OREGON
WATER RESOURCES DEPARTMENT

Surface Water Information Modeling System (SWIMS)

Missingness Patterns in Oregon Streamflow Data

Author: Cortney Cameron

Executive Summary

This item is an addendum to the draft gap tolerance report and will be appended to the final version. Missing data patterns in Oregon streamflow data were quantified for the WY 1991-2020 base period. Results showed that missing values initiate at random intervals but persist in contiguous blocks with highly variable durations. Winter months experience a disproportionate number of missing values. These patterns are broadly consistent with, though less severe than, the simulated missingness patterns in the previous gap tolerance assessment. This indicates that the gap tolerance the assessment was conservative.

Introduction

OWRD completed a gap tolerance assessment to understand the amount of missing data that can exist at a station while still producing reliable estimates of key percentiles (Huang, 2025). The gap tolerance analysis randomly removed half of years, and then 5% to 37.5% of data within those years were randomly removed.

A panel of scientific experts convened to review the work recommended that actual patterns of missingness in data be characterized to understand the representativeness of the assessment's gap generation approach relative to real data (Andrews, 2026). This memorandum provides a characterization of actual patterns of data missingness.

Methods

Patterns of missing data in mean daily streamflow records were characterized for 205 stations in Oregon with incomplete records (<10,958 days) during the base period of water years 1991–2020. Mean daily flow data were acquired from OWRD (database version January 27, 2026) with any publication status (raw, preliminary, provisional, published).

Structural Gaps

Of these 205 stations, stations with only gaps of ≥365 days represented 60% (116). These gaps typically reflect pre-installation or decommissioned periods rather than actual missing data. Accordingly, these years-long gaps are concentrated at the beginning and the end of the base period, with only 9% occurring in the middle portion. The median "structural" gap length was 15 years. Further analyses concentrated on <365-day "operational" gaps available at 60 stations with >20 gaps each <365 days in length. Methods are summarized in Table 1. Tests were completed for annual (water year) and monthly data (e.g., comparing Februaries).

Gap Onset Timing

To characterize the temporal regularity of gap onset, the burstiness coefficient was computed (Goh & Barabási, 2008). This coefficient, which can range from -1 to 1, is calculated using the standard deviation and mean of inter-gap intervals. A value of 0 indicates Poisson-random arrival where gap onsets occur at a constant average rate with random timing. Positive values indicate bursty behavior where gaps arrive in clusters separated by quiescent periods; negative values indicate periodic behavior

where gaps occur at regular intervals. An absolute value >0.3 was considered indicative of non-random onset.

Gap Sequence Randomness

The Wald-Wolfowitz (1940) runs test was applied to assess whether the sequence of missing and non-missing daily values was random. This non-parametric test counts "runs" of the same state (data present or absent) and compares the observed number to the expected number under a null hypothesis of random ordering. Fewer runs than expected indicates clustering (missing values tend to occur in contiguous blocks). The test statistic follows an approximately normal distribution, with negative Z-values indicating clustering.

Gap Length Diversity

The diversity of gap durations was quantified using Shannon entropy (Shannon, 1948). This was computed over the frequency distribution of gap lengths and normalized by the maximum possible entropy (log of the number of unique gap lengths). This normalized entropy ranges from 0, indicating all gaps have identical duration, to 1, indicating high diversity in gap lengths.

Temporal Concentration

To assess whether gaps were concentrated in specific time periods or spread evenly across the record, temporal entropy was computed over the distribution of missing days across water years. Low temporal entropy (0) indicates gaps are concentrated in specific years, while high entropy (1) indicates gaps are distributed across years.

Seasonal Uniformity

To test whether gaps were distributed uniformly across months or showed seasonal bias, a pooled chi-square goodness-of-fit test was performed pooling all stations. For each station, the proportion of missing days in each month was normalized by that station's total missing days. These proportions were averaged across stations to obtain network-level monthly patterns, which were then compared against expected proportions based on month lengths. This pooled approach accounts for varying amounts of missing data among stations and provides a single network-level test of seasonal non-uniformity.

Table 1. Summary of tests, interpretation, and annual results.

