Title

# Comment on “Five Decades of Observed Daily Precipitation Reveal Longer and More Variable Drought Events Across Much of the Western United States”



Journal
Geophysical Research Letters, 51(1)

ISSN
0094-8276

Authors
Paciorek, CJ
Wehner, MF

Publication Date
2024-01-16

DOI
10.1029/2023gl104550

# Comment on 'Five Decades of Observed Daily Precipitation Reveal Longer and More Variable Drought Events Across Much of the Western United States'

**C. J. Paciorek[1] and M. F. Wehner[2]**

[1]Department of Statistics, University of California, Berkeley, California, U.S.A.
[2]Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.

**Key Points:**

- Statistical uncertainty about trends in droughts in the southwestern US is larger than reported in a recent GRL letter.

Corresponding author: Christopher Paciorek, `paciorek@stat.berkeley.edu`

## Abstract

Changes in precipitation patterns with climate change could have important impacts on human and natural systems. Zhang et al. (2021) report trends in daily precipitation patterns over the last five decades in the western United States, focusing on meteorological drought. They report that dry intervals (calculated at the annual or seasonal level) have increased across much of the southwestern U.S., with statistical assessment suggesting the results are statistically robust. However, Zhang et al. (2021) preprocess their annual (or seasonal) averages to compute five-year moving window averages before using established statistical techniques for trend analysis that assume independence about some fixed trend. Here we show that the moving window preprocessing violates that independence assumption and inflates the statistical significance of their trend estimates. This raises questions about the robustness of their results. We conclude by discussing the difficulty of adjusting for spatial structure when assessing time trends in a regional context.


## Plain Language Summary

A recent paper reports trends in drought in the western United States, in particular increases in drought in the southwestern United States, based on changes in the lengths of time intervals without precipitation. In this ‘comment’ we note that the preprocessing approach used in the paper artificially increases the apparent statistical signal in the data and caution that the evidence for the trends reported is not as strong as presented in the paper. We conclude by discussing the difficulty of estimating trends in a statistically rigorous fashion across multiple weather stations.

## Main Text

Zhang et al. (2021) present results on daily precipitation patterns over the last five decades in the western United States using precipitation data from GHCN weather stations. A key focus of their work is on meteorological drought, quantified based on time intervals in which daily precipitation never exceeds three mm. They report that mean and longest dry intervals (calculated at the annual or seasonal level) have increased across much of the southwestern U.S. The authors use Sen's slope (also known as the Theil-Sen estimator) to quantify trends and the Mann-Kendall test to quantify statistical significance.

A key assumption of trend analysis using Sen's slope and the Mann-Kendall test is that the observations are independent about some fixed trend (Sen, 1968). It is well-known that correlation can invalidate the Mann-Kendall test, such that the distribution of p-values is not uniform under the null hypothesis, with an inflated probability of detecting a non-existent trend, and there is extensive discussion of techniques for accounting for or reducing correlation (e.g., see Hamed & Rao, 1998; Yue et al., 2002; Hamed, 2009).

We attempted to reproduce the results in Zhang et al. (2021)[Figure 3a], which presents station-specific trends in mean dry interval for 1976-2019. In discussions with the authors, we learned from them that for the trend analyses, they took yearly (or seasonal) averages and then computed the mean of those values within five-year moving windows. This was not described explicitly in the paper, although there are references to moving windows in Zhang et al. (2021)[Section 2] and the Zhang et al. (2021)[Figure 4 caption] that can be read as specifically relating to the coefficient of variation (CV) calculation. The authors apparently used the standard Mann-Kendall test, as implemented in the mkttest function from the modifiedmk R package (Patakamuri & O'Brien, 2021).

