Interactive comment on "Multi-year objective analyses of warm season ground-level ozone and PM2.5 over North America using real-time observations and Canadian operational air quality models." by A. Robichaud and R. Ménard

Anonymous Referee #1

Received and published: 12 July 2013

Robichaud and Menard create a set of objective analysis data which combine modeled and measured North American ozone and PM2.5 over the period of a decade. They then use this data to describe and explore air pollution climatology, multi-year trends, and correlations to various meteorological and economic factors. The methods for creating their objective analysis appear to be technically sound and their work provides a valuable and extensive dataset which will be of interest to ACP readers. However, the text of the manuscript needs editing and I have some concerns about the metrics that they use for analyzing climatology and trends. This paper would also benefit from including discussions of and comparisons to other relevant/similar work which is not currently cited (specific references given below). In addition, I have numerous specific comments about the authors' presentation of and interpretation of their results. After revisions, I think that this manuscript has the potential to substantially add to the scientific literature. Please see general and specific comments below.

Reply from authors:

We acknowledge and thank reviewer no 1 for his/her abundant comments and criticisms and reference to the literature which could certainly contribute to enrich and improve the revised document. Our reply and comments are given below such that every single item has been reviewed and answered carefully. But first, to fix things let us recall that the main objective of the paper was to present multi-year analyses of surface ozone and PM2.5 and show applications such as climatology and multi-year trends or changes (e.g., difference between 2012 and 2005). The objective analyses were conducted carefully integrating quality-controlled surface data. The bias of the analysis has also been corrected. A new semi-empirical procedure was found successful to tune the error statistics (background error covariance and correlation length).
The computation of trends and climatology was done in light of a multi-disciplinary point of view integrating data not only from meteorology, air quality, health science but also from the information contained in economic fluctuations which is also linked to air quality as emphasized below and explained better in the revised version. Two new Annex will be included in the revised document to make the presentation clearer. Note that the comparisons with other paper/studies as suggested by reviewer no. 1 is not always straightforward and we have avoided pushing too much these comparisons since they could be misleading because the focus of different studies under comparison and the technology used (apparatus), quality control, the time-scale and the period covered are not necessarily the same as our study. Such extensive comparison is beyond the objectives of our paper. However, we agree that qualitative comparisons are certainly possible and effort will be made to add up such comparisons in the revised version of the paper. In fact, as pointed out by another reviewer (no. 3), our paper corroborates (at least qualitatively) findings of Cooper et al, 2010 and 2012 for U.S. A qualitative

agreement with reports of IMPROVE has also been found and that will be discussed in the revised version: e.g. spatial patterns are similar.

We understand the concern of reviewer no. 1 about the choice of metrics used but this choice also depend on the focus and goals and orientation of a given paper. Rather than using a metric such as 8-h max ozone, which is strongly linked with AQ standards and regulation, we have chosen to study trends in term of the full spectrum of the distribution of a given pollutant (i.e., various percentiles, standard deviation and mean). Similarly, Cooper et al, 2010 and 2012 and Vautard et al, 2006 also used as metric the different percentile of the ozone distribution as metric rather than the metric based on U.S. national air quality standards (NAAQS) or the Canadian one. Moreover, we believe that there is no such absolute metric that everybody should use (e.g. fourth highest daily or avg. daily maximum 8-h ozone) in the science of air quality to produce trends or provide a climatology. Adding-up the latter metrics would not change any conclusions of this report.

General Comments:

Reviewer no. 1

The manuscript is in need of copy editing. There were numerous grammatical mistakes and awkward sentences.

Reply from authors

A native speaker will be reviewing the final version of the re-submitted document and we will make sure that it will be free of such kinds of “mistake” mentioned by reviewer no. 1.

Reviewer no. 1

Sections 2.1-2.3 should be re-written so that they are understandable to the general ACP audience, many of whom are not data assimilation specialists. The authors should include a concise, high-level explanation of the method. Clearer links need to be made between the mathematical language/notation and the actual atmospheric data being analyzed.

Reply from authors

This section will be reviewed and essential background information will be provided so that a reader unfamiliar with data assimilation could better follow the theory used in the paper. Thank you for pointing that out.

Reviewer no. 1

Section 2.1 seeks to give an abbreviated introduction to optimal interpolation including the notation and equations used for this data assimilation technique. Unfortunately in an attempt to summarize more thorough descriptions of these methods such as what is found in the Kalnay text, this section ends up being unnecessarily confusing for readers who are not as familiar with data assimilation techniques. Before jumping right into equation (1) the authors should begin by giving the observation equation and the state equation so that it is clear throughout this section what is meant by the observation error and the “background” error. To clarify the presentation it would also help to provide the dimensions of the vectors and matrices in equation 1.

Reply from authors
We will follow suggestions from reviewer no. 1 and introduce the equation of observation and background and make clearer the definition of observation and background error and provide dimension for vectors and matrices. However, an unfamiliar reader still needs to refer to Kalnay text of data assimilation to get a basic introduction. We believe that is the role of textbook such as Kalnay text (2003), which are introductory textbook, and not precisely the role of a scientific paper to provide a pedagogic introduction to scientific subjects. Nevertheless, efforts will be made to improve on this in the revised paper.

Reviewer no. 1
First of all, x_aˆn is a vector, correct? Describe for the reader the dimension of this vector in terms of number of grid cells and number of variables in the model output so that it is clear how this equation applies to your hourly CHRONOS and GEM-MACH output (you do not have to use specific numbers, it is just to clarify the matrix algebra in equation (1)). The state vector in OI can include multiple chemical species or meteorological variables across space, i.e. x_fˆn could be the vectorized version of gridded model output for 10s or 100s of variables. In this case the background error correlation is not just a function of distance in space but also defines how these variables are correlated with one another. From your description and the format of equations (3) and (4) it sounds like x_fˆn in your case is for a single pollutant. Again providing the dimension of x_fˆn would clarify this early on.

Reply from authors
Suggestions from reviewer no. 1 will be incorporated in the revised version.
Yes you are correct, x^f_n is a vectorized version of gridded model output.
We will make sure that the presentation will be improved in this revised version
by providing these kind of details.

Reviewer no. 1
Throughout the paper, the authors present ozone results which include all hours in the summer season. I do not think this is the proper metric for ozone which is known to have a strong diurnal cycle and is regulated based on daily maximum values (generally max of 1-h or 8-h average for each day). I strongly suggest that the authors switch their analysis to look at climatology, correlations, and trends for the daily maximum 8-h average rather than the all hours average that they have been using. There are several reasons for this suggestion:
***Both Canadian and US regulations are based on the maximum daily 8-h average and their emissions reductions are targeted at this metric not at the 24-h average that the authors analyze.*
***Most epidemiology studies use the 8-h daily max rather than 24-h average in correlating*
health outcomes to ozone.
***Daytime ozone values are likely to decrease in response to emissions control strategies*
while nighttime values are likely to increase in response to these same emissions reductions (due to ozone disbenefits from NOx which are common at night as well as during wintertime months). Consequently, any emissions-related trends which include both day and nighttime values are likely to be dampened relative to trends in daytime ozone. This behavior may also explain some of the reduction in standard deviation that

the authors report, as increasing low nighttime values while decreasing high daytime values would cause such a result.

