Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Network-Constrained Quality Control Framework for Identifying Reliability Concerns in River Discharge Records: A Case Study in the Northwest Territories

Hooman Razavi
University of Ottawa - Civil Engineering
Hamidreza Shirkhani
National Research Council Canada
Hossein Bonakdari
hbonakda@uottawa.ca

University of Ottawa - Civil Engineering

Research Article

Keywords: hydrometric data quality, data reliability, stochastic error modeling, network-based anomaly detection, mass-balance closure, streamflow uncertainty

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11563761/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Network-Constrained Quality Control Framework for Identifying Reliability Concerns in River Discharge Records: A Case Study in the Northwest Territories

Authors

Hooman Razavi[1], Hamidreza Shirkhani[2], Hossein Bonakdari[1,*]

Affiliation

[1] Department of Civil Engineering, Faculty of Engineering, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada

[2] Construction Research Centre, National Research Council Canada, 1200 Montreal Road, Ottawa, Ontario, Canada, K1A 0R6

Corresponding author

Hossein Bonakdari, Department of Civil Engineering, Faculty of Engineering, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada

Email: hbonakda@uottawa.ca

ORCID

Hooman Razavi (https://orcid.org/0009-0005-7944-672X)

Hossein Bonakdari (https://orcid.org/0000-0001-6169-3654)

Hamidreza Shirkhani (https://orcid.org/0000-0002-0893-652X)

Author email addresses

Hooman Razavi (hraza017@uottawa.ca)

Hossein Bonakdari (hbonakda@uottawa.ca)

Hamidreza Shirkhani (Hamidreza.Shirkhani@nrc-cnrc.gc.ca)

Acknowledgements

We thank the hydrographers of the Water Survey of Canada Yellowknife office, whose field records underpin every value analyzed here, and acknowledge the open publication of the WSC production-system analysis by Gharari et al. (2024), without which the framing of this study would not have been possible.

Abstract

Published river discharge is usually treated as deterministic, yet rating-curve construction, temporary shifts and overrides introduce uncertainties missed by conventional quality control (QC); only 69% of Water Survey of Canada (WSC) records can be reproduced from stage, rating curves and shifts, falling to 41% in January–February. We introduce Network-Constrained Quality Control (NCQC), which couples a stochastic error model with hydrometric redundancy to detect, attribute and assess inconsistencies that escape station-level QC. NCQC combines seven diagnostics: metadata-independent provenance detection, physical admissibility, satellite vegetation-index regime classification, regime-conditioned mass-balance closure, temporal–spatial attribution, cumulative-sum persistence detection and a network-independence test. Applied to 319 monthly observations (2000–2026) from three gauges at the Mackenzie–Liard confluence, Canada, the node closes without detectable bias at a robust relative scale of 1.67%, far tighter than published discharge uncertainty. NCQC flagged 54 months (16.9%) in 22 episodes, including a 14-month anomaly in 2015–2016, a persistent +4.79% open-water offset in 2018–2019 and provisional winter deviations reaching −38.9% in 2025–2026. In 4327 synthetic error-injection trials at an 8.0% background flag rate, detection probabilities for 5% and 10% sustained biases were 0.72 and 0.89 (0.66 and 0.87 for newly raised flags), against fewer than 6% for a station-level baseline. Correct attribution reached 0.74–0.80 upstream but only 0.30 downstream. Reported significant-digit precision independently identified the approved-to-provisional transition, and the ice-season closure scale fell from 1.46% (2000–2019) to 0.20% (2020–2024), consistent with possible dependence among estimates. NCQC thus identifies discharge inconsistencies and evaluates when mass-balance agreement stops being independent evidence of accuracy.

Keywords hydrometric data quality; data reliability; stochastic error modeling; network-based anomaly detection; mass-balance closure; streamflow uncertainty

List of symbols

Symbol	Definition
$\mathbf{a}$	Closure coefficient vector $(1, -1, -1)^{\top}$, Eq. (5)
b_ρ	Regime-specific relative bias of the closure residual, Eq. (4)
e_i	Observation error of station i (m³ s⁻¹), Eq. (2); $\mathbf{e}_t$ is the vector $(e_{\mathrm{FS}}, e_{\mathrm{SP}}, e_{\mathrm{LR}})^{\top}$
f_t	Liard partition fraction, $Q_{\mathrm{LR}}/(Q_{\mathrm{SP}} + Q_{\mathrm{LR}})$
h	Decision interval of the cumulative sum (CUSUM) test (= 10)
k	CUSUM reference value (= 0.5)
k_ρ	Regime-specific proportional error coefficient, Eq. (4)
Q_i	Monthly mean discharge at station i (m³ s⁻¹); subscripts FS, SP and LR denote Fort Simpson, Strong Point and the Liard
Q_i^{true}	True (error-free) monthly mean discharge at station i (m³ s⁻¹), Eq. (2)
q_ℓ	Net ungauged lateral inflow (m³ s⁻¹), Eq. (2)
r_t	Closure residual, $Q_{\mathrm{FS}} - (Q_{\mathrm{SP}} + Q_{\mathrm{LR}})$ (m³ s⁻¹), Eq. (4); identical to ε_t of Eq. (1)
S_s	Spatial-anomaly score of station s, Eq. (6)
$s_{0,\rho}$	Regime-specific additive tolerance floor (m³ s⁻¹), Eq. (4)
s_i	Standard deviation of the additive observation error at station i, Eq. (3)
T_s	Temporal-continuity score of station s, Eq. (6)
u_t	Combined upstream observation error, $e_{\mathrm{SP}} + e_{\mathrm{LR}}$, Sect. 6.2
z_c	Flag threshold on the standardized closure statistic (= 2), Eq. (8)

$z_{1-\beta}$	Standard normal quantile setting the detection power, Eq. (8)
Z_t	Standardized closure statistic, Eq. (4)
β	Type II error probability in Eq. (8); 1 − β is the detection power (= 0.8)
ΔS	Change in channel storage over the month ($m^3 s^{-1}$), Eq. (2)
δ	Relative error injected at a station; $\delta_{\min}$ is the smallest detectable value, Eq. (8)
ε	Closure residual of the nodal mass balance ($m^3 s^{-1}$), Eqs. (1) and (2); identical to r_t
γ_i	Standard deviation of the proportional observation error at station i, Eq. (3)
η_i	Proportional observation error at station i, Eq. (2)
κ_{ij}, κ	Correlation between the observation errors at stations i and j, Eq. (7); κ alone denotes the downstream-to-upstream error correlation in Eq. (9), κ_0 its value before 2020 and κ^* the value implied by the observed collapse
Λ_s	Implausibility score of station s, Eq. (6)
ξ_i	Additive observation error at station i ($m^3 s^{-1}$), Eq. (2)
ρ	Hydrological regime (ice-affected, transition, open-water)
σ_i	Standard deviation of the observation error at station i, Eq. (7)
σ_t	Regime-conditioned closure tolerance ($m^3 s^{-1}$), Eq. (4)
σ_u	Standard deviation of the combined upstream error u_t, Sect. 6.2
τ	Robust scale of the ice-season relative closure error over the period given by the subscript (%), Eq. (9)
φ_i	Flow share of station i, Q_i/Q_{FS}, Eq. (3)

1 Introduction

River discharge underpins water resources management, flood forecasting and hydrological modeling, and it is increasingly used to train and evaluate machine-learning models (Nearing et al. 2021; Kratzert et al. 2023). Gauged records also anchor global assessments of changing river flow (Feng and Gleason 2024), yet gauging networks are sparse and unevenly distributed (Krabbenhoft et al. 2022), so each archived record carries considerable weight. Published discharge is nevertheless a derived product rather than a direct measurement: stage is converted to discharge through rating curves, and the results are shifted, overridden and approved before release (Gharari et al. 2024). Rating shifts, hysteresis and backwater can change estimated discharge by tens of percent (Kona and Das Bhowmik 2025; Muste et al. 2026), and the choice of rating-curve model is itself a leading source of error (McMahon et al. 2025; Das and Das 2026). This uncertainty is rarely distributed with the records, which are routinely used as if they were exact.

Three problems follow for the quality control (QC) of published records. First, the station-level methods that quantify rating-curve uncertainty need gauging and rating information that is seldom released, so most data users cannot apply them (Westerberg et al. 2026; Yu and Chen 2026). Second, current QC methods, including recent machine-learning detectors, compare each series with its own history or target faults in raw sensor signals, and treat the archived values as ground truth (Rigueira et al. 2026; Badrudeen et al. 2026); errors introduced during rating-curve application and data approval therefore pass largely unexamined. Third, the mass-conservation constraint that links neighboring gauges is seldom used, although at a river confluence the downstream discharge should approximately equal the sum of the upstream contributions. Using this constraint raises two questions that existing methods do not answer: which station is responsible when closure fails, and whether close agreement is genuine evidence of accuracy or a

sign that the records are not independent. The second question is pressing for the Water Survey of Canada (WSC), where only 69 % of reported discharges can be reproduced from stage, rating curves and shifts (Gharari et al. 2024), and for cold-region gauges, where ice invalidates open-water ratings for much of the year (Usman et al. 2026).

This study develops a Network-Constrained Quality Control (NCQC) framework to address these challenges through three objectives: (i) quantify network-level discharge consistency through mass-balance closure and characterize its temporal and seasonal variability at a monitored river confluence; (ii) develop and evaluate a network-based quality-control framework for detecting, attributing and characterizing discharge anomalies; and (iii) investigate the implications of potential dependencies among discharge estimates for data reliability, particularly whether strong network agreement can be interpreted as evidence of measurement accuracy.

The proposed framework is demonstrated using long-term discharge records from three hydrometric stations at the Mackenzie–Liard confluence in northern Canada. It reveals that exceptionally strong agreement among neighboring gauges does not necessarily provide independent evidence of measurement accuracy, particularly when discharge estimates may be influenced by shared information or processing procedures.

The remainder of this paper is organized as follows. Sect. 2 reviews previous work on discharge uncertainty, hydrometric quality control and network-based consistency checks, and identifies the research gaps. Sect. 3 describes the study area and data, and Sect. 4 presents the NCQC framework and the synthetic error-injection design. Sect. 5 reports the results, Sect. 6 discusses their implications and limitations, and Sect. 7 concludes the paper.

2 Literature review

2.1 Uncertainty and provenance of published discharge

Discharge is generally estimated rather than measured continuously, using stage observations and rating curves developed from discrete field measurements (Rantz 1982; WMO 2010). Each step introduces uncertainty, particularly when rating curves require adjustments or alternative estimation procedures. Errors in gaugings and stage records propagate into discharge (Cohn et al. 2013; Horner et al. 2018), and rating-curve uncertainty itself varies in time (Jalbert et al. 2011; Reitan and Petersen-Øverleir 2011). Recent assessments confirm that the choice of rating-curve model is itself a leading source of discharge uncertainty, both where the stage–discharge relationship is highly unstable (Eleutério et al. 2025) and across classical, physically informed and machine-learning formulations (Das and Das 2026). Recent studies show that this uncertainty remains substantial and is rarely reported: published discharges are commonly treated as deterministic, although interannual rating shifts (Kona and Das Bhowmik 2025; McMahon et al. 2025), hysteresis under unsteady flow (Gharari and Razavi 2018; Muste et al. 2026) and variable backwater (Marinheiro et al. 2025) can change estimated discharge by tens of percent. Bayesian, hydraulic and drone-supported methods now allow these uncertainties to be quantified at individual stations (Westerberg et al. 2026; Yu and Chen 2026), but they require gauging and rating information that is seldom distributed with published records, and they do not evaluate the mutual consistency of records across a hydrometric network.

Gharari et al. (2024) demonstrated these challenges for the WSC, showing that only 69 % of reported discharge values could be reconstructed from stage observations, archived rating curves and temporary shifts, while the remaining 31 % involved alternative estimation procedures known as overrides. Reproducibility varied substantially by season, reaching 0.88 in June but declining to 0.41 in January and February within a 5 % tolerance. Temporary shifts and overrides can reduce discrepancies between field measurements and reported discharge, potentially suppressing residuals used in conventional rating-curve uncertainty estimation (Le Coz et al. 2014; Coxon et al. 2015), on which comparisons of uncertainty methods also rely (Kiang et al. 2018). Consequently, apparently small residuals do not necessarily indicate low measurement uncertainty.

2.2 Quality control of hydrometric and environmental data

These limitations also affect discharge QC. Conventional station-based methods, including range checks, rate-of-change thresholds, flat-line detection and comparisons with historical observations, identify anomalous values relative to individual station behavior (Lamb et al. 2003). Data-driven QC has also advanced rapidly, including explainable machine-learning anomaly detectors for water-quality time series (Rigueira et al. 2026), physics-informed fault detection across hydrological sensor networks (Mousavi et al. 2026), non-parametric change-point detection in multi-parameter environmental series (Hu et al. 2023) and statistical and machine-learning correction of high-frequency monitoring data (Badrudeen et al. 2026). These approaches, however, mainly target sensor malfunctions in raw signals and treat the derived records as ground truth; they rarely exploit conservation constraints among stations or consider errors introduced after measurement, during rating-curve application and data approval. Physics-guided learning can likewise reconstruct missing discharge records from neighboring information (Prakash and Tripura 2026), but such methods take the surviving archived values as ground truth, which is the assumption examined in this paper. Uncertainty-aware reliability frameworks developed for other networked infrastructure do detect incipient failures before they are visible at any single node (Damonti and Bernasconi 2026), yet they operate on raw sensor signals rather than on the post-processed records that agencies publish.

2.3 Spatial consistency and network constraints

Spatial consistency testing offers a complementary approach. In meteorology, spatial regression against neighboring stations has long been used to identify inconsistencies that station-level tests cannot detect (Hubbard et al. 2005; Durre et al. 2010), and it is a standard component of quality assurance in mesoscale networks (Fiebrich et al. 2010). Hydrometric networks provide an additional advantage through mass conservation. Network-theoretic criteria have been used to evaluate the design and redundancy of hydrometric monitoring networks (Mondal and Mishra 2025), but they address where gauges should be placed rather than whether the records those gauges produce are mutually consistent. At a river confluence monitored by upstream main-stem, tributary and downstream gauges, the downstream discharge should approximately equal the sum of upstream contributions, accounting for intervening inflows, travel time, channel storage and measurement uncertainty. However, identifying the responsible station and assessing whether neighboring records are statistically independent remain important challenges.

2.4 Cold-region records and remote sensing

Cold-region records add a further difficulty, because ice invalidates open-water rating curves for much of the year and the under-ice discharges used as observations are themselves highly uncertain (Usman et al. 2026), while the duration of river ice is itself changing rapidly under warming (Yang et al. 2020; Wang and Feng 2024). Satellite products are increasingly used to characterize snow and ice cover and vegetation dynamics in such data-scarce settings (Motiee et al. 2025; Rawat et al. 2026), and uncertainty-propagation analyses show that errors compound strongly along hydrological modeling chains (Wu et al. 2025). These studies, however, use remote sensing to describe environmental conditions or to drive models; they do not use it to condition the quality control of the observed discharge records.

