Response to Reviewers for manuscript 'Maximizing Ozone Signals Among Chemical, Meteorological, and Climatological Variability' (https://www.atmos-chem-phys-discuss.net/acp-2017-954/)

We would like to thank the reviewers for their valuable comments.

Below we work through each of the reviewers' comments, with the comments in black and our responses in red. We also include any alterations to the text in red after our responses with the specific additions indicated with underlines. Line references refer to the tracked changes document.

Anonymous Referee #1 Received and published: 13 December 2017

General comments:

This paper discusses the use of different temporal and spatial averaging scales to detect trends in surface ozone over the United States. This is an interesting topic that is useful to the community, and the approach is novel. However, I have two general concerns that I would like to see addressed:

1. The relevance of the particular methods discussed for detection of air quality trends should be better clarified or caveated, since the averaging time-scales suggested (10- 15 years) are comparable to the trends we seek to detect, and temporal and spatial averaging can blur localized signals of high ozone that are relevant to public health.

This is a very valid point, and one that we underemphasized in the original manuscript. We have updated portions of this manuscript to focus more on 'signals' rather than 'trends' (as we include trends as one type of signals). We add language to the discussion (Line 521-524 and 591-594) and conclusions (Lines 611-651, 615-616, 633-644, 652-656) that highlights the difficulty in balancing data availability, observation/simulation length, averaging times, and error thresholds. See our responses to the specific comments below for details on these additions.

2. Given the heavy dependence of the analysis on model simulations, I would like to see more rigorous evaluation of the model's ability to accurately predict the spatial and temporal variability of surface ozone and its response to changes in meteorology and emissions.

The CESM1.2 CAM-chem model has been extensively evaluated in previous papers mentioned in the methods section. We have added more explicit references to this evaluation throughout the manuscript (Lines 201-202 and 221-222). We have added additional evaluation of the model capabilities compared to the available observations with regard to meteorological variability (updated Figure 2 and reference to Brown-Steiner et al, in review, see following paragraph). We do not examine the impact of emissions variability in this manuscript, as this is beyond the scope of the current work, but we add additional emphasis in the conclusions that emissions variability studies are needed in future research:

Lines 618-620: *"Taking into account the complex interactions involving trends and variability between emissions, chemistry, meteorology, and climatology necessitates a variety of strategies."*

Lines 652 – 656: *"While we have detrended the CASTNET observations to compare to the constant year-2000 cycled emissions in the simulations, the CASTNET time series inherently includes the compounded variability of both meteorological and emission sources. Future studies will need to expand this analysis to include trends and variability in the emissions, as well as in the meteorology."*

In addition, these model runs (along with others) are more thoroughly compared to observations in a second paper which is now in discussion in GMD, and we have added a reference to this paper to this manuscript: Brown-Steiner, B., Selin, N. E., Prinn, R., Tilmes, S., Emmons, L., Lamarque, J.-F., and Cameron-Smith, P.: Evaluating Simplified Chemical Mechanisms within CESM Version 1.2 CAM-chem (CAM4): MOZART-4 vs. Reduced Hydrocarbon vs. Super-Fast Chemistry, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-16, in review, 2018.

In addition, a number of other statistical techniques have been applied to the problem of separating emission effects from other drivers of variability (for example, Camalier et al., Atmos. Environ., 2007, and references therein), with the potential advantage of detecting changes on shorter timescales. How do the results in this paper compare to other statistical methods? Perhaps this could be discussed in the discussion or conclusion sections.

We have added language in the conclusion that contrasts our methodologies to other methodologies, and encourages a multi-strategy approach:

Lines 615-622: "Our analysis and conceptual framework *presented here cannot solve this tension, but it does demonstrate some strategies which can* allow for a selection of spatial and temporal averaging scales*, and a consideration of the error threshold*, that can aid in this signal detection *on a case-by-case basis. Taking into account the complex interactions involving trends and variability between emissions, chemistry, meteorology, and climatology necessitates a variety of strategies. This work quantifies the impact of spatial and temporal averaging in signal detection, which can be used in conjunction with ensembles of simulations, statistical techniques, and other strategies to further out understanding of the chemical variability in our atmosphere.*"

We have also included Camalier et al., 2007 (Line 71) and other recommended citations from below (Line 69, Line 71) to the introduction in order to provide as much information to readers about the possible strategies and methodologies used for signal detection.

Specific comments:

Line 28: How is the "chemical variability" that is not related to "meteorological variability" different from an air quality signal?

Chemical variability can result from more than just meteorological variability (e.g. emissions variability, which we do not address in this paper) and also non-linear interactions between chemistry, emissions, meteorology, climatology, and surface processes. We have clarified this in the abstract:

Line 28-30: "However, the magnitude of a surface air quality signal is generally small compared to the magnitude of the underlying *chemical, meteorological, and climatological variabilities (and their interactions) that exist both in space and in time, and which include variability in emissions and surface processes.*"

Line 41: The authors state on line 31 that part of the motivation for this study is to identify the impact of emission reduction policies on e.g. ozone. Here, however, they suggest averaging over 10-15 years. This seems pretty long compared to the timescale of air quality changes and compared to the available data records, which for many CASTNET sites in only on the order of 20 years.

