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 century of monthly rainfall in the Senegal River Basin (1895–2020): automated quality control, machine learning gap-filling, and the cost of network attrition

Mbayang Thiam

mbayangthiam58@gmail.com

Université Cheikh Anta Diop de Dakar

Abdoulaye Faty

Université Cheikh Anta Diop de Dakar

Awa Niang

Université Cheikh Anta Diop de Dakar

Research Article

Keywords: quality control, gap-filling, missing data, monthly rainfall, Senegal River Basin, network attrition

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10516844/v1

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

Additional Declarations: No competing interests reported.

Abstract

Long rainfall series underpin water management in the Senegal River Basin, yet the network has contracted sharply since the 1990s and its archive carries digitisation errors. We compile, quality-control and release the most complete monthly archive yet assembled there, 138 stations and 80,837 observations, 1895 to 2020, in four climato-hydrological zones. A two-stage automated procedure, pairing deterministic screens with a robust anomaly test cross-checked against neighbours, altered 1.4% of records, yet one uncorrected error displaces a zonal rainfall index by tens of standard deviations. Reconstruction skill is quantified by Monte Carlo masking of 61 quasi-complete stations in the dense 1950 to 1990 window, crossing three gap patterns, four rates and five replicates. For random and block gaps, neighbour-based methods recover masked months with Kling-Gupta efficiency near 0.90. Skill follows the balance between inter-station coherence and relative monsoon variability, peaking in the Sudanian zone, and machine learning matches but never exceeds per-station regression. Terminal gaps, created by attrition, differ, since every per-station method acquires a wet bias reaching 9.4 mm per month, which pooled gradient boosting with geographic and time covariates cuts roughly five-fold. A satellite comparison confirms the post-drought recovery while suggesting that its gauge-derived magnitude is attrition-inflated.

1. Introduction

Rainfall observations are the primary boundary condition of every hydrological, agronomic and climate-service application in West Africa. In the Senegal River Basin, shared by Guinea, Mali, Mauritania and Senegal under the Organisation pour la Mise en Valeur du fleuve Sénégal (OMVS), rain-gauge series inform reservoir operation at Manantali and Diama, flood forecasting in the middle valley, irrigation planning in the delta, and the basin-wide assessments that support transboundary water diplomacy, that is, allocation and negotiation frameworks whose credibility ultimately rests on shared, trusted hydrometeorological records [1]. The region is also a canonical case of low-frequency rainfall variability, with a wet period in the 1950s and 1960s, the great Sahel drought of 1968 to 1993, and a partial recovery observed since the mid-1990s [2–6], with associated shifts in the regime of extreme rainfall [7] and in Sahelian runoff [8]. Detecting and interpreting such multidecadal signals places stringent demands on the completeness and accuracy of station records.

Two structural problems undermine these demands. First, the observation network has contracted sharply. In the archive assembled here, the number of active stations in the basin fell from a peak of 103 in 1975 to 65 in 2000 and only 20 by 2015, mirroring the well-documented decline of hydrometeorological monitoring across sub-Saharan Africa. Second, the historical archive, largely digitised from paper registers, contains typical transcription errors such as decimal shifts, duplicated months and impossible values. Both problems are usually handled ad hoc, station by station, with methods such as station climatology, the normal-ratio approach and inverse-distance weighting that date back to Paulhus and Kohler [9] and remain the operational default in the region’s services [10, 11].

Machine learning methods, notably Random Forest [12] and gradient boosting [13], have been proposed for precipitation infilling in various contexts [14], including Random Forest gap-filling of dense hydrometeorological networks in Chile [15], analogue-based reconstruction of forcing in the data-scarce Volta Basin [16], deep-learning precipitation and discharge modelling [17] and gradient-boosting bias correction of forecast products [18]. Applications to Senegalese synoptic stations are emerging [19]. Yet systematic evidence for West African monthly networks is scarce, and three questions remain open. First, does machine learning outperform the simple neighbour-based methods that services already use, once those are properly configured. Second, how does performance depend on the pattern of missingness, whether isolated months, whole missing years, or the terminal gaps created by station closure. Third, what happens when the training period and the reconstruction period belong to different rainfall regimes, as is unavoidable in a region with the Sahel's non-stationarity, and can the resulting bias be mitigated.

This paper addresses these questions using, to our knowledge, the most complete monthly rainfall compilation yet assembled for the Senegal River Basin, comprising 138 stations and 80,837 station-months over 1895 to 2020. Our contribution is threefold. First, we assemble and release a quality-controlled, century-long monthly database for the basin, with a fully automated, reproducible quality-control pipeline whose impact is quantified, including a demonstration that a single uncorrected decimal-shift error can distort a zonal rainfall index by tens of standard deviations. Second, we quantify reconstruction skill with replicated Monte Carlo masking experiments across gap patterns, rates and zones, and explain mechanistically why zones differ. The six-method benchmark, from the normal ratio to gradient boosting, serves as the instrument rather than the destination. Third, we isolate, quantify and partially remediate the non-stationarity bias that afflicts reconstruction of the attrition period, with direct implications for how post-1990 rainfall statistics in the basin should be interpreted.

2. Study area and data

2.1. The Senegal River Basin and its rainfall gradient

The Senegal River, about 1,800 km long, drains roughly 300,000 km² from the Fouta Djallon highlands of Guinea to its Atlantic delta at Saint-Louis. Rainfall is governed by the West African monsoon and follows a steep south to north gradient. Rather than administrative country groupings, we stratify the network into four climato-hydrological zones defined by each station's mean annual rainfall over the 1950 to 1990 reference window (Table 1). The zoning mirrors the basin's upstream to downstream structure. The Guinean zone corresponds to the headwater catchments that generate most of the river's discharge, the Sudanian zone to the upper basin of the Bafing, Bakoye and Falémé tributaries, the Sudano-Sahelian zone to the upstream valley, and the Sahelian zone to the middle and lower valley, the delta and the Mauritanian right bank. Mean annual totals range from 249 mm in the Sahelian zone to 1,449 mm in the Guinean zone, a factor of nearly six within a single basin, and the seasonal regime is unimodal, peaking in August in all zones (Fig. 2). An instructive asymmetry emerges between observation density and hydrological production, since the Sahelian zone concentrates 59 of the 138 stations but contributes

marginally to runoff, whereas the runoff-generating Guinean zone counts only 19, albeit the best-maintained of the network with 1.0% missing.

Table 1
The 138-station network by climato-hydrological zone. Gap rate is computed within each station's active period; annual rainfall is the 1950 to 1990 station mean, or the station record mean where that window is unavailable.

Climato-hydrological zone	Stations	Mean gap rate	Complete years (mean/station)	Mean annual rainfall (mm)
Sahelian, < 400 mm (lower/middle valley, delta, right bank)	59	13.3%	41	249
Sudano-Sahelian, 400–700 mm (upstream valley)	29	8.4%	59	508
Sudanian, 700–1,200 mm (upper basin Bafing–Bakoye–Falémé)	31	4.2%	47	890
Guinean, > 1,200 mm (Fouta Djallon headwaters)	19	1.0%	55	1,449

2.2. The station archive

The archive consists of 138 station files with harmonised monthly series, one date and value table per station, compiled from the national services of the four riparian countries. It totals 80,837 monthly observations, about 6,736 station-years, from January 1895 at Kayes to December 2020 at Bakel, Kidira and Saint-Louis. Figure 3 maps availability. The network grew from 16 active stations in 1920 to 72 in 1950 and a maximum of 103 in 1975, then eroded to 65 in 2000 and 20 in 2015. Within station active periods, decadal completeness never falls below 82%. Crucially for what follows, 67 stations observe at least 95% of the 492 months of the 1950 to 1990 window, and 61 do so after quality control, providing a dense training and validation ground. Companion GIS layers, a 45-station basin shapefile and a 219-entry national station directory, georeference 59 of the 138 stations (Fig. 1). Completing the georeferencing from OMVS and national directories is part of the data release.