2.5 Research gaps

Three gaps emerge from this literature, and studies published in 2026 indicate that all three remain open. First, station-level methods can quantify rating-curve uncertainty, but they require gauging and rating information that is seldom distributed with published records. The most recent hydraulic and hydrodynamic approaches still work one station at a time and do not test whether neighboring records agree (Westerberg et al. 2026; Yu and Chen 2026). Second, QC methods, whether rule-based, statistical or machine-learning, continue to focus on sensor-level faults and to treat the archived discharge values as ground truth (Badrudeen et al. 2026; Mousavi et al. 2026). Reconstruction methods likewise rely on the surviving archived values (Prakash and Tripura 2026), so errors introduced during rating-curve application, shifts, overrides and approval remain largely unexamined. Third, network-based checks implicitly assume that neighboring records are independent, yet no QC framework tests this assumption, even though agency procedures can incorporate neighboring information (Gharari et al. 2024). Reliability frameworks for other networked systems also operate on raw sensor signals rather than on published records (Damonti and Bernasconi 2026). Cold-region networks expose all three gaps most clearly, because under-ice discharge remains highly uncertain (Usman et al. 2026) and the choice of rating-curve model remains a leading source of error (Das and Das 2026). The NCQC framework described in Sect. 4 is designed to address these gaps.

3 Study area and data

The NCQC framework is evaluated at the Mackenzie–Liard confluence near Fort Simpson, Northwest Territories (Fig. 1), using 319 months (January 2000–July 2026) of discharge observations from three Water Survey of Canada (WSC) hydrometric stations, together with ERA5-Land reanalysis and Moderate Resolution Imaging Spectroradiometer (MODIS) Normalized Difference Vegetation Index (NDVI) data extracted through Google Earth Engine. Sect. 3.1 describes the nodal control volume, Sect. 3.2 the discharge records and their processing status, and Sect. 3.3 the gauge-independent covariates.

3.1 The nodal control volume

The Liard River joins the Mackenzie River near Fort Simpson, Northwest Territories, at approximately 61.8° N, 121.2° W. Three WSC stations bracket the confluence (Fig. 1c; Table 1): Mackenzie River at Strong Point (10FB006), located upstream on the main stem; Liard River near the mouth (10ED002), located upstream on the tributary; and Mackenzie River at Fort Simpson (10GC001), located downstream of the confluence. Fort Simpson gauges a drainage area of approximately 1 270 000 km², of which approximately 275 000 km² is drained by the Liard River near the mouth (ECCC 2026a; Lammers et al.

2001). Because Strong Point lies downstream of the Mackenzie River near Fort Providence (10FB001), which gauges approximately 980 000 km² (ECCC 2026a; Lammers et al. 2001), the area drained at Strong Point is at least 980 000 km². The ungauged intervening area therefore represents at most about 1% of the downstream basin. The upstream gauges lie approximately 30 km (Strong Point) and 16 km (Liard) from Fort Simpson in a straight line (Table 1). Even at a flow velocity as low as 0.2 m s^{-1}, travel times between the gauges would be less than 2 days, i.e. less than 6 % of the monthly averaging interval. The reach is unregulated, and over the reference period (Sect. 4.5) the median relative closure error is −0.05 %, with regime-specific medians between −0.70 % and +0.10 % (Table 2). This indicates that the systematic net contribution of ungauged inflow and channel storage change to monthly closure is below 1 % of the downstream discharge. These conditions allow the three gauges to define a nodal control volume governed by the discharge mass balance:

$$Q_{\mathrm{FS}}(t) = Q_{\mathrm{SP}}(t) + Q_{\mathrm{LR}}(t) + \varepsilon(t) \tag{1}$$

where $Q_{\mathrm{FS}}(t)$, $Q_{\mathrm{SP}}(t)$, and $Q_{\mathrm{LR}}(t)$ are the monthly mean discharges (m³ s^{-1}) at Fort Simpson, Strong Point, and Liard River near the mouth, respectively, and $\varepsilon(t)$ is the closure residual representing measurement inconsistencies, unmeasured local inflows, changes in channel storage, and discharge estimation or adjustment effects. At monthly resolution, the relatively small intervening drainage area and limited storage changes make this balance suitable for assessing network-level discharge consistency.

Separating the reach terms from the observation errors makes the stochastic content of the residual explicit:

$$\varepsilon(t) = q_{\ell}(t) - \Delta S(t) + e_{\mathrm{FS}}(t) - e_{\mathrm{SP}}(t) - e_{\mathrm{LR}}(t), \qquad e_i(t) = \eta_i(t)\, Q_i^{\mathrm{true}}(t) + \xi_i(t) \tag{2}$$

where $q_{\ell}(t)$ is the net ungauged lateral inflow, $\Delta S(t)$ the change in channel storage over the month, and $e_i(t)$ the observation error at station i, comprising a proportional component $\eta_i(t)$ that scales with discharge, as rating-curve and shift uncertainty does, and an additive component $\xi_i(t)$. The drainage-area, travel-time and median-closure arguments given above constrain the systematic part of $q_{\ell} - \Delta S$ to below 1 % of the downstream discharge at monthly resolution, so the closure residual is dominated by the observation errors. This decomposition is the basis of the tolerance model in Sect. 4.5, the attribution limits in Sect. 4.6 and the independence diagnostic in Sect. 4.8.

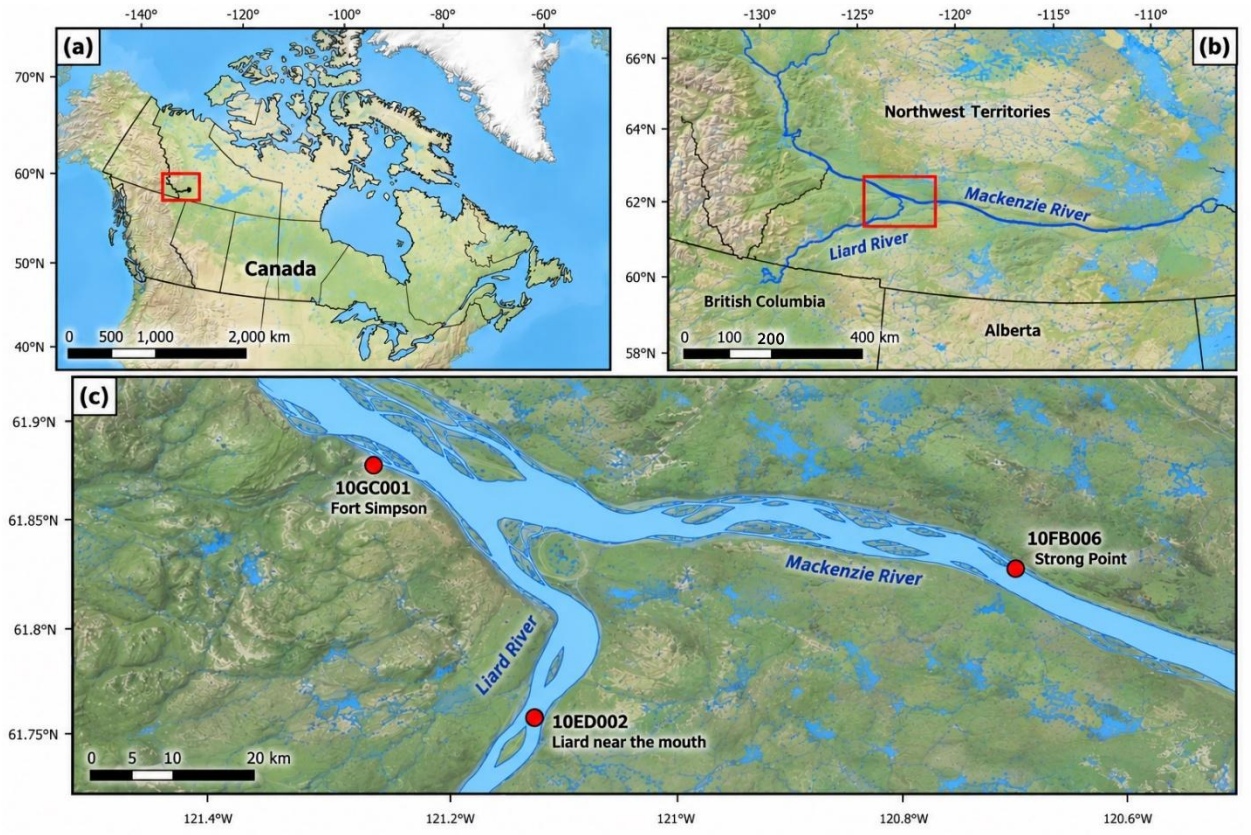


Fig. 1 Study area. (a) Location within Canada. (b) Mackenzie–Liard region; the red box marks the confluence. (c) The three hydrometric stations: Mackenzie River at Strong Point (10FB006), Liard River near the mouth (10ED002) and Mackenzie River at Fort Simpson (10GC001). Background map: Natural Earth shaded relief (naturalearthdata.com)

Strong Point contributes the larger share of the combined upstream discharge (mean 4356 $m^3 s^{-1}$; 63 %), while the Liard contributes substantially to seasonal variability (see Table 1). The median Liard fraction increases from 0.16 in March to 0.53 in June (Fig. 2c), reflecting its snowmelt-dominated regime relative to the more attenuated Mackenzie main stem (Woo and Thorne 2003). Regional hydrology is strongly influenced by a five- to six-month ice-cover season and pronounced river-ice break-up dynamics (Smith 1979; Beltaos and Prowse 2009; de Rham et al. 2020), alongside documented long-term changes in winter baseflow (St. Jacques and Sauchyn 2009; Déry et al. 2016).

3.2 Discharge records

Monthly mean discharge records for the three WSC stations were obtained for January 2000–July 2026, comprising 319 months with no missing values. All discharge data were produced by Environment and Climate Change Canada's Water Survey of Canada (Halliday 2008). Approved historical records for January 2000–December 2024 were obtained from the National Water Data Archive (HYDAT; ECCC 2026a; release of 17 July 2026, accessed August 2026), and provisional discharge estimates for January 2025–July 2026 were obtained through the WSC Wateroffice hydrometric data service (https://wateroffice.ec.gc.ca/; ECCC 2026b; accessed August 2026). For the approved period, monthly means are those reported in HYDAT, which are calculated from approved daily mean discharges. For the provisional period, monthly means were calculated as the arithmetic mean of the provisional daily mean

discharge values available in each calendar month. The dataset therefore includes two record statuses: approved values for 2000–2024, which have completed WSC quality control and approval, and provisional values from January 2025 onward, which have not. Provisional daily values were not available for every day of some months (November 2025 at Fort Simpson and Liard, May 2025 at Strong Point, and May 2026 at Liard and Strong Point), and several ice-season monthly values (e.g. December 2025–April 2026 at Fort Simpson) are round numbers, suggesting constant or estimated daily values during ice-affected periods. These features are consistent with the provisional status of the records and should be considered when interpreting the provisional winter closure errors (Sect. 5.6). This distinction is important for interpreting discharge consistency, as Sect. 5.6 demonstrates that the two periods exhibit different closure characteristics.

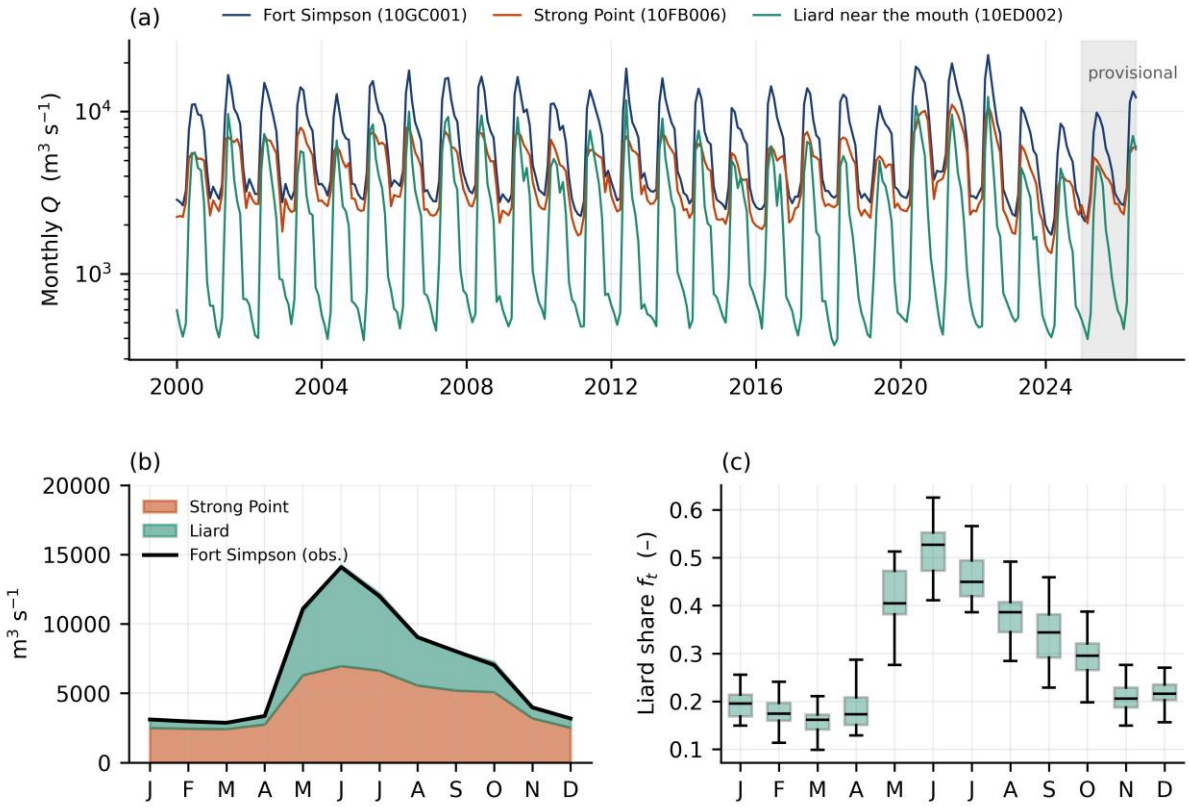


Fig. 2 Discharge at the three gauges. (a) Monthly mean discharge, January 2000–July 2026; gray shading marks provisional records. (b) Median annual regime, with the Strong Point and Liard contributions stacked against Fort Simpson. (c) Monthly climatology of the Liard partition fraction f_t

At WSC hydrometric stations, discharge is generally estimated from continuous water-level observations using stage–discharge relationships (rating curves) developed from periodic field measurements, with shifts or alternative estimation procedures applied where necessary (Rainville et al. 2002; Hamilton and Moore 2012). Daily mean discharges are derived from these higher-frequency estimates, and historical monthly means are calculated from the daily means. The distinction between open-water and ice-affected conditions is central to this process. Under open-water conditions, discharge is computed from the rating curve, adjusted by shifts where field measurements indicate a change in the stage–discharge relationship. When ice is present, backwater caused by the ice cover, frazil or anchor ice raises the water level for a given

discharge, so that discharge derived from the open-water rating curve is no longer valid. Ice-affected discharge is therefore estimated from winter field measurements, shift corrections and interpolation between measurements, supported by hydrographic and climatic information, including records from neighboring stations (Pelletier 1989; Rainville et al. 2002; Hamilton and Moore 2012; Gharari et al. 2024). Provisional records remain subject to quality-control and approval procedures, during which discharge estimates may be revised on the basis of additional field measurements, rating-curve adjustments and the identification of ice-affected periods. Because the onset and end of ice effects are often identified retrospectively, from water-level behavior during freeze-up and break-up and from subsequent winter measurements, these revisions may be applied with considerable delay. This is particularly relevant at cold-region stations such as those examined here, where ice affects the stage–discharge relationship for five to six months each year. Consequently, differences between approved and provisional records may reflect differences in processing and quality-control status rather than changes in measurement accuracy alone.