Reply from authors

The metric 8-hr max daily ozone is highly influenced by meteorology variation especially when only a decade of data is available (National Research Council, 1991; Cooper et al, 2012). Therefore, the use of that metric also requires complicated adjustment for meteorological factors which will never completely correct for these fluctuations (Thompson et al, 2001, see also more discussion below in specific comments). Moreover, we believe that the focus on daily 8-h maximum to compute trends or inter-annual variability could give the false impression that the ozone problem is solved because trends using only this metric indicate, in general, diminishing levels. Finally, the increase of night time ozone due to emissions reductions (e.g. in urban centers) does not apply or at least applies less for rural, mountainous or remote regions. Consequently, the reduction in standard deviation is not only due to emission reduction and NOx titration but also to increase of global background ozone which is a known phenomena corroborated by Cooper et al, 2010, 2012 and many others. This point is important for us.
Note that in our paper, we do take into account diurnal fluctuations. We have made several analysis in terms of different hours of the day: 1) basic verification (see table 3a,b) using different time of the day such as 00Z,06Z,12Z and 18Z, 2) analysis increments for all hours combined but also for 18UTC (about midday), figure 5a,b, 3) long term average of ozone for all hours and also for 18UTC (Midday) analysis increments, figure 4a,b., 4) cross validation for all hours of the day (figure 6,7 and 8). Now concerning trends, we believe that the high percentile 99th, 98th and 95th basically reflect highest values which essentially occur during the afternoon and are proxy for daily maximum for ozone. Also, we believe that our analysis turns out to cover the whole diurnal cycle since we have produced trends for all percentiles. As mentioned before, adding up the 8-h max daily avg. would not change any conclusions of our paper. We wanted to address more general issues such as the increase of background ozone (by using low percentiles), the evolution of the ozone distribution (through tendencies of the high, medium, low percentiles and standard deviation of the distribution) and the fact that using OA for computing trend and climatology is more adequate that using only model or observations alone. As suggested by reviewer no. 3, in the revised version of the paper the latter should be made clearer to emphasize that OA does a better job of evaluating trends that model or observation alone. Recall that the main goal of our paper is to present and make available multi-year analyses of ozone and PM2.5. Any user which wants to use our multi-year analyses can choose the metric they wish for their own application since all the multi-year analyses are available on a hourly basis for about a decade for ACP readers.

On the other hand, our paper is intended to a scientific journal which covers rather multi-disciplinary science. Our understanding is that ACP journal is not primarily intended for policy-makers nor air quality managers who have to focus on avoiding violations of air quality standards but for a much broader audience. The metrics used by different author to calculate ozone changes could vary a lot depending on the objectives of the different studies. Many authors do not even use 8-h avg daily max for ozone for trends. Cooper et al. 2010 computed trends of ozone for different percentile of the distribution whereas Cooper et al, 2012 used all 24-hours data for ozone to compute trends and they do not specifically focus on the 8-hour daily maximum. Smith and Shively, 1995 modeled trends based on exceedances of a high threshold, whereas Lelieveld et al., 2004 also used monthly mean of ozone to compute trends. Saavedra et al (2012) examined trends in Spain and have used the mean daily maximum, daily mean, and different percentiles to evaluate the trends in Spain. They observed that for the reference sites, both mean and maximum ozone trends are similar. Finally, daily 24-h mean are highly correlated with average daily 8-h maximum ozone values and also have very similar spatial patterns (Jeff Brook, Environment Canada, personal discussion, 2013). Therefore, we believe that using the average daily 8h maximum value as a metric instead of the full distribution (percentiles) would not change any conclusions of our paper.

Note that comparison of trends given in the literature (Cooper et al., 2010; Cooper et al., 2012, Chan, 2009 for example) and to a certain extent with the IMPROVE network report in U.S. give similar results qualitatively when compared to our paper as this will be emphasized in the revised version. Concerning the behavior of the standard deviation, we do not wish to exclude what reviewer no. 1 proposed but we believe that our results show two important points: a simultaneous decrease of high percentile due to better regulations and the increase of low percentile caused by the world ozone background which is rising as also mentioned by Cooper et al, 2010, 2012 and references therein. In urban centers, and we agree with reviewer no. 1, NOx titration contributes as well to the increase of low percentiles of ozone. Note that the observed data used include all kinds of sites (rural, semi-urban, urban and remote in our study) not only urban sites.
About the comment that most epidemiological studies use 8-h daily max ozone, we have to recognize that, unfortunately, health effects are not sufficiently well known to conclude that the best predictor is that metric. The choice of a single ozone metric to characterize health outcome very well is still lacking (Institute for Risk Research, 2007, thereafter IRR, 2007). Synergy between pollutants and chronic effects are certainly not taken into account by using the average ozone 8-h daily maximum. In any case, a stronger link between exposure to air pollution has been found and it is more related to PM2.5 than ozone. According to a recent study made by a panel of experts (IRR, 2007), “while air standards have historically and continue to play a central and useful role in regulating air pollutants, the findings of key epidemiological studies suggest that air quality management based on standard-setting for single pollutants is simplistic and probably suboptimal in protecting public health”. Moreover, ozone seems to be be “a primer” in many cases rather than a trigger of health outcomes as it is recognized that the major air pollution impacts on health are linked to PM2.5. Moreover, recent epidemiological studies show more consistent evidence of lung cancer effects related to chronic exposures rather than acute exposure (the later more associated with the 8-h max daily ozone). Finally, there is little to indicate a threshold concentration below which air pollution has no health outcome (IRR, 2007).

Reviewer no. 1

***Model biases are often opposite during the day and at night (for instance if the model has too little NOx, then daytime values will be underestimated while nighttime values will be overestimated). This makes interpretation of OA and model comparisons difficult.*

Reply from authors

This is true that model biases may have a pronounced diurnal signature and could change sign during the day (as indicated by the systematic error plotted in figure 3 of our paper). However, not only biases (mean O-P or systematic errors) but also standard deviation of O-P (observation minus prediction) have been used in figure 3 and elsewhere as metrics to compare model, OA and observations in our paper. The standard deviation of O-P does not change sign from day to night and is a more robust evaluation of model or objective analyses performance. Moreover, the use of FC2 (frequency of correct within a factor of 2) is a robust metric against outliers which also does not change sign with different time of the day (always above zero). FC2 is provided in tables 3 and 4 of the paper whereas the standard deviation of O-P is given in Figure 3. A review for an appropriate choice of metrics for verification is discussed for example in Chang and Hanna (2004). Therefore, we believe that this comment of reviewer no. 1 is irrelevant.

Reviewer no. 1

***Due to the regulatory focus and daily maximum 8-h ozone values, most researchers also focus on this metric. Looking at climatology and trends in 24-h ozone will not be intuitive to most ACP readers or to the scientific community as a whole.*

Reply from authors

As already mentioned above, the objective of our paper was not specifically oriented to focus to use metrics related to air quality regulations. EPA annual reports (EPA, 2010, 2012) and other similar air quality agencies address these concerns for audience such as air quality policy-makers and managers. On the other hand, a review of the literature does not confirm to us that "most researchers also focus on this metric" as pointed out by reviewer no. 1. Although the metric max 8-h ozone values for ozone is a popular metric in U.S., many reports and scientific paper do not necessarily use maximum 8-h ozone values for trends. As already mentioned in our paper (section 5.2, line 27), on the choice of metric, we were inspired by Cooper et al, 2010 (published in Nature letter) and Vautard et al, 2006 (published in J. of Geophys. Res.) which used different percentiles (such as 95th, 50th,5th) as metric to study the trends. Recently, Cooper et al., (2012) also used metrics based on the full distribution (e.g. percentiles rather than the avg. daily max 8-h ozone). On our paper, we also have added up a trend on mean and the standard deviation to complete computation of a more general trend analysis. Using only the maximum 8h values would also implicitly assume that the ozone distribution does not change with time and space (which is not true). We believe it could also lead to conclude that the ozone problem is becoming no longer serious, at least in Canada (since most of 8-h maximum daily values are decreasing below AQ standards). Our technique adopted is more general since it permits to evaluate changes of high values of ozone (high percentile change) and also values of low levels (low percentiles) as well as the mean standard deviation. It is all in there. Note also that our choice of a metric were also dictated by the need for a one that can be both apply to ozone and PM2.5.

Reviewer no. 1

Section 5.1 deals with changes in average ozone. For ozone we are much more interested in the behavior of high values than in changes of the mean. Figure 11 would be more relevant if it showed changes to 98th percentile ozone or something similar.
Section 5.1: The choice of using two distinct years to map and evaluate air pollution trends leaves the analysis vulnerable to spurious meteorology-driven changes. Even though the years were chosen so that they "both show roughly similar weather regimes over many parts of North America", it is not possible to eliminate met-driven variability

between the two years (i.e. no two years are going to have similar meteorology in all locations). It would more technically defensible to base this trends analysis on a multi-year trend to minimize the type of spurious differences that are likely to show up in comparing two distinct years. In figure 13, the authors show multi-year trend lines. It would be better if Figures 11 and 12 were based on location specific trends such as this (perhaps show gridded values as change in ppb/yr or ug/m3/ yr based on the same analysis used to create the trend lines in Fig 13). Additionally, the change in models between 2005 and 2012 adds another variable to the comparison of these two years.