3. Methods

3.1. Automated quality control

Quality control proceeds in two stages, each action being logged. Stage one applies deterministic screens. Duplicated months are collapsed at ingestion, affecting 631 records concentrated at Kayes and Saint-Louis Ville. Negative values are masked. Monthly totals above a physical ceiling of 1,500 mm, implausible anywhere in the basin, are tested for decimal-shift repair. If the value divided by 100 lies within 3.5 standard deviations of the station's monthly climatology, it is corrected, otherwise masked. Stage two applies a robust statistical screen. A modified z-score based on the monthly median and

median absolute deviation flags values above 5 as suspect. The flag is confirmed only if the median standardized anomaly of the eight most-correlated neighbour stations that month does not support an extreme of the same sign; flagged values with neighbour support are retained as genuine extremes. This two-stage design deliberately favours specificity, since the goal is to remove transcription artefacts rather than wet-season maxima.

3.2. Gap-filling methods

Six estimators are compared, all using only information available within the network. CLIM is the station's monthly climatology, the natural lower benchmark. NR is the normal-ratio method [9], a correlation-weighted mean of neighbour values rescaled by the ratio of monthly normals. LREG is ordinary least-squares regression of the target station on its eight most-correlated neighbours plus harmonic month terms. KNN is distance-weighted k-nearest-neighbour regression with k equal to ten in the same feature space. RF is a Random Forest of 150 trees. XGB is gradient-boosted trees with 350 trees, learning rate 0.05 and depth 4. For the last four, neighbour predictors are climatology-filled where themselves missing, and predictions are truncated at zero. Neighbour sets are selected by correlation of monthly standardized anomalies over 1950 to 1990, which avoids requiring coordinates for the not-yet-georeferenced stations while implicitly encoding spatial proximity. All six estimators therefore operate under strictly equal information, the network itself, which makes the comparison a test of model class rather than of predictor richness.

3.3. Covariate-augmented pooled models

To test whether machine learning benefits from the predictors that per-station designs deny it, a complementary experiment uses the 35 benchmark stations that are georeferenced. A single pooled model is trained across all station-months with features combining local information, namely the station monthly climatology, the normal-ratio estimate and the mean standardized anomaly of the eight most-correlated neighbours, geography, namely latitude, longitude, elevation where available and distance to the Atlantic coast, and time, namely seasonal harmonics and calendar year, the latter allowing the model to track low-frequency regime change. Random Forest and XGBoost are each fitted with and without the geographic block and compared against the per-station baselines on identical masks. Large-scale climate indices were deliberately left out and are reserved, together with satellite co-predictors, for a companion study.

3.4. Masking experiment design

All experiments run on the dense window of 1950 to 1990, comprising 492 months, using the 61 stations that remain at least 95% complete after quality control. For each configuration we mask a fraction of each station's observed months, refit every method on the remaining data, and evaluate on the masked truth. Three patterns are crossed with four rates of 5, 10, 20 and 40%. The random pattern masks isolated months, mimicking scattered register losses. The block pattern masks whole calendar years, mimicking lost yearbooks. The terminal pattern masks the final fraction of the window, mimicking station closure and network attrition. Random and block masks are replicated five times with

independent seeds, a Monte Carlo design yielding dispersion estimates for every score, whereas the terminal mask is deterministic and needs no replication. The terminal pattern is deliberately severe, since the window ends inside the great drought and therefore forces models trained mostly on the wetter pre-1968 regime to reconstruct dry years, which is exactly the situation faced when reconstructing the post-1990 attrition period from earlier records. Performance is measured by RMSE, MAE, bias, the coefficient of determination and the Kling-Gupta efficiency [20], overall and by zone.

4. Results

4.1. Quality control outcomes

Quality control touched 1,143 records, or 1.4% of the archive (Table 2). The headline cases are decimal shifts. Oualia in Mali reads 111,622.8 mm for July 1996 and Yélimané in Mali reads 90,120 mm for August 1936, and the division by one hundred restores plausible monsoon values wherever the station’s climatology supports it. The consequences of leaving such errors untreated are severe out of proportion to their number. Computed on raw data, the standardized rainfall index of the Sudanian zone jumps to +77 standard deviations in 1996, an aberration produced by a single station-year. After quality control, the familiar regional signal re-emerges cleanly in all four zones (Fig. 4), namely the 1950 to 1967 wet phase, the 1968 to 1993 drought, and the partial recovery since the mid-1990s, strongest in the Sahelian zone but entangled there with network attrition.

Two station-level pathologies, invisible to record-by-record screening, were uncovered by extended diagnostics and are flagged in the released base. At Saldé in the middle valley, the 1962 to 1989 portion of the series is incompatible with rainfall in millimetres, with annual sums of 26 to 42 mm and roughly half the wet-season signal of neighbouring Podor missing, and most plausibly records a different variable such as rain-day counts before reverting to credible totals after 1990. The station is excluded from climatological statistics pending archival verification and does not enter the dense-window experiments. At Saint-Louis Ville, 325 of the 576 duplicated months carry contradictory values, for example August 1951 recorded as both 0.0 and 131.5 mm, indicating a merge of two distinct sources. First occurrences were retained at ingestion, the series is flagged, and reconciliation against the original registers is required before climatological use. Both cases illustrate a category of error, inconsistent provenance, that only station-level diagnostics can catch, and both are documented in the quality-control log.

Table 2
Quality-control actions on the 138-station archive of 80,837 monthly records, of which 1.4% were touched.

Quality-control action	Count	Examples
Negative values masked	1	Aéré Lao, Sep 1996 (− 2,997 mm)
Impossible highs (> 1,500 mm) corrected by ÷ 100 after validation against station climatology	10	Oualia, May and Aug–Oct 1996; Yélimané 1936; Boghé 1995
Impossible highs not repairable, masked	3	Oualia Jul 1996 (111,622.8 mm); Yélimané Aug 1936
Robust-z outliers (|z| > 5) without neighbour support, masked	954	isolated dry-season spikes
Robust-z outliers kept (neighbours concur)	175	genuine wet extremes
Duplicate month records collapsed at ingestion	631	Kayes (55), Saint-Louis Ville (576)

4.2. Benchmark: overall performance

Figure 5 summarises the design and Table 3 details the central case of the random pattern at 10% masking, given as mean plus or minus one standard deviation over the five mask replicates. Three results stand out. First, using the network beats not using it. Every neighbour-based method improves substantially on climatology, with RMSE reductions averaging 17.5 ± 2.5% for regression across random-pattern configurations and Kling-Gupta efficiency rising from about 0.84 to about 0.90. Second, the best performers are the simplest, and the ranking is stable. Across all 40 replicated configurations, per-station linear regression achieves the lowest RMSE in 28 and the normal-ratio method in the remaining 12, while the tree-based models win in none. Third, under equal information the machine-learning models do not repay their complexity. Random Forest tracks regression to within one standard deviation and XGBoost trails consistently, with a mean improvement over climatology of 10.3 ± 3.0% against 17.5 ± 2.5%. Performance degrades only mildly from 5 to 40% masking, because the retained network is dense enough over 1950 to 1990 that even heavy thinning leaves sufficient covariance to exploit.

Table 3
Overall performance on masked months for the random pattern at 10% rate, over the 1950 to 1990 window and 61 stations, given as mean plus or minus one standard deviation over five Monte Carlo mask replicates.

Method	RMSE (mm)	MAE (mm)	Bias (mm)	R^2	KGE
Climatology	45.6 ± 1.3	22.8 ± 0.8	−0.4 ± 1.0	0.790 ± 0.007	0.843 ± 0.016
Normal ratio	37.1 ± 0.9	17.9 ± 0.4	−0.3 ± 0.8	0.860 ± 0.010	0.913 ± 0.008
Linear regression	36.6 ± 1.4	18.1 ± 0.5	+ 0.4 ± 0.7	0.864 ± 0.014	0.904 ± 0.007
k-NN	38.2 ± 1.6	18.4 ± 0.5	−2.7 ± 0.8	0.852 ± 0.014	0.866 ± 0.012
Random Forest	38.0 ± 1.2	18.4 ± 0.4	−0.3 ± 0.8	0.854 ± 0.010	0.894 ± 0.011
XGBoost	39.9 ± 1.1	19.2 ± 0.5	−0.5 ± 0.7	0.839 ± 0.011	0.894 ± 0.009

4.3. Performance by climato-hydrological zone

Zonal stratification (Fig. 6) shows that absolute and relative skill decouple along the rainfall gradient. Absolute errors scale with rainfall itself, since RMSE for linear regression ranges from 21.7 mm per month in the Sahelian zone to 61.1 mm in the Guinean zone, while explained variance peaks in the Sudanian zone at 0.886, where the network combines moderate variability with good density, and is lowest in the Sahelian zone at 0.775, where convective patchiness weakens inter-station correlation. The gain from using neighbours over climatology is largest precisely in the Sahelian and Sudano-Sahelian zones, with a gain in explained variance of about 0.14, and these are the zones with the highest missingness, an encouraging alignment of need and benefit for the archive at hand.