HYDAT daily records also carry data-quality symbols, including the symbol "B" for discharge estimated under ice-affected (backwater) conditions (de Rham et al. 2020). These symbols provide information on ice conditions that is independent of the satellite-based regime classification used here (Sect. 4.4) and could be used both to interpret winter closure discrepancies and to evaluate the NDVI-based classification. They were not used in the present analysis, which is based on monthly means, and their incorporation is identified as a priority for further work (Sect. 6.5).

Table 1 Hydrometric stations defining the Mackenzie–Liard nodal control volume. Mean discharge and upstream flow contributions are calculated over the full 319-month period

Code	Station name	Latitude	Longitude	Role	Mean Q ($m^3 s^{-1}$)
10FB006	Mackenzie River at Strong Point	61.8165° N	120.7917° W	Upstream main-stem node	4356 (63 %)
10ED002	Liard River near the mouth	61.7426° N	121.2278° W	Upstream tributary node	2520 (37 %)
10GC001	Mackenzie River at Fort Simpson	61.8685° N	121.3589° W	Downstream closure node	6875

3.3 Earth-observation and reanalysis covariates

Two gauge-independent datasets were extracted at each station location using Google Earth Engine (Gorelick et al. 2017). ERA5-Land surface runoff and total precipitation (Muñoz-Sabater et al. 2021) were used as external hydrometeorological covariates to assess discharge consistency. ERA5-Land provides hourly fields at a horizontal resolution of 0.1° (approximately 9 km) from 1950 onward; monthly means were extracted at the grid cell containing each gauge for January 2000–July 2026. These values therefore describe conditions at the gauge location rather than over the contributing drainage basins, which range from approximately 275 000 to 1 270 000 km² and span several climatic zones. In addition, the surface-runoff variable excludes subsurface runoff. ERA5-Land runoff is therefore not expected to be representative of basin-integrated discharge and is used here only as contextual information. MODIS 16-day vegetation indices at 250 m resolution (MOD13Q1; Didan 2015; Huete et al. 2002) were extracted at the pixel containing each gauge and aggregated to monthly NDVI and used as an independent indicator of seasonal surface conditions. Low NDVI values generally correspond to snow-covered or dormant vegetation, whereas seasonal increases and decreases reflect vegetation green-up and senescence, respectively. These

transitions provide contextual information on seasonal conditions associated with river-ice break-up and freeze-up. Sect. 5.3 evaluates the contribution of both datasets to anomaly detection relative to network-based discharge constraints.

4 The NCQC framework

4.1 Framework design and methodological novelty

The proposed Network-Constrained Quality Control (NCQC) framework exploits hydrometric network redundancy to detect and attribute discharge inconsistencies that may remain undetected by conventional station-based QC. Its principal novelty lies in integrating physical mass-balance constraints, hydrological regime classification, temporal and spatial anomaly detection, and probabilistic gauge attribution within a unified framework. The algorithm combines seven sequential and complementary tests (T0–T6), ranging from record-provenance assessment and physical admissibility to change-point detection and network-independence diagnostics (Fig. 3).

Four methodological contributions distinguish NCQC. First, regime-conditioned closure tolerances are estimated robustly from a reference period, accounting for seasonal differences in discharge variability rather than imposing fixed thresholds. Second, combined temporal and spatial consistency scores support identification of the most likely anomalous gauge while explicitly quantifying attribution uncertainty. Third, an independence diagnostic identifies periods of unusually small closure residuals that may indicate dependence among published discharge estimates, preventing such agreement from being interpreted automatically as measurement accuracy. Together, these components extend conventional discharge QC beyond anomaly detection to include gauge attribution, record provenance and assessment of the reliability of the network constraint itself. Fourth, the framework is accompanied by analytical results that make its own limits explicit: the identifiability of the attribution problem (Eq. 5), a closed-form expression for the smallest detectable error at each station and regime (Eq. 8), and the dependence of closure variance on cross-station error correlation (Eq. 7), which turns the independence diagnostic from a heuristic into a statement about the error covariance.

4.2 T0: provenance signature

NCQC first examines the numerical precision of reported discharge values as an indicator of record provenance. Approved WSC discharge values are conventionally reported to three significant digits, whereas provisional values may retain greater numerical precision from daily averaging. NCQC calculates a six-month running mean of the significant-digit count at each station and identifies sustained departures from the archival baseline. This provides a metadata-independent indicator of potential changes in record status or processing. Exact value repetition and sequences of identical terminal digits are additional potential signatures but are not implemented here.

4.3 T1: physical admissibility

NCQC applies two physical admissibility checks: discharge values must be positive, and downstream discharge should not fall below either upstream contribution under the assumed negligible intervening losses and storage changes. Months violating these conditions receive an N3 flag. Such violations indicate a physically inconsistent nodal balance and warrant immediate investigation. However, they do not uniquely identify the erroneous station, because an ordering violation may result from an overestimated

upstream discharge, an underestimated downstream discharge, or unrepresented reach processes. Station attribution therefore requires additional evidence from T4.

4.4 T2: hydrological regime classification

Each month is classified using the network-mean MODIS NDVI and its month-to-month change. Months with $|\Delta\text{NDVI}| \geq 0.15$ are classified as Transition; otherwise, months with NDVI < 0.15 are classified as Ice-affected, and those with NDVI ≥ 0.15 as Open-water. These criteria yield 145 ice-affected, 91 transition and 83 open-water months, broadly following the expected seasonal cycle. The classification uses satellite-derived surface conditions rather than calendar months or WSC ice-condition flags, providing an external seasonal indicator. However, NDVI-based regimes are proxies for hydrological conditions rather than direct observations of river-ice cover. Landsat-based river-ice phenology products (Yang et al. 2020; Wang and Feng 2024) could provide a more direct alternative. Comparison with WSC ice-condition symbols (Sect. 3.2) would provide an independent evaluation of this classification.

4.5 T3: regime-conditioned closure test

The form of the tolerance follows from the error decomposition in Eq. (2). If, within a regime, the proportional errors η_i are mutually independent with standard deviations γ_i and the additive errors ξ_i have standard deviations s_i, then

$$\mathrm{Var}[r_t] = \sum_i \gamma_i^2 \left(Q_i^{\mathrm{true}}\right)^2 + \sum_i s_i^2 = k_\rho^2 Q_{\mathrm{FS}}^2 + s_{0,\rho}^2, \qquad k_\rho^2 = \sum_i \gamma_i^2 \, \varphi_i^2 \tag{3}$$

where $\varphi_i = Q_i/Q_{\mathrm{FS}}$ is the flow share of station i, and $s_{0,\rho}$ is the quadrature sum of the additive error scales, estimated in practice as one quarter of the regime-specific robust residual scale. Because the flow shares vary only moderately within a regime (Table 1, Fig. 2c), k_ρ is approximately constant, so the residual scale grows linearly with discharge at high flow and approaches a floor $s_{0,\rho}$ at low flow. Equation (4) adopts the linear form $k_\rho Q_{\mathrm{FS}} + s_{0,\rho}$ rather than the square root of Eq. (3); over the observed range of downstream discharge the linear form is about 10 % to 40 % wider, so the tolerance applied here is conservative by construction and the flag counts reported in Sect. 5 are lower than a quadrature tolerance would produce.

The closure residual defined in Eq. (1) (denoted r_t below, so that $r_t \equiv \varepsilon_t$) is standardized using a regime-specific heteroscedastic tolerance:

$$r_t = Q_{\mathrm{FS}} - (Q_{\mathrm{SP}} + Q_{\mathrm{LR}}), \qquad \sigma_t = k_\rho Q_{\mathrm{FS}} + s_{0,\rho}, \qquad Z_t = \frac{r_t - b_\rho Q_{\mathrm{FS}}}{\sigma_t} \tag{4}$$

where ρ denotes the hydrological regime, k_ρ is the proportional error coefficient, $s_{0,\rho}$ is the additive tolerance floor, defined as one quarter of the regime-specific robust residual scale, and b_ρ represents the regime-specific bias. Parameters are estimated robustly using the median absolute deviation over the reference period (2000–2024), excluding 2016, 2018 and 2019 to limit the influence of identified anomalous episodes. The fitted parameters are reported in Table 2. The proportional coefficient is lowest during ice-affected months (1.11%) and highest during transition months (1.96%), reflecting differences in closure variability. Months with $|Z_t| > 2$ receive an N1 (suspect) flag, while those with $|Z_t| > 3$ receive an N2 (fail) flag.

Table 2 Hydrological regime definitions and robustly estimated tolerance parameters for Eq. (4). Parameters are fitted over 2000–2024, excluding 2016, 2018 and 2019; n denotes the number of reference months

Regime	NDVI criterion	k_ρ (%)	$s_{0,\rho}$ (m³ s⁻¹)	b_ρ (%)	n
Ice-affected	NDVI < 0.15, $\lvert\Delta\mathrm{NDVI}\rvert$ < 0.15	1.11	8.0	+0.10	120
Transition	$\lvert\Delta\mathrm{NDVI}\rvert$ ≥ 0.15	1.96	47.8	−0.46	77
Open-water	NDVI ≥ 0.15, $\lvert\Delta\mathrm{NDVI}\rvert$ < 0.15	1.39	40.8	−0.70	67

4.6 T4: attribution

Because Eq. (1) provides a single constraint involving three discharge records, closure residuals alone cannot identify the responsible station. NCQC addresses this ambiguity using two complementary diagnostics. The temporal-continuity score T_s measures deviations in each station's log-discharge from its immediate temporal neighbors, standardized by the corresponding regime-specific robust scale. The spatial-anomaly score S_s measures deviations from a regression prediction based on anomalies at the other two stations, fitted over the reference period.

This indeterminacy can be stated precisely. Collecting the station errors of Eq. (2) into a vector and neglecting the reach terms,

$$r_t = \mathbf{a}^\top \mathbf{e}_t, \qquad \mathbf{a} = (1, -1, -1)^\top, \qquad \mathbf{e}_t = (e_{\mathrm{FS}}, e_{\mathrm{SP}}, e_{\mathrm{LR}})^\top, \qquad \dim \ker(\mathbf{a}^\top) = 2 \tag{5}$$

so the closure residual is a single linear functional of a three-dimensional error vector, and the kernel of a⊤ is two-dimensional. Two consequences follow. First, any error combination lying in that kernel, in particular any triple satisfying $e_{\mathrm{FS}} = e_{\mathrm{SP}} + e_{\mathrm{LR}}$, leaves r_t unchanged and is invisible to T3 however large its components. Second, an error added to the downstream record and an error of the same size subtracted from an upstream record generate identical residuals, so closure cannot distinguish them. Attribution therefore requires information from outside the mass balance, which is the role of the two scores defined below, and the ambiguity that remains is a property of the constraint rather than a deficiency of the scores. Sect. 5.7 quantifies it: correct attribution is weakest at the downstream gauge, as Eq. (5) implies.

Both diagnostics are combined into the implausibility score defined in Eq. (6), with an additional penalty assigned to stations implicated in T1 ordering violations.

$$\Lambda_s = 0.6\,|T_s| + 0.4\,|S_s|, \qquad \mathrm{source} = \underset{s}{\mathrm{argmax}}\,\Lambda_s \tag{6}$$

The weights were set a priori, giving moderately greater weight to temporal continuity, and were not fitted to the flagged episodes; their influence on attribution is evaluated in Sect. 5.7. Attribution is assigned to the station with the highest score, while its share of the total score expresses relative attribution strength. Values near 1/3 indicate weak discrimination among stations and should not be interpreted as calibrated probabilities. Episodes with a dominant share below 0.40 are reported as having inconclusive attribution. This threshold is supported by the synthetic experiments (Sect. 4.10), in which attributions with a dominant share below 0.40 were correct in 50 % of detected trials, compared with 65–77 % for higher shares.

4.7 T5: episode aggregation and change-point detection

Consecutive flagged months, allowing a single-month interruption, are grouped into episodes. Station attribution is based on scores accumulated across each episode rather than individual months, improving temporal consistency. In parallel, a two-sided cumulative sum (CUSUM) test is applied to Z_t using a reference value of $k = 0.5$ and a decision interval of $h = 10$. This identifies persistent shifts or gradual changes that may remain below individual monthly thresholds. Such changes can be consistent with rating-curve adjustments, datum shifts or other systematic alterations in discharge estimation, although their causes require independent verification. The decision interval was set from average-run-length (ARL) considerations rather than by tuning. Under the Siegmund approximation, a two-sided CUSUM with $k = 0.5$ and $h = 10$ has an in-control ARL of the order of 7×10^4 months, so no false alarm is expected over a 319-month record, while a sustained shift of one standard deviation is signalled after approximately 20 months and a shift of about three standard deviations, the standardized size of the 2018–2019 open-water offset (Sect. 5.2), after approximately four months.

4.8 T6: network independence diagnostic

The final test assesses whether the discharge records provide sufficiently independent information for network-based QC. WSC procedures may incorporate upstream or downstream observations when estimating discharge through overrides (Gharari et al. 2024). Such adjustments can reduce closure residuals without necessarily reducing the underlying measurement uncertainty. NCQC therefore evaluates the proportion of ice-affected months satisfying $|r_t/Q_{\mathrm{FS}}| < 0.5\%$ within a moving window of twelve consecutive ice-affected months. An N5 flag is raised when this proportion exceeds 0.70 over a sustained window. The flag identifies unusually strong agreement that may reflect dependence among published estimates. It does not establish that overrides occurred or that the discharge values are incorrect; rather, it indicates that closure alone may provide limited independent evidence of their accuracy.

