We thank the reviewer for his/her comments. Our responses to specific comments are below (bold, indented).

The authors investigate O3 trends in the Northern Front Range Metropolitan Area of Colorado, a region which has exhibited ongoing issues with O3 exceedances in spite of significant reductions in NOx emissions. In addition to examining overall trends over time, the authors use weekday/weekend comparisons of NOx and O3 to help explain features of local chemistry, and also compare O3 vs. temperature over time. Overall this paper is clear, well-organized, and represents a solid, if incremental addition to the existing air-pollution literature. I recommend publication, following improvements in a few areas.

First, and most importantly, I have concerns over the authors' use of binned temperatures as a preliminary step to linear regression. While I understand that this methodology has been utilized for similar purposes in the past, there are clear statistical flaws related to the practice that should be addressed before these results can be considered robust. Specific issues in the context of this paper include the following: • At relatively small sample sizes (n = 64-92 per summer), terms such as "95th percentile" become somewhat problematic. Dividing this already thin sample size into even smaller 3◦C temperature bins must have, I assume, resulted in some bins with observations in the single digits. What methodology was used to determine percentiles from such small sample sizes? • By choosing uniformly spaced bin widths (years, in the case of this paper's temporal analysis, and uniform 3◦C temperature widths in the case of the O3/T comparisons) information regarding sample sizes within each bin is lost completely. A bin containing more observations clearly should be weighted more heavily than a bin with fewer, but as written I see no indication that this kind of weighting was performed. This issue will be especially consequential for the temperature bins, since the relatively sparse temperature extremes will be incorrectly given weights equal to those of the middle bins, most likely exaggerating the resulting slopes. See Wasco and Sharma, 2014 for a description of how evenly spaced bins can produce exaggerated slopes as a result of this bias. Two methods that could correct this bias are equal number bins (with variable temperature widths based on the frequency distribution) and quantile regression (Koenker and Bassett, 1978). I think either of these would be superior to the current "equal distance bin" approach, with quantile regression also having the benefit of simultaneously addressing the small sample size issue. Wasko C, Sharma A. Quantile regression for investigating scaling of extreme precipitation with temperature. Water Resour Res 2014;50:3608–14. Koenker R, Bassett Jr G. Regression Quantiles. Econometrica 1978;46:33– 50. Further examples of this technique applied specifically to similar air-quality questions may be found elsewhere in the literature.

Thank you to the reviewer for a detailed explanation of the issues with uniformly spaced temperature bins, and the suggestion of weighting the yearly trends. We will address both topics below:

1) **Temporal trends and weighting of years: The EPA ozone, NO_2, and temperature data are available at an hourly time resolution. For the temporal trends of ozone and NO_2 we calculated daily averages for 10:00 am – 4:00 pm for summer data (Jun-Aug). To determine the percentiles for each summer at a site we aggregated the daily averages and applied the Tukey method to find the 5th, 33rd, 50th, 67th, and 95th percentiles (figure 2a, figure 3a). As the reviewer noted relatively small sample sizes can be problematic when calculating high**

or low percentiles (95th and 5th). We believe that the reviewer is referring to the tendency for the percentile calculations at the 5th or 95th to be skewed by low and high outliers, which becomes more problematic as the sample size decreases. As the sample size becomes sufficiently small the 5th and 95th percentiles will tend to equal the minimum and maximum values of the data, which can be outliers. We went back through the yearly trends to investigate the influence of outliers on the percentiles and found that only 1 year at 2 sites (Welby and Carriage 2004) exhibited 1 day of unrealistically low ozone (<5 ppbv), which is lower than typical background ozone, and were removed as outliers to not skew the 5th percentile values. Below is a table summarizing the number of daily average points for each year used in the percentile calculations.

	Number of points in long term ozone trend daily averages						NO_2 trends	
Year	Welby	Rocky Flats	Greeley	Fort Collins	Carriage	CAMP	CAMP	Welby
2000	90	88		89	91			
2001	89	90		91	90		89	89
2002	88	85	87	91	91		85	78
2003	86	91	91	91	91		74	
2004	87	91	91	91	85		80	81
2005	91	91	91	91	89	63	91	91
2006	90	91	91	91	88	91	82	
2007	91	89	91	91	86	90	89	91
2008	90	91	87	91	91			90
2009	84	91	91	91	91			
2010		89	91	77	90		91	78
2011	91	91	91	91	91		71	86
2012	87	91	90	90	80	91	90	71
2013	86	90	91	75		91	91	86
2014	91	91	90	78		91	91	91
2015	90	91	91	91		90	90	84

The reviewer suggests weighting the yearly trends by the number of data points to correct for differences in the number of points in different years. However, we note that >90% of the years for all sites with available data have 80-92 daily averages, and we thus expect a negligible effect on the analysis from weighting based on the number of data points.