4.4. Why zones differ: coherence versus variability

The zonal contrasts of Fig. 6 are not accidental, since they follow from two measurable and competing properties of the rainfall field (Table 4). What an estimator can exploit is spatial coherence, the median correlation between a station and its eight best neighbours, and what it must explain is relative variability, summarised by the wet-season coefficient of variation over July to September. Coherence is actually highest in the Sahelian and Sudano-Sahelian zones, at 0.45 to 0.48, where interannual forcing is strongly shared, but the Sahelian zone couples it with the basin's most erratic monsoon, with a coefficient of variation of 0.70 and occasional rainless wet-season months. Much of the month-to-month variance is therefore convective noise that no neighbour can see, which caps explained variance at 0.78 despite the smallest absolute errors. The Guinean headwaters present the mirror image, combining the steadiest regime with a coefficient of variation of 0.35 and the weakest coherence at 0.39, a signature of orographic heterogeneity in the Fouta Djallon sampled by only ten dense-window gauges. The optimum lies between, since the Sudanian zone combines moderate variability with solid coherence, yielding the basin's best reconstructions. The practical corollary inverts the network's historical logic. The zones that pose the hardest reconstruction problem per gauge, Guinean and Sahelian, are respectively the basin's

runoff engine and its most-instrumented but noisiest region, so densification pays most exactly where the archive is thinnest relative to the field's structure.

Table 4
Mechanisms of zone-wise reconstructibility for dense-window stations, contrasting the spatial coherence available to estimators with the relative variability to be explained. Explained variance comes from the random pattern at 10% masking with linear regression.

Zone	Stations (dense window)	Median top-8 neighbour corr.	Wet-season CV	R^2 (random 10%)
Sahelian	20	0.45	0.70	0.775
Sudano-Sahelian	14	0.48	0.53	0.830
Sudanian	17	0.43	0.40	0.886
Guinean	10	0.39	0.35	0.794

4.5. Terminal gaps and the non-stationarity penalty

The terminal pattern breaks the comfortable picture. All methods lose 13 to 28 points of Kling-Gupta efficiency relative to the random pattern at equal rates, converging toward, and at some rates barely exceeding, the climatology baseline (Fig. 5, right column). More importantly, every method acquires a systematic wet bias that grows with gap length, from 8.5 mm per month averaged across methods at 5% masking to 15.0 mm per month at 40%. The mechanism is plain. With the masked span falling in the 1968 to 1990 drought, models calibrated predominantly on the wetter preceding decades transfer a wet climate into the reconstruction. Neighbour predictors mitigate but do not remove the effect, because neighbour records within the same climatic regime carry the drought signal only where those neighbours remain active. Figure 7 illustrates the case of Sélibaby in the Sudano-Sahelian zone, where the reconstruction tracks interannual variability credibly yet accumulates a mean annual excess of about 41 mm over the masked span from 1974 to 1990. This is the configuration in which the real archive finds itself after 1995, with few surviving stations and a shifting regime, and it argues for explicit uncertainty language and, where possible, satellite-era co-predictors in any reconstruction of the attrition period.

The covariate experiment of Section 3.3 shows that this penalty is not irreducible. It is here, not in routine infilling, that machine learning finally earns its keep (Table 5). On the 35-station georeferenced subset under the terminal mask at 40%, per-station regression carries a wet bias of 9.4 mm per month with a Kling-Gupta efficiency of 0.756. Pooling the stations into a single model and giving it a calendar-year covariate lets the learner infer the ongoing regime shift from the stations that remain active. XGBoost cuts the bias to 2.0 mm per month and lifts the efficiency to 0.865, with Random Forest close behind, and stripping the geographic block degrades RMSE only marginally, which shows that the essential ingredients are pooling and time-awareness rather than geography as such. For random gaps, by contrast, the same covariate-augmented models still fail to beat per-station regression, at 32.9 against 32.4 mm per month on this subset. The two regimes of the problem could hardly be drawn more sharply.

Table 5
The attrition scenario is where model class matters. Results are for the terminal mask at 40% on the georeferenced subset, where pooled, time-aware learners cut the non-stationarity wet bias several-fold.

Method (terminal 40%, 35 georeferenced stations)	RMSE (mm)	Bias (mm)	R^2	KGE
Linear regression (per station)	39.5	+ 9.4	0.738	0.756
Normal ratio (per station)	41.1	+ 2.7	0.717	0.848
Random Forest, pooled with covariates	38.7	+ 3.6	0.748	0.852
XGBoost, pooled with covariates	37.3	+ 2.0	0.767	0.865
XGBoost, pooled, no geographic block	37.5	+ 0.9	0.764	0.855

4.6. External consistency check against satellite estimates

As an independent, satellite-era check, the quality-controlled network is compared with CHIRPS version 2.0 annual rainfall aggregated over a Sahel-centred window of the basin, using the regional mean of annual accumulations from 1981 to 2025 [21]. Standardized anomalies of the gauge-network annual mean and of the satellite regional mean correlate at 0.72 over 1981 to 2016 (Fig. 8), despite the imperfect match between the satellite averaging window and the gauge set, and despite a network that shrinks to 15 reporting stations by 2016. Both records confirm the post-drought recovery, with the satellite product showing 24.6 mm per decade, or about 0.5 standard deviations per decade, over 1981 to 2014. Instructively, the gauge-network trend over the same window is steeper at about 0.8 standard deviations per decade. The satellite record thus corroborates the direction of the recovery while suggesting that its gauge-derived magnitude is inflated, which is independent external support for the attrition-bias mechanism quantified in Section 4.5. A pixel-scale confrontation at the georeferenced stations, using the full satellite grids, is the designed next step of this check.

5. Discussion

Three findings deserve emphasis, each with practical consequences for the basin's data stewardship.

Machine learning helps where the problem is non-stationary, not where it is routine. The replicated experiments locate the boundary precisely. For stationary gap types such as random months and missing years, the covariance structure of a monthly network is essentially linear, so per-station regression and the normal ratio reach the information ceiling, and neither tree ensembles under equal information nor covariate-augmented pooled learners surpass them. But for the terminal gaps created by station closure under regime change, model class matters, since pooled, time-aware gradient boosting cuts the systematic wet bias roughly five-fold relative to per-station regression. The operational translation is two-sided. For routine infilling, the normal-ratio and regression tools the region's services already master, applied with correlation-based neighbour selection and proper cross-validation, are competitive with the state of the art, so investment should target data rescue and network density rather

than model sophistication. Recent comparative studies point the same way once scale is accounted for. Sandoval-Quilodrán et al. [15] obtain their best Random Forest gap-filling performance in south-central Chile at monthly aggregation with dense neighbour networks, conditions under which our results show linear methods already saturate the attainable skill, while the settings where tree-based and deep models demonstrably pay off involve heterogeneous or high-frequency predictors, namely multi-product forecast bias correction [18], multi-temporal-scale precipitation and discharge transformation [17], or spatially complex sub-daily extremes where uncertainty-aware regionalization is required [22]. Our contribution is to establish, by controlled experiment, on which side of that divide the monthly infilling problem of a Sahelian network falls, and it falls on the linear side. Toure et al. [19] reach a compatible conclusion for Senegalese synoptic stations.

Attrition-period reconstructions are biased, not merely uncertain. The wet bias of 8 to 15 mm per month under terminal gaps is a structural effect of regime change rather than a defect of any particular estimator, since all six methods exhibit it in the same direction and magnitude. Studies that infill post-1990 gaps by regression on longer-record neighbours, or that extend climatologies across the attrition period, risk overstating the Sahelian recovery. Conversely, the strong post-2000 recovery signal in our Sahelian zonal index (Fig. 4), computed on a shrinking station set, should itself be read with this caveat, a caution now supported externally since the satellite record confirms the recovery's direction but at a markedly gentler slope than the shrinking gauge network. Disentangling genuine recovery from attrition artefacts is a natural next application of the reconstructed database, ideally anchored on satellite-era products from 1981 onward [21]. The anchoring must itself be chosen with care, since recent evaluations in African and Mediterranean basins show that gauge-calibrated products perform best at monthly scale but degrade for high-intensity events and in data-sparse terrain [23–25], which argues for using them as large-scale constraints on the reconstruction rather than as point-scale truth. Where even that is impossible, analogue-based reconstruction offers a complementary route recently demonstrated in the neighbouring Volta Basin [16].