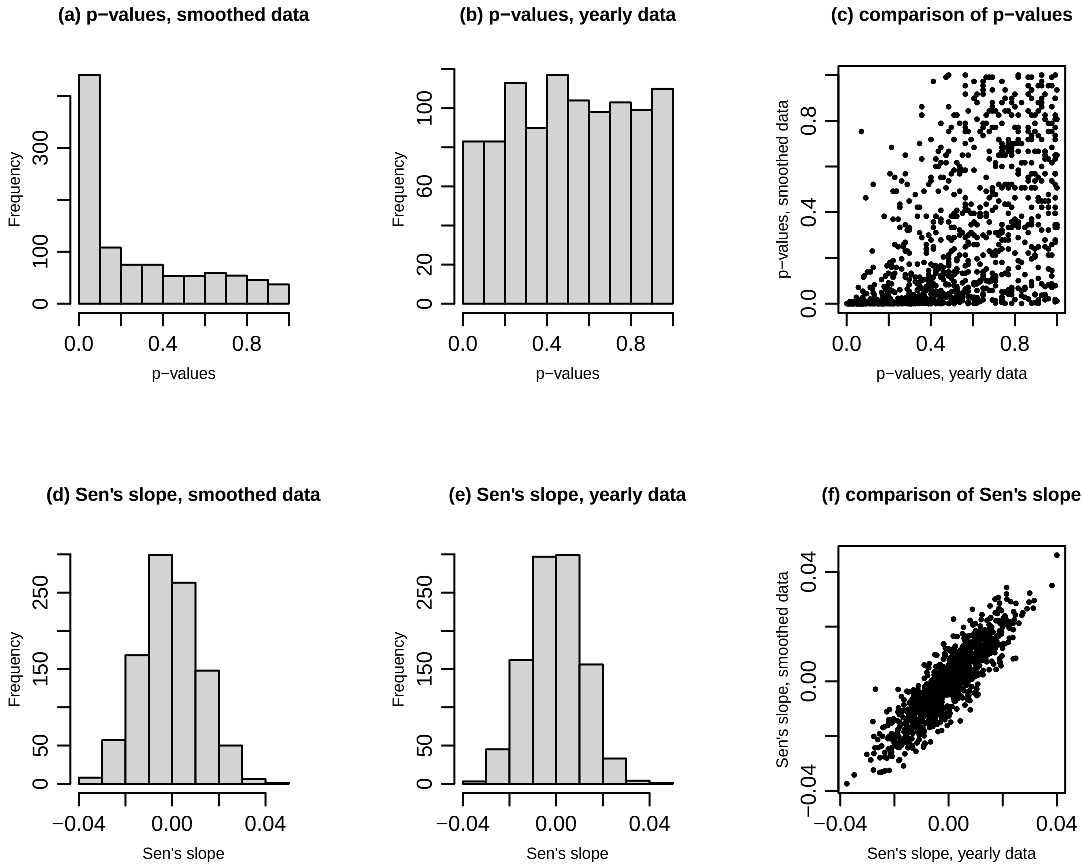


**Figure 1.** Comparison of p-values (top row) and Sen's slope values (bottom row) for simulated independent data and smoothed (five-year overlapping moving windows) data.

What is the impact of using moving averages of the observations as opposed to the observations directly? Conceptually, it's clear that using moving averages (i.e., averaging with blocks of multiple values where the blocks overlap) introduces correlation by construction. So this raises the concern that the results on statistical significance may be distorted relative to direct use of the observations (and that the Sen's slope estimates may differ as well, although the impact of correlation on the slope is less clear conceptually).

We can see the impact of smoothing the raw data before trend analysis by simple simulation. We simulated 1000 'time series' of 44 'years' of completely independent data, with no trend. We then used the standard Mann-Kendall test applied to five-year moving window averages, which gives 40 'years' of smoothed data. Figure 1 shows the p-values from the test and Sen's slope values with and without smoothing. It's clear that the p-values from the test when using smoothed data are not uniformly distributed and are bunched near zero compared to the uniformly-distributed p-values that we expect and see when applying the test to the unsmoothed data. Second, we note that under the null hypothesis, there does not seem to be a systematic effect on the Sen's slope values, although the estimates differ before and after smoothing.

To assess the impact of smoothing on the results of Zhang et al. (2021), we focus here on Zhang et al. (2021)[Figure 3a] as an example analysis. With gracious assistance from the authors regarding the details of handling missing observations and handling dry intervals that overlap two (water) years, we were able to essentially reproduce the results of Zhang et al. (2021)[Figure 3a] (with minor quantitative differences) when using five-year moving windows. Here we show the Sen's slope values and statistical significance ($p < 0.05$) when reproducing the Zhang et al. (2021) approach (Figure 2a) compared to using the original yearly values without smoothing (Figure 2b). We see that, as expected based on the simulations shown in Figure 1, the p-values are generally larger when using unsmoothed data. While the trends in the Southwest (particularly Arizona, southern California, and New Mexico) are generally positive, the results at many stations are no longer statistically significant at the 0.05 level (Figure 2b). Figure 3 shows how the p-values and Sen's slope values compare. Many of the p-values under the smoothed data are bunched near zero, as also seen in the simulation. For the Sen's slope values, there seems to be a systematic pattern that the values are larger when using smoothed data. This was not seen in the simulation, so it's not clear if this is a systematic effect of smoothing, but given the assumptions behind the Sen's slope estimator, we have more confidence in the slope estimates from the unsmoothed data.