Reply from authors

In section 5.1, our interest was to map differences from two similar years (2005 and 2012). We will define better below and in the revised version what do we mean by "two similar years". Concerning the need to use many different metrics including 98th percentile, we would like to point out that this is done in Figure 13A and table 6. In fact, we evaluate how the whole distribution of ozone is changing with time. We believe our analysis is fairly complete by both showing geographical differences of mean ozone (figure 11) and also trends describing temporal changes of the whole distribution (including percentile 98th) (figure 13). Both figures are complimentary. Finally, when reviewer no. 1 mention 'We", it is difficult to know to whom is he/she referring to. People from EPA, US investigators, air quality policy –makers and managers ?

We have used differences between 2012 and 2005. Now, as reviewer no. 1 pointed out, it is not possible to completely eliminate met-driven variability between two specific years. However, it is possible to select two years which have minimum differences. We are providing here, as additional material (Annex B in the revised paper), a principal component analysis (PCA) demonstrating that the summer of year 2005 and 2012 are similar with respect to factors influencing ozone inter-annual variability (e.g. 2005 and 2012 are neighbor in the PCA plot, see Figure B.1 reproduced below). More details will be given in the revised paper but let us mention that this PCA analysis includes meteorological factors (temperature anomalies of US and Canada), economic fluctuations as form of gross domestic product growth rate and area burned in US and Canada by wildfires. Tele-connection indices such as NAO (North Atlantic Oscillation) and ENSO/MEI (the atmospheric multivariate expression of El Nino) were also included in the PCA analysis but found having little impact. Temperature and economic fluctuations were found to be the main drivers as revealed by table 8 (original manuscript). We apologize that details were lacking about the rationale for selection of the two years in the original paper but we now provide these details in the revised paper.

We have avoid plotting multi-year trends as a form of geographical map because it would involve projecting on grid points (for mapping purposes), multi-year trend found locally in the "observed space". The interpolation error of mapping these trends would be difficult to control so mapping multi-year trends could end up being noisy and lead to artifacts. As mentioned by EPA (2012, p 9) on ozone trends; "one site may show increases in ozone levels while nearby sites show decreases". Note that our results for trends of ozone are consistent with results of Cooper et al, (2010 and 2012).

Reviewer no. 1

The year-to-year changes in Figure 13 make it look like perhaps the model switch (between 2009 and 2010) had noticeable impact on ozone but not on PM2.5. A *comparison between the time series in Fig 13 and one based only on monitoring data might reveal whether the jump in ozone values between 2009 and 2010 is based on a real air quality change or is due to the switch in modeling systems.*

Reply from authors

Of course model switch could produce such changes (figure 3). However, we did not use model values in any computation of trend or production of a climatology or any other applications precisely to avoid artifacts and spurious effect that could be caused from model change. We rather use values from the objective analysis where the bias correction have almost eliminated the OA systematic errors (see figure 3, and figures 6 to 8). There is no such switch in analysis bias for OA. The consistency is maintained since the model bias is corrected for the analysis OA. We believe that changes from 2009 and 2010 in figure 13 are rather likely due to changes of temperatures (anomaly +0.172 C in 2009 and + 0.522 in 2010 in U.S.) and economic conditions in both Canada and US (gross domestic product growth rate jumped from strong negative values in 2009 to moderate positive values in 2010 in both Canada and U.S.). As mentioned in our paper, economic short term fluctuations have an impact on short term pollutant fluctuations (Friedlingstein et al., 2010; Granados et al, 2012; Castellanos and Boersma, 2012) and disregarding this factor could lead to incomplete or wrong conclusions.

Reviewer no. 1

page 13986, lines 16-19 and Sections 5.1 and 5.2: Although the analysis described here provides some advantages over past trends analyses, there are multiple reports and journal articles that are not mentioned in this article which evaluate trends in ozone and/or PM2.5 based on measurements made at monitors across North America. The US EPA regularly releases trends reports for ozone, PM2.5 and other pollutants (http://www.epa.gov/airtrends/reports.html) as does the corporate institute for research in the atmosphere (CIRA) at Colorado State for PM2.5 (http://vista.cira.colostate.edu/improve/Publications/improve_reports.htm). The article should acknowledge these other efforts and should compare results found here to other studies' findings:

***Our Nation's Air: Status and trends through 2010, US EPA, Office of Air Quality Planning and Standards, Research Triangle Park, NC, February 2012, EPA-454/R-12-001: http://www.epa.gov/airtrends/2011/report/fullreport.pdf*

***Our Nation's Air: Status and trends through 2008, US EPA, Office of Air Quality Planning and Standards, Research Triangle Park, NC, February 2010, EPA-454/R-09-002: http://www.epa.gov/airtrends/2010/report/fullreport.pdf*

***Interagency Monitoring of Protected Visual Environments, Spatial and seasonal patterns and temporal variability of haze and its constituents in the United States, Report V, June 2011, ISSN 0737-5352-87: http://vista.cira.colostate.edu/improve/Publications/Reports/2011/PDF/IMPROVE_V_FullReport.pdf*

***Murphy, D.M., Chow, J.C., Leibensperger, E.M., Malm, W.C., Pitchford, M., Schichtel, B.A., Watson, J.G., White, W.H., Decrease in elemental carbon and fine particle mass in the United States, Atmospheric Chemistry and Physics, 11, 4679-4686, 2011.*

***Hand, J.L., Schichtel, B.A., Malm, W.C., Pitchford, M.L., Particulate sulfate ion concentration and SO2 emission trends in the United States from the early 1990s through 2010, Atmospheric Chemistry and Physics, 12, 10353-10365, 2012.*

Reply from authors

We agree with reviewer no. 1 and we will acknowledge by making references with the above reports in a revised version of our paper as much as possible. However, note that the comparisons with other paper/studies as suggested by reviewer no. 1 is not always straightforward since the methodology, network used, and metrics and the period covered, the quality control, the technology used (instruments artifacts), the spatio-temporal scale, period of the year, season, etc. often are not the same as our study. Nevertheless, we agree to provide a qualitative comparison as much as possible in a revised version with relevant paper such as Cooper et al, 2010 and 2012 for ozone and with the IMPROVE report for PM2.5. This is beyond the scope of this paper to provide an extensive comparison with all existing papers on the subject.

Specific Comments:

Reviewer no. 1

Page 13970, line 10 and line 13: Technically the term "aerosol" refers to the mixture of gas and particles, not just the particles themselves. Also, particle comprising PM2.5 are often not just "solid" but include a liquid phase as well (both aqueous and organic phases may be liquid).

Reply from authors

In the book of Seinfeld and Pandis, 1998 who are world authority in the field of atmospheric science and air quality, it is stipulated on that aspect the following, on p.97; "whereas an aerosol is technically defined as a suspension of fine solid or liquid particles in a gas, common usage refers to the aerosol as the particulate component only". On our study, measured PM2.5 using mostly TEOM (Tapered element oscillating microbalance monitors) tend to evaporate most of the liquid phase on the aerosol at the heated inlet of the apparatus so that only the particulate matter is measured, not the water (Allen et al, 1997). On the other hand, model value of PM2.5 (CHRONOS or GEM-MACH) used in the objective analysis only include solid matter as well so that there is no inconsistency there. Finally, particulate matter affects health, the solid part, not the water around.

Reviewer no. 1

Page 13970, line 11-13: Secondary formation is also a major source. This sentence implies that all particles are primary in nature.

Reply from authors

In the revised text we will mention about secondary formation. In the original text, we wanted to refer to sources of primary emissions. We will correct that accordingly for completeness.