We recognize that our suggestion of averaging over 10-15 years is challenging, but it is consistent with recent literature (e.g. Barnes et al. 2015; Garcia-Menendez et al. 2017). We hope with this manuscript to demonstrate some of the difficulties that arise when trying to detect the impact of, for instance, emissions reduction policies on ozone. In particular, we hope to demonstrate that the variability in atmospheric chemistry needs to be quantified, examined, and addressed in a direct manner when identifying signals, and that the temporal and spatial context of the particular signal needs to be provided as supporting evidence that a particular signal is robust. We have added the following sentences to the conclusion section to emphasize this point.

Lines 633-641: *"We recognize that achieving a 10 – 15 year temporal averaging window is difficult, but this recommendation is consistent with recent literature (e.g. Barnes et al., 2015; Garcia-Menendez et al., 2017). For studies where 10 – 15 years of averaging is impractical, we recommend that some spatial and temporal context is provided that demonstrates that the signals being examined are robust and not the result of internal variability or noise."*

We also add language on Lines 43-46 and 611-615, in response to additional reviewer comments below.

Line 42: If you average over several hundred kilometers, do you risk missing policy or health-relevant ozone exceedences that occur at more local scales?

Absolutely! Again, we hope to demonstrate the challenges of identifying chemistry signals at small spatial scales. In particular, if you are examining signals and the smallest spatial scales, it is likely a longer temporal period will be required to 'escape' the variability at that scale. We address this in the Discussion Section, in particular Line 513 in asking "What is the magnitude of ozone variability due to meteorology alone at the smallest spatial scale?" To further clarify this in the manuscript, we added the following to the abstract:

Lines 43-46: *"If this level of averaging is not practical (e.g. the signal being examined is* *at a local scale), we recommend some exploration of the spatial and temporal variability to provide context and confidence in the robustness of the result."*

Line 66: For signal detection, see also Weatherhead et al., Physics & Chemistry of Earth, 2002; Strode and Pawson, JGR, 2013; Deser et al., Climate Dynamics, 2011

These citations have been added as further examples of the history and difficulties in signal detection into the introduction, as Camalier et al., 2007 (Lines 69 and 71).

Lines 96-110: While it is true that the 4th highest MDA8 criteria includes some averaging, it is also aimed at capturing the high end of the distribution rather than just the long-term mean. Isn't this lost by simply averaging over longer periods?

It is, and the 4th highest MDA8 metric has been designed for a practical legal purpose. It is also a standard metric used throughout the literature, so we felt that examining the impact of spatial and temporal averaging on this metric would be an appropriate way of adding to the literature, and the broader context of this particular metric. We add the following sentence to address this:

Lines 286-289: "*Some of the averaging strategies we present can average away the high ozone behavior this MDA8 O_3 metric is intended to quantify, but it is such a well-reported metric that focusing our analysis on it allows for ready comparisons to other studies."*

Line 148: Since you are interested in different spatial scales, why not include urban air quality sites as well as CASTNET?

Since the CASNTET observations are from more rural sources, they are generally accepted as more appropriate to compare to coarse-grid cell models such as CAM-Chem. We add additional references to other studies that used CASTNET observations in this way:

Line 240: *"(e.g. Brown-Steiner et al., 2015; Phalitnonkiat et al., 2016)"*

Future research should extend analysis like that presented here to models of different resolutions (and the associated observations). We have added this suggestion to the last paragraph of the Discussion Section:

Lines 591-594: *"Furthermore, future research examining the impact of spatial and temporal averaging using regional-scale models, models with different resolutions, and the inclusion of urban observations could provide additional insight into understanding chemical variability and averaging techniques."*

Line 205: Please highlight the key differences between this and the earlier model version.

The Tilmes et al. (2015) reference (and references therein) fully documents CESM1.2, although we had previously omitted the reference from this sentence. It has been added in Lines 191, 198, and 221.

Line 252: The assertion that the spatial variability is well-captured is not really evident in Figure 2. Maybe overplot the observations on top of the model map, or report the spatial correlation between the model and the observations.

We have updated Figure 2b to better compare model/observations. We also add a reference in the caption to Brown-Steiner et al. (in review, GMDD), which performs additional model-observation comparisons of CAM-Chem.

Section 3.1 and Fig. 2: It would be helpful to show the temporal variability of the observations along side that of the model

Since we only compare the year 2000 in this figure (this was not clear in the caption and this has been updated), there is not enough data for a full comparison of temporal variability, but Figure 2b has been updated and additional references to Brown-Steiner et al. (in review, GMDD) on Line 898, which extends model-observation comparisons for MOZART-4 (and other mechanisms).

Line 255: Clarify that it is the standard deviation in the model.

We have now clarified this as follows:

Line 301: "…The standard deviation *of the simulated MDA8 O_3*…"

Line 272: What is the correlation between the modeled and observed timeseries? Figure 3e suggests a lot of mismatches between the observations and model. What does this mean in terms of the uncertainty in your model-based findings?

Model correlations to observations depend on the region (with R^2 values ranging from 0.42 – 0.88, see the updated Figure 2b), and Figure 3e compares the average over the entire Eastern US. Comparisons of the seasonal correlations to observations are available in Brown-Steiner et al. (in review, GMDD) and are generally high for MOZART CAM-chem (0.8 – 0.9). References to this paper are added to this manuscript (Lines 198, 201-202, Figure 2 Line 898). Since we include cycled year 2000 emissions for out simulations, we do not expect a high correlation for the entire time series, even when compared to the detrended CASTNET observations since we do not simulate the real-world emissions variability, especially when comparing individual sites to model grid boxes. This additional uncertainty that comes from assuming cycled emissions has been noted in other comments, and additional language has been put in the Discussion and Conclusions to explore the implications (Lines 521-524 (see below), 611-615 (see above), 652-656 (see below)).

