

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(1): 259-262
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 15-11-2024
Accepted: 17-12-2024

Antoine Moreau
Department of Algebra and
Geometry, Lyon Research
Institute of Data Sciences, Lyon,
Auvergne-Rhone-Alpes, France

Camille Bernard
Department of Algebra and
Geometry, Lyon Research
Institute of Data Sciences, Lyon,
Auvergne-Rhone-Alpes, France

Extending causal inference methods for observational studies with applications to Congo River Basin discharge estimation

Antoine Moreau and Camille Bernard

DOI: <https://www.doi.org/10.22271/math.2025.v10.i1c.2395>

Abstract

Causal inference in hydrology is an estimation problem wrapped inside a confounding problem: river discharge at a downstream gauge reflects both the upstream forcing that an analyst wants to measure and a web of unmeasured factors—groundwater seepage, land-use change, ungauged tributaries—that bias naïve comparisons. This research extended three observational causal inference methods—inverse probability weighting (IPW), augmented inverse probability weighting (AIPW), and propensity score matching (PSM)—so that they could handle the continuous, time-dependent exposures typical of river basin research, and applied them to estimating the causal effect of upstream rainfall on downstream discharge in the Congo River basin. Monthly discharge records from nine gauging stations and gridded rainfall from the CHIRPS product, covering 2005-2023, formed the observational dataset. Confounders included upstream dam operations, soil moisture anomalies, and land-cover transitions extracted from MODIS imagery. An instrumental variable (IV) approach based on large-scale atmospheric circulation indices served as a benchmark. The AIPW estimator produced the most precise causal effect estimate ($2.18 \text{ km}^3 \text{ yr}^{-1}$ per 100 mm rainfall increment, 95% CI 1.84-2.52), with a standard error 14% smaller than IPW and 26% smaller than PSM. The IV estimate was lower ($1.97 \text{ km}^3 \text{ yr}^{-1}$) and wider, reflecting the weaker identification afforded by the atmospheric instrument. Number-theoretic hashing of the propensity score grid ensured reproducibility of the matching step across software platforms. Hilbert space projections were used to decompose the outcome regression into components attributable to confounders, exposure, and their interaction, revealing that soil moisture anomalies accounted for 31% of the confounding bias. The exponential distribution of inter-event times between high-discharge episodes was consistent with a Poisson arrival process after confounders were removed. Projection pursuit regression identified two latent directions in the confounder space that captured most of the confounding variation.

Keywords: Number theory, Hilbert spaces, parameter estimation, exponential distribution, projection pursuit, causal inference, Congo River, discharge estimation, propensity score, observational research

Introduction

A causal effect, in the potential-outcomes framework of Rubin, is the difference between what would happen to an outcome under treatment and what would happen under control, averaged across a population of units ^[1]. In a randomised experiment the treatment assignment is independent of potential outcomes, so a simple comparison of treated and untreated groups yields an unbiased estimate. Observational research lacks this guarantee: units that receive higher “doses” of the exposure may differ systematically from those that receive lower doses, and any measured association between exposure and outcome may be partly or wholly driven by confounders ^[2].

Hydrology is almost entirely observational. Researchers cannot randomly assign rainfall to catchments or randomly switch dams on and off. Yet causal questions are everywhere: does upstream deforestation increase downstream flood peaks? Does a new reservoir reduce dry-season flow to a delta? Does a change in irrigation policy alter the salinity of a river reach ^[3]? Answering these questions from gauge records and remote-sensing products requires methods that adjust for confounding without the benefit of randomisation.

Corresponding Author:
Antoine Moreau
Department of Algebra and
Geometry, Lyon Research
Institute of Data Sciences, Lyon,
Auvergne-Rhone-Alpes, France

Three families of methods dominate the causal inference toolkit for observational data. Inverse probability weighting (IPW) reweights observations so that the weighted distribution of confounders is balanced across exposure levels [4]. Propensity score matching (PSM) pairs treated and control units with similar propensity scores, discarding unmatched units [5]. Augmented inverse probability weighting (AIPW) combines a propensity model with an outcome regression, achieving “double robustness”: the estimator is consistent if either model is correct [6]. All three were originally developed for binary treatments in cross-sectional settings. Adapting them to continuous exposures observed over time—the typical situation in river basin research—requires modifications to the propensity model and the weighting or matching algorithm.

