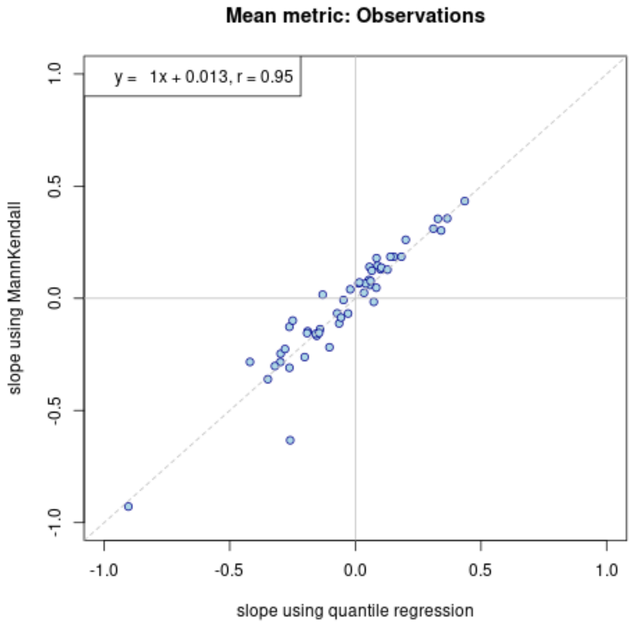
Comment: This is a very good study, written by experts at EPA and it will be a welcome addition to the literature. I find the research question to be important and the conclusions are clearly supported by the analysis. I don't have any concerns regarding the analysis or conclusions, my only recommendation is that the authors update their terminology regarding the reporting of statistical findings, to be more consistent with the Tropospheric Ozone Assessment Report and current thinking regarding the limitations of the expression "statistically significant", as described below.

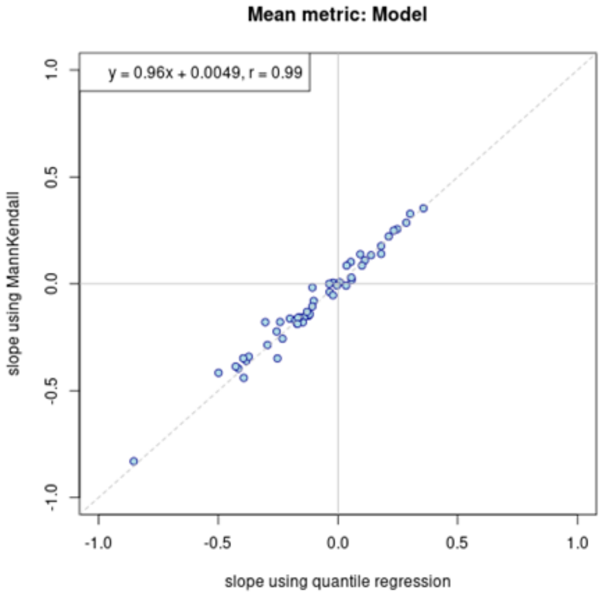
Response: Thank you for this comment.

Comment: Regarding the use of the Theil-Sen/Mann-Kendall method for calculating trends, the authors state that they chose this method because of the small sample sizes and because it does not require assumptions about the distribution of the residuals. Another reason that is often given for the choice of this method is that it is resistant to outliers. The problem is that in order to remove the impact from outliers, this method automatically ignores up to 29% of the data points in a sample (see Section 2 of Chang et al., 2021). This would be fine if the analyst believed that the outliers are due to instrument errors, but there is no reason to throw out data if they are believed to be reliable. In your case there is no reason to believe that your samples contain erroneous data points that should be ignored. For this reason, the Tropospheric Ozone Assessment Report has abandoned the Theil-Sen/Mann-Kendall method that was used in the first phase of TOAR (2014-2019). A further problem with the Theil-Sen method is that it produces unrealistically narrow 95% confidence intervals. This is shown in Figure 1 of the TOAR-II Recommendations for Statistical Analyses (available at https://igacproject.org/activities/TOAR/TOAR-II). Figure 1 compares the trend and 95% confidence interval calculated by 10 different methods for the ozone time series at Mace Head, Ireland. The Theil-Sen method has the narrowest 95% confidence interval by far, and the reason is that this method ignores 29% of the data; by throwing out all of the extreme values the sample has very little variability and therefore a straight line can be fit through the remaining data within a very narrow range. The second phase of TOAR-II is now recommending the use of quantile regression, as described in the TOAR-II Recommendations for Statistical Analyses. Quantile regression was used to good effect in the very nice paper by co-author B. Wells (Wells et al., 2021), and it could easily be applied to your current analysis.