Reviewer no. 1

Page 13970, line 17-21: The authors leave off the most serious health effect that has been linked to PM2.5, death. As they state later in section 5.1.2, a 10 ug/m3 change in PM2.5 has been associated with a 6% change in death rate. It would be appropriate to mention that health outcome here.

Reply from authors
We are very well aware that PM2.5 excesses are linked to mortality and morbidity since we mention that in section 5.1.2 as pointed out by the reviewer no. 1. We will just move this information on line 17-21 as requested.

Reviewer no. 1
Table 1 is unnecessary.

Reply from authors
Reviewer no. 1 does not give any reason why Table 1 is unnecessary. This table (which is now revised to include the information about death rate and other relevant info) is now quite complete and turns out to be a quick reference in the paper for our readers not familiar with the environmental and health issues related to surface ozone and PM2.5 and who would like a quick summary as a form of a table. Table 1 is the result of an extensive effort to make a multi-disciplinary summary. We rather believe that ACP readers would be pleased about it.

Reviewer no. 1
Page 13973, lines 2-13, page 13986, lines 16-19, and Page 14000, line 27-page 14001 line 2: The US EPA in collaboration with the US CDC have undertaken a similar project in which they have used a hierarchical bayesian model to combine air quality model results and measured concentrations of ozone and PM2.5 at 12km and 36km resolutions to look at multi-year trends (2001-2008 with more years in progress). Information about and data from this project is publically available (http://www.epa.gov/heasd/research/cdc.html). Several journal articles have been published
on the hierarchical bayesian model developed for this purpose:
***Berrocal, V.J., Gelfand, A.E., Holland, D.M, A Spatio-temporal downscaler for output from numerical models, Journal of Agricultural, Biological, and Environmental Statistics, 15 (2), 167-197, 2010.*
***McMillan, N.J., Holland, D.M., Morara, M., Feng, J. Combining numerical model output and particulate data using Bayesian space-time modeling, Environmetrics, 21, 48-65, 2010*
***Berrocal, V.J., Gelfand, A.E., Holland, D.M. Space-Time data fusion under error in computer model output: an application to modeling air quality, Biometrics, 68, 837-848, 2012*

Reply from authors
Thanks for the information. But, it is not within the scope of our paper to review all the possible projects which are being undertaken. Our technique also differs from the above work. However, in the revised version of our paper, we agree that for ACP readers it would be interesting to mention references to hierarchical Bayesian as an alternative method of producing analyses. Note however that our interest was both Canada and U.S. whereas the above studies only focus on U.S. territory.

Reviewer no. 1
Page 13976, line 23: References to Wikipedia are not acceptable in a peer-reviewed article. Please find a more reliable reference.

Reply from authors

The Cholesky decomposition to invert matrices is something quite standard and well explained on the Wikipedia site which could be appreciated by unfamiliar reader. However, we understand that for peer-review process, a more academic reference is needed and will be provided in the revised version.

Reviewer no. 1

Page 13979, line 27 – Page 13980, line 4: What are the implications of assuming homogeneous background error given that ozone and PM2.5 have large spatial gradients in urban areas which cannot be fully captured by a 15km grid resolution? In addition to the rural versus urban difference it seems likely that across such a large domain the correlation structure of the errors probably changes in coastal areas vs interior locations, high elevation vs flat etc. A brief discussion of the implications of the assumption of homogenous background error would be appropriate.

Reply from authors

This is a very good question. But the fact that the density of the data in urban centers and near coastline (California and U.S. Eastern seabord) is high enough in our case (see Fig. 2 of our paper) make the assumption of homogeneity and isotropy not critical, at least for these locations. We refer here to the paper of Frydendall et al. (2009) for the treatment of anisotropy and the influence of network density on correlation length. On our side, we have made experiment with non-homogeneous correlation background error but without success. However, this is part of our own future work to investigate more on this aspect. We will mention this in the discussion section in the revised paper. Thanks for pointing out this interesting issue. However, capturing local scales features is beyond the scope of the present study.

Reviewer no. 1

Page 13984, line 18: Please add the following references:

***Herring, S. and Cass, G. The magnitude of bias in the measurement of PM2.5 arising from volatilization of particulate nitrate from Teflon filters, J. Air Waste, Manage., 49, 725-733, 1999*

***Frank, N.H. Retained nitrate, hydrated sulfates, and carbonaceous mass in Federal Reference Method fine particulate matter for six eastern US cities, J. Air Waste Manage.,56, 500-511, 2006.*

Reply from authors

We do not quite understand why these two references are required. Our study have not used any measurements made up with Teflon nor nylon filters. The real-time measurement data from AIRNOW and Canadian NAPS and rural stations (CAPMON network) from near-real time monitoring networks have been used which do not directly use Teflon nor nylon filters, as far as we know. Rather, in the measurement of PM2.5, many TEOM have been used (without FDMS technology). These tend to underestimate PM2.5 but mostly in winter months (Allen et al, 1997). But our study only deals with summer cases. So these two papers, although very interesting seem irrelevant to us.

Reviewer no. 1

Page 13985, line 27: By using 18 UTC instead of the same local time everywhere, the authors have picked a time when ozone concentrations are typically higher on the East Coast (2pm LST) than on the West Coast (11am LST). This artificially inflates *eastern ozone compared to western ozone in this figure and is visually misleading. It would be more appropriate to show this comparison either using local time or using daily maximum ozone values.*

Reply from authors

We agree with reviewer no. 1 to modify 18 UTC for local time (2pm LST) avoiding any inconsistencies and possible misleading interpretation. We will replace the bottom images of figures 4A and 4B with day time ozone and PM2.5 as requested. Thanks for pointing out this inconsistency to us.

Reviewer no. 1

I suggest that you move Section 4 to come before Section 3. Validation of results should come before any interpretation of those results.

Reply from authors

We agree to do so and the revised version will be reflecting that. Note that reviewer no. 3 also requested that change.

Reviewer no. 1

Page 13991, line 8: OA is not a monitoring system.

Reply from authors

This typo will be corrected.

Reviewer no. 1

Page 13991, lines 10-12: Other likely explanations include:
***Lack of dense monitors in these areas making OA more uncertain*
***Many studies have shown that coarse model resolution (15km and 21km) cannot capture meteorology in complex terrain. Models often need to be resolved at 4km, 1km or even finer resolutions to capture complex airflows and phenomena such as cold pools in mountainous regions.*
***The satellite seems to be able to better capture certain air pollution features than OA, such as the high PM2.5 concentrations in Salt Lake City which are due to cold pool meteorology that regional air quality models have trouble simulating.*

Reply from authors

Cold pool meteorology is certainly common throughout the Western US and Canada and models tends to not provide good results in such cases. Thank you very much for the useful remarks. There will be included in the revised version.

Reviewer no. 1

Page 13991, line 25 – Page 13992, line 1: I can understand why it is desirable to filter out outlier when they are due to erroneous measurements, but sometime the outliers may be real and important. The US ozone standard is based on the annual 4th highest ozone concentration (_98th percentile) and the US daily PM2.5 standard is based on the 98th percentile. So it does not seem desirable to filter out the very types of events that are the main focus of regulation and research.

Reply from authors

In data assimilation and objective analysis field, it is necessary to remove outliers before performing any data analysis (Gauthier et al., 2003; Ingleby and Lorenc, 1993). This is a standard practice. For example, in Canada, some data from few stations have shown that they were at times contaminated with the zero-span auto calibration for ozone so it is necessary to remove this artifact. The main quality control algorithm (so-called background check) is based on the difference O-P (observation minus model Prediction). Note that the percentage of data screened out by our quality control is less than 2% for most of the time. So it filters extreme values and outliers, not necessarily high values (e.g., high percentiles). There is always a risk that a good data would be excluded by the quality control. But on the other hand, without quality control, bad data could have a tremendous impact and gives spurious results and contaminate trends.