Quality control is cheap, and its absence is expensive. The entire automated pipeline touches 1.4% of records, yet a single untreated record was shown to displace a zonal index by tens of standard deviations. The asymmetry argues for making logged, reversible, automated quality control a standard preprocessing layer for every historical archive in the region, particularly ahead of frequency analyses whose extreme-value estimates are acutely sensitive to spurious maxima. Our neighbour-confirmation rule, which retains flagged extremes when correlated neighbours concur, preserved 175 genuine wet extremes that a naive z-score filter would have destroyed, a point of direct relevance to regional frequency analysis where quantile estimates inherit the full weight of every retained maximum [26, 27]. The same logic increasingly governs climate-projection work, since multi-site assessments now treat the critical evaluation of historical station records as a precondition for downscaling and bias correction [28], and our logged, reversible layer is designed to slot into exactly that workflow.

A basin stake, not merely a methodological one. Reliable, gap-free monthly rainfall underlies the questions that currently animate West African hydrology. The attribution of the Niger's recurrent Niamey

floods to evolving mesoscale convective systems rests on trustworthy rainfall and runoff baselines [29]. Water-allocation bargaining among riparian states presupposes records all parties accept [1]. And the OMVS's reservoir and irrigation planning inherits every bias present in the underlying climatology. By quantifying exactly where the reconstruction of this basin's record is trustworthy, namely random and block gaps in all zones with a Kling-Gupta efficiency near 0.90, and where it is structurally biased, namely terminal gaps under regime change, this study gives those downstream uses an explicit error budget rather than an implicit act of faith.

Limitations. Neighbour selection by correlation, adopted because 79 stations still await georeferencing, ignores physical distance and orography. A fully georeferenced rerun with kriging-type methods is planned, and the covariate experiment is necessarily confined to the 35 georeferenced dense-window stations, with elevation available for only part of them. Large-scale climate indices and satellite or reanalysis co-predictors were deliberately excluded, so the covariate set tests geography and time rather than the full predictor space, and the satellite check of Section 4.6 is regional and annual, so a pixel-scale monthly confrontation at the georeferenced stations is the natural next step. The benchmark is confined to the 1950 to 1990 window and the monthly scale, so daily-scale conclusions may differ. Two flagged stations, Saldé and Saint-Louis Ville, await archival reconciliation. Finally, exploiting the reconstructed base for trend, variability and extremes analysis, the questions that motivate it, is reserved for a companion study.

6. Conclusions

We assembled, quality-controlled and benchmarked the most complete monthly rainfall database yet compiled for the Senegal River Basin, comprising 138 stations over 1895 to 2020 and spanning a six-fold rainfall gradient. An automated two-stage quality-control procedure touched 1.4% of records while repairing spectacular digitisation errors whose neglect would corrupt any regional analysis. Replicated Monte Carlo masking experiments established that neighbour-based infilling recovers random and block gaps with high, stable skill at rates up to 40%, that zone-wise skill is governed by the balance between spatial coherence and relative monsoon variability, that linear methods reach the information ceiling for these stationary gap types, and that terminal gaps, the pattern that actually afflicts the archive, carry a non-stationarity penalty whose systematic wet bias pooled, time-aware machine learning reduces several-fold without eliminating the loss of skill. The quality-controlled database, quality-control log and pipeline code are released openly, providing the basin's four riparian countries with a common, citable foundation for climate services, frequency analysis and transboundary water management, and a template transferable to the neighbouring Gambia and Niger basins.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Author contributions

Mbayang Thiam conceived the study, curated the data, designed and implemented the quality-control and gap-filling methods, performed the analysis, prepared the figures and wrote the original draft. Abdoulaye Faty contributed to the hydrological interpretation of the results and to the review and editing of the manuscript. Awa Niang contributed to the framing of the study, to the interpretation of the zonal analysis and to the review and editing of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable. The study uses only historical instrumental rainfall records and involves no human participants, human data or animals.

Consent for publication

Not applicable.

Data availability

The quality-controlled monthly database, the complete quality-control log, the benchmark results and the Python pipeline will be made available through an open-access repository upon publication, and from the corresponding author upon reasonable request. Raw station files originate from the national meteorological and hydrological services of Guinea, Mali, Mauritania and Senegal and are redistributed within their respective data policies.

Dual publication

The manuscript is original, has not been published previously and is not under consideration by another journal. A companion manuscript analysing rainfall change on the reconstructed database is in preparation and is cited as such; it does not overlap with the present work.

Third-party material

All figures are original and produced by the authors. No third-party material requiring permission is reproduced.

Acknowledgements

The authors thank the national meteorological and hydrological services of Guinea, Mali, Mauritania and Senegal, whose observers and archivists produced the records analysed here.

References

1. Yuan L, He W, Wu X, Kong Y, Yang Y, Ramsey TS, Degefu DM. Allocating water resources in transboundary river basins: a sequential Rubinstein bargaining approach with risk discounting. J Hydrol Reg Stud. 2026;63:102989.
2. L'Hôte Y, Mahé G, Somé B, Triboulet JP. Analysis of a Sahelian annual rainfall index from 1896 to 2000: the drought continues. Hydrol Sci J. 2002;47(4):563–72.
3. Ali A, Lebel T. The Sahelian standardized rainfall index revisited. Int J Climatol. 2009;29(12):1705–14.
4. Lebel T, Ali A. Recent trends in the Central and Western Sahel rainfall regime (1990–2007). J Hydrol. 2009;375(1–2):52–64.
5. Nicholson SE. The West African Sahel: a review of recent studies on the rainfall regime and its interannual variability. ISRN Meteorol. 2013;2013:453521.
6. Bodian A, Diop L, Panthou G, Dacosta H, Deme A, Dezetter A, Ndiaye PM, Diouf I, Vischel T. Recent trend in hydroclimatic conditions in the Senegal River Basin. Water. 2020;12:436. https://doi.org/10.3390/w12020436.
7. Panthou G, Vischel T, Lebel T. Recent trends in the regime of extreme rainfall in the Central Sahel. Int J Climatol. 2014;34(15):3998–4006.
8. Mahé G, Paturel JE. 1896–2006 Sahelian annual rainfall variability and runoff increase of Sahelian rivers. C R Geosci. 2009;341(7):538–46.
9. Paulhus JLH, Kohler MA. Interpolation of missing precipitation records. Mon Weather Rev. 1952;80(8):129–33.
10. Teegavarapu RSV, Chandramouli V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol. 2005;312(1–4):191–206.
11. World Meteorological Organization. Guide to climatological practices. WMO-No. 100. Geneva: WMO; 2011.
12. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

13. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. pp. 785–94.
14. Kim JW, Pachepsky YA. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J Hydrol. 2010;394(3–4):305–14.
15. Sandoval-Quilodrán P, Barrientos G, Valdés J, León-Soto I, Mardones P, Duarte E, Arriagada P, Aguayo M, Rubilar R. Machine learning gap-filling of hydrometeorological time series in south-central Chile: streamflow, temperature and precipitation. J Hydrol Reg Stud. 2026;66:103659.
16. Gerber L, Dembélé M, Mariéthoz G. Analogue-based reconstruction of hydrological forcing for modelling in data-scarce regions: application to the Volta River Basin. J Hydrol Reg Stud. 2026;66:103571.
17. Ma C, Ma S, Hao Y, Zhu J, Lei Q, Sang J, Hao H. Deep learning based multi-temporal scale precipitation modeling for spring discharge prediction, Shentou springs, China. J Hydrol Reg Stud. 2026;63:103097.
18. Javadi M, Alizadeh H. Gradient boosting machine for enhancing global forecast data products of seasonal precipitation: Hirmand river basin experience. J Hydrol Reg Stud. 2026;66:103561.
19. Toure M, Klutse NAB, Sarr MA, Bhuiyan MAE, Kenne AD, Thiaw WM, Badiane D, Gaye AT, Ndiaye O, Mbow C. Machine learning approaches for imputing missing meteorological data in Senegal. Appl Comput Geosci. 2025;27:100281.
20. Gupta HV, Kling H, Yilmaz KK, Martinez GF. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J Hydrol. 2009;377(1–2):80–91.
21. Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, Husak G, et al. The climate hazards infrared precipitation with stations: a new environmental record for monitoring extremes. Sci Data. 2015;2:150066. https://doi.org/10.1038/sdata.2015.66.
22. Malamos N, Mazzoglio P, Iliopoulou T. Uncertainty-informed regionalization of sub-daily rainfall extremes in northwestern Italy using bilinear surface smoothing with credible-interval-based data screening. J Hydrol Reg Stud. 2026;66:103720.
23. Desta T, Zewdie M, Assefa T, Tesfaye T. Statistical evaluation of multi-satellite precipitation products: performance, uncertainty, and sensitivity analysis against ground observations and climatic trends in the Omo-Gibe transboundary basin, Ethiopia. J Hydrol Reg Stud. 2026;66:103683.
24. Yıldız MB, Di Nunno F, de Marinis G, Granata F. Evaluating global precipitation datasets over Sicily: from daily estimates to extreme events. J Hydrol Reg Stud. 2026;63:103062.
25. Aschale TM, Palazzolo N, Buonacera G, Cancelliere A, Peres DJ. An evaluation of ERA5-Land reanalysis data for drought monitoring: comparison with drought indices from observational datasets in Sicily. Hydrol Res. 2026;57:100016.
26. Hosking JRM, Wallis JR. Regional frequency analysis: an approach based on L-moments. Cambridge: Cambridge University Press; 1997.