2) Uniformly spaced temperature bins versus temperature bins with the same number of data points: The reviewer suggests redoing the ozone-temperature analysis using temperature bin widths dictated by a constant number of data points in a bin instead of using uniform temperature bins. As the reviewer noted we were dividing an already small sample size of 80-90 daily averages into temperature bins, some of which contained <10 data points for the high and low temperature bins. Applying the percentile calculations to such small sample sizes was not statistically robust, and tended to only yield the minimum and maximum values for those temperature bins. To increase the number of data points for a more robust statistical analysis we used the hourly ozone and temperature data. For a full 92-day summer data set we are now working with 552 data points (10:00am – 4:00pm, 6

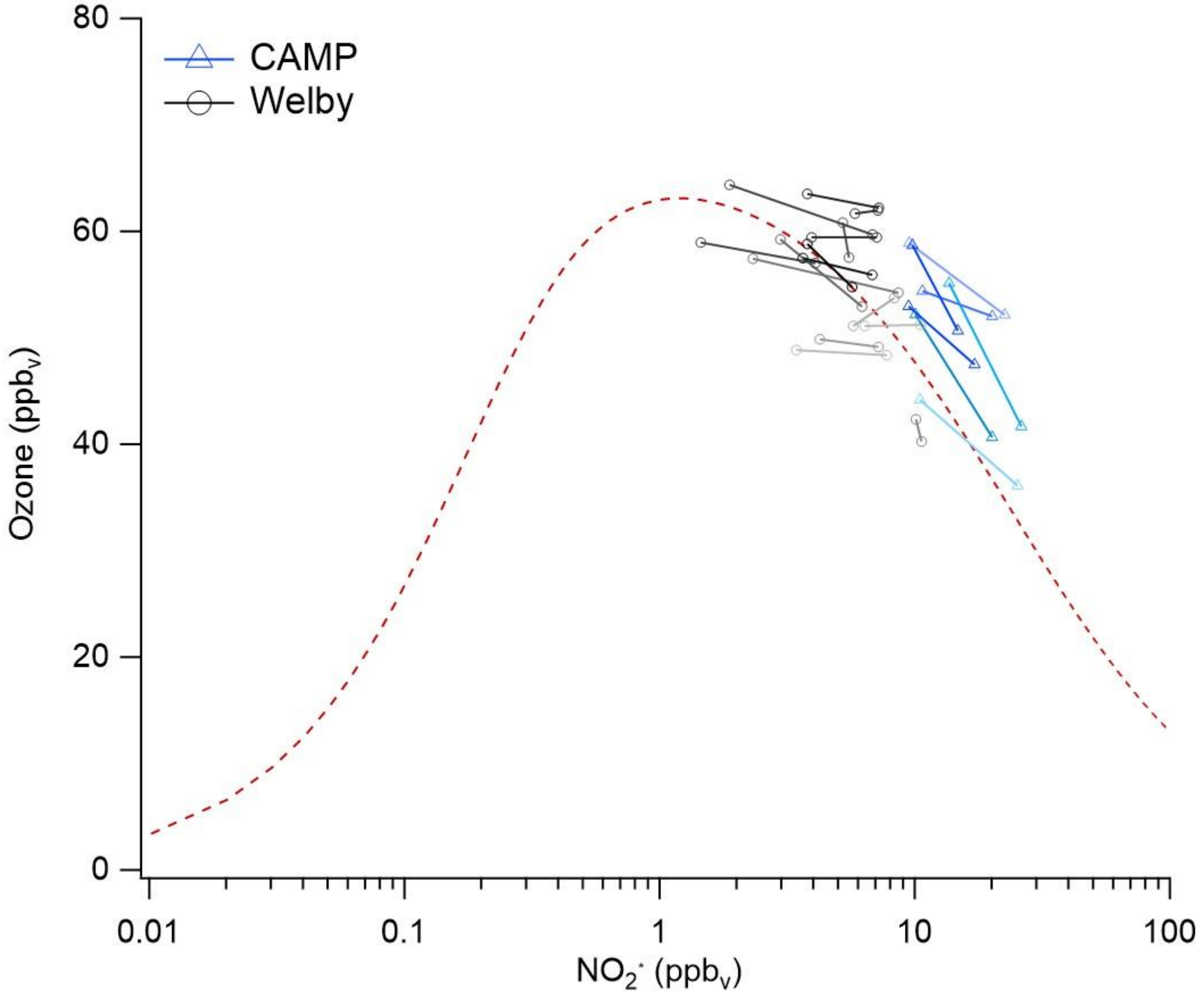
hours per day). The 552 data points were split into 5 temperature bins with 110 data points each, with the two extra points disregarded. Due to missing data, the smallest number of data points for a single temperature bin was 51 (CAMP 2005), but >90% of bins contained 100-110 data points. Due to the scarcity of bins with <100 data points we did not weight the ozone-temperature relationships by the number of points in each bin. We have updated figures 8 and 9 with this improved analysis. Below are summary tables of the number of ozone points in each temperature bin for each site and year. We note that this has no substantive effect on the interpretation of the data, nor conclusions drawn, but does make for a more robust analysis.