We do not buy that "extreme events are the main focus of regulation and research". Extreme values analysis is the focus of certain researcher in air quality but not the main focus of atmospheric chemistry research community. By definition extreme values in air quality occur on a small spatio-temporal domain therefore affecting only limited population over a small amount of time and thus having limited health impact (at least for ozone and PM2.5). For example, an exceptional extreme event of PM2.5 could be produced by a local firework which plume hit a monitor by accident. Such extreme values if retained by the analysis could mislead the interpretation of results: for example, one may conclude that regulations had a worse effect in the specific year or month when that single extreme event happened. It is well known that in statistical analysis, outliers have a considerable weight and should be removed since it could contaminate regression analysis (Weisberg, 1985). Note that the fact that all high percentiles have basically the same trends (see figure 13 of our paper for both ozone and PM2.5) suggests that our approach and methodology is somehow robust and that using other metric such as the average daily 8-h maximum value is unlikely to change conclusions. Moreover, our trend computation qualitatively compares well with other studies such as Cooper et al, 2010, 2012 (as pointed out by reviewer no. 3). Note also that using the 98th percentile or annual 4th highest ozone concentration (as mentioned by reviewer no. 1) is in some sense a metric which removes outliers (the upper 2% of the distribution in the former and the first 3 highest values in the latter). So the same argument could apply here: e.g. replacing a metric using explicitly a quality control of outliers by another one which also does, but indirectly, in the sense that the 98th percentile or 4th highest value does not include the extreme or outliers values either.

Reviewer no. 1

Page 13992, lines 4-5: Again, locations that are close to primary PM sources (factories, agricultural burning etc) are the very locations where health effects from PM are of concern. Why is it desirable to have a technique that "disregard[s] data influenced too heavily by the proximity of local strong sources of PM2.5"? These are the very locations that we are most interested in for protecting public health.

Reply from authors

There is no systematic procedure to disregard in any sense observations near local strong sources in our study. The quality control of objective analysis will reject data having 10 standard deviation of model residual (O-P) for PM2.5 and 5 standard deviations for ozone or will reject hourly data having too much temporal swing. We have to recall that the model (P) will presumably have the knowledge of real source of

emissions (through emission inventories). Therefore, there will not be systematic rejection of data near major sources. In normal situation, only few percent of data is rejected by our background check algorithm (based on O-P). It is only when the source does not exist in the emission inventory but is real and the difference between the model and the observation is beyond several std. dev., that it will be rejected in the quality control. From experience of our system, we have seen some good values rejected at times but they were from special events such as from fireworks or stations too close to wildfires and certainly not from factories or agricultural emissions since the latter are included in emission inventories and hence included in models so that O-P will not be extreme.
We agree to review the text there so that any possible confusion is eliminated.

Reviewer no. 1

Page 13993, lines 10-20:
***Figure 11 also shows increases in mean ozone in several urban areas (Detroit, Chicago and others). These ozone increases in urban areas may be due to reductions in NOx titration from mobile-source NOx decreases.*

Reply from authors

We agree with that. Thanks for pointing out this information. It will be included in the revised version.

Reviewer no. 1

***Fire activity is highly variable from year-to-year, so by comparing two discrete years (2005 and 2012) you are more likely to see spurious effects from wildfires that may or may not represent a longer term trend in ozone from fires. Since models pick up strong localized fire effects and wildfires in 2005 and 2012 are not likely to have occurred in the exact same locations, it seems likely that the authors would see both increases and decreases in ozone due to differences in wildfires. If Figures 11 and 12 showed multiyear trends, as previously suggested, any conclusions about ozone due to increasing fire frequency would be more robust.*

Reply from authors

First, let us say that we have computed difference of averages for a long period (the whole summer June-July-August 2005 and 2012). The differences are not snapshot or for a particular event or episode but for a whole summer period. Averaging over three months period smooth out local effects. Second, the area burned for both 2005 and 2012 were both above the decadal average (2002-2012). In Canada, the situation was reversed (below the 10-year average) but it turns out that it was also similar for both years 2012 and 2005 (see table B.1 attached in the revised document for figures concerning forest area burned). Reviewer no. 1 is asking to change figures 11 and 12 for multi-year trends. But why changing figures (e.g. fig. 11C) which corroborate other papers on important results such as Cooper et al, 2010, 2012 ?

Reviewer no. 1

***The authors state that ozone decreased in the intermountain West, but from Figure 11 it appears that ozone increased in large areas of the intermountain West (CO, UT). The authors should be more specific and list the US states for which ozone decreased.*

Reply from authors

This paragraph will be re-written in terms of the state name rather than the geographical mountain region (e.g. intermountain West which could cover many states). Thanks for pointing this out. This will be revised and corrected.

Reviewer no. 1

Page 13993, line 19: The authors give PM2.5 changes in ug/m3 per year based on two years of data. They should base these numbers on multiyear trends such as those shown in Fig 13.

Reply from authors

We will correct that and remove any reference to the word "trend" when we are dealing with differences (e.g., in Figure 11)

Reviewer no. 1

Page 13994, lines 8-10: by "growing local socio-economic and industrial activities" do the authors mean increases in oil and gas drilling? If so, they should state this explicitly. If not, they should consider this as a potential driving force due to the rapid increase of emissions from these operations in the Western US.

Reply from authors

Oil and gas drilling in U.S. and oil sands activities in Western Canada have increased in the recent decade. From 5 million barrels production per day in 2008, U.S. crude oil production increased to 6.5 million barrels per day in 2012 (US. Energy information Administration, 2013). We agree that we need to be more specific in here and will add up this information.

Reviewer no. 1

Section 5.2: When creating inter-annual trends, it is important to use a consistent set of trends monitors with long-term observation records. Including monitors that operated for only part of the trends period may introduce bias. Have the authors filtered their monitoring data using this criterion before creating their OA surfaces?

Reply from authors

Reviewer no 1 is absolutely right for the case when only observations are used to produce trends. Usually, if a monitor has not too much missing data (usually in the range 10-25% missing) for a specific year, it is appropriate to use for trends and vice-versa.
However, our study has used objective analyses to produce trend not observations, nor models. This has the advantage that if a station has missing values, the OA fills it up. The value provided for missing data by the OA system has been shown to yield very good results in areas of missing observations (but not too far from reporting monitoring sites) as demonstrated in section 4.1 (cross-validation). This advantage of OA was also pointed out by reviewer no. 3. What reviewer no. 1 is referring to does not apply in the case of trends made from objective analyses although it is a serious concern for trend made up from observation alone and we agree with that.

Reviewer no. 1

Page 13994, lines 11-12: The authors use speculative arguments here about the expectation that wildfires will increase in the West with warming.

Reply from authors

This is not a speculation. Climate warming is expected to increase the likelihood and severity of wildfires especially in Western U.S. We are aware that climate change deniers object to that kind of projection and claim this is speculation, but these projections were obtained from many model studies (IPCC, 2007) and applied in the context of Western US (National Academy of Science, 2011).

Reviewer no. 1

Data is available on fire activity and burned area in 2005 and 2012. Instead of speculating, the authors should use this data to determine whether/where wildfires were more prevalent in 2012 compared to 2005.They could then make more definitive conclusions on whether wildfire activity actually contributed to the changes that they estimate.

Reply from authors

We agree to remove this statement which was imprecise. As mentioned before, 2005 and 2012 are similar for temperature (anomalies above average in both cases) and also for the amount of area burned which are also both above average in U.S. (http://www.nifc.gov/fireInfo/fireInfo_stats_totalFires.htm). Instead, a new table (now called table B.1) will be included as part of Annex B of the revised paper. We agree that perhaps the text was confusing but will be changed to something more explicit and certainly clearer for ACP readers.

Reviewer no. 1

Page 13993, lines 10-20 and Page 13994, lines 21-22: Why do the authors think that the frequency or magnitude of stratospheric intrusions changed between 2005 and 2012? I am unaware of any studies which suggest that STE is increasing. If there is published literature on this phenomenon then the authors should cite it. If not, it seems very speculative to suggest STE is changing without any reason or evidence to suggest this might be the case. Since intrusions often happen at frontal boundaries, if the weather were quite different in these two years one might expect a difference in stratosphere-troposphere-exchange (STE), but the authors stated previously that these two years were specifically chosen for comparison due to similar weather patterns.

