

Evidence review · Land and freshwater

Choosing a reference catchment for the flow that matters

Elín Bjarnadóttir · 12 October 2026

Fictional author and institute · Editorial demonstration, not peer reviewed

ABSTRACT

A neighbouring catchment is not automatically a credible hydrological reference. This evidence review synthesises paired-catchment research, reference-by-reference diagnostics, and evidence on catchment-dependent water-balance performance. It argues that selection should begin with the response variable and timescale, distinguishing annual yield, summer low flow, and snowmelt timing. A fictional calibration shows how a reference-gauge change can inflate an apparent target effect even when target observations remain unchanged. The review separates calibration representativeness from later stability and considers multiple references, meaningful temporal validation, extrapolation, stewardship, and monitoring history. It performs no site selection and claims no disturbance effect. Its practical implication is to state exactly which behaviour a reference predicts and how that relationship will be tested, rather than treating proximity or a high overall correlation as sufficient evidence.

Keywords: reference catchment · paired catchments · counterfactual flow · calibration stability · low-flow comparison

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

The nearest neighbour is a candidate, not a control

A reference catchment is useful because it helps estimate what might have happened in another catchment without a particular disturbance or intervention. Proximity can make the pair share weather, but it does not guarantee similar snow storage, groundwater response, geology, or seasonal runoff. Two neighbouring rivers can track each other closely in annual totals and diverge during the low-flow weeks that matter for habitat or water supply. Reference selection must follow the quantity and timescale of the decision.

This evidence review synthesises a published review of paired-catchment experiments, a recent reference-by-reference analysis at the H. J. Andrews Experimental Forest, and a multi-catchment study of water-balance performance. It translates their methodological lessons into a hypothetical northern catchment decision. It is not a site-selection exercise, a new experiment, or evidence that a specific northern disturbance changed flow. All comparison values are invented teaching examples.

A reference estimates a relationship, not an identical river

Paired-catchment analysis commonly uses a pre-disturbance calibration period to relate flow in the target catchment to flow in the reference. That relationship is then used with post-disturbance reference observations to estimate the target's counterfactual flow. The difference between observed and predicted target flow is interpreted in light of disturbance and other evidence. This is stronger than comparing one before value with one after value, because shared climatic variability can be partly accounted for.

Brown and colleagues' review describes matching catchments on attributes including slope, aspect, soils, area, climate, and vegetation. It also shows that annual, seasonal, and low-flow responses are not interchangeable and that adjustment after vegetation change can take time. Their synthesis is largely about vegetation manipulations across diverse climates. It provides design principles, not a numerical transfer function for permafrost terrain.

A reference need not have the same discharge magnitude. Flow can be expressed as water depth per catchment area, and a calibrated relationship can accommodate some systematic differences. Nevertheless, normalising by area does not erase differences in storage, connectivity, or precipitation. A very large basin may integrate storms and snowmelt differently from a small target even when their annual runoff depths appear similar.

Choose the response variable before ranking candidates

For an annual-yield question, compare annual runoff and its response to wet and dry years. For summer low flow, compare the low-flow season specifically, including recession behaviour, zero-flow periods where relevant, and measurement quality near the lower limit. For a snowmelt-timing question, assess melt onset, elevation distribution, aspect, and routing. A single all-season correlation can be dominated by the seasonal cycle and conceal poor agreement in the relevant subset.

The multi-catchment analysis by Michailovsky and colleagues found that water-balance performance varied with characteristics including snow cover and latitude. Although that study concerns remote-sensing products rather than paired experiments, it reinforces a practical restriction: a broadly successful hydrological relationship

is not automatically successful under every climate and storage regime. Candidate matching should consider mechanisms, not only the highest summary score.

Some differences should lead to rejection rather than a more elaborate regression. A reference downstream of a regulating lake may poorly represent a flashy head-water target. A reference whose groundwater receives water across a topographic divide may have a different effective contributing area. A catchment already affected by road building or drainage may not provide an unaffected comparison for those processes. These are physical incompatibilities, not merely inconvenient residuals.

Worked example: the reference changes the estimated effect

Fictional worked example. Before a proposed disturbance, a target catchment's annual runoff is related to reference runoff by: expected target runoff = 20 millimetres + $0.8 \times$ reference runoff. Assume the model was fitted appropriately and has a separately assessed prediction uncertainty. In one later year, reference runoff is 300 millimetres, so predicted target runoff without disturbance is 260 millimetres.

If observed target runoff is 285 millimetres, the model difference is +25 millimetres. It is not automatically a demonstrated disturbance effect. The difference must be compared with prediction uncertainty, instrument history, other changes, and the credibility of the counterfactual assumption. Reporting a percentage without these checks would make a model residual sound more causally certain than the design allows.