	Number of Points in Welby temperature bins					Number of Points in Rocky Flats temperature bins					Number of Points in Greeley temperature bins				
Year	**Bin 1**	**Bin 2**	**Bin 3**	**Bin 4**	**Bin 5**	**Bin 1**	**Bin 2**	**Bin 3**	**Bin 4**	**Bin 5**	**Bin 1**	**Bin 2**	**Bin 3**	**Bin 4**	**Bin 5**
2000	104	110	110	110	110	103	105	107	107	110					
2001	106	108	109	105	110	107	107	108	110	110					
2002	105	106	107	109	102	102	98	99	96	101					
2003	97	96	104	110	106	109	104	110	109	109					
2004	96	108	105	105	104	107	109	108	108	105					
2005	108	107	110	110	109	110	110	108	110	110					
2006	109	105	106	109	100	109	109	108	107	110					
2007	110	110	110	108	108	110	107	108	109	98					
2008	104	103	106	110	109	107	110	105	110	110					
2009	102	93	99	92	103	109	110	109	109	109					
2010						110	108	102	103	96					
2011	109	107	105	108	110	106	110	110	110	110					
2012	106	106	110	110	62	110	110	110	108	108	110	109	109	109	107
2013	110	109	106	108	72	106	110	110	110	105	110	110	103	108	109
2014	110	110	109	110	109	110	110	110	110	110	108	109	108	108	104
2015	103	108	110	107	109	107	110	110	110	108	108	105	108	108	108

	Number of Points in Fort Collins temperature bins					Number of Points in Carriage temp bins					Number of Points in Camp temp bins				
Year	**Bin 1**	**Bin 2**	**Bin 3**	**Bin 4**	**Bin 5**	**Bin 1**	**Bin 2**	**Bin 3**	**Bin 4**	**Bin 5**	**Bin 1**	**Bin 2**	**Bin 3**	**Bin 4**	**Bin 5**
2000	104	109	108	107	107	90	94	95	90	91					
2001	77	90	91	93	96	109	103	109	109	109					
2002	81	88	98	93	72	105	108	110	109	110					
2003	107	106	107	109	104	106	105	110	109	110					
2004	110	110	108	110	105	109	109	108	108	108					
2005	70	89	102	108	108	109	105	104	101	94	51	74	70	74	103
2006	107	107	110	110	110	92	109	109	104	105	110	109	108	107	107
2007	109	107	108	108	110	106	98	105	109	110	108	104	108	109	110
2008	109	109	108	107	110	90	103	104	100	107					
2009	105	110	110	109	110	107	110	109	110	110					
2010	104	110	110	110	110	109	110	109	110	110					
2011	110	110	108	108	110	108	106	109	110	104					
2012	110	108	105	108	100	108	108	110	110	108	108	107	109	110	109
2013	110	108	108	109	109						108	107	110	110	109
2014	109	110	110	110	110						109	110	110	110	110
2015	95	108	110	109	105						110	110	109	108	105

2. Figure 6: While I appreciate the attempt to use many symbols to distinguish years, I think the end result just doesn't work. The dense area around 10 ppb NO2 in particular is nearly impossible to interpret easily. I suggest either abandoning the symbols entirely, and using shaded dots to represent different years, or else zooming in on the data to create more whitespace in this concentrated region.

We have revised this figure to minimize the visual interference and clustering of the symbols. The revised figure is below:



3. The usage of "standard deviation" in several figure captions seems unclear. For example, on Figure 9 it seems to suggest that this is a standard deviation of many regression slopes. Is this the standard error of a single regression? Was bootstrapping performed, leading to many regression coefficients?

We have revised and updated most of the figures per a suggestion from reviewer 1 to be more consistent with the error analysis. The updates are as follows;

Figure 2b. The error bars are now the 95% confidence intervals around the reported ozone/year slopes.

Figure 3b. We included an additional figure similar to Figure 2b to show the NO_2/year slopes for the 5th, 50th, and 95th percentiles with the error bars representing the 95$^{\%}$ confidence intervals around the slopes.

Figure 5 was updated with suggestions from reviewer 1 comment 7 to show the weekday and weekend averages with the 95% confidence intervals.

Figure 7a was updated and shows the average weekday minus weekend ozone for each year for the six sites. The solid grey line represents the aggregated average of the six sites with the shading representing the 95% confidence interval.

Figure 7b was updated and shows the average weekday minus weekend NO_2 for each year for the CAMP and Welby sites. The error bars represent the 95% confidence interval of the averages.

Figure 8a was updated with the new equal bin size approach, and the averages of those temperature bins for each year are displayed. The 95% confidence intervals for the O_3 bin averages were not included in the figure for clarity purposes, but are typically <8 ppb$_v$.

Figure 9 was updated with the new equal bin size approach suggested, and the 95% confidence intervals around the yearly O_3/temperature slopes are included.