Reply from authors

The text should read "possible change of vertical transport" instead of "increase of vertical transport". We apologize for this mistake. However, It must be clear that inter-annual fluctuations of tropopause folding do exist and that stratospheric ozone can change tropospheric background. In fact, recent evidence suggests that that the contribution of STE on surface ozone in the Western USA may be greater than from Asian emissions. STE contributes to the inter-annual variability (Ambrose et al., 2011; Lin et al., 2012a,b, Oltmans, 2013) but it is not clear how it contributes to the trend. The correlation between ozone changes in the lower stratosphere and in the tropospheric ozone is also very significant according to Tarasick et al, 2005 leading to the conclusion that at least a portion of the changes in troposphere results from changes in the lower stratosphere. What the authors had in mind is that changes of surface ozone between two years could possibly be affected by changes of the lower stratosphere linked with changes of frequency and strength of tropopause folding simply because of the inter-annual variability which seems to increase with climate change (IPCC, 2007). In the revised version, however, we agree to remove reference to this point because this would need further investigation which are without the scope of our paper.

Reviewer no. 1

Page 13995, line 2-3: All the more reason to look at peak daily values instead of ozone from all hours, so that the authors can tease out different trends during the day (when ozone is of most concern from a health perspective) and at night.

Reply from authors

Reviewer no. 1 seems to focus on only one possibility of presenting the data using the avg. daily max 8-h ozone. We reiterate that the metric adopted here (a bit similar to Vautard et al, 2006, Cooper et al, 2010, 2012) consists of examining the trends of many percentiles, e.g. the full distribution. Again, it is all in there. High percentiles are linked with high values (mostly afternoon peak values) which impact more acute health effects and low percentile associated with increase of the background ozone (rural regions) or NO titration (urban regions). Low concentration cannot be excluded as having impacts on health either as there is no evidence to indicate a threshold concentration below which air pollution has no effect on population health (NRI, 2007). As mentioned before, we claim that our methodology is more powerful than just examining the trend of maximum daily 8-h values of ozone which will reflect only the trend of high values (where nothing can be deduced about the evolution of background ozone, about the changes of the ozone distribution). EPA regulators focus on metric linked to air quality standards for obvious reasons, however atmospheric scientists need to look at other aspects as well as the trend of background ozone which is of great scientific interest. Ozone is also a greenhouse gas and its mean variation is interesting to know as well, not only in terms of a metric dictated by NAAQS regulation.

Reviewer no. 1

Page 13995, line 20 - page 13998, line 2: I don't see the merit in trying to create a statistical model from datasets consisting of 9 and 11 obs. It might make sense if they were looking at a 30+ year trend, but the authors are really stretching their data too thin here. I would drop this entire section, and tables 7/8.

Reply from authors

Our focus was to analyze changes over the past decade. Over a period of 30 years or more, other factors come into play and should be taken into account such as climate change, land use changes, population movements, and other socio-economic factors which could drastically change over such a long period. Studying long term trends is out of the scope of this study but has been done by many authors already. Moreover, it was not possible in our context to do so since our OA could not be produced before 2001 (AQ models started in operation in 2001 in Canada). Also, our interest was to put the most recent period (the previous decade) in perspective. For example, explaining decrease of pollutant concentrations in the year 2007-2009 with better regulations and met fluctuations only is not appropriate as one might be tempted to conclude. Economic fluctuations have also contributed and this is the goal of pointing that important point in tables 7/8.

There is no such thing in statistics saying that when N is under 30, a statistical study should not be performed. What will determine if the period has appropriate length is not the number of years but rather the following: beside emissions regulations, ozone and

PM2.5 to a great extent depends on weather cycles and economic factors. Weather cycles and also short economic fluctuations are usually taking place over a period of less than 10 years or so. Therefore, our study well cover a complete cycle (2002-2012 for ozone or 2004-2012 for PM2.5) so there is indeed a merit to make such an analysis which should be view as an update of long-term trends made by numerous other authors. The impact of changes of emission due to new regulations over the past decade, the expansion of gas and drilling in US and oil sands in Canada both expanded in the last decade needs to be pointed out. Removing both tables 7 and 8 and dropping entire section would mean ignoring the impact of economic fluctuations on short term fluctuations of primary pollutants such as primary PM2.5. We believe this is a mistake. As shown by numerous author (Granados et al., 2012 and Friedlingstein et al., 2010 for CO_2 emissions; Castellanos and Boersma, 2012 for NO_2 total column) economic fluctuations do play a significant role in explaining short term emission and concentration of air pollutants in the troposphere. For example, the economic crisis of year 2007-2009 produces a slowdown of anthropogenic activities in North America which affected construction, vehicule-miles traveled, energy consumption and CO_2 emissions diminished during 2007-2009 in phase with the decrease of the U.S. gross domestic product (see EPA, 2010, Figure 3). The correlations in tables 7 and 8 are not coincidences even if the p-value is not always less than 0.05 (the latter is mostly due to the fact that the sample is small, $N \sim 10$). Now concerning the question about if whether or not the data is too thin, let us point out that the number of degree of freedom (n-k-1 where n=nb of obs, k =number of parameters for a statistical model) only needs to be greater or equal than 1 to perform statistical analyses. Most statistical tables for significance level in most textbooks give values for significance level from 1 to 100 degrees of freedom (e.g., for very small to medium size samples). Not only large samples of data analysis are making science advancement. In fact, in a small sample, a given p-value could provide more evidence of significance than the same p-value for a larger sample (Freeman, 1993). In many journals, number of sample with $N < 30$ are often presented. In conclusion, we agree to perhaps drop table 7 but table 8 is essential to the paper (also recognized by reviewer no. 3).

Reviewer no. 1

Page 13995, line 29: What is the source of the data on the wikimedia site? This site appears to be a data site in the vein of Wikipedia with no quality checks or peer-review. Can the authors find a more reputable sour for their data?

Reply from authors

This is very easy to correct and change to a more academic reference and will be done in the revised version. All values were now taken from www.tradingeconomics.com in the revised version and found not to change anything on the result. Note that the Dow Jones index is available from the same source but also from the following https://research.stlouisfed.org/fred2/series/DJA/downloaddata?cid=32255/ and was also found to give identical values.

Reviewer no. 1

Page 13995, lin23 – Page 13996, line 8: Other studies have also been conducted that relate various meteorological parameters to air quality. How do your findings compare to theirs?

***Camalier, L., Cox, W., Dolwick, P., The effects of meteorology on ozone in urban*

areas and their use in assessing ozone trends, Atmospheric Environment, 41, 7127-7137, 2007.
***Zheng, J., Swall, J.L., Cox, W.M., Davis, J.M., Interannual variation in meteorologically adjusted ozone levels in the eastern United States: A comparison of two approaches, Atmospheric Environment, 41, 705-716, 2007.*
***Davis, J., Cox, W., Reff, A., Dolwick, P., A comparison of CMAQ-based and observation-based statistical models relating ozone to meteorological parameters, Atmospheric Environment, 45, 3481-3487, 2011.*

Reply from authors

The literature about statistical methods for the meteorological adjustment of surface ozone is abundant and not only limited to the references given above by reviewer no. 1. For example, Thompson et al. (2001) made a review of these kinds of statistical methods. The conclusion of the authors is that no one is most appropriate for all purposes and all meteorological scenarios. Moreover, according to these authors, a comprehensive and reliable methodology for space-time analysis seems to be lacking despite recent work including the above references mentioned by reviewer no.1. Correcting ozone for meteorological fluctuations has to be done on the right spatio-temporal scale. If such corrections occur on time scales much shorter or larger than required, serious adjustment problem for ozone would result. For example, hourly corrections are most relevant for short-term predictions and photochemical model evaluation (Thompson et al, 2001). Also applied linear regression statistics may overlook nonlinear complexity of ozone with their precursors (NOx and VOC). In any case, finding and applying the most appropriate correction techniques is beyond the scope of this paper and would require sensitivity tests from model. Instead, our goal was to provide trends based on a simple methodology without complicated adjustments which we believe are incomplete in any case.