This use of smoothing before statistical analysis of trends appears to occur throughout Zhang et al. (2021). Clearly this introduces questions about the station-specific trend and significance results in Zhang et al. (2021)[Figures 1 and 3] and related figures in the supplemental materials. In addition, the authors report apparently similar analyses at the regional level, e.g., Zhang et al. (2021)[Figures 2 and 4]. (We suspect the regional analyses average over all stations in each region at the annual/seasonal level, but we are not sure.) For example, Zhang et al. (2021)[Figure 4] apparently uses the five-year moving window averages of the mean values and five-year moving window CV values and then computes Sen's slope and uses the Mann-Kendall test to compute p-values. Of course some time window is needed to compute the CV, but doing this using overlapping windows as opposed to adjacent, non-overlapping windows introduces the same concerns about inducing correlation.

Given the clear inflation of significance caused by smoothing, and somewhat increased slope estimates, what can we conclude about the scientific results presented in Zhang et al. (2021)? First, the station-specific uncertainty is clearly quite a bit larger than presented. This may not be surprising given we would expect a low signal to noise ratio in estimating dry intervals (and related quantities) from precipitation values, which are of course quite variable at the daily level. Second, in much of the western U.S. there are not clear patterns in trends of mean dry interval, apart from the Southwest and possibly the northern Great Plains (Figure 2b). The northern Great Plains show consistent decreases in dry interval lengths, although only a limited number of stations are individually significant. Stations in Arizona show statistically significant increases in dry interval lengths, but similar increases elsewhere in the Southwest are generally not significant.

To draw more robust conclusions, ideally one would adjust the p-values in light of the multiple testing from doing analyses at multiple stations (e.g., the well-known false discovery rate procedure of Benjamini and Hochberg (1995)), or carry out a joint statistical analysis of all the stations simultaneously in a way that accounts for the spatial correlation structure. These are not easy tasks given the strong spatial correlation, complicated by the real-world effects of topography and weather patterns that produce nonstationary spatial correlation structure. There is consistent evidence from the multiple testing literature that when there is positive correlation, if one uses adjustment procedures such as Benjamini and Hochberg (1995) that assume independent p-values, the number of tests found to be significant is conservative (i.e., one should flag more tests as being significant than the procedure does) (Fithian & Lei, 2020). While there is statisti-

