

Methods note · Observation and data

A rank is a distribution

Amina Okafor · 7 October 2026

Fictional author and institute · Editorial demonstration, not peer reviewed

ABSTRACT

Sorting point estimates creates an order even when observations cannot reliably resolve the differences. This methods note distinguishes descriptive ranks from uncertainty in rank, drawing on league-table methodology, the JCGM uncertainty framework, and NASA's GISTEMP ensemble documentation. A fictional two-estimate comparison holds marginal uncertainties fixed while changing error correlation, showing why individual uncertainty bars cannot determine support for a lead. The note then explains joint propagation through ranking, finite-ensemble limits, comparison-set definitions, and public displays that retain magnitude alongside position. It calculates no actual climate ranking and does not extend NASA's documented ensemble beyond its stated period. The principal implication is to report an estimated order separately from its model-dependent stability, especially when ordinal cutoffs carry decision consequences.

Keywords: rank uncertainty · correlated errors · ensembles · scientific communication

A source-based editorial text, not a report of original Still North research. Examples and imagery are illustrative.

Ordering estimates is not the same as resolving differences

Sorting three reported values always produces an order, even when the evidence cannot reliably distinguish the first two. A ranking interface can therefore manufacture apparent precision without making a single numerical error. The problem is especially visible in environmental records whose annual differences are small relative to observational uncertainty. First place may be clear, ambiguous, or dependent on a processing choice. The sorted column alone cannot tell a reader which situation applies.

This methods note examines uncertainty in ranks using an explicitly hypothetical comparison, the measurement-uncertainty framework of the JCGM Guide, NASA's GISTEMP uncertainty ensemble documentation, and Goldstein and Spiegelhalter's work on league tables. These sources address different applications. They are brought together to analyse a shared statistical problem, not to treat environmental records as institutional performance scores. No new temperature ranking or empirical probability is calculated here.

Define the quantity and the contest

A rank requires a comparison set. Warmest among which years, over which region, for which season, and in which dataset? Changing from a calendar year to a water year changes the measured period. Changing spatial coverage changes the target. Extending a record can change an old year's rank without changing its estimated value. A claim such as "third highest" is incomplete unless the reader can recover the set within which third place was assigned.

It is also necessary to distinguish descriptive and inferential ranks. A descriptive rank orders the point estimates in a fixed release. An inferential rank asks how that order might change under the stated uncertainty model. Both can be useful. Reporting the former as if it proves the latter is the mistake. Rounding creates another distinction:

two displayed values can tie even when the underlying estimates differ. Alphabetical sorting is not a scientific tie-breaker.

Goldstein and Spiegelhalter show how plausible parameter realisations can be ranked repeatedly to obtain distributions of ranks. Their institutional examples also warn that defining an appropriate comparison matters before uncertainty is calculated. An exquisitely estimated rank of an unsuitable quantity remains an unsuitable summary.

Fictional worked comparison: a narrow lead

Suppose two invented annual index estimates, A and B, are 1.20 and 1.17 units. Each has standard uncertainty 0.04 units. These are teaching quantities with no connection to an actual year or climate product. For this first calculation, assume normally distributed errors, correctly specified uncertainties, and independence between the two errors. The difference A minus B is 0.03 units, with standard uncertainty given by the square root of 0.04 squared plus 0.04 squared, approximately 0.057.

Under that illustrative joint uncertainty distribution, about 70% of the probability mass for the difference lies above zero. The point estimate places A first, but the narrow lead is not overwhelming evidence that A exceeds B. The 70% figure is conditional on the invented model. It is not a confidence label imported from a climate assessment, nor a probability that any particular real year holds a record.

Now retain the same marginal uncertainties but assume correlation of 0.8 between the errors. The variance of the difference becomes 0.04 squared plus 0.04 squared minus twice 0.8 times 0.04 times 0.04. Its standard uncertainty is approximately 0.025. Under the same normal model, about 88% of the distribution for the difference lies above zero. Nothing about the displayed values or their

Illustrative assumption	Standard uncertainty of A minus B	Model probability that A exceeds B
Independent errors	About 0.057 units	About 70%
Error correlation of 0.8	About 0.025 units	About 88%

individual uncertainty bars has changed. Their relationship has.

