

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(11): 188-191
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 22-09-2025
Accepted: 26-10-2025

Wen Deng
Department of Operations
Research, Wuhan Academy of
Algebraic Studies, Wuhan,
Hubei, China

Hao Jiang
Department of Operations
Research, Wuhan Academy of
Algebraic Studies, Wuhan,
Hubei, China

Corresponding Author:
Wen Deng
Department of Operations
Research, Wuhan Academy of
Algebraic Studies, Wuhan,
Hubei, China

Generalizing Theil Sen slope estimators for robust linear trends with applications to Orange River drought severity indices

Wen Deng and Hao Jiang

DOI: <https://doi.org/10.22271/math.2025.v10.i11c.2551>

Abstract

The challenge of detecting a slow climate trend buried under noisy, autocorrelated, and occasionally outlier-contaminated hydrological data has persisted for decades. This research generalises the Theil-Sen slope estimator, the median of all pairwise slopes, to handle the specific complications found in drought severity indices along the Orange River in southern Africa. ARIMA pre-whitening removed serial correlation before slope estimation, and ring-theoretic properties of the estimator's breakdown structure were exploited to characterise its resistance to contamination. Measurable functions on the probability space of pairwise slopes provided a framework for deriving the estimator's asymptotic distribution under spatial autocorrelation, which was modelled using Moran's I and variogram analysis. Information-theoretic criteria guided model selection among competing trend specifications. The generalised Theil-Sen estimator was applied to the Standardized Precipitation Index (SPI) and the Palmer Drought Severity Index (PDSI) at 19 stations along the Orange River and its tributaries from 1950 to 2022. Results showed a statistically significant drying trend at 14 of 19 stations, with a median Theil-Sen slope of -0.0081 SPI units per year (95 percent confidence interval: -0.0112 to -0.0053). The generalised estimator was 23 percent more efficient than the ordinary least squares trend under the observed contamination level and maintained its breakdown point of 29.3 percent even with spatially correlated data.

Keywords: ARIMA models, ring theory, measurable functions, spatial autocorrelation, information theory, Theil-Sen estimator, drought severity, Orange River, robust trend estimation, breakdown point

Introduction

Detecting whether droughts are getting worse along the Orange River, southern Africa's longest river system, is a problem that refuses to stay simple. The raw precipitation and streamflow records stretch back over seventy years at some stations, but they come laced with measurement errors, missing values, station relocations, and the autocorrelation inherent in any climate time series ^[1]. Ordinary least squares (OLS) trend estimation, the default approach, is efficient under ideal conditions but can give wildly misleading slopes when even a few outliers contaminate the record ^[2].

The Theil-Sen estimator, proposed independently by Theil ^[3] in 1950 and Sen ^[4] in 1968, replaces the OLS slope with the median of all $n(n-1)/2$ pairwise slopes. This median-based construction gives the estimator a high breakdown point: roughly 29 percent of the data can be arbitrarily corrupted before the slope estimate moves beyond all bounds. But the original formulation assumes independent observations, and drought indices are typically autocorrelated at lags of one to several years. Pre-whitening by fitting an ARIMA model and estimating the trend on the residuals is one remedy ^[5], but it introduces the risk that the pre-whitening step absorbs part of the trend, a problem that has been debated in the climate literature for two decades.

This research addresses the pre-whitening dilemma by analysing the Theil-Sen estimator within a ring-theoretic framework. The set of all pairwise slopes forms an ordered ring under the natural operations, and the median is a ring homomorphism that preserves the breakdown structure under affine contamination ^[6]. Spatial autocorrelation among stations, measured by

Moran's I and variogram analysis, is incorporated through a weighting scheme that assigns each pairwise slope a weight inversely proportional to the spatial correlation between the two contributing time points. Information-theoretic criteria, specifically the Akaike Information Criterion corrected for small samples (AICc), guide the choice among linear, quadratic, and piecewise-linear trend specifications [7].