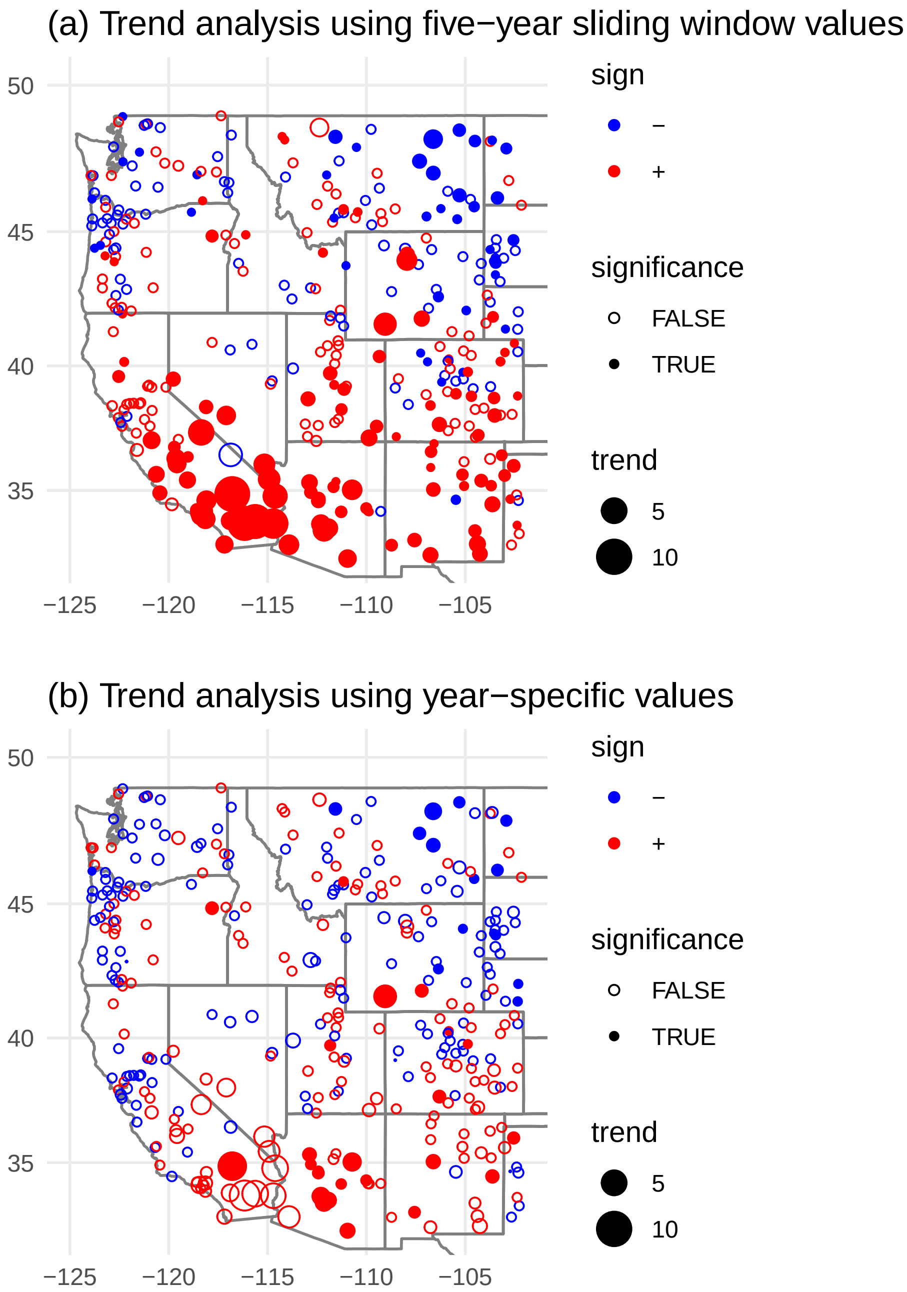


**Figure 2.** Trend analysis of annual mean dry interval length using Sen's slope (days per decade) and Mann-Kendall test significance ($p < 0.05$), reproducing Zhang et al. (2021)[Figure 3a] using (a) five-year sliding window values as in Zhang et al. and (b) original year-specific values. Trends whose absolute Sen's slope value is less than 0.5 are set to 0.5 (or -0.5 for negative trends) to avoid having points that cannot be seen.