The JCGM Guide, section 5.2, requires significant correlations to be included when combining input uncertainties. In a ranking problem, this is not a technical footnote. A common calibration offset may move all estimates together and largely cancel from their differences. Conversely, errors associated with different coverage patterns can affect comparisons differently. Individual error bars do not contain this joint information.

Propagate uncertainty through the ranking

For more than two estimates, a rank depends on every competitor. A transparent approach is to generate or use joint realisations of the entire comparison set, rank the values within each realisation, and summarise how often each item occupies each position. The procedure must preserve relevant shared errors and dependencies. Drawing every year independently from its displayed interval can create combinations that the original measurement model would never produce.

NASA's GISTEMP documentation supplies an operational example of propagation rather than a ready-made answer to every rank question. Its documented observational ensemble contains 200 reconstructions for 1880 to 2020. NASA recommends rerunning an analysis across ensemble members and summarising the resulting distribution, using the appropriate gridded or large-scale series. Those documented dates matter: the ensemble described there must not be casually presented as covering later years.

A real rank analysis would need an ensemble or other uncertainty model suitable for its actual period and region. It would state what sources of uncertainty are included and what are not. NASA also recommends checking conclusions across surface-temperature analyses, since uncertainty within one product does not represent every difference between methodological choices. A high probability under one model is not protection against an omitted bias shared by all its realisations.

Finite ensembles also limit numerical resolution. With 200 members, simple counts change in steps of half a percentage point, and the empirical proportions have simulation variability. Reporting six decimal places would not add knowledge. If an estimate never loses first place within that finite ensemble, describe the observation faithfully rather than asserting mathematical certainty that it cannot lose.

Display the distinction the reader needs

A useful public table can retain the descriptive rank while placing the estimated value, definition, and uncertainty beside it. Where rank uncertainty materially affects the headline, show a distribution or an interval of ranks and explain the method. Pairwise differences can be clearer than a dense display when the actual question concerns

only two adjacent years. A rank interval should not be mistaken for an interval in the measured units.

Do not turn overlapping conventional confidence intervals into an automatic declaration that two quantities are indistinguishable, or non-overlap into a universal significance test. The relevant comparison concerns the difference, with its dependence structure and inferential method. Likewise, failure to resolve a difference does not prove equality. A report can say that the leading estimates are close without declaring that all values are the same.

The original scale often matters more than the ordinal position. A first-to-second change can correspond to a tiny numerical revision, while a tenth-ranked value can remain far above a historical baseline. Give readers the magnitude and the wider series. Replacing a robust long-term pattern with an argument over neighbouring ranks can conceal the more consequential evidence.

When a rank becomes a decision

Some uses attach consequences to an order: selecting sites for further study, allocating resources, or choosing which indicators receive attention. A hard cutoff between positions five and six can then amplify an unresolved difference. The decision rule should state why rank is relevant and whether small differences justify sharply different treatment. Evidence does not determine that value judgement.

A practical alternative may be a group of plausible candidates, a magnitude-based threshold with explicit uncertainty, or additional observation where the comparison is consequential. None is automatically preferable. The important step is to make the decision's dependence on the ranking visible, rather than allowing a sorting function to act as an unacknowledged policy.

Limits and implication

The worked comparison assumes normal errors and known covariance, conditions that need not hold in sparse environmental records. Rank distributions can be asymmetric, multimodal, or highly sensitive to the comparison set. This note does not prescribe a universal statistical model, a significance threshold, or an uncertainty label. It explains why a model must be attached to claims about rank stability.

A precise summary can therefore make two statements together: this release gives A the largest point estimate, and its lead over B has the stated degree of support under the stated model. That is not a retreat from evidence. It preserves the useful distinction between an order that can be printed and a difference that the observations can actually resolve.

References

1. Goldstein, H., and Spiegelhalter, D. J. (1996). League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society A*, 159, 385–443.

2. Joint Committee for Guides in Metrology (2008, corrected 2010). Evaluation of Measurement Data: Guide to the Expression of Uncertainty in Measurement. JCGM 100:2008.
3. NASA Goddard Institute for Space Studies (2025 documentation). GISTEMP Uncertainty Analysis: Observational Uncertainty Ensemble.