The rationale for this diagnostic is that the closure variance depends not only on the size of the individual errors but on their correlation across stations. From Eq. (2), neglecting the reach terms,

$$\mathrm{Var}[r_t] = \sigma_{\mathrm{FS}}^2 + \sigma_{\mathrm{SP}}^2 + \sigma_{\mathrm{LR}}^2 - 2\kappa_{\mathrm{FS,SP}}\,\sigma_{\mathrm{FS}}\sigma_{\mathrm{SP}} - 2\kappa_{\mathrm{FS,LR}}\,\sigma_{\mathrm{FS}}\sigma_{\mathrm{LR}} + 2\kappa_{\mathrm{SP,LR}}\,\sigma_{\mathrm{SP}}\sigma_{\mathrm{LR}} \quad (7)$$

where σ_i is the standard deviation of e_i and κ_{ij} the correlation between the errors at stations i and j. The expression shows that $\mathrm{Var}[r_t]$ can be made arbitrarily small by positive correlation between the downstream error and the upstream errors, with no reduction whatsoever in the individual σ_i. Closure agreement is therefore not by itself evidence of accuracy, and a sustained collapse of the residual scale is as consistent with growing error dependence as with improving measurement. T6 detects the collapse; Sect. 6.2 evaluates Eq. (7) for the observed collapse and reports the correlation it would require.

4.9 Flag taxonomy

NCQC assigns six network-level flags (N0–N5; Table 3) to complement existing WSC quality symbols. Flags N0–N3 describe monthly discharge consistency, whereas N4 and N5 identify persistent closure changes and potential loss of network independence, respectively. N4 and N5 are reported alongside the monthly flags rather than replacing them. This classification provides additional information on discharge reliability and helps prioritize records for further review.

Table 3 NCQC flag taxonomy, diagnostic criteria, interpretation and recommended actions. Network-level flags complement existing station-level quality symbols

Flag	Condition	Interpretation	Action
N0	$\lvert Z_t \rvert \leq 2$	Consistent with network constraints	No additional network-based review
N1	$2 < \lvert Z_t \rvert \leq 3$	Suspect closure anomaly	Review and annotate
N2	$\lvert Z_t \rvert > 3$	Significant network inconsistency	Prioritize for review before approval
N3	T1 violation	Physical admissibility violation	Investigate and verify affected records
N4	CUSUM excursion	Persistent shift or drift in closure	Examine rating and processing history
N5	sustained $\lvert r_t / Q_{\mathrm{FS}} \rvert < 0.5$ % in more than 70% of the T6 window	Potential loss of independent network information	Assess record dependence and provenance

4.10 Skill assessment by synthetic error injection

Because the true error state of the discharge records is unknown, NCQC detection and attribution performance are evaluated through synthetic error injection. Four error types are considered: sustained multiplicative bias (2–6 months), single-month spikes, persistent steps (6–12 months), and frozen values repeated from the preceding month. Errors ranging from 2% to 50% are introduced at randomly selected stations and starting months within the reference record. The tolerance parameters of Eq. (4) are fitted once using the uncorrupted reference record and held fixed throughout the experiments, preventing recalibration to injected errors. A total of 4327 trials were conducted. The probability of detection (POD) is defined as the proportion of trials with at least one flagged month within the injected period. Because approximately 8 % of reference months are already flagged before any error is injected, a trial may register a detection through a pre-existing flag that is unrelated to the injected error. POD is therefore also reported for newly raised flags only, i.e. for trials in which at least one month within the injected window is flagged that was not flagged in the uncorrupted record. The false-alarm rate is estimated from the corresponding flag frequency in the uncorrupted reference record, referred to as the background flag rate. Because the reference record may itself contain genuine but undocumented errors, the background flag rate is an upper bound on the false-alarm rate; it cannot be verified against observations without independent knowledge of the true discharge, such as WSC revision histories or field-measurement records. Receiver operating characteristic (ROC) curves use the maximum $\lvert Z_t \rvert$ within each injected window as the detection score. For a 10% error, this approach yields a POD of 0.88, compared with 0.89 using the window-based flag criterion. To quantify the added value of the network constraint, the same trials were also screened with a conventional station-level QC baseline applied to each record independently. The baseline combines (i) a climatological test on the log-discharge anomaly relative to the reference-period median and robust scale for the same calendar month, (ii) a temporal-continuity (rate-of-change) test equivalent to the T4 temporal score, and (iii) a flat-line test for exactly repeated monthly values. A common threshold for tests (i) and (ii) (4.5 robust standard deviations) was chosen so that the combined baseline produced the same 8.0 % background flag rate as NCQC, providing a like-for-like comparison of detection performance.

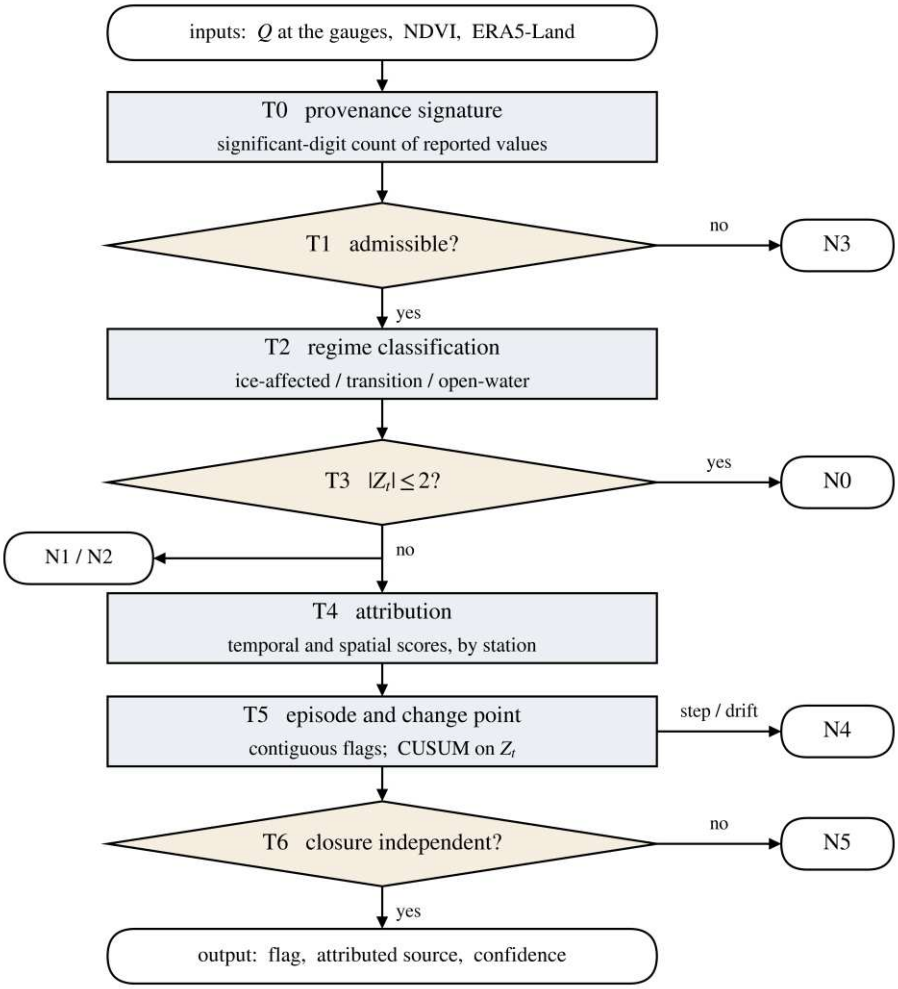


Fig. 3 NCQC decision framework, from provenance assessment (T0) to the network-independence diagnostic (T6). N0–N5 are the network flags of Table 3

The simulated detection rates can be compared with the detection limit implied by the tolerance model. A relative error δ at station s displaces the closure residual by δQ_s and the standardized statistic by $\delta Q_s/\sigma_t$. Requiring that displacement to exceed the flag threshold $z_c = 2$ with probability $1 - \beta$ gives

$$\delta_{\min}(s,\rho) = \left(z_c + z_{1-\beta}\right)\frac{\sigma_t}{Q_s} \approx \left(z_c + z_{1-\beta}\right)k_\rho \frac{Q_{\mathrm{FS}}}{Q_s} \tag{8}$$

with $z_{1-\beta} = 0.84$ for a power of 0.8. Because $\sigma_t \approx k_\rho Q_{\mathrm{FS}}$, the smallest detectable error at a station scales with the regime tolerance coefficient and inversely with that station's share of the flow. Evaluated with the regime-median flow shares and discharges of the reference period and the parameters of Table 2, using the approximate form of Eq. (8), which neglects the additive tolerance floor, $\delta_{\min}$ is 3.1 %, 3.9 % and 17.2 % at Fort Simpson, Strong Point and the Liard respectively under ice, 3.9 %, 6.6 % and 9.6 % in open water, and 5.6 %, 8.9 % and 15.1 % during transition months. Retaining the floor raises these limits to 3.9 %, 4.8 % and 20.8 % under ice, 5.1 %, 8.7 % and 12.6 % in open water, and 7.0 %, 11.8 % and 18.6 % during transition months. These values are compared with the simulated detection rates in Sect. 5.7. Both forms are computed by the script 06_analytical_checks.py in Online Resource 1.

5 Results

5.1 Mass-balance closure at the confluence

The three-gauge mass balance exhibits strong agreement over the 319-month record (Fig. 4). The coefficient of determination between Q_{FS} and $Q_{\mathrm{SP}} + Q_{\mathrm{LR}}$ is 0.9953, while the Nash–Sutcliffe efficiency is 0.9952. The median relative closure error is 0.00%, and its robust scale over the reference period is 1.67%. This closure variability is substantially smaller than published estimates of individual discharge uncertainty, typically around 10% under open-water conditions and potentially higher during ice-affected periods (Pelletier 1989; Coxon et al. 2015; Kiang et al. 2018; Westerberg et al. 2016; McMahon et al. 2025). Such agreement may reflect relatively accurate discharge estimates, correlated errors among stations, or both. Shared information used during discharge estimation could reduce closure residuals without a corresponding reduction in individual measurement uncertainty. This possibility is examined in Sect. 6.2. The reference-period closure residuals exhibit a leptokurtic distribution, characterized by a pronounced central peak and heavier tails than a Gaussian distribution with the same robust scale (Fig. 4c). This pattern is consistent with a mixture of tightly constrained and more variable discharge estimates, potentially associated with temporary shifts and overrides described by Gharari et al. (2024). However, the residual distribution alone cannot establish the underlying processing mechanism. Its non-Gaussian structure also indicates that a conventional Gaussian error model may inadequately represent both small and extreme closure deviations. The departure from normality can be quantified directly from the flag rates. Under a Gaussian model the thresholds $|Z_t| > 2$ and $|Z_t| > 3$ would flag 4.6 % and 0.3 % of months, whereas the observed background rates on the reference record are 8.0 % and 3.0 % (Sect. 4.10), larger by factors of 1.8 and 11 respectively. The tolerance model of Eq. (4) therefore reproduces the scale of the residual but understates the frequency of extreme closure deviations, which is why the thresholds are treated as operating points tied to a known background flag rate rather than as significance levels.

5.2 Regime dependence and temporal non-stationarity

Closure variability differs substantially across hydrological regimes (Fig. 5). Over the reference period, the robust scale of the absolute residual increases from 32 $m^3 s^{-1}$ during ice-affected months to 163 $m^3 s^{-1}$ under open-water conditions, supporting the proportional tolerance formulation in Eq. (4). However, the relative tolerance is lowest during ice-affected months (1.11%) and highest during transition months (1.96%). This contrasts with the greater measurement uncertainty generally expected under ice and suggests that closure consistency does not necessarily reflect individual discharge accuracy, as discussed in Sect. 6.2.

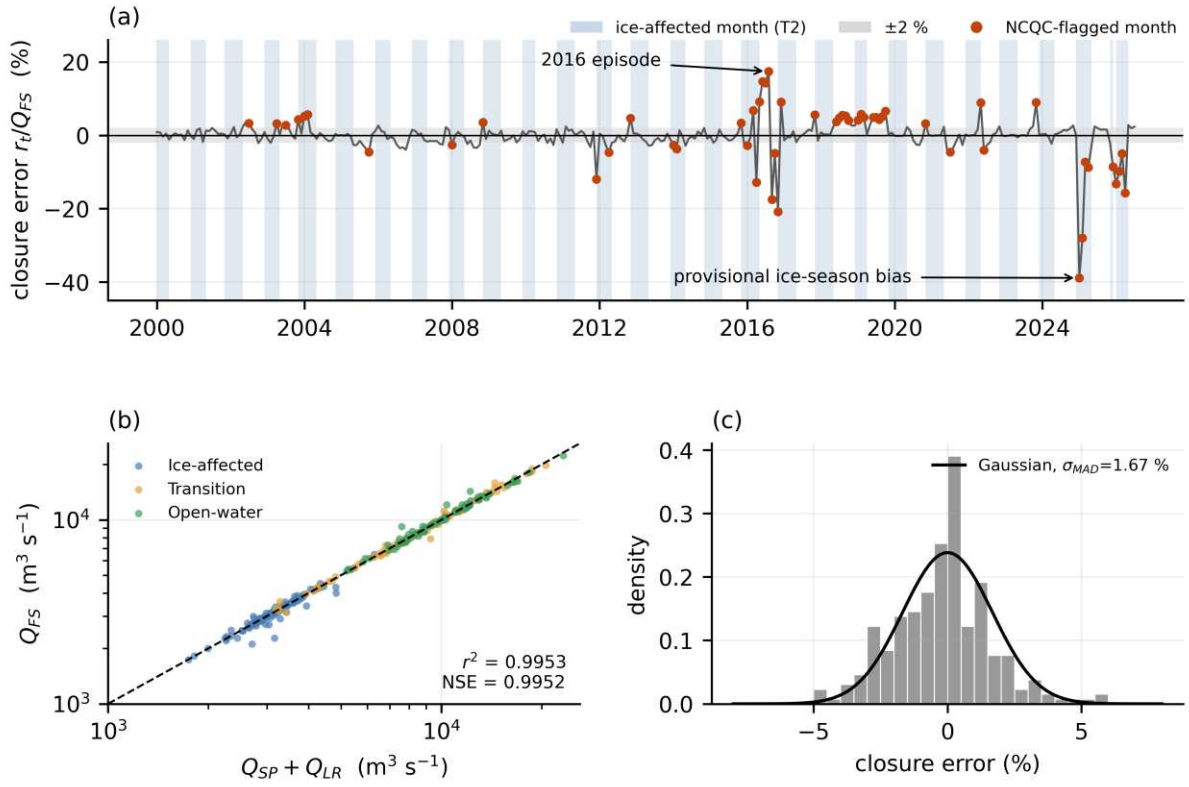


Fig. 4 Mass-balance closure. (a) Relative closure error r_t/Q_{FS}, 2000–2026; gray band ±2 %, blue shading ice-affected months, red points NCQC flags. (b) Downstream discharge versus the sum of upstream discharges, by regime. (c) Reference-period closure errors compared with a Gaussian distribution of equal robust scale

The closure error also exhibits pronounced temporal non-stationarity (Fig. 5c), with five distinct epochs. During 2000–2009 and 2010–2017, relative closure errors remained centered near zero, with robust scales of approximately 1–2%. In 2018–2019, the median open-water closure error shifted to +4.79%, with a robust scale of only 0.76%, indicating a persistent systematic offset rather than increased random variability. Such behavior is consistent with a change in discharge estimation or rating-curve application, although its cause cannot be established from closure alone. During 2020–2024, ice-season closure variability decreased markedly, with a robust scale of 0.20% compared with 1.46% during 2000–2019; 85% of ice-affected months fell within ±0.5%. From January 2025, the provisional ice-season records exhibited a median closure error of −9.8% and a robust scale of 7.1%, indicating a substantial deterioration in network consistency.