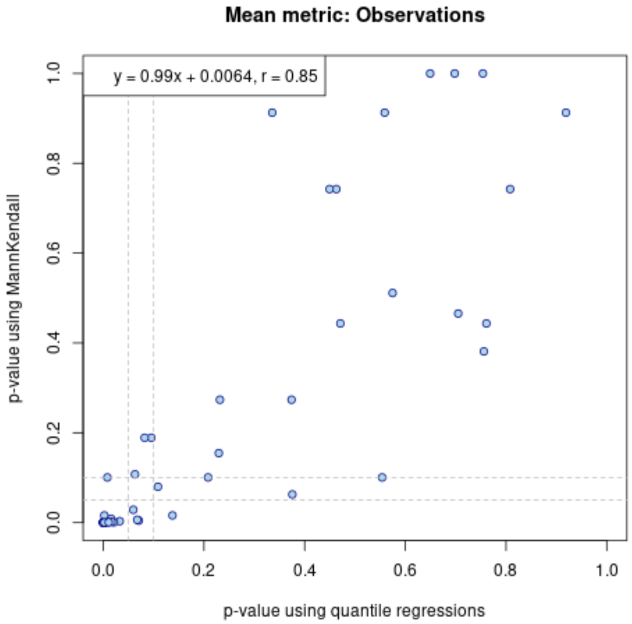
Response: Thank you for this comment. As stated in the manuscript, we believe that Theil-Sen/Mann-Kendall methods are appropriate due to the relatively small sample sizes in contrast to the large ozone datasets used in the TOAR analysis. However, based on the reviewer's suggestions we repeated our analysis using quantile regression. We found that regression slopes between Theil-Sen and quantile regression were nearly identical. When comparing P-values we found that they did differ between methods but did not find any systematic bias towards higher or lower P-values with one method versus the other. Importantly, most regressions that had significant slopes in our original analysis using either a P-Value cutoff of 0.05 or 0.1 still had significant slopes when using quantile regression. Similarly, most areas that had insignificant slopes in our original analysis also had insignificant slopes with the quantile regression method. Based on these results, we have opted not to update the regression methods used in this manuscript. Full results from this comparison are provided below.