Section 3.2, first paragraph: Some of this could go in the methods section.

We moved up the more technical description to the newly added Section 2.4 (Line 257-274).

Lines 374-375: Can you explain why? Do these regions have higher variability?

Yes, this has been clarified and the reader is pointed to Figure 2d:

Lines 446-448: "*Shorter* windows (or smaller thresholds) are needed in the Western US *(where variability is smaller, see Figure 2d)* than in the Eastern *US (where variability is larger) as well* as over coastal and highly populated regions."

Line 430: The relationship between chemical and meteorological variability also depends on emission levels (e.g. Bloomer et al., GRL, 2009), and these are unlikely to remain constant over a decadal averaging window. Thus the real situation will be more complicated than the constant-emission model-based analysis shown here. The model-based analysis is still useful, but should be more carefully caveated.

We add additional text at the end of this paragraph to caveat the limits of our methodology and highlight the complexities that arise when considering trends and variability in emissions, meteorology, and climate:

Lins 521-524: "*A more comprehensive analysis of chemical variability will need to account for both meteorological and emission variability, which is complicated by temporal trends in both the emissions of ozone precursor species and the climate."*

Technical: Line 374: "Shorter" not "short"

This has been corrected Line 446).

Anonymous Referee #2 Received and published: 15 February 2018

General Comments ——————-

This manuscript describes an evaluation of the variability of surface ozone concentrations over the United States during summer. In particular, the authors analyze the effects of meteorological variability on ozone concentrations, and the dependence of this variability on temporal and spatial averaging scales. The goals is to use averaging to provide a more robust estimate of the uncertainty in the "true" ozone concentration, independent of the influence of meteorological "noise". The idea that spatial or temporal averaging can reduce meteorological variability is not a new one, but this paper presents a useful and innovative framework for analyzing the choice of time and space scales, depending on the uncertainty threshold required for a particular application. This writing in this paper could be improved significantly to clarify the methods used and the basis for the recommendations being made. I list below some such suggestions for ways the manuscript can be improved. With revisions, this paper would be appropriate for publication in ACP, and would be a helpful contribution to the literature on detecting robust signals in ozone over a noisy background.

Specific Comments ——————

Abstract

line 41 – This 10-15 year time period pertains to detecting a robust estimate of mean ozone concentrations. What are the implications for detecting trends (e.g., driven by emission changes) in ozone? For instance, large robust trends in ozone were detected in observations as a result of emission reductions following the NOX SIP Call. This manuscript claims to provide information on estimating trends in ozone, but does not really provide specific information on trend detection methodologies.

We explore some of the literature on ozone trends in the introduction (Cooper et al., 2012, Barnes et al., 2016, and others), and although we do not provide specific trend detection methodologies, we feel that we have demonstrated the potential risks of calculating trends based on an individual selection of years. You are correct in that we use the word 'trend' in many places where we really mean 'signal,' so we have changed the word 'trend' to 'signal' in several of these places throughout the manuscript to better reflect our intended message: the description of signals that we present in the introduction.

We have also added language (addressing other comments) that address the implications of the 10 – 15 year time period throughout the manuscript (Lines 46 and 633-644, addressed in previous comments, and Lines 611-615):

Lines 611-615: "*In particular, it would be impractical to delay interpreting observations for 10 – 15 years, or alternatively to expand the spatial averaging such that small-scale features are smoothed away. Nonetheless, it is unwise to over-interpret trends and signals based on observations from a limited spatial area and over a short temporal* *period.*”

lines 44-46 – For which other quantities might these results be applicable? What features of the spatiotemporal distribution dictate the choice of optimal spatial and temporal averaging periods.

Those are excellent questions and we intentionally left this open to the reader. Naturally, this analysis could apply to other chemical species, but also chemistry-meteorology interactions (e.g. ozone-temperature relationship), surface features (land use cover, plant functional type, surface roughness, albedo, cloud and boundary layer variables, etc). We add the following to the discussion section, indicating some quantities that this strategy may apply to:

Lins 580-584: *“*In particular, low-frequency oscillations (e.g. ENSO, and others) and other forms of internally or externally forced trends (e.g. anthropogenic and natural changes in emissions) are readily adaptable to this type of analysis, *which could address signals pertaining to precipitation, biogenic emissions, boundary layer variables, cloud properties, and many others.”*

1. Introduction

lines 93-95 – Mention also internal (unforced) variability.

Added:

Line 99-101: “This approach cannot address structural uncertainties *and internal (unforced) variability* between models, but is capable of identifying parametric uncertainties within a single model.”

lines 91-97 – There is not a clean distinction between running ensembles of model runs with different initial conditions versus "expand[ing] the temporal averaging window". In the case of "climatological" runs such as those done here with CAM-Chem, running more years in a single simulation is nearly identical in practice to running more years of a single simulation.