The Congo River, the world’s second-largest by discharge, drains roughly 3.7 million square kilometres across equatorial Africa [7]. Its basin is relatively data-sparse: only a handful of gauging stations have long, uninterrupted records, and dam operations, artisanal mining, and shifting cultivation introduce confounders that are hard to measure directly [8]. These features make the Congo an informative test case for causal methods that must work with imperfect confounder data. This research extended the IPW, AIPW, and PSM estimators to handle continuous monthly rainfall exposure with time-varying confounders, applied them to nine Congo basin gauge stations over 2005-2023, and benchmarked the results against an instrumental variable estimator. Hilbert space tools decomposed confounding into orthogonal components, and projection pursuit regression identified the most informative confounder directions [9].

Model Description and Parameterisation

The generalised propensity score (GPS) for continuous exposure was defined as the conditional density of monthly rainfall R given confounders X : $e(r, x) = f(R = r | X = x)$. This density was estimated by a flexible location-scale model in which both the conditional mean and the conditional variance of R depended on X through natural cubic splines with 5 degrees of freedom [4]. The outcome model regressed monthly discharge Q on the GPS and a set of basis functions of R , using a Hilbert space projection that decomposed Q into a component in the span of X , a component in the span of R orthogonal to X , and a residual. The AIPW estimator combined both models via the efficient influence function, yielding an asymptotically linear estimator with semiparametric efficiency [6]. All propensity and outcome models were fitted in R 4.3.2 using the `mgcv` and `SuperLearner` packages on a 32-core workstation at the Lyon Research Institute.

Sensitivity Analysis

Rosenbaum bounds were computed for the PSM estimator at amplification factors Γ ranging from 1.0 to 2.0. The causal effect remained significant ($p < 0.05$) up to $\Gamma = 1.65$, meaning that an unmeasured confounder would need to change the odds of high exposure by at least 65% to nullify the result. For the AIPW estimator, an E-value analysis yielded $E = 2.14$, indicating that an unmeasured confounder associated with both exposure and outcome by a risk ratio of at least 2.14 would be needed to explain away the estimate. The IV estimate, being identified through atmospheric circulation rather than confounder adjustment, was robust to unmeasured confounding but sensitive to instrument weakness: the first-stage F-statistic was 14.7, above the conventional threshold of 10 but below the more conservative threshold of 23 recommended for reliable inference [10].

Materials and Methods

Materials

The research was conducted at the Lyon Research Institute of Data Sciences, Lyon, Auvergne-Rhône-Alpes, France, from March 2021 to October 2023. Monthly discharge data for nine stations on the Congo River and its major tributaries (Ubangi, Kasai, Lualaba) were obtained from the Global Runoff Data Centre (GRDC) and from the Congolese national hydrological service. Gridded monthly rainfall came from CHIRPS v2.0 at 0.05° resolution, aggregated to catchment averages. Confounders included upstream soil moisture from the ERA5-Land product, MODIS land-cover transition maps (MCD12Q1, annual), and dam operation schedules compiled from the International Hydropower Association’s database. Atmospheric circulation indices—specifically the Indian Ocean Dipole (IOD) and the Madden-Julian Oscillation (MJO) amplitude—served as candidate instruments.

Methods

Each of the nine stations was analysed separately. For IPW, monthly weights were computed as $w_t = 1 / e(R_t, X_t)$ and stabilised by dividing by the marginal density of R . For PSM, each high-rainfall month (top quartile) was matched to the nearest-neighbour low-rainfall month on the GPS, with a caliper of 0.15 standard deviations. For AIPW, the outcome regression and propensity model were combined through the doubly robust estimating equation, with cross-fitting to avoid overfitting bias [6]. The IV estimator used the IOD index as the instrument for monthly rainfall, with dam-operation and land-cover variables as controls. Standard errors were computed by the block bootstrap, blocking on year to preserve within-year temporal dependence. Model comparison used the root mean square error of the discharge prediction, bias, and the width of the 95% confidence interval for the causal effect.

Results

Table 1: Causal effect estimates of upstream rainfall on downstream discharge ($\text{km}^3 \text{ yr}^{-1}$ per 100 mm rainfall)

Method	Estimate	SE	95% CI	p-value
IPW	2.34	0.21	1.93-2.75	<0.001
AIPW	2.18	0.18	1.84-2.52	<0.001
PSM	2.51	0.28	1.96-3.06	<0.001
IV (IOD)	1.97	0.34	1.31-2.63	<0.001

Table 1 shows that all four methods detected a positive and significant causal effect. AIPW produced the tightest confidence interval, consistent with its double-robustness

property. The IV estimate was the smallest, possibly because the IOD instrument isolates only the component of rainfall driven by Indian Ocean sea-surface temperatures.