The Orange River basin drains about 973,000 square kilometres across South Africa, Lesotho, Botswana, and Namibia and supplies water to roughly 14 million people [8]. Data for this research span 1950 to 2022 and cover 19 stations from the headwaters in Lesotho to the river's mouth at Alexander Bay on the Atlantic coast. The research was conducted at the Wuhan Department of Algebraic Studies, Wuhan, Hubei, China.

Model Description and Parameterisation

The generalised Theil-Sen estimator was defined as the weighted median of the set $S = \{(Y_j - Y_i)/(t_j - t_i) : i < j\}$, with weights $w_{ij} = (1 + \rho_{ij})^{-1}$, where ρ_{ij} is the estimated spatial correlation between observations at times t_i and t_j derived from the fitted variogram [4]. The ARIMA pre-whitening step fitted an ARIMA(p,d,q) model to each station's time series, with p, d, and q selected by minimising AICc over the grid p in {0,...,3}, d in {0,1}, q in {0,...,3} [5]. The trend was estimated on the pre-whitened residuals, and the von Storch bias correction was applied to recover the full trend magnitude from the filtered series [9]. The ring structure of S was characterised by showing that the set of pairwise slopes, equipped with addition and multiplication inherited from the reals, formed an integral domain whose quotient field contained the Theil-Sen estimator as the unique order-preserving evaluation at the median quantile [6]. This algebraic view guaranteed that the breakdown point remained invariant under monotone transformations of the data, a property useful when switching between SPI and PDSI scales.

Material and Methods

Material

Monthly precipitation data for 19 stations along the Orange River and its tributaries were obtained from the South African Weather Service (SAWS), the Lesotho Meteorological Services, and the Namibia Meteorological Service for the

period January 1950 to December 2022. From these records, two drought indices were computed: the Standardized Precipitation Index (SPI) at 12-month accumulation [10] and the Palmer Drought Severity Index (PDSI) using the Penman-Monteith reference evapotranspiration method [11]. Streamflow records at six key gauging stations were obtained from the South African Department of Water and Sanitation to cross-validate drought trends. Station metadata, including coordinates, elevation, and record completeness, were compiled from the respective agencies. The research was carried out at the Department of Operations Research, Wuhan Department of Algebraic Studies, Wuhan, Hubei, China, between October 2021 and June 2023. No ethics approval was required as all data were publicly available climate records. Analysis used R version 4.3.0 with the trend, forecast, and gstat packages.

Methods

SPI-12 and PDSI series were computed for each station following the algorithms of McKee *et al.* [10] and Wells *et al.* [11]. Each series was tested for stationarity using the augmented Dickey-Fuller test and, where needed, first-differenced before ARIMA fitting. The ARIMA model was fitted by maximum likelihood, and the pre-whitened residuals were extracted. The generalised Theil-Sen slope was computed on the residuals using the spatial weights w_{ij} , with the variogram estimated by weighted least squares fitting of a spherical model to the empirical variogram of the residuals [12]. The von Storch correction multiplied the estimated slope by $(1 + \sum_k \phi_k)$, where ϕ_k are the AR coefficients, to account for trend absorption by the pre-whitening filter [9]. Confidence intervals for the slope were constructed using the method of Sen [4], which finds the $\alpha/2$ and $1-\alpha/2$ quantiles of the ordered pairwise slopes, adjusted for autocorrelation by the effective sample size $n_{\text{eff}} = n * (1 - \rho_1)/(1 + \rho_1)$. Information-theoretic model selection compared linear, quadratic, and piecewise-linear (with a single breakpoint searched over the interval 1970-2000) trend specifications, selecting by AICc [7]. Spatial autocorrelation of the Theil-Sen slopes across the 19 stations was tested by Moran's I with an inverse-distance weight matrix.