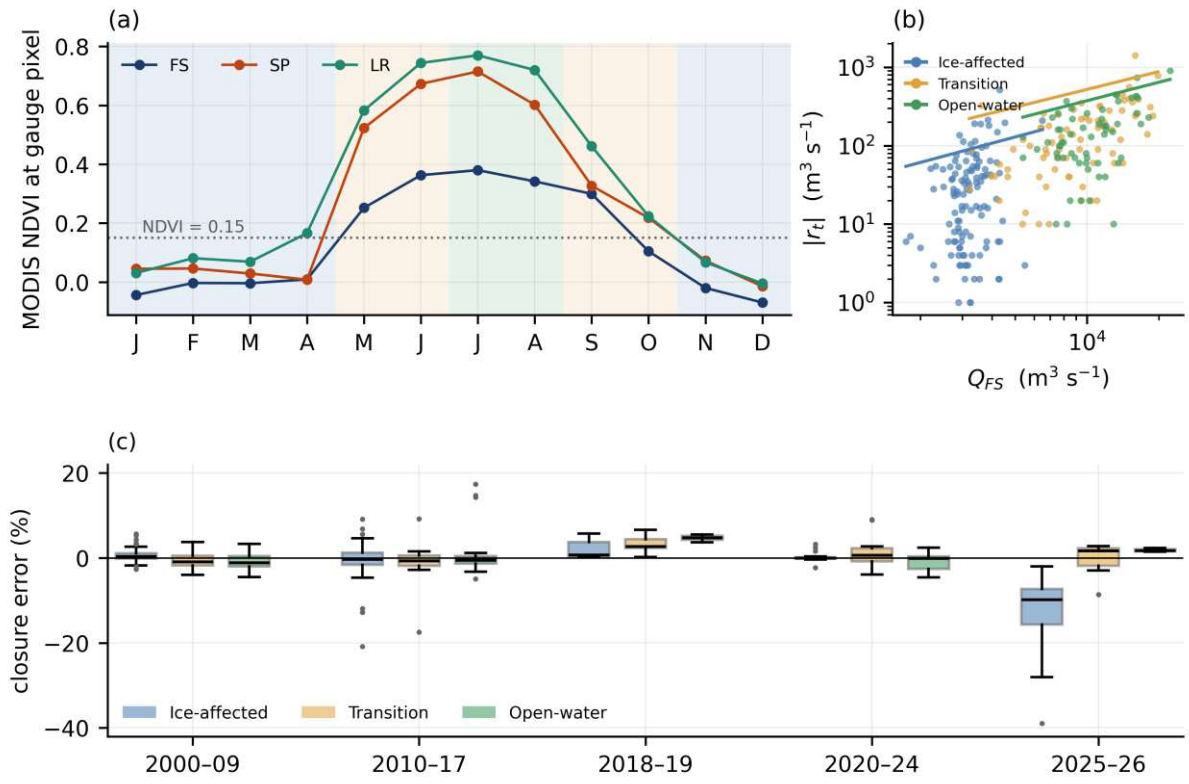


Fig. 5 Regime dependence and non-stationarity of closure. (a) NDVI climatology at the gauges, shaded by modal regime. (b) Absolute closure residual versus downstream discharge, with the $\pm 2\sigma$ tolerance of Eq. (4) for each regime. (c) Relative closure error by epoch and regime

5.3 Contribution of individual QC constraints

Fig. 6 compares the information provided by the different QC constraints. Mass-balance closure shows strong agreement between upstream and downstream discharge (Fig. 4). Inter-station log-anomaly correlations are highest between Fort Simpson and Strong Point (r = 0.91) and lower for the Liard River (r = 0.69), while temporal persistence explains 59% of downstream anomaly variance. In contrast, the external covariates provide limited explanatory power. Correlations between ERA5-Land runoff anomalies at individual gauge pixels and discharge anomalies range from 0.12 to 0.26. At Fort Simpson, runoff explains only 7% of discharge anomaly variance, compared with 3% for precipitation and 1% for NDVI. These weak relationships partly reflect the use of point-scale reanalysis estimates rather than basin-integrated hydrological variables. ERA5-Land may therefore provide supporting meteorological context for flagged anomalies but is not a reliable substitute for neighboring discharge observations at this spatial scale. Network mass balance and inter-station consistency provide the strongest QC constraints, while temporal persistence offers complementary information for anomaly attribution. The limited explanatory power of the external covariates supports their use as supplementary diagnostics rather than primary discharge references.

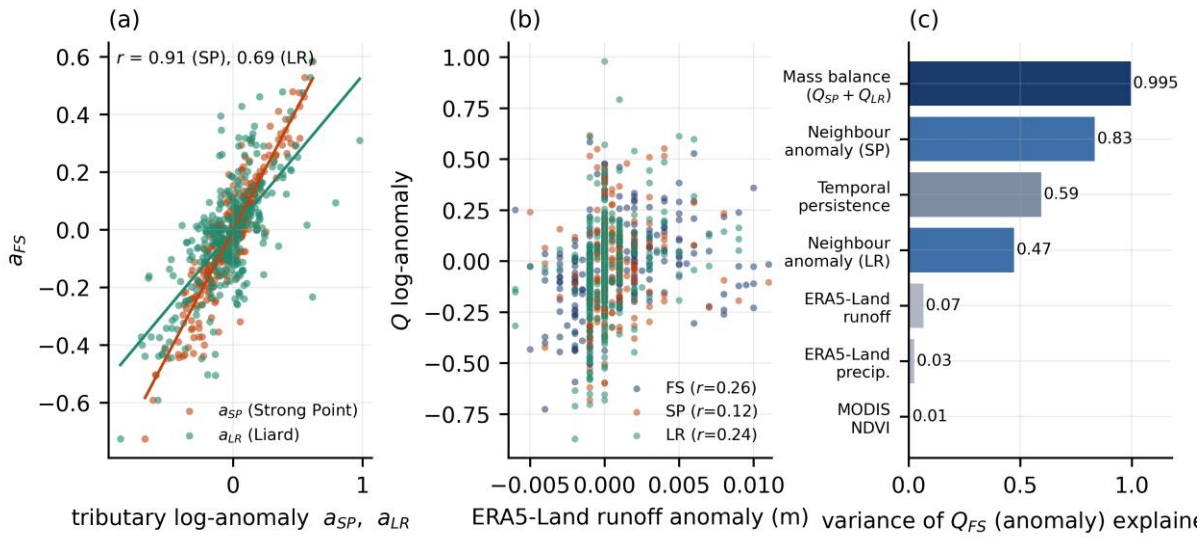


Fig. 6 Information content of candidate QC constraints. (a) Inter-station log-anomaly relationships at Fort Simpson. (b) Discharge log-anomalies versus ERA5-Land runoff anomalies. (c) Share of downstream anomaly variance explained by each constraint

5.4 Network flags and detected episodes (2000–2026)

Applying NCQC to the full 319-month record yields 265 N0, 22 N1, 29 N2 and 3 N3 classifications. Overall, 54 months (16.9%) are flagged for network inconsistency. These months form 22 episodes (Fig. 7), comprising 13 single-month and nine multi-month episodes. The latter account for most of the persistent anomalies (Table 4). The CUSUM analysis reveals additional temporal structure (Fig. 7c). A pronounced positive excursion occurs in 2016, followed by a sustained increase during 2018–2019, reaching 41.4 and corresponding to the +4.79% open-water closure offset. This shift develops largely through moderate monthly deviations that are less evident from individual threshold exceedances. From January 2025, the CUSUM declines sharply, reaching −80.7 in April 2026 before recovering slightly to −75.1 by July 2026. Gauge attribution remains uncertain for many episodes (Fig. 7d). Eleven of the 22 episodes have a dominant attribution share below 0.40, only slightly above the non-discriminating value of 1/3. Four episodes exceed 0.55: November 2008 (Liard, 0.64), December 2011 (Strong Point, 0.56), November 2020 (Liard, 0.58) and January–April 2025 (Strong Point, 0.60). The latter also includes two T1 violations, providing additional evidence of physical inconsistency, although not definitive identification of the responsible gauge.

Table 4 Multi-month NCQC episodes detected during 2000–2026. Attribution share is the dominant station's proportion of the total implausibility score (1/3 indicates no discrimination among stations). Maximum $|Z_t|$ denotes the largest absolute standardized closure statistic within each episode.

Period	Months	Attributed to	Confidence	max $\lvert Z_t \rvert$	Character
Nov 2003–Feb 2004	4	Liard[a]	0.39	4.1	Positive ice-season anomaly
Jan–Feb 2014	2	Strong Point[a]	0.37	2.8	Negative ice-season anomaly
Nov 2015–Dec 2016	14	Strong Point	0.43	16.0	Sign-reversing anomaly with N3 violation
Jun–Oct 2018	5	Strong Point[a]	0.39	3.4	Persistent ice-free-season anomaly
Jan–Mar 2019	3	Liard[a]	0.35	4.1	Positive ice-season anomaly
Jun–Oct 2019	5	Strong Point[a]	0.39	3.1	Persistent ice-free-season anomaly
May–Jun 2022	2	Strong Point[a]	0.35	4.2	Freshet-transition anomaly
Jan–Apr 2025	4	Strong Point	0.60	26.8	Provisional ice-season anomaly with two N3 violations
Dec 2025–Apr 2026	5	Liard[a]	0.35	11.8	Provisional ice-season anomaly

a Dominant share below 0.40; attribution is inconclusive (Sect. 4.6)

5.5 Case study 1: the 2015–2016 episode

The longest detected episode spans 14 months, from November 2015 to December 2016 (Fig. 8a, c). Its most distinctive feature is the reversal of the closure error. From May to August 2016, downstream discharge exceeded the sum of upstream contributions by 9.2%, 14.7%, 14.3% and 17.4%, respectively. The error subsequently reversed to −17.5% in September and −20.8% in November. This pronounced sign reversal is inconsistent with a simple, constant rating-curve offset and suggests time-varying inconsistencies in one or more discharge records. In November 2016, Strong Point reported 4110 $m^3 s^{-1}$, exceeding the 4000 $m^3 s^{-1}$ recorded downstream at Fort Simpson, even before accounting for the Liard contribution. This T1 violation indicates a physically inconsistent discharge configuration under the assumed negligible intervening losses and storage changes. Between September and November, Strong Point discharge remained relatively elevated (5190, 5330 and 4110 $m^3 s^{-1}$), while Fort Simpson declined from 7900 to 6900 and then 4000 $m^3 s^{-1}$. Strong Point discharge subsequently decreased by approximately half between November and December. The episode-level attribution assigns the highest implausibility share to Strong Point (0.43), although this provides only limited discrimination among stations. The episode therefore identifies a substantial network inconsistency within the approved discharge record, warranting further examination of the original measurements and processing history.

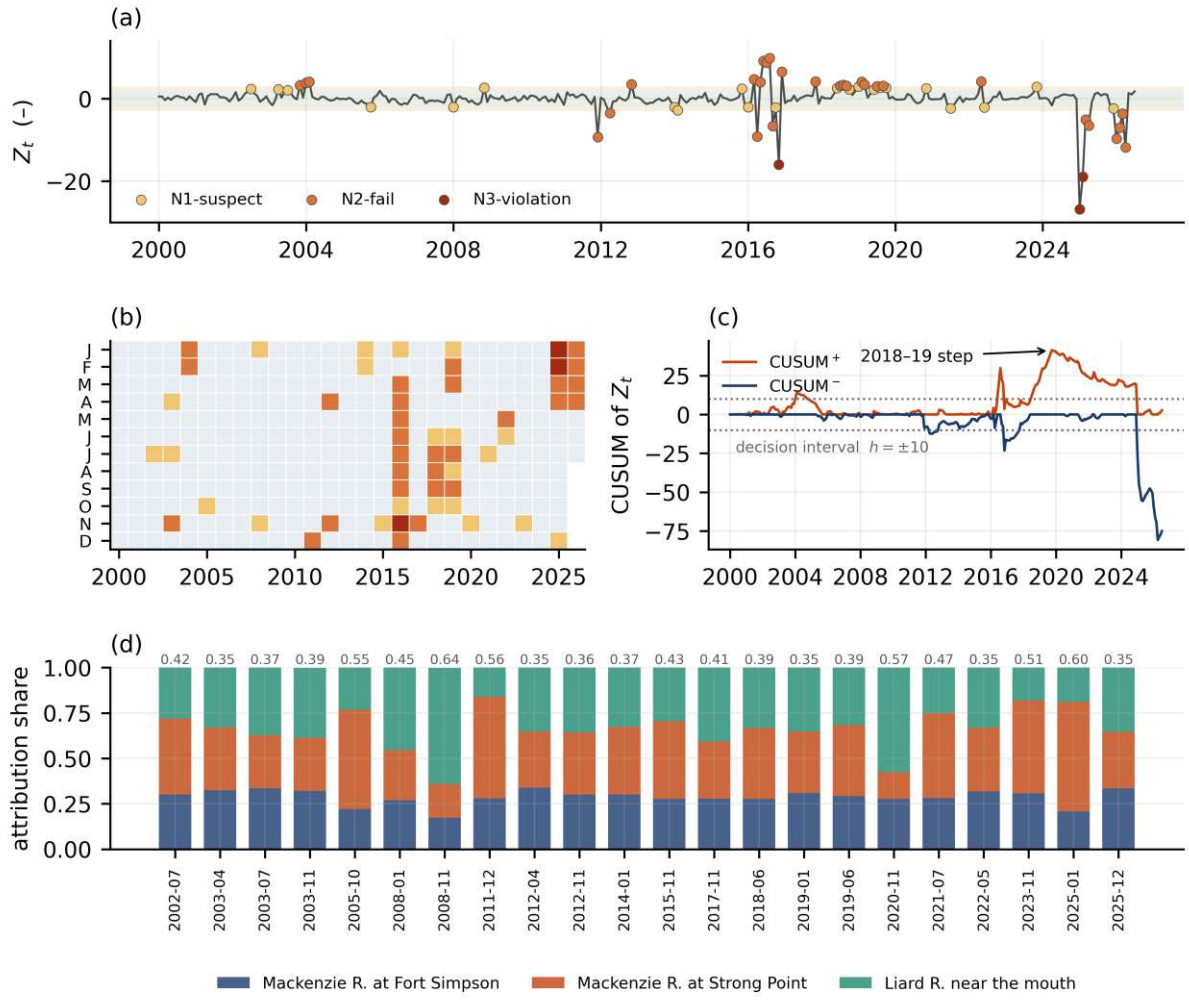


Fig. 7 NCQC results, January 2000–July 2026. (a) Standardized closure statistic Z_t with the N1 and N2 thresholds. (b) Monthly flag matrix. (c) Two-sided CUSUM statistic. (d) Episode attribution shares; 1/3 indicates equal attribution among the three stations