Test	Research Question	Interpretation	Annual Result
Burstiness	Do gaps come at regular intervals (periodic), at random times, or in bursts?	-1 = periodic; 0 = random; +1 = bursty	Median = 0.25
Runs Test	Are missing values clustered together (persistence) in time or scattered randomly?	Significant value ($p < 0.05$) indicates non-random clustering (missing days tend to follow each other)	100% significant
Gap Length Entropy	Are gaps of similar or varied lengths?	0 = uniform gap lengths; 1 = diverse gap lengths	Median = 0.96
Temporal Entropy	Are gaps concentrated in specific water years or spread evenly throughout the record?	0 = uniform (gaps clustered in same years); 1 = diverse (gaps spread across different years)	Median = 0.95
Chi Square	Are gaps distributed uniformly across months or do certain months have more gaps?	Significant chi-square ($p < 0.05$) indicates non-uniform seasonal distribution	$p < 0.0001$

Results

Annual results are summarized in Table 1. Monthly results are summarized in Figure 1.

Operational Gaps

Among the 60 stations with ≥20 operational missing days, a total of 21,213 missing days were distributed across 1,088 gaps. Stations had a median of 228 missing days (range: 21-2,432) occurring in a median of 5 gaps (range: 1-331). The median gap length was 24 days (range: 1-332 days).

Gap Onset Timing

The annual burstiness coefficient had a median of 0.25 (IQR: -0.07 to 0.35), indicating gap onsets approximated random Poisson-like timing rather than clustered bursts or periodic intervals. Monthly patterns were consistent, with burstiness coefficients ranging from -0.11 to 0.03 across months, all near zero (Figure 1a).

Gap Sequence Randomness

All stations (100%) showed significant non-random clustering of missing values annually (median Z = -101.69, IQR: -103.96 to -98.61). This pattern was consistent across all months (100% significant monthly), with median Z-statistics ranging from -30.07 to -28.32, indicating that once gaps begin, missing values tend to persist in contiguous blocks (Figure 1b).

Gap Length Diversity

Gap lengths showed high diversity, with a median normalized entropy of 0.96 (IQR: 0.85-1.00), indicating gaps varied substantially in duration rather than having uniform lengths. Monthly analyses showed maximum diversity (median entropy = 1.00) across all months (Figure 1c).

Temporal Concentration

Gaps were distributed relatively evenly across water years, with a median temporal entropy of 0.95 (IQR: 0.90-0.99). Monthly temporal entropy ranged from 0.87 to 0.99, indicating gaps were not concentrated in specific years but spread throughout the 30-year record (Figure 1d).

Seasonal Uniformity

Based on a pooled chi-square test, the distribution of missing days across months was significantly non-uniform ($p < 0.0001$). Winter months (January, February, December)

were overrepresented in missing days (1.47-1.63× expected values), while spring and summer months (April through August, plus October) were underrepresented (0.63-0.82× expected values).