We agree. There are many modeling choices (ensembles with different initial conditions and internally simulated meteorology, ensembles with internally simulated meteorology and different emissions (either transient or cycling a single year), ensembles with forced meteorology and different emissions, ensembles with different sets of online/offline forcing datasets (oceans, ice, land, etc.). What we have done in this paper is one strategy, and we hope that future studies will select other strategies. We have added the following sentence to the conclusion to indicate that what we present is one strategy among many:

Lines 641-644: *“We also recognize that our analysis is just one strategy for enhancing signal detection capabilities, and will ideally be used alongside others, such as perturbed initial condition ensembles, running simulations with either internal or forced meteorology, and examining a region or time period with different models or parameterizations.”*

lines 123-125 – You mention here that the objective is to "limit the likelihood of over-confidence in an estimate of surface ozone". Presumably, the goal is more than that. Rather than just providing an improved (large) estimate of local variability, the averaging method suggested here also aims to reduce the underlying uncertainty due to meteorological variability.

Yes, this has been added:

Lines 130-132: "Our objective in this study is to provide a framework for selecting spatial and temporal averaging scales *that reduces the uncertainty in analyzing ozone signals and* limits the likelihood of over-confidence in an estimate of surface ozone that arises from meteorological variability."

lines 154-155 – Model resolution is not addressed in this study. How would varying model resolution compare with the other "parametric" changes in the model discussed here?

That is an excellent question that was outside of the scope of this paper, but we have added this as a path for future research at the end of the Discussion Section:

Line 585: "*Furthermore, future research examining the impact of spatial and temporal averaging using regional-scale models, models with different resolutions, and the inclusion of urban observations could provide additional insight into understanding chemical variability and averaging techniques.*"

2.1 CAM-Chem

In this section and throughout the paper, the model name "MOZART" seems to be used interchangably with "CAM-chem", including in the names of the simulations. This is confusing, since MOZART and CAM-chem, although closely related, are distinct models. Please clarify throughout the paper.

Throughout the manuscript, we have updated the descriptions. We leave in the name MOZART when we are specifically referencing the chemical mechanism and CAM-chem when we are more generally talking about the simulation. This has been made explicit in the methods section:

Lines 196-198: "We conduct our simulations using the MOZART-4 chemical mechanism (Emmons et al., 2010), *which is a full tropospheric chemical mechanism integrated into CAM-Chem (e.g. Brown-Steiner et al., in review).*"

line 200 – Here and elsewhere throughout the paper, clarify that you are only considering the effect of future *climate*, not actually fully simulating future conditions (e.g., future emissions).

We have clarified this on Line 215 ("…*We also include two reference simulations of the future climate,* …") and throughout the manuscript.

2.3 Telescoping Regional Definitions lines 230-232 – This sentence is repetitive of Intro.

This sentence has been removed.

3.1 Spatial and Temporal Comparisons

line 248 – Throughout the paper, the notation "DM8H" is used for the daily maximum 8-hour ozone concentration. Elsewhere in the literature, this seems to be referred to as "MDA8".

DM8H has been changed to MDA8 throughout the manuscript.

line 248 – "MOZART" –> "CAM-chem"

We have corrected this here and throughout the manuscript.

lines 255-259, Figure 2 – Show standard deviation and/or variability from the observations as well. If the standard deviation were similar between the model and observations, would the model ozone bias cause the (relative) variability to differ significantly?

Figure 2b has been updated with a direct comparison between the model and the observations. The standard deviation comparison between the model and the observations again depends on the region. Table 1 summarizes both standard deviation and the variability (standard deviation / mean) to demonstrate the impact of the different magnitudes of ozone that result from model bias on both the absolute standard deviation (ppbv) and the relative standard deviation as represented by variability (%). We have also added a clarification:

Lines 305-307: "*We include this relative standard deviation metric since the CAM-chem biases make it difficult to compare standard deviations directly.*"

line 283 – Add "(Figure 2, Table 1)" after "Here".

This has been added, Lines 300-301.

lines 283-285 – This sentence is repetitive of the first paragraph in this section.

We have removed this sentence (and the insertion from the previous comment has been moved to the first paragraph of this section).

line 289 – Add "from continental to a single NE U.S. grid box" after "telescoping regions".

This has been added, Line 342.

line 290 – Add "albeit with lower overall variability" after "captures this trend".

This has been added, Line 343.

3.2 Variability, Averaging Windows, and Thresholds

line 314 – Add "underlying variability at the" before "particular choice of spatial and temporal scale".

This has been added, Line 358.

line 328 – Does "variability" here refer to standard deviation (as suggested by the ppbv thresholds) or as previously used, the relative variability (s.d./mean)? Confusing. Make sure to define the quantities being discussed.

We do not mean the previously defined definition of variability, so we have clarified this on Line 395, where we replaced "variability" with "anomaly for any selection of averaging window".

line 329 – Clarify what is meant here by "This difference".

This has been clarified on Line 395, replacing the word "difference" with "potential error."

3.3 Selection of Temporal Averaging Scales

line 358-359 – Add "meteorological variability causing ozone anomalies" before "exceeding particular thresholds", if this is the intended meaning.

This interpretation is the intended meaning, so "meteorological variability causing ozone anomalies" has been added to line 430.

line 363 – "Increas[ing] the threshold" is not really a strategy for "filtering out the noise". It is more like accepting the higher level of noise.