27. Barna DM, Engeland K, Kneib T, Thorarinsdottir TL, Xu CY. Regional flood frequency analysis at multiple durations: a comparison of duration consistency in quantile and parameter regression techniques. Hydrol Res. 2026;57:100002.
28. Secci D, Todaro V, D'Oria M, Jomaa S, Tanda MG. Multi-site evaluation of historical climate records and CMIP6 projections across the Mediterranean. J Hydrol Reg Stud. 2026;66:103713.
29. Guehi FJ, Jucá Oliveira RA, Gosset M, Kacou M, Zahiri EP, Fiolleau T. A revisit of Niger river's major flood events and the role of convective systems with satellite data and hydrological modeling. J Hydrol Reg Stud. 2026;63:103025.

Figures

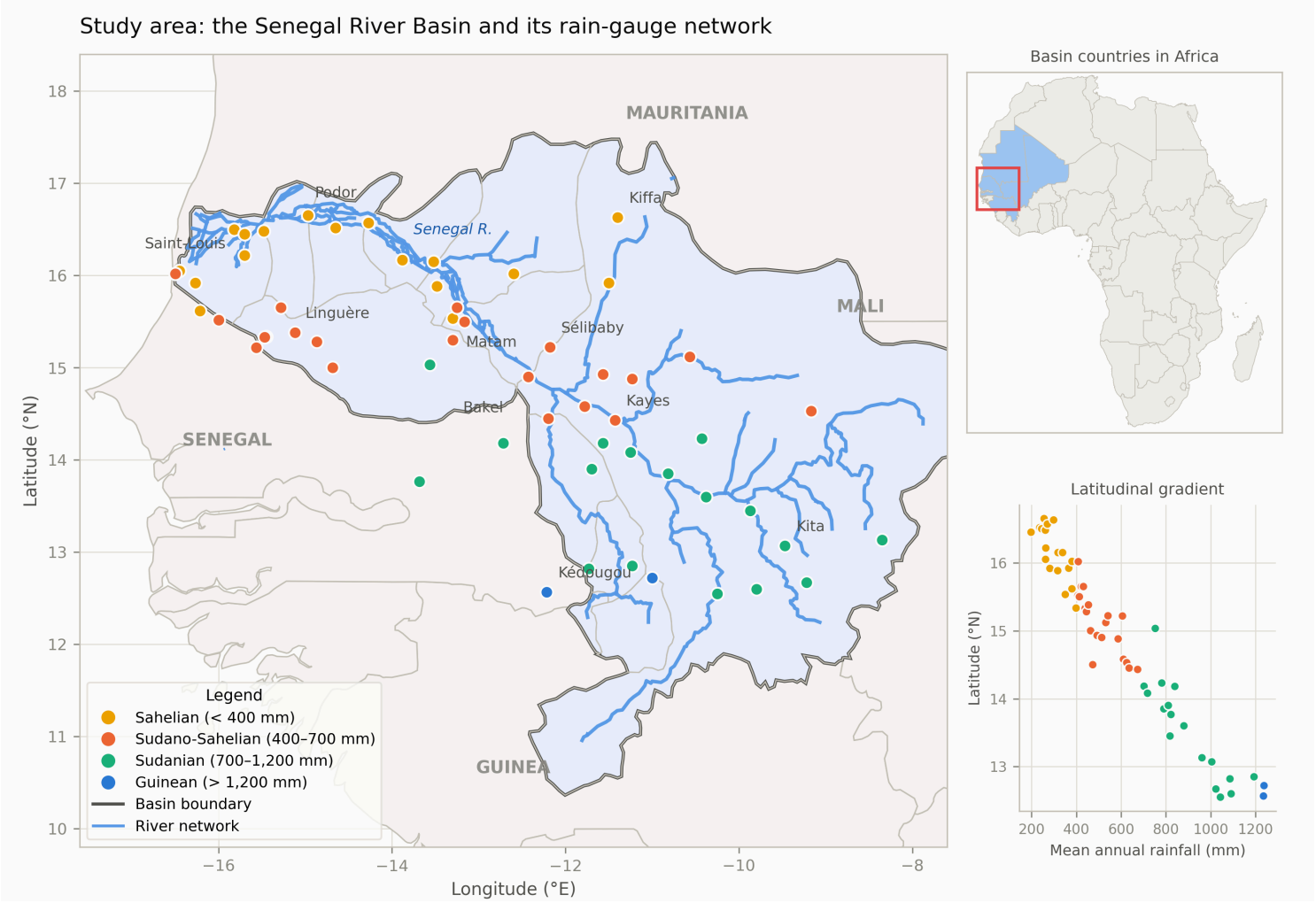


Figure 1

Study area. Main panel: basin boundary, sub-catchments and river network, with the 59 georeferenced stations of the 138-station network coloured by climato-hydrological zone. Inset: the four basin countries within Africa. Lower-right panel: the latitudinal rainfall gradient across the georeferenced stations.