What reviewer no. 1 is suggesting would require tremendous amount of resources because not only the influence of weather has to be removed and in any case will never be removed completely since the atmospheric variability is present at all scales and is different for different sites. For example, Camalier et al (2007) seems to apply only for urban regions. Moreover, all 3 papers mentioned by reviewer no. 1 do not take into account short term influence of the economy nor forest fires inter-annual variability either.

Remember that we are analyzing inter-annual trend of summer ozone. The merit of figure 11 in our paper is that similar years difference are used taking into account various factors including economical fluctuations and wildfires (see Figure B.1 where a principal component analysis suggests similar year for summer 2005 and summer 2012, see more details in the revised paper).

Not taking into account the impact of economic fluctuations could bring to incomplete conclusions (Castellanos and Boersma, 2012; Granados et al, 2012). For example, in EPA(2012) report, on page 8, it is mentioned: “Some areas in the eastern U.S. experienced more unhealthy days in 2010 compared to 2009, mostly due to weather conditions being more conducive to ozone formation in these areas in 2010”. We agree that meteorological conditions influence surface ozone. However, from 2009 to 2010, U.S. and Canada were getting out of the recession 2007-2009 and the impact is that the gross domestic product growth rate jumped from strong negative values both in US and Canada (quarterly minimum of about -3.8 and minus -3.7 respectively) in 2009 to moderate positive values (+2 and +3.8 respectively) in 2010 which is a quick and very significant economic re-adjustement. It is known that anthropogenic emitted chemical species concentrations are linked with GDP changes (see reference given above) which agrees also with the finding of tables 7 and 8 of our paper. On the other hand, trying to correct trends for all various factors could become extremely complex. Therefore, for simplicity we adopted the method of differences based on similar year as an alternative for mapping geographically multi-year trends.

Reviewer no. 1

Page 13996 and 13997: All of the abbreviations make this section hard to read. I suggest eliminating the abbreviations in this section.

Reply from authors

By removing table 7 as mentioned above, this will solve most of the problem of abbreviations.

Reviewer no. 1

Tables 5a and 5b: Please describe the averaging time (all hours, daily max etc.) and units (ppb or %) in the table caption. Is this for summertime only? If so, state this as well.

Reply from authors

We will change that with pleasure. It is a good idea to repeat in the table legend and this will be done in the revised version. Thank you to point that out to us.

Reviewer no. 1

Table 6: What statistical test was used to determine the p-value? Why is $P < 0.25$ considered significant? Standard practice in the research community is to consider $P < 0.05$ (or 0.01) significant. Please use the 0.05 cutoff or provide a compelling reason for this break from standard practice.

Reply from authors

The p-value given in our paper corresponds to a standard F-statistic (Fisher tests) which is computed by taking the ratio of the mean square regression to the mean square residual. On the other hand, one should not confuse p-value with the concept of cutoff values (α) for making decision to reject a particular data or not. The latter is rarely different from 0.05 or 0.01 in practice as noted by reviewer no. 1. However, reporting results with their level of significance p-value is something else. And in fact, reporting the p-value (whether or not it is bigger than 0.05) provides much more information than just claiming significance or not based on the 5% level (Freeman, 1993). Also, showing or reporting results with larger p-values than 0.05 is a matter of choice and is somewhat arbitrary although the most popular values for p-values are 0.1, 0.05, 0.01, 0.001 and not only and strictly 0.05 or 0.01. The over-abuse of the 5% level (p-value <0.05) is discussed in Stigler (1988). In our study, we adopted the following: $p < 0.01$ extremely significant, $p < 0.05$ highly significant, $0.05 < p < 0.15$ weakly to moderately significant and $0.15 < p < 0.25$ only marginally significant. Now this choice was done in view of the limited sample size (N=11 or N=9) in Table 7. Note that there is also a dependence of the p-value on the size of the sample. A p-value from a small sample could provide a stronger evidence than the same p-value for a large sample (Freeman, 1993). In fact, for large samples ($N > 200$), $p < 0.001$ is rather desirable to achieve high significance more than the 0.05 criteria (Freeman ,1993). On the opposite, for small size sample, reporting higher p-values has the merit of giving information as well about the confidence to be expected.

Note also that often in statistical tables for t-Student's test or Fisher's test, p-value up to 0.25 are often given in many text books of statistics to be able to analyze small population sample (for example, tables C.3 in Neter et al., 1988). That was the base of choosing the value of 0.25 in our paper. But associating levels of confidence to p-values is to a certain extent a matter of choice. For example, in the IMPROVE report mentioned by reviewer no. 1, insignificant trends are defined when $p > 0.15$ http://www.epa.gov/airtrends/2010/report/fullreport.pdf . Also the paper of Cooper et al (2012), (figure 4) reports p-values as high as 0.16 for some data. At that level of significance, we can say that this association contains information but not statistically significant at the 0.05 level but at the 0.16 level. There is no such good and bad results (p < or > 0.05) and absolute threshold in the field of statistics.

Reviewer no. 1

It would also be useful if this table were broken out by region like Tables 5a and 5b.

Reply from authors

This will be done with pleasure in the revised version.

Reviewer no. 1

Figures 4-14: Increase the font size. The scales on these tables are completely unreadable using the current font size. Please print out a copy of the article pdf before it is resubmitted to make sure that numbers/scales/axes on your figures are readable.

Reply from authors

The original size of the figures, numbers/scales and axes were all readable in the version and a copy was originally printed before submitted to the editor. However, in the editing process, figures got reduced and sometimes surprises appear which are out of the control of the authors. However, in the revised version, we will take into account the final size of the figures in ACP journal to re-dimension the legend so that it is readable.

Reviewer no. 1

Figures 11/12: Make the areas labeled "Unreliable zone" white or gray. The red is distracting.

Reply from authors

This will be changed to another color as suggested in the revised version.

Reviewer no. 1

Figure 12: Are these annual or summertime values?

Reply from authors

It is clear from the figure legend what it is (summer, e.g. June July August).

Reviewer no. 1

Figure 13: How were trend lines created? Please include a brief explanation in the text of what methods were used to create the decadal trend. Add y-axis labels to the right edges of these plots. This makes it easier to see the magnitude of the increases and decreases in the trend lines.

Reply from authors

Trends were calculated using statistical package called SAS (Statistical Analysis Software). Details will be added up in the revised version and changes requested done accordingly.

Reviewer no. 1

Figure 14: Consider using black and white (or gray) for the unmasked land/water to eliminate colors which are similar to those included in the color bar.

Reply from authors

We believe that there is no conflict between land/water colors and the color bar. We do not provide analysis over oceans so this is somewhat irrelevant. Changing that will produce unnecessary extra work for the authors and no added value for readers.

Additional references (for other references mentioned above but not listed here. see the original submitted paper):

Ambrose, J. L., Reidmiller, D. R., Jaffe, D. A.: Causes of high O3 sin the lower free troposphere over the Pacific Northwest as observed at the Mt. Bachelor Observatory, Atmos. Environment, 45, 5302-5315.

Seinfeld, J. H., and Pandis, S. N.: Atmospheric chemistry and physics. From air pollution to climate change, John-Wiley & Sons, Eds., USA, 1998.

Cooper, O. R., Gao, R. S., Tarasick, D., Leblanc, T., and Sweeney, C.: Long-term ozone trends at rural ozone monitoring sites across the United States, 1990-2010, J. Geophys. Res., 117 ,D22307, doi:10.1029/2012JD018261, 2012.

Cooper, O. R.: Global surface ozone trends,a synthesis of recently published findings. Presented at the NOAA-GMD Global Monitoring Annual Conference, May 21-22, 2013, Boulder, Colorado. http://www.esrl.noaa.gov/gmd/annualconference/slides/33-130409-A.pdf

EPA: *Our Nation's Air: Status and trends through 2010, US EPA, Office of Air Quality Planning and Standards, Research Triangle Park, NC, February 2012, EPA-454/R-12-001: http://www.epa.gov/airtrends/2011/report/fullreport.pdf, 2012.*

EPA: Our Nation's Air: Status and trends through 2008, US EPA, Office of Air Quality Planning and Standards, Research Triangle Park, NC, February 2010, EPA-454/R-09-002: http://www.epa.gov/airtrends/2010/report/fullreport.pdf, 2010.