This has been clarified:

Line 435: "…either average over longer periods, or *acknowledge the level of noise and* increase the threshold."

lines 367 -370 – Confusing as written. Separate out the mention of Fig.S3 to a second sentence, e.g., "Similarly, in Supplemental Figure S3, one column (the 5-year averaging window) is selected."

We agree that these sentences were confusing as written. They have been updated and clarified:

Lines 439-442: "Supplemental Figure S3 *extends the analysis of Figure 5 by comparing* the MOZ_2000, MOZ_2050, and MOZ_2100 simulations *across the four thresholds for* the 5-year averaging window. *Figure 6 similarly compares* the 1 ppbv ozone *threshold* *across the four averaging windows for* MOZ_2000, MOZ_2050, and MOZ_2100."

line 369 – "Figure 6" –> "Figure 5"

We have clarified this section, Lines 439-442.

line 369 – Add "compare with" before "equivalent plots".

We have clarified this section, Lines 439-442.

line 370 – "Figures 7" –> "Figures 6".

We have clarified this section, Lines 439-442.

4. Discussion

line 434 – Add "variability" after "surface ozone".

We have added this text, Line 525.

line 460 – Cut comment in parentheses about future simulations. It is not known whether the future simulations will/would exhibit biases.

We agree with the reviewer, and have deleted this text.

5. Conclusions

line 502 – Add "and" after "configurations".

We have added this text, Line 603.

line 506 – Add "summertime" before "surface ozone". Clarify throughout conclusions that the analysis presented here is restricted to summer.

We have added the phrase "summertime" before references to ozone throughout the conclusion section (Lines 598, 607, 623, 628, and 645).

line 513 – Add "summertime" before "ozone variability".

We have added this text, Line 623.

line 523 – As mentioned earlier, the discussion of trend detection in the manuscript is very weak. Much more could (and should) be said about the application of the averaging methods presented here for trend detection. For instance, what are the implications of needing 10-15 year averaging windows for the length of timeseries needed to detect ozone trends (e.g., forced by climate change or emissions changes)?

In addition to additional examination of the implications of the 10 – 15 year averaging window ((Lines 43-46, Lines 611-615), we add the following text:

Lines 652-656: "*While we have detrended the CASTNET observations to compare to the constant year-2000 cycled emissions in the simulations, the CASTNET time series inherently includes the compounded variability of both meteorological and emission sources. Future studies will need to expand this analysis to include trends and variability in the emissions, as well as in the meteorology.*"

lines 524-530 – Mention here the compounding of (meteorological) variability in the observations with changes caused by variability/trends in emissions.

We address this along with the previous comment (Lines 633-644).

Figure 2 – Add the standard deviations plotted here standard deviations of daily ozone concentrations? If so, then for comparison with Figure 5, it would be useful also to show the interannual standard deviation of seasonal mean ozone.

These are for MDA8 O_3 mixing ratios, and is clarified in the caption (Line 899). Because the value of standard deviation would be different for every time and spatial scale, we don not think that it is practical to include interannual standard deviations here. We focus much of this manuscript on the variability and thresholds at the smallest spatial scales, which is represented in Figure 2 and Table 1.

Figure 3 – Explain that the CAM-chem simulation has fixed year-2000 emissions and SST, but time-varying meteorology. Why are the CASTNET values for 2000 "de-trended", instead of showing raw 2000 values? Change "MOZART" to "CAM-chem". In legend text in panel (a), also change "MOZART" to "CAM-chem".

Explanation added, terms updated. The detrending is centered at the year 2000, so the raw and detrended values are the same. This has been clarified in the caption, Lines 922-924.

Figure 4 – Define what is meant here by "variability". Is it the standard deviation, or the relative variability (s.d./mean)? Mention in caption that this plot shows summer ozone only. This is confusing from how the vertical axis is plotted.

It has been clarified that this is a plot of summertime MDA8 O_3 anomaly, Line 940.

Figure 8 – Change panel titles to the names of the regions. Keep the description of the regimes for filtering effectiveness in the text instead.

The panel titles have been updated in Figure 8 and the descriptions of the regions have been moved to the Caption of Figure 8 (Lines 980-982).

Maximizing Ozone Signals Among Chemical, Meteorological, and Climatological Variability

Benjamin Brown-Steiner[1,2,3], Noelle E. Selin[2,4,5], Ronald G. Prinn[1,2,5], Erwan Monier[1,2], Simone Tilmes[6], Louisa Emmons[6], Fernando Garcia-Menendez[7]

Correspondence to: Benjamin Brown-Steiner (bbrownst@aer.com)

1. Center for Global Change Science, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139
2. Joint Program on the Science and Policy of Global Change, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139
3. Now at Atmospheric and Environmental Research, 131 Hartwell Avenue, Lexington, Massachusetts, 02421
4. Institute for Data, Systems, and Society, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139
5. Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139
6. Atmospheric Chemistry Observations and Modeling Lab, National Center for Atmospheric Research, 3450 Mitchell Lane, Boulder, CO 80301
7. Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC 27695