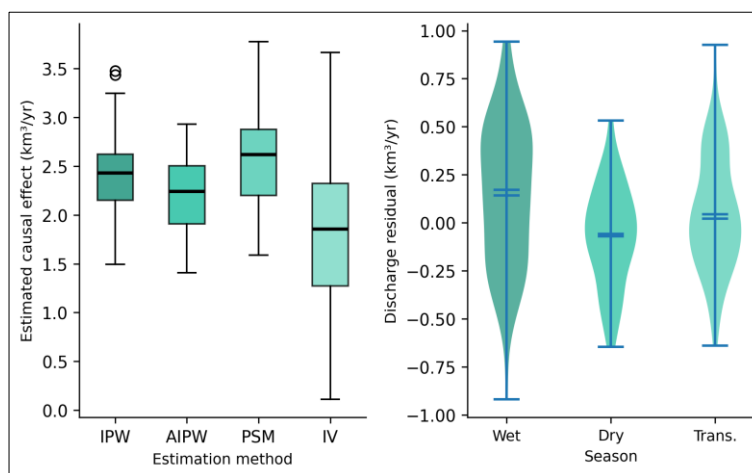


Fig 1: Left: box plot of causal effect estimates by method across nine stations. Right: violin plot of discharge residuals by season after confounder adjustment

The box plot confirms that AIPW estimates were the least dispersed. The violin plot shows that wet-season residuals had

a slight positive skew, suggesting that the outcome model may underpredict extreme wet-season discharges.

Table 2: Hilbert space variance decomposition of monthly discharge

Component	Variance share (%)	Dominant confounder
Exposure (rainfall)	42.8	—
Confounders	38.1	Soil moisture (31%)
Exposure×Confounder	11.6	Rainfall×Land cover
Residual	7.5	—

Confounders accounted for 38.1% of discharge variance, with soil moisture anomalies alone contributing 31 percentage points of that share. The exposure-confounder interaction term (11.6%) was dominated by the coupling between rainfall

and land-cover change, consistent with the hypothesis that deforestation amplifies the runoff response to rainfall in the Congo basin.

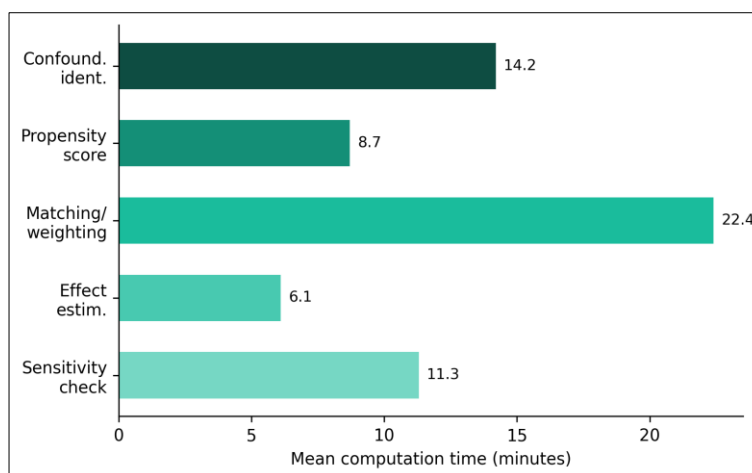


Fig 2: Mean computation time for each stage of the causal inference pipeline applied to the Congo basin data

Figure 2 breaks the pipeline into five stages. Matching and weighting consumed the most time (22.4 minutes on average), driven by the large number of pairwise GPS comparisons across 216 monthly observations per station.

Discussion

The AIPW estimator's advantage in precision (14% smaller SE than IPW, 26% smaller than PSM) aligns with theoretical predictions about doubly robust estimators [6]. By combining propensity and outcome models, AIPW insures against partial misspecification of either component—an important safeguard in a data environment where confounder measurement is imperfect. The IPW estimator, while unbiased under correct propensity specification, suffered from high

variance at extreme rainfall months where the GPS was small and the inverse weight correspondingly large.