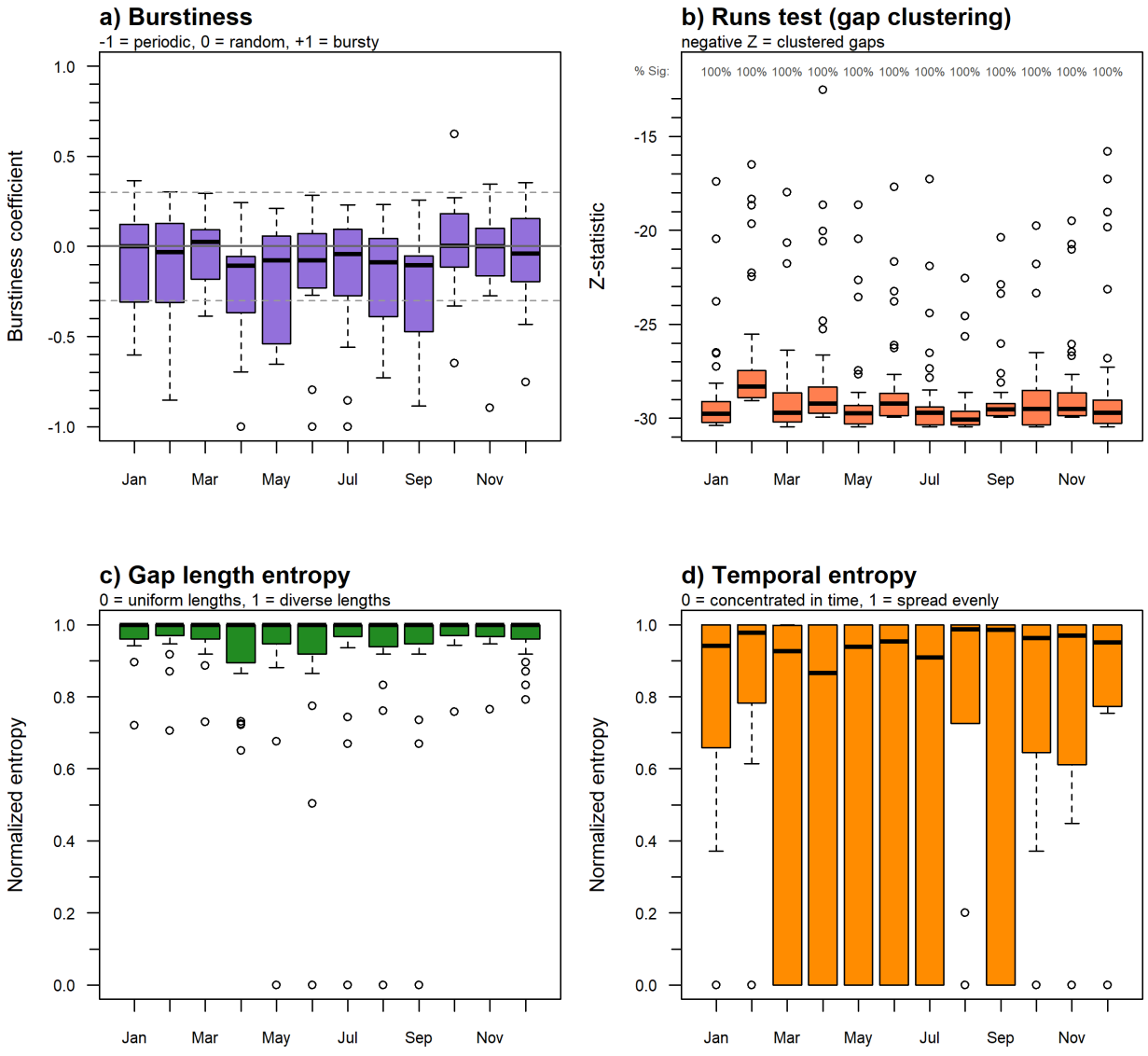

Figure 1. Distributions of information-theoretic metrics characterizing operational gaps (<365 days) by month for 60 Oregon streamflow stations (WY 1991-2020): a) burstiness coefficient for gap onset timing, b) runs test Z-statistic for sequence randomness (percentage labels indicate proportion of stations with significant clustering), c) normalized Shannon entropy for gap length diversity, and d) normalized temporal

entropy for concentration of gaps across water years.

Discussion

The tests showed that every station exhibited significant clustering of missing values: when a gap begins, it tends to persist. In contrast, the burstiness coefficients near zero indicate that gap onset timing approximates a Poisson process: new gaps begin at roughly random intervals rather than clustering in time or occurring at regular periodic intervals. The high length entropy indicates that gaps vary substantially in duration.

Taken together, these results characterize streamflow data gaps as arising from approximately random-onset events that, once initiated, persist for variable durations before resolution. This pattern is consistent with random equipment issues that take variable amounts of time to resolve. The winter concentration of gaps likely reflects challenging field conditions that complicate data collection and maintenance activities during these months.

Relative to actual patterns of missingness, the gap tolerance assessment approach approximates some characteristics but may overestimate impacts. The gap tolerance analysis removed entire years plus 5-37.5% additional random data within remaining years (Huang, 2025). While the random onset of removal approximates the Poisson-like gap initiation observed in actual data, and the year-long removals capture the clustering tendency, the scale differs: actual gaps persist for a median of 24 days. This conservative approach likely overestimates the impact of realistic gap patterns on percentile calculations. Despite the relatively extreme simulated missingness, percentile calculations were robust to missing data (Huang, 2025).

Conclusion

Operational streamflow gaps exhibit random onset timing, but once initiated, temporal clustering of missing data is significant, with highly variable durations and strong winter seasonality. This pattern is consistent with random issues requiring variable repair times. Gap tolerance assessments using year-long removals likely overestimate impacts relative to the shorter, more typical operational gaps observed in practice. Additional work is needed to evaluate potential benefits of record extension to fill short operational gaps.

References

Andrews, R. 2026. TAG Meeting #2 – Feedback Summary. Oregon Water Resources Department.

Goh, K. I., & Barabási, A. L. (2008). Burstiness and memory in complex systems. Europhysics Letters, 81(4), 48002. https://doi.org/10.1209/0295-5075/81/48002

Huang, C. (2025). Gap tolerance analysis for streamflow data. Oregon Water Resources Department.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. The Annals of Mathematical Statistics, 11(2), 147–162. https://doi.org/10.1214/aoms/1177731909