Granados, J. A. T., Ionides, E. L., Carpintero, O.: Climate change and the world economy : short-run determinants of atmospheric CO2, Env. Sci. & Policy, 21, 50-62, 2012 http://dx.doi.org/10.1016/j.envsci.2012.03.008

Oltmans, S. J., Lefohn, A. S., Shadwick, D., Harris, J. M., Scheel, H. E., Galbally, I., Tarasick, D. W., Johnson, B. J., Bunke, E.-G., Claude, H., Zeng, H., Nichol, S., Schmidlin, F., Davies, J., Cuevas, E., Redondas, A., Naoe, H., Nakano, T. and Kawasato, T.: Recent tropospheric ozone changes – A pattern dominated by slow or no growth, Atmospheric Environment, 67 (2013) 331-351, http://dx.doi.org/10.1016/j.atmosenv.2012.10.057.

Freeman, P. F.: The role of p-values in analyzing trial results, Statistics in medicine, vol. 12, 1443-1452, 1993.

Friedlingstein, P., Houghton, R. A., Marland, G., Hackler, J., Boden, T. A., Conway, T. J., Canadell, J.G., Raupach, M. R., Ciais, P., Le Quere, C.: Update on CO2 emissions, Nat. Geosci. 3, 811-812, 2010.

Frydendall, J., Brandt, J., and Christensen, J.H.: Implementation and testing of a simple data assimilation algorithm in the regional air pollution forecast model, DEOM, Atmos. Chem. Phys., 9, 5475-5488, 2009.

Ingleby, N.B., and Lorenc, A.C.: Bayesian quality control using multivariate normal distributions, Quart. J.R. Meteor. Soc, 119, 1195-1225, 1993.

Lelieveld, J., van Aardenne, J., Fischer, H., de Reus, M., Williams, J., and Winkler, P.: Increasing ozone over the Atlantic Ocean, Science, vol 304, 1483, 2004, doi:10.1126/science.1096777.

Lin, M., Fiore, A. M., Cooper, O. R., Horowitz, L. W., Langford, A. O., Levy II, H., Johnson, B. J., Naik, V., Oltmans, S. J., and Senff, J. C.: Springtime high surface ozone events over the western United States: Quantifying the role of stratospheric intrusions, J. Geophys. Res., 117, D00V22, doi:10.1029/2012JD018151, 2012a.

Lin, M., Fiore, A.M., Horowitz, L. W., Cooper, O. R., Naik, V., Holloway, J., Johnson, B. J., Middlebrook, A. M., Oltmans, S. J., Pollack, I. B., Ryerson, T. B., Warner, J. X., Wiedinmyer, C., Wilson, J., and Wyman B.: Transport of Asian ozone pollution into surface air over the western United States in spring, J. Geophys. Res. ,117, D00V07, doi:10.1029/2011JD016961, 2012b.

National Research Council, 1991. Rethinking the ozone problem in urban and regional air pollution. Committee on tropospheric ozone formation and measurement, Natl. Acad. Press, Washington. D.C.

Neter, J., Wasserman, W., and Whitmore, G.A.: Applied Statistics. Allyn and Bacon Inc., 1988, U.S.A.

Institute for Risk Research: Air pollution and public health: A guidance document for risk managers. May 2007, University of Waterloo, ISBN 978-0-9684982-5-5.

Saavedra, S., Rodriguez, A., Souto, J. A., Casares, J. J., Bermúdez, and Soto, B. : Trends of rural tropospheric ozone at the northwest of the Iberian Peninsula, The Scientific Word Journal, article ID 603034, doi:10.1100/2012/603034, 2012.

Stigler, S. Fisher and the 5% level. Chance. vol 21, no. 4 2008.
http://link.springer.com/article/10.1007%2Fs00144-008-0033-3

U.S. Energy Information Administration: Annual energy outlook 2013, Rep. DOE/EIA-0383(2013), Dept. of Energy, Washington, D.C., available at www.eia.gov/forecasts/aeo/

Thompson, M. L., Reynolds, J., Cox, L. H., Guttorp, P., and Sampson, P. D.: A review of stistical methods for the meteorological adjustment of tropospheric ozone, Atmos. Env., 35 (2001), 617-630, 2001.

Weisberg, S.: Applied Linear regression. John Wiley & Sons., 1985. U.S.A.

ANNEX B (will be added in the revised version)

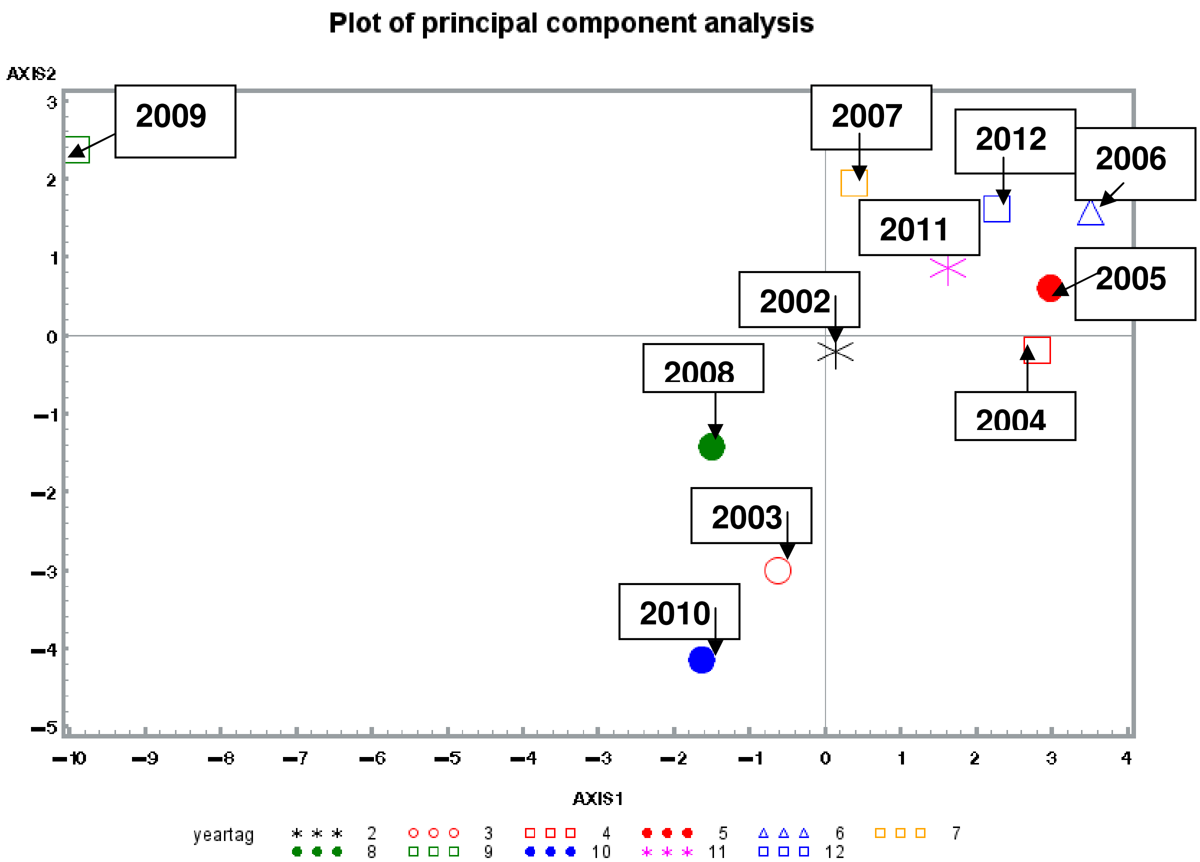

FIGURE B1. Principal component analysis based on summer statistics for each year in both U.S. and Canada (based on temperature anomalies, gross domestic product growth rate, total area burned by wildfires, NAO and ENSO indices). Note that the strongest loading factor of the first principal component analysis was found to be the U.S. gross domestic product growth rate.