Abstract

The detection of meteorological, chemical, or other signals in modeled or observed air quality data – such as an estimate of a temporal trend in surface ozone data, or an estimate of the mean ozone of a particular region during a particular season – is a critical component of modern atmospheric chemistry. However, the magnitude of a surface air quality signal is generally small compared to the magnitude of the underlying chemical, meteorological, and climatological variabilities (and their interactions) that exist both in space and in time, and which include variability in emissions and surface processes.~~,~~ This can present difficulties for both policy-makers and researchers as they attempt to identify the influence or 'signal' of climate trends (e.g. any pauses in warming trends), the impact of enacted emission reductions policies (e.g. United States NO_x State Implementation Plans), or an estimate of the mean state of highly variable data (e.g. summertime ozone over the Northeastern United States). Here we examine the scale-dependence of the variability of simulated and observed surface ozone data within the United States and the likelihood that a particular choice of temporal or spatial averaging scales produce a misleading estimate of a particular ozone signal. Our main objective is to develop strategies that reduce the likelihood of overconfidence in simulated ozone estimates. We find that while increasing the extent of both temporal and spatial averaging can enhance signal detection capabilities by reducing the 'noise' from variability, a strategic combination of particular temporal and spatial averaging scales can maximize signal detection capabilities over much of the Continental US. We recommend temporal averaging of at least 10 - 15 years combined with regional spatial averaging over several hundred kilometer spatial scales. If this level of averaging is not practical (e.g. the signal being examined is at a local scale), we recommend some exploration of the spatial and temporal variability to provide context and confidence in the robustness of the result. These results are consistent between simulated and observed data, and within a single model with different sets of parameters. The strategies selected in this study are not limited to surface ozone data, and could potentially maximize signal detection capabilities within a broad array of climate and chemical observations or model output.

1 Introduction

The capability to detect air quality signals – be they meteorological, chemical, or of some other type – is a fundamental component of modern climate science and atmospheric chemistry. The debate over the existence or length of a global warming hiatus (Lewandowski et al., 2015; Roberts et al., 2015; Medhaug et al., 2017) and research examining the time of emergence of climatological (Weatherhead et al., 2002; Deser et al., 2012; Hawkins and Sutton, 2012; Elía et al., 2013; Schurer et al., 2013), meteorological (Giorgi and Bi, 2009; King et al., 2015), chemical (Camalier et al., 2007; Strode and Dawson, 2013; Barnes et al., 2016; Garcia-Menendez et al., 2017), and other sectoral signals (e.g. Monier et al., 2016) embody an accumulation of techniques and strategies for filtering noise (due to natural variability) and maximizing the capability to detect statistically significant signals and trends in noisy data. It is well established that temporal averaging (e.g. Lewandowski et al., 2015) and spatial averaging (e.g. Frost et al., 2006; Hawkins and Sutton, 2012; Barnes et al., 2016) can enhance signal detection capabilities in atmospheric data. Here we extend this research by quantifying the impact of both spatial and temporal averaging – individually and in combination – of surface ozone on the magnitude of the calculated variability, which is largely driven by the influence of meteorological variability on ~~the~~ atmospheric chemistry (e.g. Jacob and Winner, 2009). We offer recommendations for strategically averaging in space and time to maximize signal detection capabilities. In particular, we examine estimates of mean ozone and of the ozone variability that results from meteorology, although our approach can be generalized to other air quality applications.

For observed ozone data, strategies for reducing spatial and temporal noise are limited: a longer time series is needed, more observations need to be made, or the spatial region over which the ozone observations are being averaged ~~over~~ needs to be enlarged. For surface ozone estimates using models, however, there exist a variety of strategies for reducing the noise (due to chemical and meteorological variability) relative to the strength of the signal, although they cluster into three main types. The first strategy is to average or combine multiple runs of structurally different models under the assumption that errors, biases, and uncertainties within the individual models are reduced and the multi-model or multi-dataset mean is a best estimate of the actual, aggregated ozone field. This is most notably done with multi-model ensembles within the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP) framework (Lamarque et al., 2013; Young et al., 2013; Stevenson et al., 2013), and this approach tends to assume that all members in the ensemble are independent and equally skillful. This assumption, however, may result in a loss of some valuable information (Knutti, 2010). Another form of this strategy is to run multiple model runs within a single model, but under different initial conditions or sets of parametric assumptions (e.g. Deser et al, ~~2010~~2012; Monier et al., 2013, 2015; Kay et al., 2015; Garcia-Menendez et al., 2015, 2017). This approach cannot address structural uncertainties and internal (unforced) variability between models, but is capable of identifying parametric uncertainties within a single model.

The second strategy to reduce ozone variability is to expand the temporal averaging window, which can influence the interpretation of the determined ozone value (e.g. Brown-Steiner et al., 2015). The Environmental Protection Agency (EPA) National Ambient Air Quality Standard (NAAQS) for ozone (US EPA, 2015) explicitly takes this into account, both in the length of the averaging period (daily maximum 8-hour average) and the selection criteria for the standard (fourth-highest over the previous 3 years). The calculated ozone variability can be further reduced by utilizing even longer averaging periods, such as monthly (e.g. Rasmussen et al., 2012), seasonal (e.g. Fiore et al., 2014; Barnes et al., 2016), annual, or decadal mean values (e.g. Garcia-Menendez et al., 2017). This strategy is analogous to the averaging of meteorological data to derive a climate signal, and just as Lewandowsky et al. (2015) recommend averaging 17 or more years in order to achieve climatological estimates of temperature trends, there is a growing body of literature recommending averaging short time scale chemical variability (what could be called chemical weather, see Lawrence, 2005) for 15 or more years (e.g. Garcia-Menendez et al, 2017) in order to achieve an estimate of the what could be called the chemical climate (see Möller, 2010).