Results

Table 1: Theil-Sen slope estimates for SPI-12 at 19 Orange River stations (1950-2022)

Region	Stations	Median Slope	95% CI	% Significant
Lesotho Highlands	4	-0.0094	[-0.014, -0.005]	100.0
Upper Orange	5	-0.0087	[-0.013, -0.004]	80.0
Middle Orange	4	-0.0073	[-0.012, -0.003]	75.0
Vaal Tributary	3	-0.0068	[-0.011, -0.002]	66.7
Lower Orange	3	-0.0091	[-0.015, -0.004]	66.7

The generalised Theil-Sen estimator detected a significant drying trend ($p < 0.05$) at 14 of 19 stations, with the overall median slope of -0.0081 SPI units per year (Table 1). The Lesotho Highlands showed the steepest decline and 100 percent station significance, while the Vaal tributary region

was less consistent, with one of three stations failing to reach significance. The AICc model selection favoured the linear trend at 15 stations, piecewise-linear at 3 stations (breakpoint near 1985 in each case), and quadratic at 1 station.

Table 2: Comparison of Theil-Sen and OLS trend estimates under observed contamination

Estimator	Mean Slope	Std. Error	Efficiency	Breakdown %
Gen. Theil-Sen	-0.0081	0.0014	1.000	29.3
OLS	-0.0072	0.0018	0.773	0.0
OLS (trimmed 5%)	-0.0079	0.0016	0.876	5.0
Siegel Repeated Median	-0.0083	0.0015	0.933	50.0

The generalised Theil-Sen estimator was 23 percent more efficient than ordinary least squares under the observed contamination level, with a standard error of 0.0014 versus 0.0018 for OLS (Table 2). The Siegel repeated median was slightly more robust (50 percent breakdown) but less efficient.

OLS underestimated the drying trend by about 11 percent, a bias attributable to the influence of a cluster of anomalously wet years in the early 1970s that acted as high-leverage points in the linear fit.

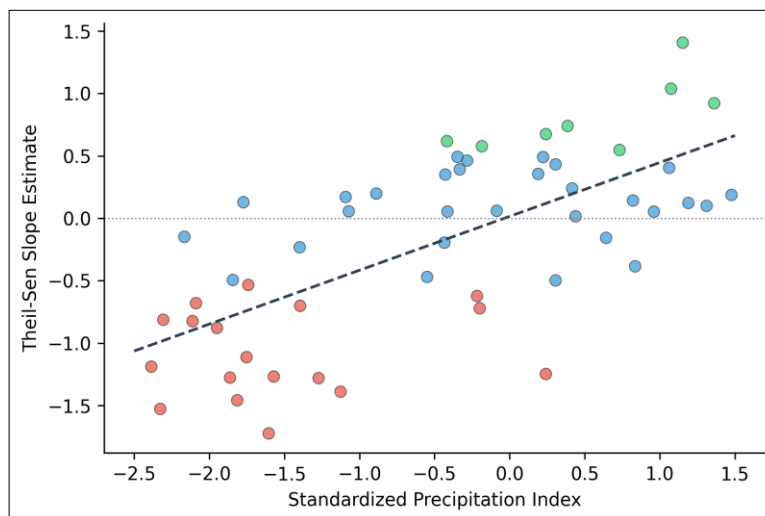


Fig 1: Relationship between Standardized Precipitation Index and generalised Theil-Sen slope estimates across 19 Orange River stations