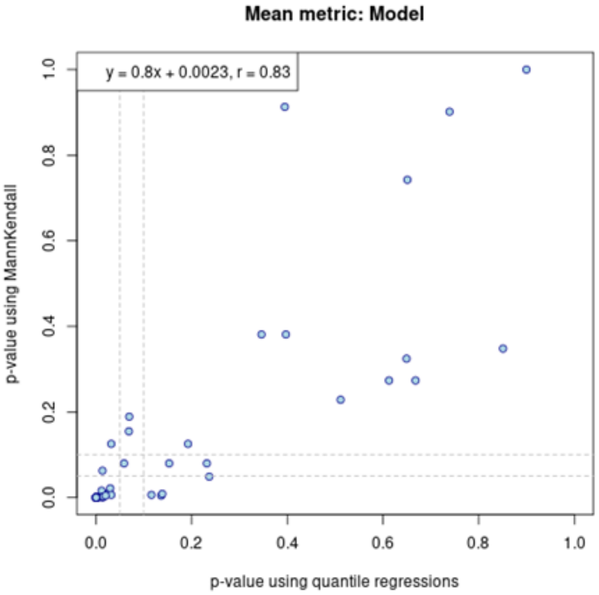
Mean Metric: slopes





Mean Metric: p-values

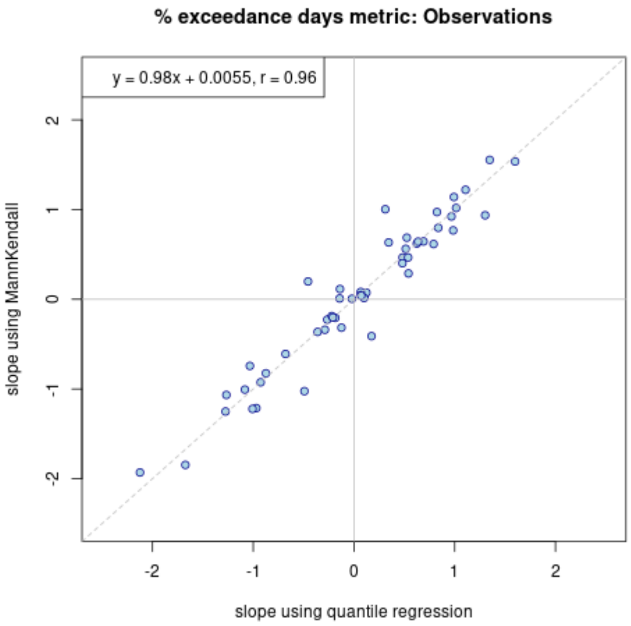


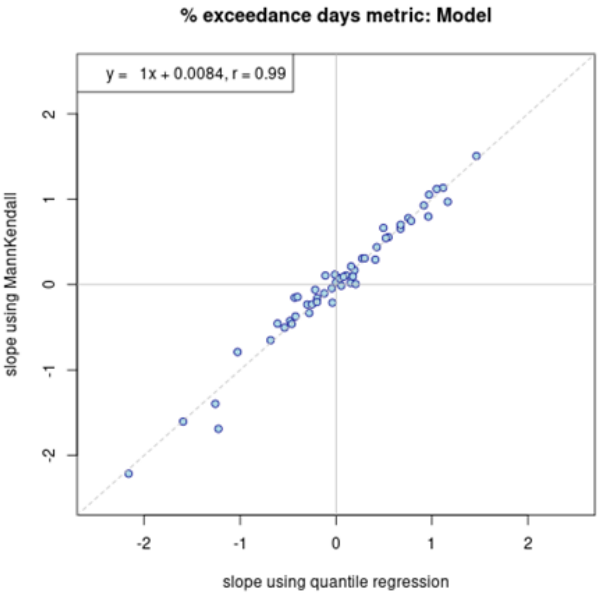


Mean Metric: p-values

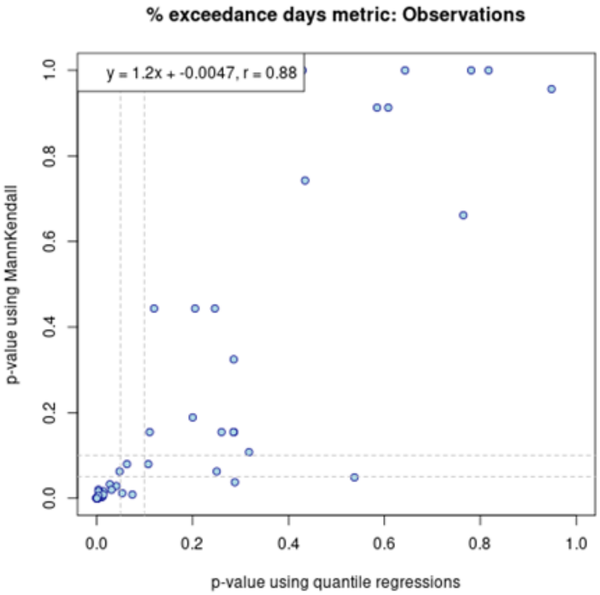
	p-value	N- model	N -obs
Number of areas with regressions	N/A	52	52
Number of areas with significant slopes in both regression types	0.05	28	22
	0.1	30	25
Number of areas with insignificant slopes in both regression types	0.05	17	24
	0.1	12	19
Number of areas with significant MK slope and insignificant QR slope	0.05	4	4
	0.1	6	3
Number of areas with insignificant MK slope and significant QR slope	0.05	2	1
	0.1	3	4