The third strategy to reduce ozone variability is to average surface ozone values over larger spatial regions, and while there is a significant body of literature discussing the capability and interpretation of coarse resolution model representations of the sub-grid scale heterogeneity (Pyle and Zavody, 1990; Searle et al., 1998, Wild et al., 2006), there are few that strategically expand the spatial scale over which averaging is applied in order to maximize signal detection capabilities. This strategy has been applied in other fields of the atmospheric sciences as well as for general gridded datasets (e.g. Pogson and Smith, 2015), and spatial averaging has been suggested as a means of reducing temperature variability and smoothing biases at the smallest spatial scales within a single model run (Räisänen and Ylhäsi, 2011). This “scale problem” has also been noted as an important consideration when analyzing aerosol indirect effects (McComiskey and Feingold, 2012) and for the detection and attribution of extreme weather events (Angélil et al., 2017).

Our objective in this study is to provide a framework for selecting spatial and temporal averaging scales that reduces the uncertainty in analyzing ozone signals and limits the likelihood of over-confidence in an estimate of surface ozone that arises from meteorological variability. This type of framework can be useful from two different research perspectives. The first research perspective has a priori an ozone estimate (either observed or modeled) at a certain spatial and temporal scale (e.g. a 3-year simulation of surface ozone over the Northeastern US) and wants to quantify the likelihood that this estimate is representative of the long-term ozone behavior (rather than overly sensitive to meteorological variability of that particular 3-year period). Since ozone is strongly influenced by natural fluctuations in meteorology (Jacob and Winner, 2009; Jhun et al., 2015) and since extremes in surface ozone and temperature tend to co-occur (Schnell and Prather, 2017), atypically hot or cold periods can strongly influence ozone behavior over short time scales.

The second research perspective is to identify an ozone signal of a certain magnitude (or threshold) and ~~needs to~~ decide what spatial and temporal averaging scales are needed to best identify that signal. The ozone signal could be large (e.g. determining the effectiveness or compliance with a 5 ppbv incremental reduction of the EPA NAAQS for ozone (US EPA, 2015)) or small (e.g. identifying annual ozone trends within the US, which Cooper et al. (2012) show can be on the order of 0.10 – 0.45 ppbv), and can be highly sensitivity to spatial and temporal heterogeneity and meteorological variability. Barnes et al. (2016) found that surface ozone trends over 20-year periods can vary by ± 2 ppbv due solely to climate variability, while interannual variability can be on the order of ± 15 ppbv (Fiore et al., 2003; Tilmes et al., 2012; Lin et al., 2014) and day-to-day variability can be even larger, extending regularly from near-background levels of 40 – 50 ppbv up to 100 ppbv during the summertime (Fiore et al., 2014).

In this study, we quantify the impact of both temporal and spatial averaging on the calculated ozone variability – due solely to meteorological variability – in order to maximize the capability to detect ~~trends~~signals. We use simulated ozone (with the Community Atmosphere Model with Chemistry, CAM-chem) and observational data (with the EPA's Clean Air Status and Trends Network, CASTNET) within the United States in order to answer the following four questions:

(1) Within a given dataset (model or observations), with both spatial and temporal coverage, what is the magnitude of the ozone variability due to meteorology at the smallest scale, and how does spatial and temporal averaging reduce this variability? (2) Are there combinations of temporal and spatial averaging scales that maximize the signal detection capability for surface ozone data? (3) How sensitive are the above strategies to different configurations (i.e. emissions, meteorology, and climate) of the CAM-chem modeling framework? And (4) How could they be applied to other datasets (chemical, meteorological, or climatological)? We limit our focus to spatial scales within the United States as it has high spatial and temporal variability and numerous observations, and since averaging over larger regions (e.g. the Northern Hemisphere, or the globe) would produce a smaller calculated variability.

In Section 2, we describe the CAM-chem model and our simulations, as well as the CASTNET observational database and the regional definitions used throughout this paper. In Section 3 we quantify the temporal and spatial variability of surface ozone, show how temporal and spatial averaging reduces the calculated ozone variability, and demonstrate the spatial heterogeneity of the calculated ozone variability. In Section 4, we discuss the potential strategies that could be used to maximize ozone ~~trend~~ signal detection due to meteorological variability, explore uncertainties, and make recommendations for future research.

2 Methods

We examine both present-day (one simulation and one observed dataset) and future (two simulations) surface ozone in this study. For present-day analysis, we simulate surface ozone using CAM-chem, a component of the Community Earth System Model (CESM) and available observations within the US from the EPA CASTNET database. For future analysis, and in order to examine the potential for patterns of variability to change in the future, we utilize two existing simulations of CAM-chem conducted by Garcia-Menendez et al. (2017). Much of this analysis is conducted using the R language (R-Project, www.r-project.org). Here we summarize each of the three datasets and our approach to our analysis in Section 3.