5.6 Case study 2: the 2025–2026 provisional record

The second major period of inconsistency coincides with the transition from approved to provisional discharge records in January 2025 (Fig. 8b, d). During the 2025–2026 ice seasons, the median relative closure error was −9.8%, reaching −38.9% in January 2025 and −28.0% in February. In both months, Strong Point discharge exceeded that at Fort Simpson, producing two T1 violations. Together with November 2016, these constitute the three physical admissibility violations identified in the full record. In contrast, open-water closure errors during 2025–2026 generally remained within approximately ±2%. The T0 provenance diagnostic independently identifies a change in numerical reporting precision (Fig. 8e). During 2000–2024, annual mean significant-digit counts remained near 2.8–2.9 across the three stations, consistent with archival rounding conventions. In 2025, these increased to 7.0–9.4, remaining at 4.9–5.4 in 2026. This abrupt change provides a metadata-independent signature of altered reporting precision, coinciding with the transition to provisional records and the onset of pronounced winter closure inconsistencies. These findings demonstrate that record status is an important factor in network-level discharge consistency. The larger winter residuals in provisional records are consistent with differences in processing and subsequent

adjustments associated with WSC approval procedures, including temporary shifts and overrides (Gharari et al. 2024). In addition, several provisional ice-season monthly values are based on incomplete or apparently constant daily estimates (Sect. 3.2), which may contribute to the larger winter residuals. However, the available data do not establish which adjustments were applied or their magnitudes. The results therefore reveal substantial pre-approval network inconsistencies, reaching approximately 39%, without demonstrating that equivalent corrections are routinely imposed during final approval.

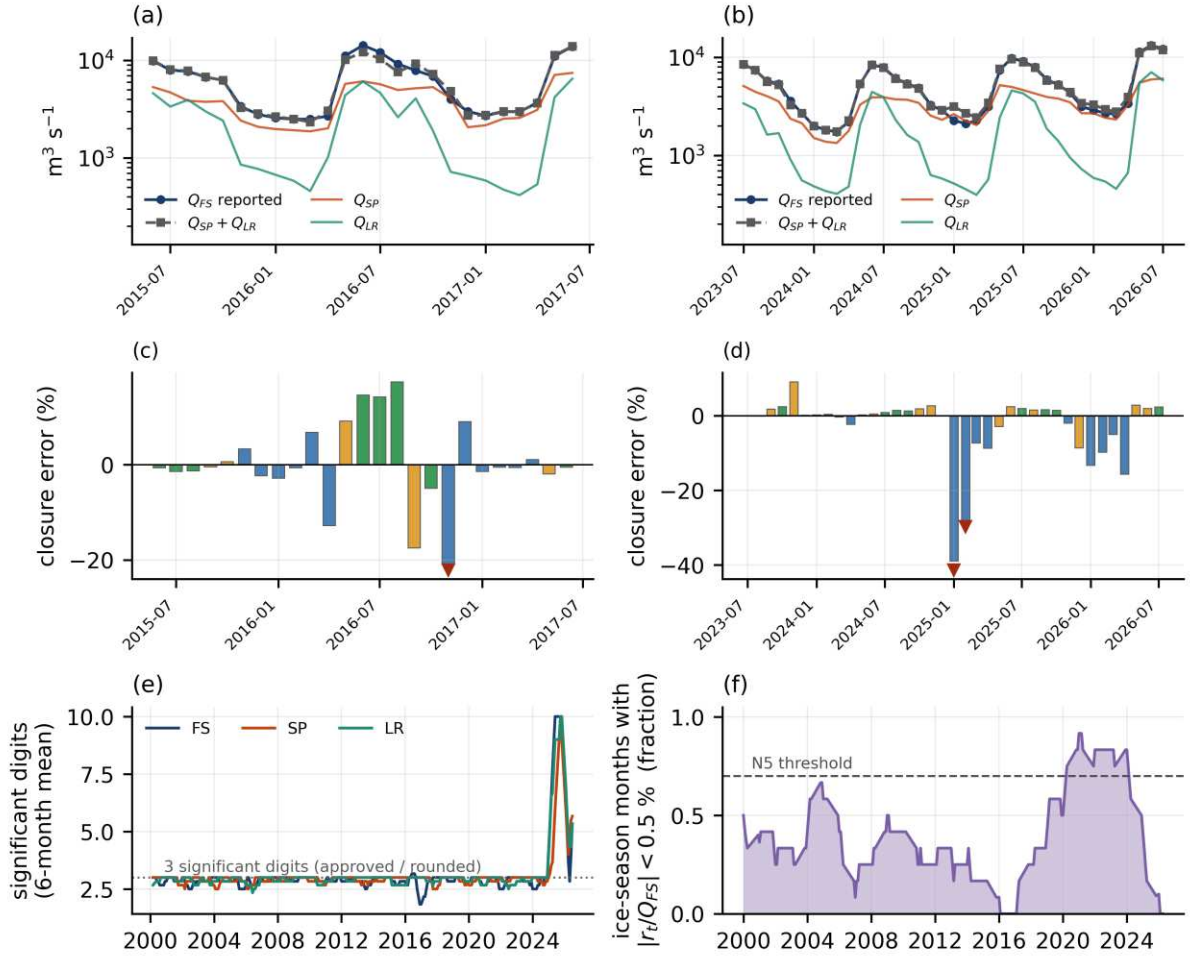


Fig. 8 Major episodes and supporting diagnostics. (a, c) Hydrographs and relative closure errors for the 2015–2016 episode; ▼ marks the November 2016 T1 violation. (b, d) The same for the 2025–2026 provisional record. (e) T0: reported significant digits. (f) T6: share of ice-affected months with near-zero closure, with the N5 threshold

5.7 Detection and attribution skill

Synthetic error-injection experiments demonstrate that NCQC detection performance depends on error magnitude, duration and hydrological regime (Fig. 9). At the operating threshold $|Z_t| > 2$, corresponding to an 8.0% background flag rate on the uncorrupted reference record, the probability of detection (POD) for sustained multiplicative biases increases from 0.41 at 2% to 0.72 at 5%, 0.89 at 10%, 0.97 at 15% and 0.99 at 20% (Table 5). Longer steps are detected more reliably (POD = 0.91 at 5%), whereas single-month spikes are more difficult to identify (0.54 at 5%). Frozen-value errors achieve a POD of approximately 0.90 across the tested magnitudes, reflecting their distinctive temporal structure. Part of this detection

performance arises from flags already present in the uncorrupted record: 21–31 % of sustained-bias windows, 4–13 % of spike windows and 47–54 % of step windows contained at least one pre-existing flag. Counting only newly raised flags, POD for sustained biases is 0.23, 0.66 and 0.87 at 2 %, 5 % and 10 %, respectively, 0.87 for 5 % steps and 0.46 for 5 % single-month spikes, confirming that the detection of errors of 5 % or more is driven primarily by the injected perturbations rather than by pre-existing inconsistencies. The conventional station-level baseline performed considerably worse at the same 8.0 % background flag rate. Its window-based POD for sustained biases of 2–20 % remained between 0.30 and 0.36, close to the rate expected from pre-existing flags alone, and newly raised flags occurred in fewer than 8 % of these trials, rising to 51 % only for 50 % biases. Steps and spikes behaved similarly, with new-flag POD of at most 0.15 for errors up to 20 %. The exception was frozen values, which the flat-line test detected in 96–99 % of trials because the injected repetitions are exact, compared with 86–90 % for NCQC. Thus, at this site, the network constraint provides most of the sensitivity to systematic errors of 2–20 %, whereas exact-repetition checks remain a useful complement (Sect. 4.2). Detection performance also varies by regime. For a 5% injected error, POD reaches 0.84 during ice-affected months, compared with 0.59 in open water and 0.50 during transition periods (Fig. 9b). This pattern reflects the narrower ice-season closure tolerance rather than necessarily greater accuracy of winter discharge estimates. The lower detection performance during transition months is particularly relevant given the hydrological complexity of river-ice break-up and freeze-up. The ROC analysis (Fig. 9c) indicates that $|Z_t| > 2$ provides a practical balance between detection sensitivity and review burden, particularly for errors of 10% or greater. Increasing the threshold to $|Z_t| > 3$ reduces the background flag rate from 8.0% to 3.0%, but decreases POD for a 10% error by approximately nine percentage points. Attribution performance is strongly station-dependent (Fig. 9d). Among detected trials, the correct-attribution rate reaches 0.80 for Strong Point and 0.74 for Liard, exceeding the 1/3 random-assignment benchmark. In contrast, attribution accuracy at Fort Simpson is only 0.30. This limitation reflects the underdetermined nature of the three-gauge mass balance: downstream errors can produce closure deviations similar to those caused by compensating upstream errors. The temporal and spatial diagnostics therefore provide useful discrimination at the upstream stations but remain insufficient for reliable downstream attribution. Attribution performance is also sensitive to the weights in Eq. (6). Increasing the weight of the spatial score slightly improved correct-attribution rates, to 0.84 at Strong Point and 0.78 at Liard when the spatial score was used alone, whereas relying on temporal continuity alone reduced them to 0.66 and 0.54, respectively. Fort Simpson remained between 0.22 and 0.32 for all weightings tested (temporal weights of 0–1). At episode level, the attribution of the November 2003–February 2004, 2015–2016, June–October 2019 and January–April 2025 episodes was unchanged for all weightings, whereas the other multi-month episodes, all with dominant shares below 0.40, changed attribution with the weighting. Attribution shares were informative mainly for upstream errors: for errors injected at Fort Simpson, dominant shares above 0.50 almost always identified an upstream station, because a downstream error produces the same closure deviation as compensating upstream errors. Attribution shares should therefore be interpreted as relative evidence rather than calibrated probabilities, particularly when the downstream gauge may be affected. Potential improvements are discussed in Sect. 6.5.

These regime and station patterns are reproduced by the analytical detection limit of Eq. (8). For a 5 % injected error, $\delta_{\min}$ is exceeded in its approximate form at two of the three stations under ice, at one station in open water and at none during transition months, matching the observed ordering of regime POD (0.84, 0.59 and 0.50 respectively). With the floor retained, a 5 % error no longer exceeds the limit at any station in open water, but the open-water limits (5.1–12.6 %) remain below those of transition months (7.0–18.6 %), so the ordering is preserved. The same expression explains the station ordering: the Liard carries the

smallest share of the flow, so its $\delta_{\min}$ is the largest in every regime: about 4.4–5.5 times that of the main-stem gauges under ice, but only 1.5–2.5 times in open water and 1.7–2.7 times during transition months. It also accounts for the counter-intuitive result that detection is most sensitive during the ice season, despite under-ice discharge being the least certain, because $\delta_{\min}$ is proportional to k_ρ, which is smallest in that regime; the sensitivity therefore reflects the narrowness of the closure tolerance rather than the accuracy of the winter estimates. The agreement between the closed-form limit and the simulations indicates that the detection behaviour of NCQC is governed by the tolerance model rather than by details of the injection design.

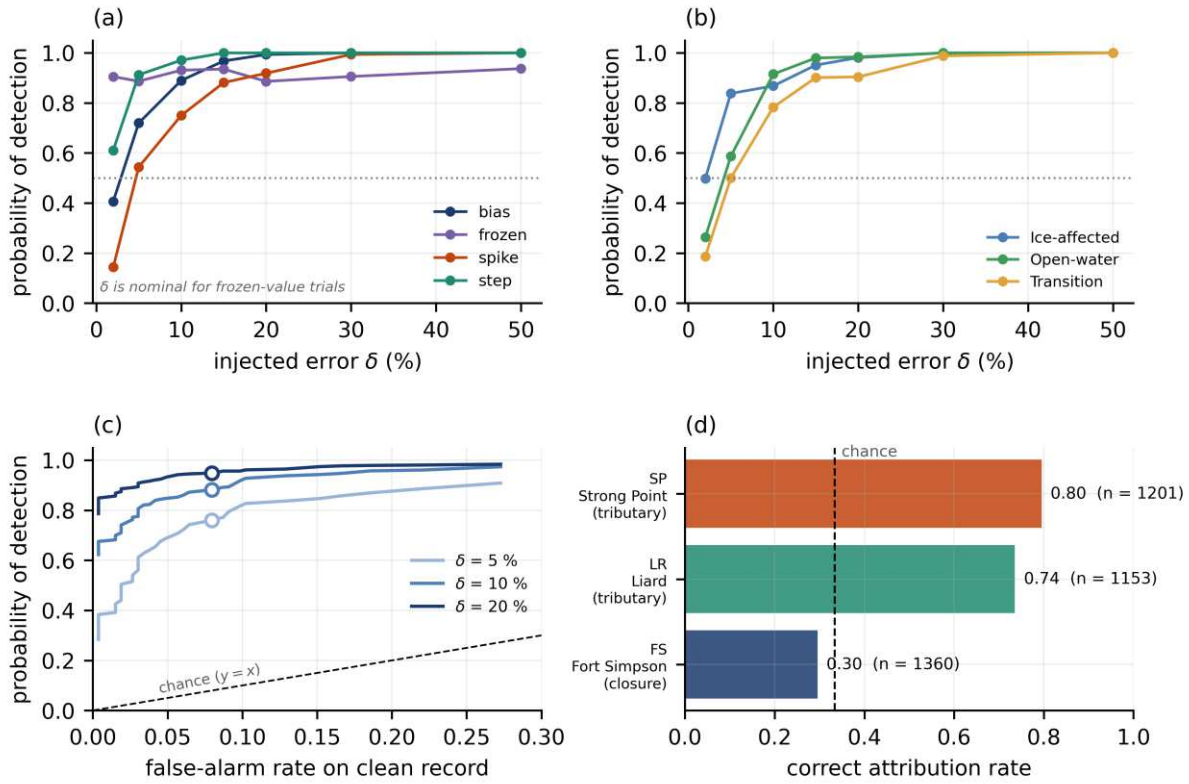


Fig. 9 Detection and attribution skill in 4327 error-injection trials. (a) Probability of detection (POD) by error magnitude and type. (b) POD by regime, excluding frozen-value trials. (c) Receiver operating characteristic (ROC) curves; open circles mark the $|Z_t| > 2$ threshold. (d) Correct-attribution rate by station; the dashed line marks the 1/3 random benchmark

Table 5 Probability of detection (POD) by error type and injected magnitude at the $|Z_t| > 2$ threshold. The background flag rate on the uncorrupted reference record is 8.0%

Error type	2 %	5 %	10 %	15 %	20 %	30 %	50 %
Sustained bias	0.41	0.72	0.89	0.97	0.99	1.00	1.00
Step (6–12 months)	0.61	0.91	0.97	1.00	1.00	1.00	1.00
Single-month spike	0.14	0.54	0.75	0.88	0.92	0.99	1.00
Frozen value	0.91	0.89	0.93	0.94	0.89	0.91	0.94

6 Discussion

6.1 Added value of network-based quality control

The results demonstrate that network-based QC can identify discharge inconsistencies that may remain undetected by conventional station-level tests. The 2015–2016 episode includes values within the historical range of individual stations, while the persistent 2018–2019 offset is difficult to distinguish from normal discharge variability in individual hydrographs. Similarly, the 2025–2026 provisional winter values become suspect primarily when evaluated against neighboring gauges. These findings highlight the principal advantage of NCQC: mass-balance redundancy provides information unavailable from individual station records. The injection experiments further demonstrate sensitivity to relatively small systematic errors, with POD values of 0.72 and 0.89 for sustained biases of 5% and 10%, respectively. A conventional station-level baseline operated at the same background flag rate raised new flags in fewer than 6 % of these trials. This performance is notable given that published uncertainty estimates for individual discharge values are often of comparable or greater magnitude. Network redundancy therefore provides a complementary means of identifying inconsistencies without requiring any individual station to serve as an error-free reference.