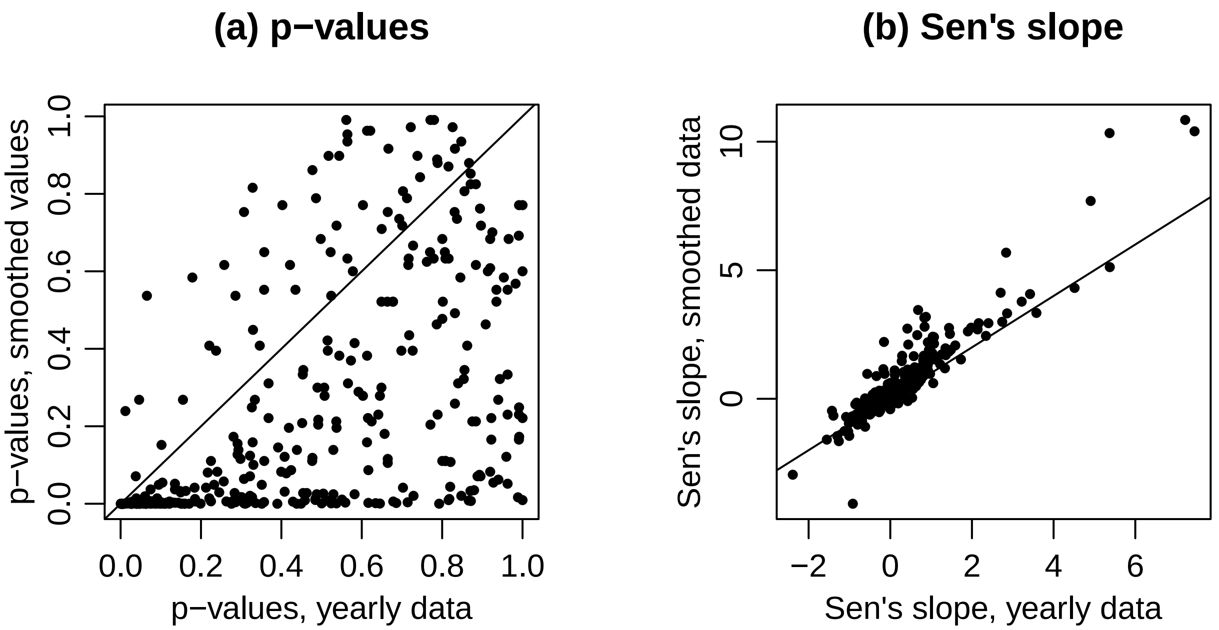


**Figure 3.** Comparison of (a) p-values and (b) Sen's slope values based on year-specific data and smoothed (five-year overlapping moving windows) data.

cal literature on spatial multiple testing (e.g., Sun et al., 2015; Risser et al., 2019), there is not a well-developed general methodology for doing so with spatially-correlated p-values. In this example, applying the Benjamini and Hochberg procedure flags no locations as significant, which is not helpful, given it is likely conservative but to an unknown degree. However, 11.5% of the locations are individually significant, greater than expected under the full null hypothesis of no trend anywhere. One could do a formal field significance test, but that would not allow us to make any formal inference about where the trends are notable. From a general perspective, the spatial clustering of the trend estimates gives some indication that there may be a more robust signal of trend in the Southwest U.S. than simply considering the raw p-values would indicate. Furthermore, such increases are consistent with future projections of decreased precipitation in the Southwest U.S. and in Mexico due to the poleward shift in the Hadley Circulation (Easterling et al., 2017) and associated changes in weather types (Prein et al., 2016). Given the limitations of p-values for making affirmative claims about hypotheses (in contrast to being able to reject a null hypothesis) (Wasserstein & Lazar, 2016), the reduced statistical significance presented here is not grounds for rejecting the hypothesis that Southwest drought is already increasing. At the same time, without a clear statistical procedure that takes account of the spatial context, it's not clear how robust the observed trends in the Southwest are.

# Open Research

## Availability Statement

Datasets used in this study were downloaded from `https://www.ncdc.noaa.gov/ghcnd-data-access`.

Code for reproducing this analysis can be obtained from `https://github.com/paciorek/grl-comment`.

# Acknowledgments

This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.



# References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*(1), 289–300.

Easterling, D. R., Arnold, J., Knutson, T., Kunkel, K., LeGrande, A., Leung, L. R., … Wehner, M. (2017). Precipitation change in the United States. In D. Wuebbles, D. Fahey, K. Hibbard, D. Dokken, B. Stewart, & T. Maycock (Eds.), *Climate science special report: A sustained assessment activity of the U.S. Global Change Research Program* (pp. 301–335). U.S. Global Change Research Program.

Fithian, W., & Lei, L. (2020). *Conditional calibration for false discovery rate control under dependence.*

Hamed, K. (2009). Exact distribution of the Mann–Kendall trend test statistic for persistent data. *Journal of Hydrology*, *365*(1-2), 86–94.

Hamed, K., & Rao, A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, *204*(1-4), 182–196.

Patakamuri, S. K., & O'Brien, N. (2021). modifiedmk: Modified versions of Mann Kendall and Spearman's rho trend tests [Computer software manual]. (R package version 1.6)

Prein, A. F., Holland, G. J., Rasmussen, R. M., Clark, M. P., & Tye, M. R. (2016). Running dry: The US Southwest's drift into a drier climate state. *Geophysical Research Letters*, *43*(3), 1272–1279.

Risser, M. D., Paciorek, C. J., & Stone, D. A. (2019). Spatially dependent multiple testing under model misspecification, with application to detection of anthropogenic influence on extreme climate events. *Journal of the American Statistical Association*, *114*(525), 61–78.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*(324), 1379–1389.
Sun, W., Reich, B. J., Cai, T. T., Guindani, M., & Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *77*(1), 59.
Wasserstein, R. L., & Lazar, N. A. (2016). *The ASA statement on p-values: context, process, and purpose* (Vol. 70) (No. 2). Taylor & Francis.
Yue, S., Pilon, P., Phinney, B., & Cavadias, G. (2002). The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes*, *16*(9), 1807–1829.
Zhang, F., Biederman, J. A., Dannenberg, M. P., Yan, D., Reed, S. C., & Smith, W. K. (2021). Five decades of observed daily precipitation reveal longer and more variable drought events across much of the western United States. *Geophysical Research Letters*, *48*(7), e2020GL092293.