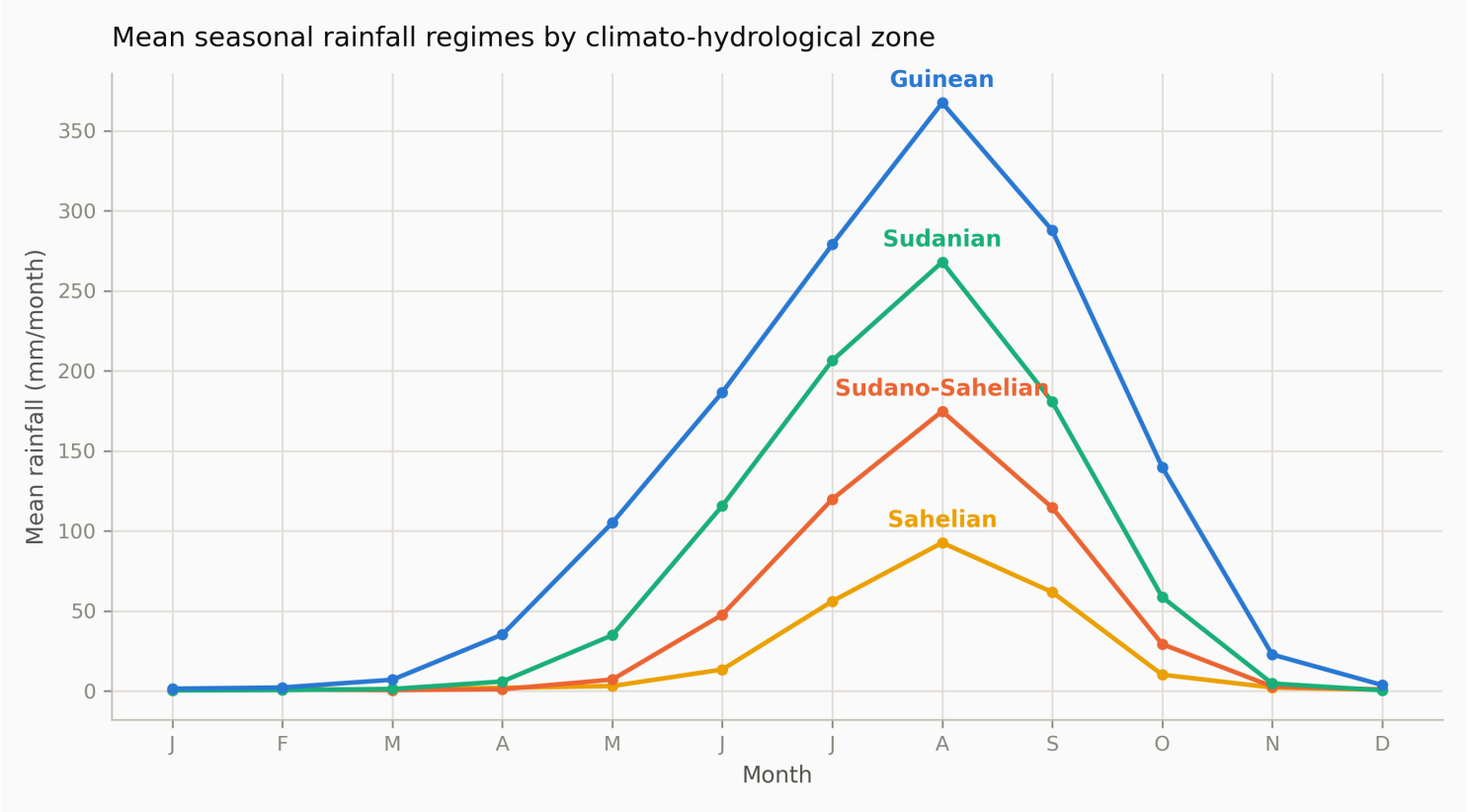


Figure 2

Mean seasonal rainfall regimes by zone, showing the Sahelo-Guinean gradient with a common August peak.

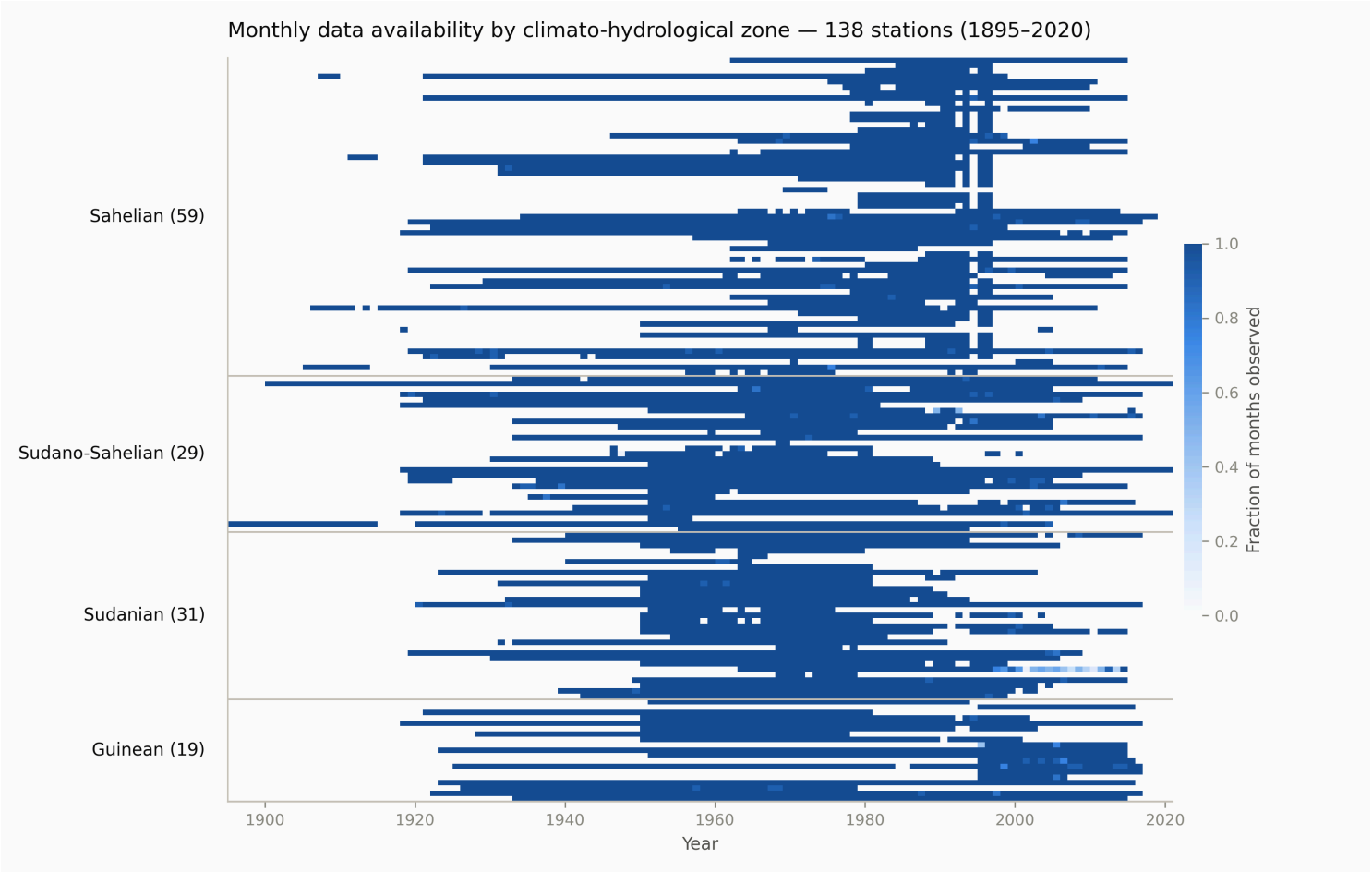


Figure 3

Monthly data availability by station and year, grouped by climato-hydrological zone, with blue indicating months observed. Note the sharp post-1995 attrition.