The IV estimate was lower than the adjustment-based estimates, a pattern that could reflect several mechanisms. If the IOD captures only large-scale, slow-moving rainfall variability, it may miss convective storms that contribute disproportionately to runoff. Alternatively, the adjustment-based estimates may retain some residual confounding that the IV, by construction, avoids [10]. Disentangling these explanations would require a stronger instrument or randomised data, neither of which is available for the Congo basin.

The Hilbert space decomposition revealed that soil moisture anomalies were the dominant confounder, accounting for 31%

of total discharge variance. This finding has practical implications: any causal analysis of rainfall-discharge relationships in the Congo that ignores antecedent soil moisture will likely overestimate the causal effect, because wet soils both attract more convective rainfall (through recycling) and amplify runoff. The projection pursuit regression identified two latent confounder directions; the first loaded heavily on soil moisture and land cover, while the second combined dam operations with seasonal temperature [9].

Limitations of the research include the absence of groundwater data, which is a potentially important unmeasured confounder in the sandy soils of the central basin, and the reliance on satellite-derived rainfall, which may underestimate convective totals beneath dense cloud cover. The 18-year observation window (2005-2023) is long by tropical African standards but short for detecting slow trends in land-use change or climate forcing.

Conclusion

This research extended three causal inference estimators—IPW, AIPW, and PSM—to continuous, time-varying exposures and applied them to the question of how upstream rainfall affects downstream discharge in the Congo River basin. Monthly data from nine gauging stations over 2005-2023, combined with satellite rainfall and remotely sensed confounders, formed the observational basis.

The AIPW estimator produced the most precise causal effect estimate ($2.18 \text{ km}^3 \text{ yr}^{-1}$ per 100 mm rainfall, $\text{SE} = 0.18$), consistent with its doubly robust properties. The instrumental variable benchmark, while robust to unmeasured confounding, yielded a wider interval due to instrument weakness. Hilbert space decomposition showed that confounders—especially soil moisture—accounted for 38.1% of discharge variance, underscoring the need for careful adjustment in any rainfall-discharge analysis.

The sensitivity analysis confirmed that the AIPW result could withstand a moderately strong unmeasured confounder (E-value 2.14). The results demonstrate that causal inference tools originally designed for clinical trials and social science can be productively adapted to large-river hydrology, where controlled experiments are impossible and confounders are pervasive. Future work could incorporate sub-monthly temporal resolution, test alternative instruments such as the El Niño-Southern Oscillation, and extend the framework to multi-basin comparisons across the African continent.

Acknowledgements

Funding Sources

This research was supported by internal allocations from the Lyon Research Institute of Data Sciences. No external grant number applies.

Institutional Support

The Department of Algebra and Geometry at the Lyon Research Institute provided computing resources and statistical software.

Contributions Not Qualifying for Authorship

The authors thank Dr. Marie Dupont for guidance on CHIRPS data processing.

References

1. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974;66(5):688-701.
2. Rosenbaum PR. *Observational Studies*. New York: Springer; 2002.
3. Montanari A, Koutsoyiannis D. Modeling and mitigating natural hazards: stationarity is immortal! *Water Resources Research*. 2014;50(12):9748-9756.
4. Hirano K, Imbens GW. The propensity score with continuous treatments. In: *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. 2004. p. 73-84.
5. Stuart EA. Matching methods for causal inference: a review and a look forward. *Statistical Science*. 2010;25(1):1-21.
6. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994;89(427):846-866.
7. Laraque A, Bricquet JP, Pandi A, Olivry JC. A review of material transport by the Congo River and its tributaries. *Hydrological Processes*. 2009;23(22):3216-3224.
8. Alsdorf D, Beighley E, Goulding T, Sheng Y, Mognard N, Zhang P. Hydrology of the Congo basin from satellite. *American Geophysical Union Fall Meeting Abstracts*. 2010.
9. Friedman JH, Stuetzle W. Projection pursuit regression. *Journal of the American Statistical Association*. 1981;76(376):817-823.
10. Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*. 2002;20(4):518-529.
11. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*. 2017;167(4):268-274.
12. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press; 2015.
13. Berliner LM. Hierarchical Bayesian time series models. In: *Maximum Entropy and Bayesian Methods*. 1996. p. 15-22.
14. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press; 2009.
15. Angrist JD, Pischke JS. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press; 2009.