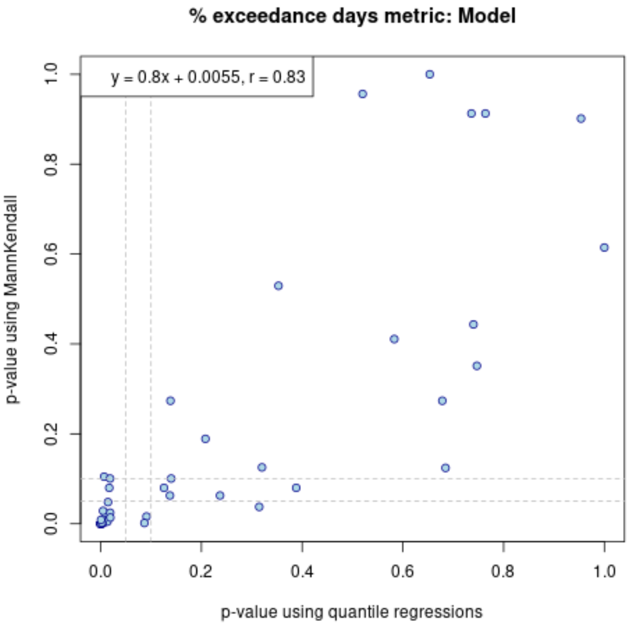
% exceedance days metric: slopes





% exceedance days metric : p-values


% exceedance days metric: Observations


% exceedance days metric: Model

% exceedance days metric: p-values

	p-value	N- model	N -obs
Number of areas with regressions	N/A	52	52
Number of areas with significant slopes in both regression types	0.05	26	25
	0.1	29	29
Number of areas with insignificant slopes in both regression types	0.05	20	22
	0.1	16	19
Number of areas with significant MK slope and insignificant QR slope	0.05	3	4
	0.1	5	4
Number of areas with insignificant MK slope and significant QR slope	0.05	3	1
	0.1	2	0

Comment: Throughout the paper the authors use the expression “statistically significant”, however this expression is now recognized as being problematic and it should be abandoned and replaced by the more useful method of reporting all trends (with uncertainty) and all p-values, followed by a discussion of the trends and the author’s opinion regarding their confidence in the trend values. This advice comes from a highly influential paper by Wasserstein et al. (2019), published in the journal, The American Statistician, that has already been cited over 1300 times (according to Web of Science). This advice was adopted by the first phase of TOAR (Tarasick et al., 2019) and will also be used by TOAR-II. Some other recent papers on ozone trends that have taken this advice are: Chang et al., 2020; Cooper et al., 2020; Gaudel et al., 2020; Chang et al., 2022; Wang et al., 2022; Mousavinezhad et al., 2023. Because these papers report all trend values, uncertainties, and all p-values, and also discuss the trend results, there is no confusion regarding the findings, and one does not even notice that the term “statistically significant” is not used at all.

The authors describe a trend as “no trend” when the p-value is greater than 0.05. There are two problems with this approach:

1) as described above the expression “statistically significant” which is tied to the p-value of 0.05 should be abandoned. Just because a trend has a p-value of 0.06, it does not mean that there is absolutely no trend, it just means that there is a gray area and the trend is not as robust as one that has a p-value of 0.02. Chang et al. (2017) provide a nice demonstration of the useful information that can be gleaned from a trend with a p-value greater than 0.05 (see their Figure 13). They calculated a regional ozone trend for the eastern USA using all available ozone monitors (in summer the trend was strongly negative for the period 2000-2014). They then conducted an exercise to see what would happen to the regional trend if they threw out all time series with a p-value less than 0.05. The result was almost the same because the time series with p-values greater than 0.05 still reflected the overall regional decrease of ozone.

2) The authors are using the Theil-Sen method to calculate trends and p-values. As described above the 95% confidence intervals are unrealistically narrow using this method, and therefore the p-values are also too low. This means that too many sites are classified as having a real trend, according to the 0.05 p-value threshold. If the authors use another method for calculating trends (like quantile regression) the p-values will increase and they would then have to classify more sites as having “no trend”. Given the gray area around p-values, and given that trends with p-values greater than 0.05 can still be reliable, there is no justification for dichotomizing ozone time series as “trend” or “no trend” based on a p-value.

When I look at the maps in Figure 4 and 5 I am left wondering about the non-attainment regions labeled as “no trend”. Is there really no trend here, i.e. a flat line, or is there still a decrease, but it just doesn’t reach the arbitrary threshold of $p<0.05$? A good example is Tuscan Buttes. Table S-1 shows the observed and modelled trend is the same (0.14) but because the model has a p-value of 0.02 this trend is considered to be real, while the observations have a p-value of 0.06 and are classified as “no trend”. The TOAR papers report all trends and all p-values and the trend values in their map plots are colored according to p-value (Fleming et al., 2018). This allow the reader to see if a trend is still notable (e.g. a p-value between 0.05 and 0.10) or if there really and truly is no trend (e.g. a p-value > 0.33). It would be very helpful to the reader if the authors can color their maps according to p-value, in a manner similar to TOAR.

Response: Thank you for this comment. We have revised the maps in Figures 4 and 5 to show the P-value ranges from the TOAR assessment: $P <= 0.05$, $0.05 < P <= 0.1$, $0.1 < P <= 0.33$, and $P > 0.33$. We have also revised the timeseries symbols in figures 1, 2, 3, 6, 7, 8 and 9 to use symbols representing these four P-value ranges. We have attempted to remove the term “statistically significant” wherever possible and instead just report the P-value ranges. We now define the “no trend” areas using a threshold of $P > 0.33$. We have also removed the symbols indicating statistical significance from Figures 11 and 12.