6.2 Closure variance collapse and network independence

One of the most important findings is the pronounced reduction in ice-season closure variability during 2020–2024. The robust scale decreased from 1.46% during 2000–2019 to 0.20%, with 85% of ice-affected months closing within ±0.5%. Such agreement is striking given the difficulties of under-ice discharge estimation, including ice-induced backwater, infrequent field measurements and rating-curve uncertainty. Gharari et al. (2024) also reported substantially lower winter reproducibility of WSC discharge estimates from stage and rating-curve information. One possible explanation is the use of neighboring discharge observations during estimation procedures. WSC overrides and temporary shifts can incorporate additional information when conventional stage–discharge relationships are unreliable (Gharari et al. 2024). If neighboring observations influence multiple records, closure residuals may become artificially small relative to the underlying measurement uncertainties. Consistent with this possibility, the T6 diagnostic (Fig. 8f) exceeds the N5 threshold during 2020–2024, compared with an average near-zero closure fraction of 0.30 during 2000–2019 (0.38 in 2000–2009 and 0.24 in 2010–2019). This result is not specific to the selected thresholds. For closure tolerances of 0.25–0.75 % and N5 thresholds of 0.70–0.80, with windows of 12 or 16 ice-affected months, N5 is raised only within 2020–2024, and never in the provisional 2025–2026 record. The proportion of ice-affected months within ±0.25 % was 0.74 during 2020–2024 compared with 0.18 during 2000–2019. Isolated earlier exceedances occurred only for less restrictive settings (a threshold of 0.60, an eight-month window or a tolerance of 1.0 %), mainly in 2000–2005. Nevertheless, the diagnostic establishes only that agreement during 2020–2024 is unusually close relative to the earlier record; it cannot by itself distinguish dependence among discharge estimates from a genuine improvement in measurement or processing practice. This finding has two implications. First, network-based QC must assess whether its physical constraint provides independent information; otherwise, unusually small residuals may be misinterpreted as evidence of high accuracy. Second, uncertainty estimates derived directly from closure residuals may be underestimated when the discharge records contain correlated errors or neighbor-informed adjustments. The contrast between the 2020–2024 approved records and the substantially larger provisional winter residuals in 2025–2026 reinforces this concern, although differences in record status alone cannot establish the underlying cause. These results support the recommendation of Gharari et al. (2024) that discharge-processing provenance should be more accessible. Large-sample hydrological datasets, including CAMELS, HYSETS and Caravan, would also benefit from retaining relevant quality and provenance information alongside discharge observations (Addor et al. 2017; Arsenault

et al. 2020; Kratzert et al. 2023). Ultimately, distinguishing genuinely accurate independent measurements from agreement introduced during data processing requires access to rating-curve adjustments, temporary shifts and override histories.

Equation (7) allows the magnitude of the implied dependence to be stated. Writing $u_t = e_{\mathrm{SP}} + e_{\mathrm{LR}}$ for the combined upstream error and assuming, for this calculation, that the individual error scales are unchanged between the two periods and that $\sigma_{\mathrm{FS}} \approx \sigma_u \equiv \sigma$, Eq. (7) reduces to $\mathrm{Var}[r_t] = 2\sigma^2(1-\kappa)$, where κ is the correlation between the downstream error and the combined upstream error. The ratio of the ice-season robust scales τ of the relative closure error then gives

$$\kappa^* = 1 - \left(\frac{\tau_{2020\text{–}2024}}{\tau_{2000\text{–}2019}}\right)^2 (1-\kappa_0) \tag{9}$$

If the errors were uncorrelated before 2020, so that $\kappa_0 = 0$, reproducing the fall from 1.46 % to 0.20 % with no improvement in the individual estimates requires $\kappa \approx 0.98$; starting from $\kappa_0 = 0.5$ the requirement is $\kappa \approx 0.99$. Correlations of this size are difficult to reconcile with three independently derived discharge estimates. The calculation is illustrative rather than an estimate of κ, because it assumes stationary individual error scales, which closure alone cannot verify, and a genuine reduction in the σ_i remains equally consistent with the residuals. It does, however, bound the alternatives: either the individual discharge uncertainties at these three stations fell by a factor of about seven between 2019 and 2020, or the estimates ceased to be independent.

6.3 Implications for discharge uncertainty estimation

During periods without evidence of network dependence, closure residuals provide an empirical measure of combined discharge inconsistency across the three gauges. The regime-specific parameters in Table 2 quantify how this inconsistency varies with hydrological conditions, without requiring an explicit rating-curve error model. However, closure residuals represent the combined effects of errors across stations and cannot directly quantify the uncertainty of any individual discharge record. Correlated errors may also reduce closure variability without improving individual accuracy. NCQC therefore complements, rather than replaces, established discharge uncertainty frameworks (Coxon et al. 2015; Le Coz et al. 2014; Mansanarez et al. 2019). Its principal advantage is that it evaluates published discharge records, including the combined effects of measurement, rating-curve application and subsequent processing. This is particularly relevant to the uncertainty categories discussed by Gharari et al. (2024), for which conventional residual-based methods may not fully represent WSC production practices. Network constraints could also inform the construction of physically consistent discharge ensembles (Wong et al. 2021), provided that cross-station error dependence is explicitly represented.

6.4 Operational implementation

NCQC offers a computationally efficient complement to existing hydrometric QC procedures and could be implemented in three operational settings. First, real-time screening could apply monthly or more frequent network-closure checks to provisional discharge records, identifying N1–N3 conditions for hydrographer review. The 2025–2026 episode illustrates the potential value of such screening, particularly where substantial winter inconsistencies and physical admissibility violations occur. More frequent implementation would require appropriate treatment of travel time and channel storage. Second, retrospective auditing could use CUSUM diagnostics to identify persistent changes in archived discharge consistency. The 2018–2019 offset illustrates how small but sustained deviations may remain difficult to

identify through individual station inspection. Such changes warrant examination of rating histories and processing records, particularly when discharge data are used for long-term trend assessment (Hamilton and Moore 2012; Whitfield 2012). Third, quality metadata integration could incorporate NCQC flags alongside existing WSC quality symbols. This would allow downstream users to distinguish ordinary network consistency, significant closure anomalies and periods of potentially reduced network independence. In this respect the approach parallels uncertainty-aware reliability monitoring of other networked infrastructure, where the aim is to flag incipient failures before they become visible at any single node (Damonti and Bernasconi 2026). In particular, N5 would indicate that closure agreement provides limited independent evidence of discharge accuracy, rather than establishing that neighbor-informed processing occurred.

6.5 Limitations and generalization

First, gauge attribution remains underdetermined at a single three-station confluence. Although attribution accuracy reaches 0.80 at Strong Point and 0.74 at Liard, it is only 0.30 at Fort Simpson. Additional gauges could provide complementary spatial constraints. Mackenzie River at Jean Marie River (10FB007) and Mackenzie River near Fort Providence (10FB001) are potential candidates for extending the analysis, subject to verification of their locations and hydraulic connectivity. Network-theoretic placement criteria (Mondal and Mishra 2025) could indicate which additional stations would most improve attribution. More generally, nested control volumes could improve error localization where the resulting closure equations provide independent information. However, additional equations do not automatically guarantee unique attribution, particularly when errors are correlated or multiple stations are affected. Second, monthly averaging reduces the influence of travel time and short-term storage changes but limits sensitivity to brief anomalies. Application at daily resolution would require explicit consideration of routing, storage and river-ice dynamics, particularly during break-up and ice-jam periods (Beltaos and Prowse 2009). Higher temporal resolution could improve the detection of short-lived errors, for which the current method achieves a POD of only 0.54 at a 5% single-month spike.

Third, the reference period excludes 2016, 2018 and 2019 based on identified anomalies, introducing potential selection bias. Robust parameter estimation and fixed tolerances during synthetic injection reduce sensitivity to extreme observations but do not eliminate this limitation. Independent calibration and validation periods would provide a stronger assessment of generalization. Fourth, the framework has been evaluated at a single confluence in a large, ice-influenced northern basin. Application to smaller catchments, regulated rivers or confluences with substantial intervening drainage would require explicit treatment of additional inflows, storage changes and operational effects. Although the underlying mass-balance principle is general, detection thresholds and attribution performance must be evaluated for each network configuration. The broader relevance of the approach is supported by similarities in hydrometric data-processing practices across North America, including the use of Aquarius and related discharge-adjustment procedures (Gharari et al. 2024).

7 Conclusions

This study demonstrates the potential of hydrometric network redundancy as an operational quality-control instrument. Applied to the Mackenzie–Liard confluence over 2000–2026, NCQC led to six main findings. First, the confluence mass balance achieved r^2 = 0.9953, a median relative closure error of 0.00% and a

robust relative scale of 1.67%; this strong agreement provides a useful baseline for detecting network inconsistencies, although it does not directly quantify individual station accuracy. Second, by integrating provenance, physical admissibility, NDVI-based regime classification, regime-conditioned closure, temporal–spatial attribution, CUSUM and network-independence diagnostics, NCQC flagged 54 months (16.9%) grouped into 22 episodes, including persistent anomalies in approved records and pronounced inconsistencies in provisional records. Third, synthetic experiments yielded detection probabilities of 0.72 and 0.89 for sustained biases of 5% and 10%, respectively, at an 8.0% background flag rate, whereas a conventional station-level baseline operating at the same background rate raised new flags in fewer than 6% of these trials. Correct-attribution rates were 0.80 at Strong Point and 0.74 at Liard, but only 0.30 at Fort Simpson, highlighting the need for additional independent network constraints.

Fourth, changes in reported significant-digit precision provided a metadata-independent signature of the January 2025 transition from approved to provisional records, demonstrating the value of numerical provenance as a complementary QC diagnostic. Fifth, during 2025–2026 the median ice-season closure error reached −9.8%, with individual deviations approaching −39% and two physical admissibility violations; these results reveal substantial network inconsistencies in provisional winter discharge estimates, without establishing their causes or the magnitude of subsequent corrections. Sixth, ice-season closure variability decreased to a robust scale of 0.20% during 2020–2024, compared with 1.46% during 2000–2019. This unusually close agreement raises concerns about potential dependence among discharge estimates and shows why small closure residuals should not automatically be interpreted as high measurement accuracy.

NCQC provides a physically grounded framework for detecting discharge inconsistencies, prioritizing their investigation and identifying periods when network closure may offer limited independent information. Its principal contribution is not to replace station-level QC or quantify individual gauge uncertainty, but to introduce complementary network-level evidence into hydrometric quality assessment. Incorporating network flags, regime-specific closure statistics and processing provenance into published discharge datasets would improve transparency and support more reliable hydrological analyses.