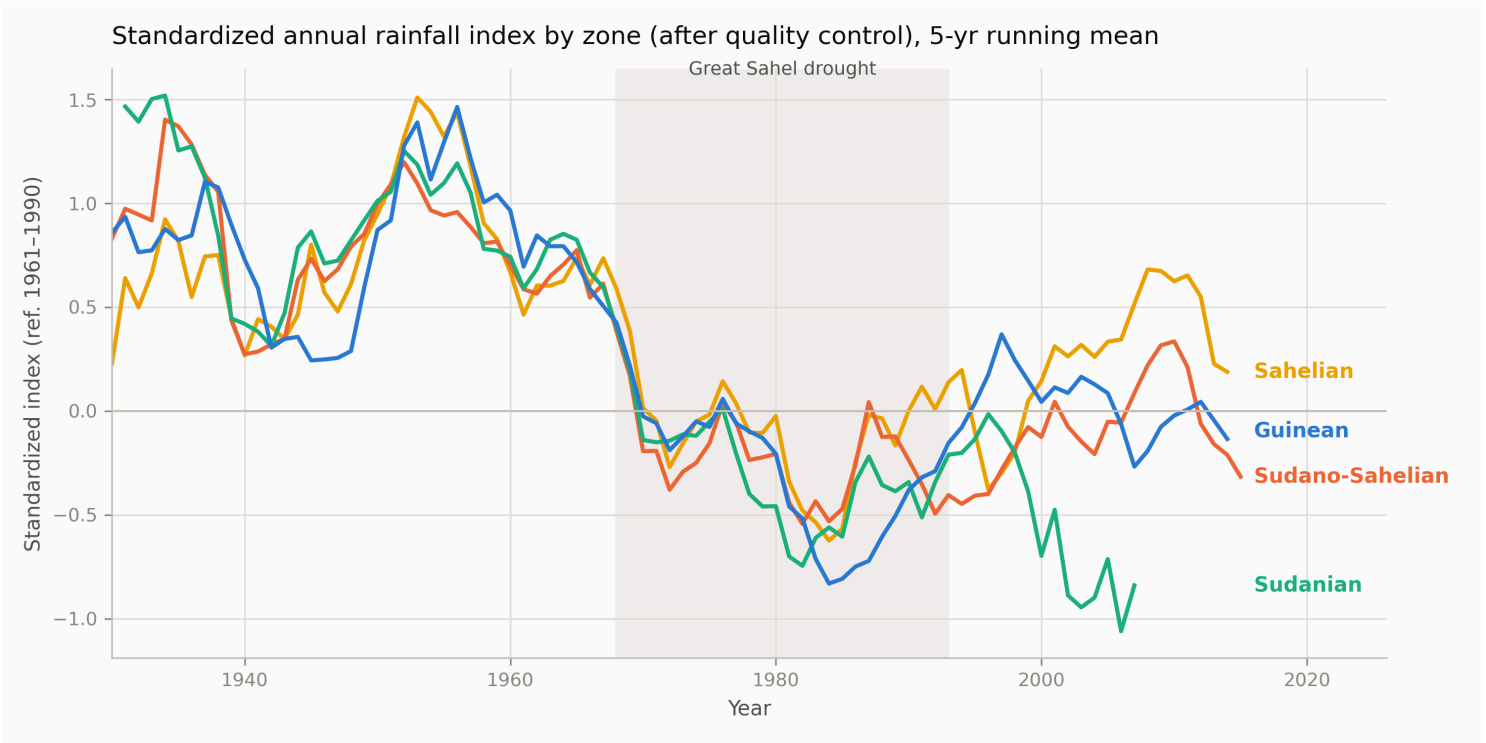


Figure 4

Standardized annual rainfall index by zone, with the 1961 to 1990 reference, after quality control and smoothed by a five-year running mean. Shading marks the great Sahel drought of 1968 to 1993.

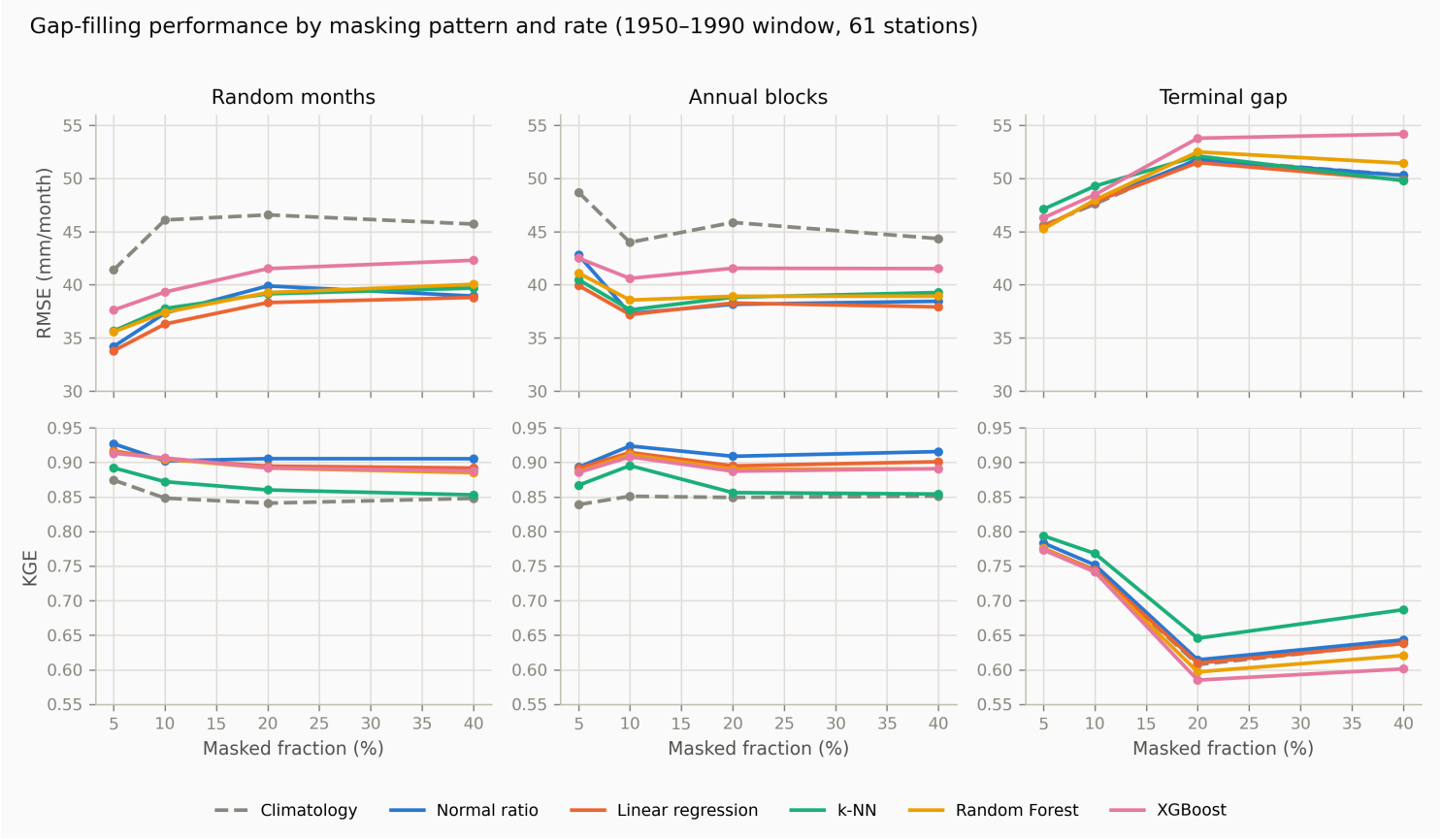


Figure 5

RMSE, upper row, and Kling-Gupta efficiency, lower row, versus masked fraction for the three gap patterns across 61 stations. The dashed grey line is the climatology baseline.

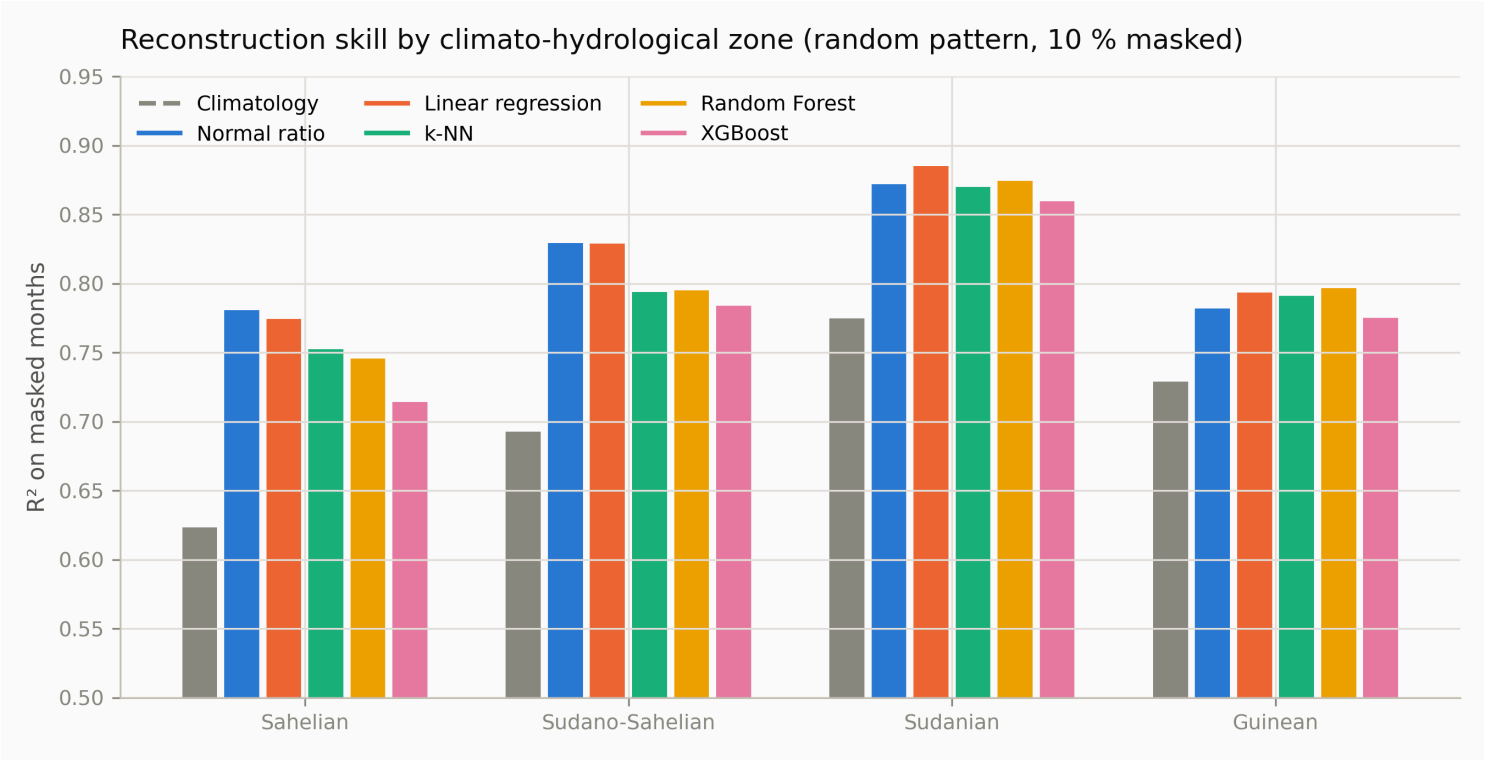


Figure 6

Coefficient of determination on masked months by zone for the random pattern at 10 % masking.