Now suppose an unrecorded change in the reference gauge causes its reported runoff to fall by 40 millimetres, while target observations are unchanged. The model prediction becomes $20 + 0.8 \times 260 = 228$ millimetres. The apparent target difference grows to +57 millimetres. The additional 32 millimetres arises entirely from the reference record. A reference gauge is therefore part of the causal evidence chain, not a neutral background variable.

A different failure occurs if both gauges work correctly but the reference develops a new drainage pathway. The consequence for the estimate can be similar, yet its diagnosis is hydrological rather than instrumental. A calibration model cannot distinguish these causes from the target residual alone. Field history, independent precipitation and storage evidence, and other reference catchments provide ways to investigate them.

Representativeness and stability are separate assumptions

Wondzell and colleagues distinguish whether a calibration period represents the underlying relationship from whether that relationship remains stable afterwards. A short run of unusually similar years can produce a tight regression that extrapolates poorly. A long calibration does not prevent subsequent gauge changes, disturbances, or divergent ecological trajectories. A high coefficient of determination establishes neither assumption by itself.

Calibration should cover the conditions under which the counterfactual will be used. A model fitted only during wet years may be weak for drought assessment.

If the post-disturbance climate lies outside the calibration range, flag the extrapolation. Test predictive performance by withholding meaningful blocks of time rather than individual days scattered among closely related observations. Temporal dependence and unequal residual variance should be considered when estimating uncertainty.

There is no universal calibration length that establishes representativeness. Several ordinary years may constrain a relationship less than a shorter period spanning relevant extremes, while a short record may simply lack those extremes. A study can report what is covered and which effect sizes it could reasonably detect. It should not disguise an underpowered comparison as evidence of no impact merely because the model difference is not statistically significant.

More than one reference can test the background

Wondzell and colleagues used comparisons among untreated reference catchments to examine apparent changes that could not have been caused by the target treatment. This reference-by-reference logic can reveal background drift, measurement limitations, and non-treatment variability. Their findings arise from a particular long-term forest experiment, not a universal numerical threshold for northern rivers. The transferable idea is to test the stability assumption using evidence beyond the treated-reference pair.

Multiple references do not solve everything. They may share the same precipitation bias, disturbance, or gauge method. They may also differ physically from the target and from each other. Their role should be specified: independent counterfactual predictors, diagnostics for regional change, or estimates of background variability. Averaging several weakly justified references can obscure incompatibility rather than improve inference.

Selection should occur before inspecting the post-disturbance outcome wherever possible. Choosing whichever reference produces the preferred effect introduces an avoidable analytical bias. If a reference later becomes unsuitable, document the reason and show the implications of the change. A replacement series needs overlap and a rationale, not simply a new name in an unchanged model formula.

The monitoring agreement is part of comparability

A viable reference must be maintainable and legitimately accessible. Long records depend on permissions, local governance, stable stewardship, and the ability to document changes. Similar technical protocols help, but the reference does not become available merely because it is hydrologically attractive. Research arrangements must respect the relevant authorities and cannot treat neighbouring catchments as interchangeable land for an experiment.

Maintain comparable gauge metadata, time conventions, ice corrections, gap flags, and version histories. Record reference-side disturbances with the same care given to the target. A shared processing pipeline can improve consistency, yet common processing error remains possible. Independent checks at consequential low and

high flows are especially valuable where the stated question depends on those conditions.

Limits and a defensible choice

This review does not establish causal certainty from observational pairing, nor recommend intervention in any real catchment. Unmeasured differences can remain, and broad climatic change can undermine historical relationships. The cited forest experiments and global product evaluations must be adapted cautiously to permafrost, lakes, glaciers, and sparse northern networks.

A defensible selection states which flow behaviour the reference predicts, why its mechanisms are comparable, what conditions calibrated the relationship, and how stability will be checked. If no candidate supports the required counterfactual, report that limitation and use a narrower descriptive comparison. A reference catchment is most useful when its own capacity to change remains visible. That is what makes the comparison testable rather than merely convenient.

References

1. Brown, A. E., Zhang, L., McMahon, T. A., Western, A. W., and Vertessy, R. A. (2005). A review of paired catchment studies for determining changes in water yield resulting from alterations in vegetation. *Journal of Hydrology*, 310, 28–61.
2. Wondzell, S., Johnson, S., Grant, G., Henshaw, D., and Ward, A. (2025). Rethinking paired-catchment studies: Should we be replicating our controls? *Water Resources Research*, 61, e2024WR038981.
3. Michailovsky, C. I., Coerver, B., Mul, M., and Jewitt, G. (2023). Investigating sources of variability in closing the terrestrial water balance with remote sensing. *Hydrology and Earth System Sciences*, 27, 4335–4354.