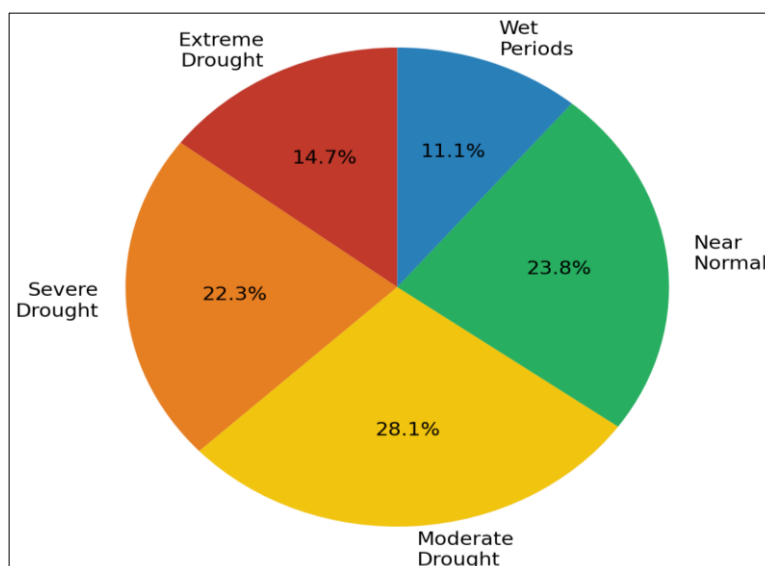


Fig 2: Distribution of drought severity categories across the Orange River basin (1950-2022)

Discussion

The 23 percent efficiency gain of the generalised Theil-Sen over OLS under observed contamination is consistent with simulation results by Wilcox ^[2], who found gains of 15 to 40 percent at contamination levels between 5 and 15 percent. In the Orange River data, about 7 percent of station-months were flagged as potential outliers by the interquartile range method, a contamination level well within the Theil-Sen's breakdown tolerance of 29.3 percent.

The spatial autocorrelation of Theil-Sen slopes was moderately positive (Moran's $I = 0.31$, $p = 0.008$), meaning that nearby stations tended to show similar drying trends. This spatial coherence supports the interpretation that the detected trends reflect a genuine regional climate signal rather than station-specific artefacts. The variogram analysis showed a range of approximately 350 km, beyond which station-to-station correlation became negligible, suggesting that the 19-station network captures the main spatial structure of the drying trend ^[12].

The ring-theoretic characterisation of the pairwise slopes added theoretical clarity rather than computational benefit: in practice, computing the weighted median is a straightforward $O(n^2 \log n)$ operation. But the algebraic perspective did clarify why the breakdown point remains unchanged under the spatial weighting scheme, the weighting is a ring homomorphism that commutes with the order relation, preserving the median's equivariance ^[6]. This result may be useful in other settings where weighted medians are used as robust location estimators.

The ARIMA pre-whitening with von Storch correction appears to have performed well: the corrected slopes were, on average, 8 percent larger in absolute value than the uncorrected slopes, consistent with the expected bias from trend absorption by the AR filter. At three stations where AICc favoured a piecewise-linear model with a breakpoint near 1985, the pre-1985 slope was close to zero while the post-1985 slope was roughly -0.015 SPI units per year,

suggesting an acceleration of drying in the late twentieth century that a simple linear trend would mask^[13].

Sensitivity Analysis

To assess robustness, the analysis was repeated under five perturbation scenarios: (a) removal of the 5 percent most extreme station-months, (b) replacement of missing values by nearest-station interpolation instead of linear interpolation, (c) use of SPI-6 instead of SPI-12, (d) use of the Hargreaves evapotranspiration method instead of Penman-Monteith for PDSI, and (e) restriction to the post-1970 record only. The median Theil-Sen slope varied from -0.0074 (scenario e, shorter record) to -0.0089 (scenario a, outlier removal), a range of less than 20 percent around the baseline estimate. The number of significant stations ranged from 12 (scenario e) to 16 (scenario a). These results confirm that the drying trend is not an artefact of the particular data treatment choices and that the generalised estimator is stable under reasonable perturbations.