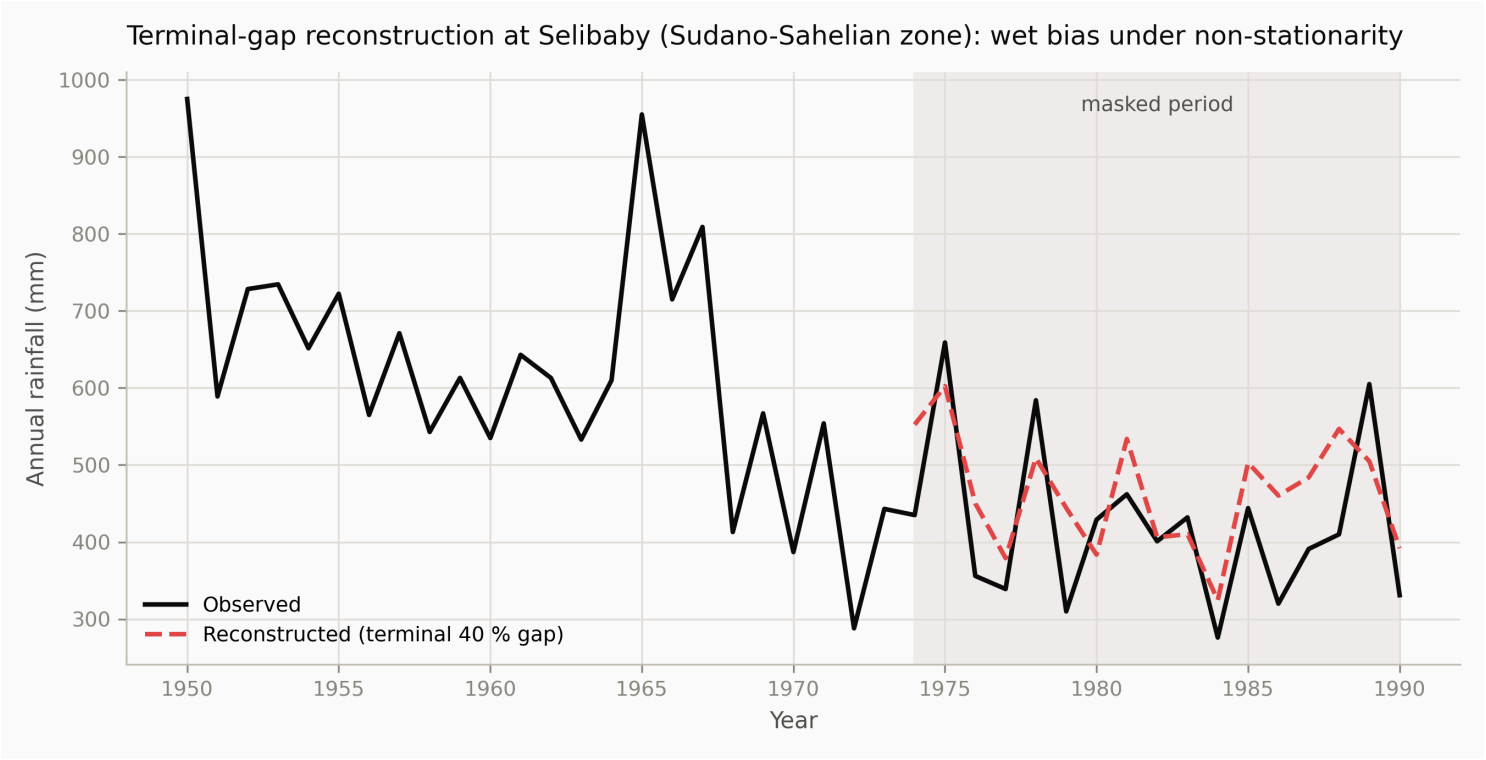


Figure 7

Terminal-gap experiment at Sélibaby, comparing observed annual totals, solid line, with the reconstruction over the masked final 40 % of the window, dashed line.

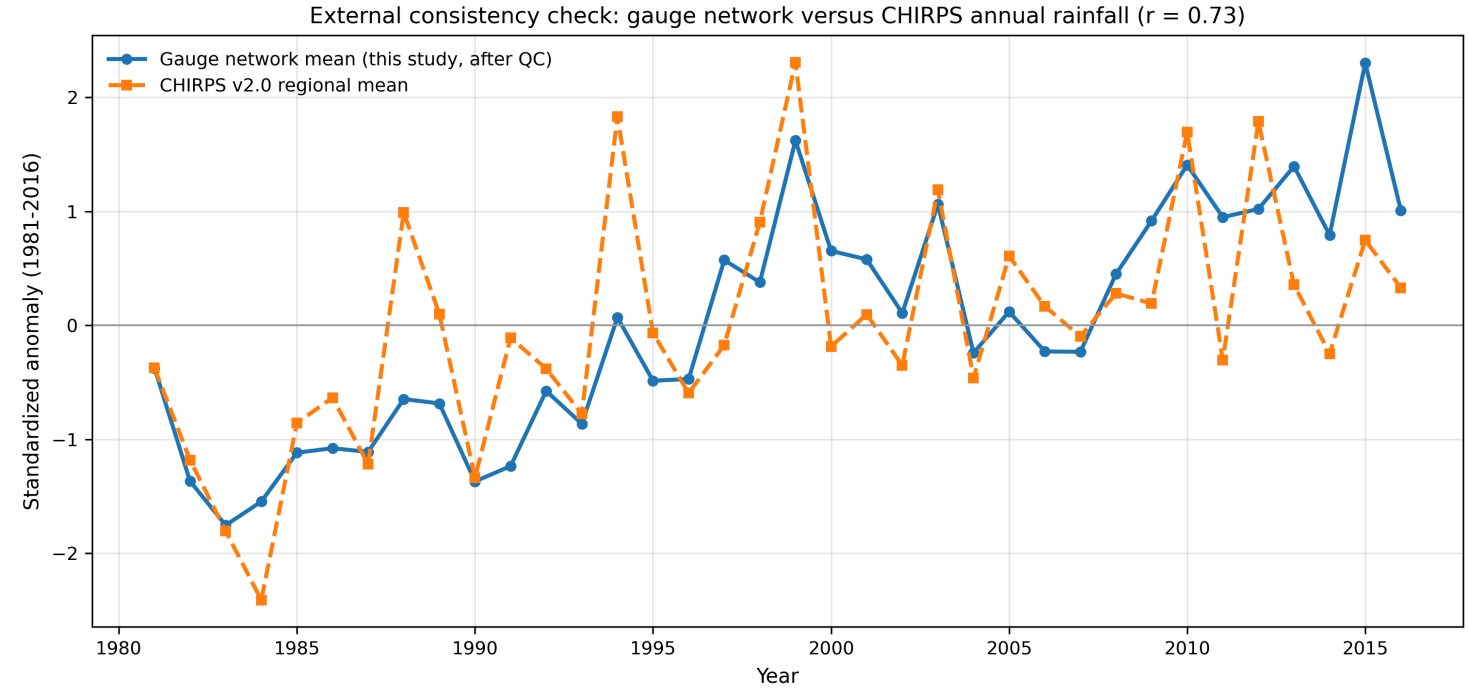


Figure 8

Standardized annual rainfall anomalies from 1981 to 2016, comparing the quality-controlled gauge network mean with the satellite regional mean. Both show the recovery, but the gauge trend is steeper, consistent with attrition inflation.

Supplementary Files

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

- SupplementarymaterialDAS.docx
- SupplementarydataDAS.zip