References

Addor N, Newman AJ, Mizukami N, Clark MP (2017) The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol Earth Syst Sci 21:5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Arsenault R, Brissette F, Martel J-L, Troin M, Lévesque G, Davidson-Chaput J, Gonzalez MC, Ameli A, Poulin A (2020) A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. Sci Data 7:243. https://doi.org/10.1038/s41597-020-00583-2

Badrudeen AT, Sahoo D, Sawyer CB, Pike JW, Harmel RD (2026) A critical review of statistical, signal processing and machine learning methods for continuous and high-frequency water quality data improvement. Ecol Inform 94:103619. https://doi.org/10.1016/j.ecoinf.2026.103619

Beltaos S, Prowse T (2009) River-ice hydrology in a shrinking cryosphere. Hydrol Process 23:122–144. https://doi.org/10.1002/hyp.7165

Cohn TA, Kiang JE, Mason RR (2013) Estimating discharge measurement uncertainty using the interpolated variance estimator. J Hydraul Eng 139:502–510. https://doi.org/10.1061/(ASCE)HY.1943-7900.0000695

Coxon G, Freer J, Westerberg IK, Wagener T, Woods R, Smith PJ (2015) A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. Water Resour Res 51:5531–5546. https://doi.org/10.1002/2014WR016532

Damonti E, Bernasconi G (2026) An uncertainty-aware reliability framework for incipient failure detection in critical infrastructure: application to water distribution networks. Stoch Environ Res Risk Assess 40:113. https://doi.org/10.1007/s00477-026-03239-6

Das T, Das S (2026) Comparative assessment of classical, physically informed, machine learning, and hybrid rating curve models for streamflow estimation. Stoch Environ Res Risk Assess 40:262. https://doi.org/10.1007/s00477-026-03400-1

de Rham L, Dibike Y, Beltaos S, Peters D, Bonsal B, Prowse T (2020) A Canadian river ice database from the national hydrometric program archives. Earth Syst Sci Data 12:1835–1860. https://doi.org/10.5194/essd-12-1835-2020

Déry SJ, Stadnyk TA, MacDonald MK, Gauli-Sharma B (2016) Recent trends and variability in river discharge across northern Canada. Hydrol Earth Syst Sci 20:4801–4818. https://doi.org/10.5194/hess-20-4801-2016

Didan K (2015) MOD13Q1 MODIS/Terra vegetation indices 16-day L3 global 250 m SIN grid V006. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13Q1.006

Durre I, Menne MJ, Gleason BE, Houston TG, Vose RS (2010) Comprehensive automated quality assurance of daily surface observations. J Appl Meteorol Climatol 49:1615–1633. https://doi.org/10.1175/2010JAMC2375.1

ECCC (Environment and Climate Change Canada) (2026a) National Water Data Archive (HYDAT), release of 17 July 2026 (Hydat_sqlite3_20260717). Water Survey of Canada. https://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/national-archive-hydat.html. Accessed August 2026

ECCC (Environment and Climate Change Canada) (2026b) Wateroffice: real-time and historical hydrometric data. Water Survey of Canada. https://wateroffice.ec.gc.ca/. Accessed August 2026

Eleutério I, Silva F, Costa V (2025) Assessing the stage-period-discharge framework for modeling highly unstable rating curves: a case study at the UHE Peti Carrapato streamflow gauging station, Brazil. Stoch Environ Res Risk Assess 39:829–845. https://doi.org/10.1007/s00477-024-02888-9

Feng D, Gleason CJ (2024) More flow upstream and less flow downstream: the changing form and function of global rivers. Science 386:1305–1311. https://doi.org/10.1126/science.adl5728

Fiebrich CA, Morgan CR, McCombs AG, Hall PK, McPherson RA (2010) Quality assurance procedures for mesoscale meteorological data. J Atmos Ocean Technol 27:1565–1582. https://doi.org/10.1175/2010JTECHA1433.1

Gharari S, Razavi S (2018) A review and synthesis of hysteresis in hydrology and hydrological modeling: memory, path-dependency, or missing physics?. J Hydrol 566:500–519. https://doi.org/10.1016/j.jhydrol.2018.06.037

Gharari S, Whitfield PH, Pietroniro A, Freer J, Liu H, Clark MP (2024) Exploring the provenance of information across Canadian hydrometric stations: implications for discharge estimation and uncertainty quantification. Hydrol Earth Syst Sci 28:4383–4405. https://doi.org/10.5194/hess-28-4383-2024

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ 202:18–27. https://doi.org/10.1016/j.rse.2017.06.031

Halliday R (2008) Water Survey of Canada – a history. Technical Report En37-436/2008E-PDF, Environment Canada, Ottawa

Hamilton AS, Moore RD (2012) Quantifying uncertainty in streamflow records. Can Water Resour J 37:3–21. https://doi.org/10.4296/cwrj3701865

Horner I, Renard B, Le Coz J, Branger F, McMillan HK, Pierrefeu G (2018) Impact of stage measurement errors on streamflow uncertainty. Water Resour Res 54:1952–1976. https://doi.org/10.1002/2017WR022039

Hu YM, Yang CX, Liang ZM, Luo XY, Huang YX, Tang C (2023) A non-parametric approach for change-point detection of multi-parameters in time-series data. J Environ Inform 42:65–74. https://doi.org/10.3808/jei.202200478

Hubbard KG, Goddard S, Sorensen WD, Wells N, Osugi TT (2005) Performance of quality assurance procedures for an applied climate information system. J Atmos Ocean Technol 22:105–112. https://doi.org/10.1175/JTECH-1657.1

Huete A, Didan K, Miura T, Rodriguez EP, Gao X, Ferreira LG (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ 83:195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Jalbert J, Mathevet T, Favre A-C (2011) Temporal uncertainty estimation of discharges from rating curves using a variographic analysis. J Hydrol 397:83–92. https://doi.org/10.1016/j.jhydrol.2010.11.031

Kiang JE, Gazoorian C, McMillan H, Coxon G, Le Coz J, Westerberg IK, Belleville A, Sevrez D, Sikorska AE, Petersen-Øverleir A, Reitan T, Freer J, Renard B, Mansanarez V, Mason R (2018) A comparison of methods for streamflow uncertainty estimation. Water Resour Res 54:7149–7176. https://doi.org/10.1029/2018WR022708

Kona SV, Das Bhowmik R (2025) Understanding the interannual variation in the river stage–discharge relationship and its impact on flood discharge estimation. J Hydrol Eng 30:04025020. https://doi.org/10.1061/JHYEFF.HEENG-6484

Krabbenhoft CA, Allen GH, Lin P, Godsey SE, Allen DC, Burrows RM, DelVecchia AG, Fritz KM, Shanafield M, Burgin AJ, Zimmer MA, Datry T, Dodds WK, Jones CN, Mims MC, Franklin C, Hammond JC, Zipper S, Ward AS, Costigan KH, Beck HE, Olden JD (2022) Assessing placement bias of the global river gauge network. Nat Sustain 5:586–592. https://doi.org/10.1038/s41893-022-00873-0

Kratzert F, Nearing G, Addor N, Erickson T, Gauch M, Gilon O, Gudmundsson L, Hassidim A, Klotz D, Nevo S, Shalev G, Matias Y (2023) Caravan – a global community dataset for large-sample hydrology. Sci Data 10:61. https://doi.org/10.1038/s41597-023-01975-w

Lamb R, Zaidman M, Archer D, Marsh T, Lees M (2003) River gauging station data quality classification (GSDQ). R&D Technical Report W6-058/TR, Environment Agency, Bristol

Lammers RB, Shiklomanov AI, Vörösmarty CJ, Fekete BM, Peterson BJ (2001) Assessment of contemporary Arctic river runoff based on observational discharge records. J Geophys Res 106:3321–3334. https://doi.org/10.1029/2000JD900444

Le Coz J, Renard B, Bonnifait L, Branger F, Le Boursicaud R (2014) Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: a Bayesian approach. J Hydrol 509:573–587. https://doi.org/10.1016/j.jhydrol.2013.11.016

Mansanarez V, Renard B, Le Coz J, Lang M, Darienzo M (2019) Shift happens! Adjusting stage–discharge rating curves to morphological changes at known times. Water Resour Res 55:2876–2899. https://doi.org/10.1029/2018WR023389

Marinheiro R, Matos JP, Matos J, Sousa V, Meireles I, Mahunguana M, Francisco H, Zunguze N (2025) Estimating streamflow from stage measurements affected by downstream water levels: the hyperbolic discharge rating curve. Int J River Basin Manag. https://doi.org/10.1080/15715124.2025.2553801

McMahon TA, Peel MC, Amirthanathan GE (2025) Assessing rating curve uncertainty. Hydrol Sci J 70:687–694. https://doi.org/10.1080/02626667.2024.2442736

Mondal S, Mishra AK (2025) A complex network-based framework for evaluating hydrometric monitoring networks. Stoch Environ Res Risk Assess 39:3979–3993. https://doi.org/10.1007/s00477-025-03045-6

Motiee H, Ahrari S, Motiee S, McBean E (2025) Assessment of climate change with remote sensing data on snow and ice cover in the Rocky Mountains glaciers. J Environ Inform 46:1–13. https://doi.org/10.3808/jei.202500544

Mousavi F, Torabi Haghighi A, Silander J, Monemi M, Rasti M (2026) Physics-informed fusion learning for nowcasting fault detection of hydrological multi sensor network. Water Resour Res 62:e2026WR043436. https://doi.org/10.1029/2026WR043436

Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, Boussetta S, Choulga M, Harrigan S, Hersbach H, Martens B, Miralles DG, Piles M, Rodríguez-Fernández NJ, Zsoter E, Buontempo C, Thépaut J-N (2021) ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst Sci Data 13:4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Muste M, Kim K, Kim D, Demir I (2026) A hysteresis-based framework for evaluating and selecting streamflow monitoring approaches under unsteady gradually varied flow conditions. Water Resour Res 62:e2026WR043867. https://doi.org/10.1029/2026WR043867

Nearing GS, Kratzert F, Sampson AK, Pelissier CS, Klotz D, Frame JM, Prieto C, Gupta HV (2021) What role does hydrological science play in the age of machine learning?. Water Resour Res 57:e2020WR028091. https://doi.org/10.1029/2020WR028091

Pelletier PM (1989) Uncertainties in streamflow measurement under winter ice conditions – a case study: the Red River at Emerson, Manitoba, Canada. Water Resour Res 25:1857–1867. https://doi.org/10.1029/WR025i008p01857

Prakash R, Tripura J (2026) A physics-guided machine learning framework for accurate reconstruction of missing discharge records: case study of the Gandak River at Hajipur, India. Stoch Environ Res Risk Assess 40:135. https://doi.org/10.1007/s00477-026-03264-5

Rainville F, Hutchinson D, Stead A, Moncur D, Elliott D (2002) Hydrometric manual – data computations: stage–discharge model development and maintenance. Environment Canada, Ottawa

Rantz SE (1982) Measurement and computation of streamflow. US Geological Survey Water Supply Paper 2175, US Government Printing Office, Washington, DC. https://doi.org/10.3133/wsp2175

Rawat M, Nguyen-Huy T, Ali A, Ghahramani A (2026) Investigating MODIS-derived vegetation metrics integration with modelling to predict plant available water in a data-scarce humid subtropic paddock of Australia. J Environ Inform 47:1–20. https://doi.org/10.3808/jei.202600554

Razavi H (2026) NCQC: Network-Constrained Quality Control for river discharge records (Mackenzie–Liard confluence, 2000–2026), version 1.2.0. Zenodo. https://doi.org/10.5281/zenodo.23632705

Reitan T, Petersen-Øverleir A (2011) Dynamic rating curve assessment in unstable rivers using Ornstein–Uhlenbeck processes. Water Resour Res 47:W02524. https://doi.org/10.1029/2010WR009504

Rigueira X, Olivieri D, Araujo M, Pazo M (2026) An explainable AI model for anomaly detection in water quality data. Water Resour Manag 40:305. https://doi.org/10.1007/s11269-026-04676-2

Smith DG (1979) Effects of channel enlargement by river ice processes on bankfull discharge in Alberta, Canada. Water Resour Res 15:469–475. https://doi.org/10.1029/WR015i002p00469

St. Jacques J-M, Sauchyn DJ (2009) Increasing winter baseflow and mean annual streamflow from possible permafrost thawing in the Northwest Territories, Canada. Geophys Res Lett 36:L01401. https://doi.org/10.1029/2008GL035822

Usman KR, Montero RA, Ghobrial T, Anctil F, van Loenen A (2026) Development of an under-ice river discharge forecasting system in Delft-Flood Early Warning System (Delft-FEWS) for the Chaudière River based on a coupled hydrological-hydrodynamic modelling approach. Geosci Model Dev 19:1559–1580. https://doi.org/10.5194/gmd-19-1559-2026

Wang X, Feng L (2024) Patterns and trends in Northern Hemisphere river ice phenology from 2000 to 2021. Remote Sens Environ 313:114346. https://doi.org/10.1016/j.rse.2024.114346

Westerberg IK, Wagener T, Coxon G, McMillan HK, Castellarin A, Montanari A, Freer J (2016) Uncertainty in hydrological signatures for gauged and ungauged catchments. Water Resour Res 52:1847–1865. https://doi.org/10.1002/2015WR017635

Westerberg IK, Mansanarez V, Lyon S, Lam N, Hallén J, Karlsen RH (2026) Rapid rating curve uncertainty estimation using hydraulic modeling: use of UAV-derived data and impacts of gauging strategies on uncertainty in hydrological signatures. Front Water 8:1833490. https://doi.org/10.3389/frwa.2026.1833490

Whitfield PH (2012) Why the provenance of data matters: assessing fitness for purpose for environmental data. Can Water Resour J 37:23–36. https://doi.org/10.4296/cwrj3701866

WMO (2010) Manual on stream gauging, volumes I and II. WMO-No. 1044, World Meteorological Organization, Geneva

Wong JS, Zhang X, Gharari S, Shrestha RR, Wheater HS, Famiglietti JS (2021) Assessing water balance closure using multiple data assimilation- and remote sensing-based datasets for Canada. J Hydrometeorol 22:1569–1589. https://doi.org/10.1175/JHM-D-20-0131.1

Woo M-K, Thorne R (2003) Streamflow in the Mackenzie Basin, Canada. Arctic 56:328–340. https://doi.org/10.14430/arctic630

Wu HJ, Ye XD, Zhang BY, Chen B (2025) Assessment of uncertainty propagation from climate modeling to hydrologic forecasting under changing climatic conditions. J Environ Inform 45:27–41. https://doi.org/10.3808/jei.202500530

Yang X, Pavelsky TM, Allen GH (2020) The past and future of global river ice. Nature 577:69–73. https://doi.org/10.1038/s41586-019-1848-1

Yu C-W, Chen B-H (2026) Uncertainty assessment of non-stationary stage-discharge relationship with hydrodynamic extensions across flow regimes. J Hydrol Reg Stud 67:103839. https://doi.org/10.1016/j.ejrh.2026.103839

Statements and Declarations

Funding This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant (RGPIN-2026-05029).

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Use of artificial intelligence During preparation of this manuscript, the authors used AI tools solely for language editing and grammatical correction. The authors subsequently reviewed and revised all affected text and take full responsibility for the manuscript's content.

Author contributions Hooman Razavi: conceptualization, methodology, software, formal analysis, data curation, writing of the original draft. Hossein Bonakdari: supervision, validation, funding acquisition, writing - review and editing. Hamidreza Shirkhani: conceptualization, methodology, writing - review and editing. All authors read and approved the final manuscript.

Data availability All data needed to reproduce this study are archived on Zenodo (https://doi.org/10.5281/zenodo.23632705; Razavi 2026) under CC BY 4.0 and are provided as Online Resource 1, together with the three station workbooks and all processed outputs. The discharge records are the property of the Water Survey of Canada and are available through the National Water Data Archive (HYDAT; ECCC 2026a) and the Wateroffice hydrometric data service (ECCC 2026b). ERA5-Land and MODIS products were extracted through Google Earth Engine and are openly available from their respective providers.

Code availability The NCQC algorithm (ncqc.py) and the workflow scripts 01_load_data.py, 03_error_injection.py, 04_make_figures.py, 05_sensitivity_and_baselines.py and 06_analytical_checks.py are archived in the same Zenodo record under the MIT license and are included in Online Resource 1. Every number quoted in Sects. 5 and 6 and in Tables 2 to 5 is written by 04_make_figures.py, 05_sensitivity_and_baselines.py and 06_analytical_checks.py to summary.json, episodes.csv, sensitivity_summary.json and analytical_summary.json. The error-injection experiments use a fixed random seed and are bit-reproducible.

Supplementary information Online Resource 1 (ESM_1.zip): the NCQC code, input data, processed outputs and figures needed to reproduce all results of this article (NCQC version 1.2.0; see the README file in the archive).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.zip