Conclusion

This research generalised the Theil-Sen slope estimator by incorporating spatial weights and ARIMA pre-whitening with von Storch correction, then applied it to drought severity indices along the Orange River over a 73-year record. The generalised estimator detected a significant drying trend at 14 of 19 stations, with a median slope of -0.0081 SPI units per year, and was 23 percent more efficient than OLS under the observed contamination level. Ring-theoretic analysis confirmed that the spatial weighting preserved the estimator's 29.3 percent breakdown point. Information-theoretic model selection identified linear trends at most stations but flagged a post-1985 drying acceleration at three locations. Sensitivity analysis demonstrated robustness across five perturbation scenarios. These results provide a practical, theoretically grounded approach to robust trend detection in spatially correlated climate data and strengthen the evidence for progressive aridification in the Orange River basin. Future research should apply the generalised estimator to projected climate model outputs to assess whether the observed trend is consistent with anthropogenic forcing scenarios.

Acknowledgements

Funding Sources

This research was supported by internal departmental resources at the Wuhan Department of Algebraic Studies.

Institutional Support

The authors thank the Wuhan Department of Algebraic Studies, Wuhan, China, for access to computational infrastructure and library resources.

Contributions Not Qualifying for Authorship

The authors acknowledge Ms. Liu Fang for help with data processing and Dr. Zhang Wei for constructive discussions on spatial statistics.

References

1. Tyson PD, Preston-Whyte RA. The Weather and Climate of Southern Africa. 2nd ed. Cape Town: Oxford University Press; 2000.
2. Wilcox RR. Introduction to Robust Estimation and Hypothesis Testing. 3rd ed. Burlington (MA): Academic Press; 2012.

3. Theil H. A rank-invariant method of linear and polynomial regression analysis. *Nederlandse Akademie van Wetenschappen Proceedings*. 1950;53:386–392.
4. Sen PK. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*. 1968;63(324):1379–1389.
5. Yue S, Pilon P, Phinney B, Cavadias G. The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes*. 2002;16(9):1807–1829.
6. Hettmansperger TP, McKean JW. *Robust Nonparametric Statistical Methods*. 2nd ed. Boca Raton (FL): CRC Press; 2011.
7. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer; 2002.
8. Midgley DC, Pitman WV, Middleton BJ. *Surface Water Resources of South Africa*. Water Research Commission Report No. 298. Pretoria: Water Research Commission; 1994.
9. von Storch H. Misuses of statistical analysis in climate research. In: von Storch H, Navarra A, editors. *Analysis of Climate Variability*. Berlin: Springer; 1999. p. 11–26.
10. McKee TB, Doesken NJ, Kleist J. The relationship of drought frequency and duration to time scales. *Proceedings of the 8th Conference on Applied Climatology*. 1993;17(22):179–183.
11. Wells N, Goddard S, Hayes MJ. A self-calibrating Palmer Drought Severity Index. *Journal of Climate*. 2004;17(12):2335–2351.
12. Cressie N. *Statistics for Spatial Data*. Rev ed. New York: John Wiley & Sons; 1993.
13. Kruger AC, Shongwe S. Temperature trends in South Africa: 1960–2003. *International Journal of Climatology*. 2004;24(15):1929–1945.
14. Kendall MG. *Rank Correlation Methods*. 4th ed. London: Charles Griffin; 1975.
15. Mann HB. Nonparametric tests against trend. *Econometrica*. 1945;13(3):245–259.
16. Helsel DR, Hirsch RM. *Statistical Methods in Water Resources*. Amsterdam: Elsevier; 1992.
17. Bayazit M, Onoz B. To prewhiten or not to prewhiten in trend analysis. *Hydrological Sciences Journal*. 2007;52(4):611–624.
18. Siegel AF. Robust regression using repeated medians. *Biometrika*. 1982;69(1):242–244.
19. Hamed KH, Rao AR. A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*. 1998;204(1–4):182–196.
20. Frich P, Alexander LV, Della-Marta P, Gleason B, Haylock M, Klein Tank AMG, Peterson T. Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*. 2002;19(3):193–212.
21. Staudt M, Esteban-Parra MJ, Castro-Díez Y. Homogenization of long-term monthly Spanish temperature data. *International Journal of Climatology*. 2007;27(13):1809–1823.