzy data treated as functional data: A one-way ANOVA test approach*, Computational Statistics & Data Analysis, vol. 56, no. 4, pp. 943–955, 2012. doi:10.1016/j.csda.2010.06.013.
62. B. Phipson and G. K. Smyth, *Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn*, Statistical Applications in Genetics and Molecular Biology, vol. 9, no. 1, Article 39, 2010. doi:10.2202/1544-6115.1585.
63. Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society: Series B, vol. 57, no. 1, pp. 289–300, 1995. doi:10.1111/j.2517-6161.1995.tb02031.x.
64. S. Holm, *A simple sequentially rejective multiple test procedure*, Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70, 1979. https://www.jstor.org/stable/4615733.
65. P. Bertrand and F. Goupil, *Descriptive statistics for symbolic data*, in Analysis of Symbolic Data, edited by H.-H. Bock and E. Diday, Springer, Berlin, pp. 106–124, 2000. doi:10.1007/978-3-642-57155-8_6.
66. E. de A. Lima Neto and F. de A. T. de Carvalho, *Centre and range method for fitting a linear regression model to symbolic interval data*, Computational Statistics & Data Analysis, vol. 52, no. 3, pp. 1500–1515, 2008. doi:10.1016/j.csda.2007.04.014.
67. E. de A. Lima Neto and F. de A. T. de Carvalho, *Constrained linear regression models for symbolic interval-valued variables*, Computational Statistics & Data Analysis, vol. 54, no. 2, pp. 333–347, 2010. doi:10.1016/j.csda.2009.08.010.
68. P. Filzmoser and R. Viertl, *Testing hypotheses with fuzzy data: The fuzzy p-value*, Metrika, vol. 59, no. 1, pp. 21–29, 2004. doi:10.1007/s001840300269.
69. R. L. Berger, *Multiparameter hypothesis testing and acceptance sampling*, Technometrics, vol. 24, no. 4, pp. 295–300, 1982. doi:10.1080/00401706.1982.10487790.
70. J. P. Romano and M. Wolf, *Stepwise multiple testing as formalized data snooping*, Econometrica, vol. 73, no. 4, pp. 1237–1282, 2005. doi:10.1111/j.1468-0262.2005.00615.x.
71. B. Schmeiser, *Batch size effects in the analysis of simulation output*, Operations Research, vol. 30, no. 3, pp. 556–568, 1982. doi:10.1287/opre.30.3.556.
72. N. M. Kiefer and T. J. Vogelsang, *A new asymptotic theory for heteroskedasticity-autocorrelation robust tests*, Econometric Theory, vol. 21, no. 6, pp. 1130–1164, 2005. doi:10.1017/S0266466605050565.
73. R. Ibragimov and U. K. Müller, *t-statistic based correlation and heterogeneity robust inference*, Journal of Business & Economic Statistics, vol. 28, no. 4, pp. 453–468, 2010. doi:10.1198/jbes.2009.08046.
74. A. E. Dudek, J. Leśkow, E. Paparoditis and D. N. Politis, *A generalized block bootstrap for seasonal time series*, Journal of Time Series Analysis, vol. 35, no. 2, pp. 89–114, 2014. doi:10.1002/jtsa.12053.
75. Y. Benjamini and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, The Annals of Statistics, vol. 29, no. 4, pp. 1165–1188, 2001. doi:10.1214/aos/1013699998.
76. Z. K. Bakirov and G. J. Székely, *Student's t-test for Gaussian scale mixtures*, Journal of Mathematical Sciences, vol. 139, no. 3, pp. 6497–6505, 2006. doi:10.1007/s10958-006-0366-5.
77. Open-Meteo, *Historical Forecast API documentation* and *Previous Runs API documentation*, https://open-meteo.com/en/docs/historical-forecast-api (accessed 11 October 2026).