2.1 CAM-chem

The present-day simulation (MOZ_2000) was conducted using CAM-chem model version 1.2.2, with the CAM4 atmospheric component (see Tilmes et al., 2015; 2016 for model description and evaluation). The model has been used extensively for a wide range of atmospheric chemistry research and is included in the ~~Atmospheric Chemistry and Climate Model Intercomparison Project (~~ACCMIP~~,~~ (Lamarque et al., 2012; Young et al., 2012 and references therein). We conduct our simulations using the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4) chemical mechanism (Emmons et al., 2010), which is a full tropospheric chemical mechanism integrated into CAM-Chem (e.g. Lamarque et al., 2012; Tilmes et al., 2015; Brown-Steiner et al., in review). ~~with o~~Offline forced meteorology from the Modern-Era Retrospective analysis for Research and Applications (MERRA) reanalysis product (Rienecker et al., 2011) for 26 meteorological years (1990 – 2015). Additional model evaluation and comparisons to surface and ozonesonde observations can be found in Brown-Steiner et al. (in review). This simulation has 56 vertical levels – adopted from MERRA meteorology – and 96 latitudinal and 144 longitudinal grid cells. We aim to isolate the variability to the meteorologically-driven impact on atmospheric chemistry so we repeat year-2000 anthropogenic emissions from the ACCMIP (Atmospheric Chemistry and Climate Model Intercomparison Project) inventory (Lamarque et al., 2012) and all non-biogenic emissions for all meteorological years, and include specified long-lived stratospheric species (O_3, NO_x, HNO_3, N_2O, N_2O_5) as in MOZART-4 (Emmons et al., 2010), an online biogenic emissions model MEGAN (Guenther et al., 2012), and forced sea ice and sea surface temperatures to year 2000 historical conditions. Like many state-of-the-art chemical tracer models, the CAM-chem exhibits some biases, most notably for our purposes a high bias in simulated surface ozone in the Eastern US (e.g. Lamarque et al., 2012; Brown-Steiner et al., 2015; Travis et al., 2016; Barnes et al., 2016). Recent efforts have been successful in partially reducing these biases (e.g. Sun et al., 2017).

We also include two reference simulations of the future climate, MOZ_2050 and MOZ_2100 (simulating the meteorological years 2035 – 2065 and 2085 – 2115, respectively) using the CESM CAM-chem simulations described in detail by Garcia-Menendez et al. (2017) with one set of initial condition data, and a climate sensitivity of 3.0 ˚C. These simulations do not include projections of any changes in future emissions. Compared to the present-day simulation (MOZ_2000)~~s~~, these future simulations (MOZ_2050 and MOZ_2100) have several parametric differences: the model version is 1.1.2 (see Tilmes et al., 2015 and references for information on model development), the atmospheric component is CAM3, the emissions (which are held constant at year-2000 levels) are from the Precursors of Ozone and their Effects in the

Troposphere database (see Garcia-Menendez et al., 2017), and the meteorology is derived from a linkage between the Massachusetts Institute of Technology Integrated Global System Model (MIT IGSM) and the CESM CAM model (Monier et al., 2013), and as such has 26 vertical levels. For a full description of these simulations, see Garcia-Menendez et al. (2017).

2.2 CASTNET

The observational database comes from the EPA Clean Air Status and Trends Network (CASTNET), which has more than 90 surface observational sites within the United States and has been collecting hourly surface meteorological and chemical data since 1990 (US EPA, 2016 and https://www.epa.gov/castnet). We collected data from all sites that reported complete ozone data from each year and removed data that was marked invalid within the downloaded EPA files. The number of sites that matched these criteria varied from year to year, but generally we have between 55 and 94 sites throughout the 1991 – 2014 period. The CASTNET observational network is located primarily in rural sites, and thus is considered to be a reasonable comparison to coarse grid cell model output (e.g. Brown-Steiner et al., 2015; Phalitnonkiat et al., 2016). Since a notable trend in observed ozone data exists, especially in the Northeastern US (Frost et al., 2006), and since the simulations have no change in anthropogenic emissions, and thus no ozone trend, we detrended the CASTNET data for each of the four averaging regions (described below) using a simple linear regression.

2.3 Telescoping Regional Definitions

In order to isolate the impact of the size of the spatial scale over which ozone data is averaged, we analyze ozone data at different spatial scales. The largest region considered is the entire Continental US, while the smallest regions considered are at the individual grid cell level of the CESM CAM-chem model (1.9˚x2.5˚ latitude/longitude). ~~We focus on the US since there are CASTNET observations that provide adequate coverage in both space and time, and since the US has significant temporal and spatial variability.~~ Data and statistics for the other regions (i.e. the Midwestern and Southeastern US) are included in the Supplemental Material, but do not alter the conclusions we draw from the Northeastern US. For CESM CAM-chem data, we averaged all grid cells within each region, while for the CASTNET data we first average sites within each corresponding CESM CAM-chem grid cell, and then averaged these data together. These telescoping regions are shown in Figure 1.

2.4 Temporal Averaging Windows

To explore the impact of temporal averaging, we examine ozone across a range of temporal averaging windows, ~~that range~~ from 1 day up to the full 26 years for the CESM data (1990-2015), the full 24 years for the detrended CASTNET data (1991 – 2014), and the 30 years available from the future scenarios of Garcia-Menendez et al. (2017). Each averaging window, therefore, can be considered to be a "sample" of possible realizations of meteorology. For instance, a selection of an averaging window of 1 year has 26 possible slices within the 1990 – 2015 MOZ_2000 data, while a selection of an averaging window of 10 years has 17 possible slices within the CESM data (N = # years – length of window +1). In this study, we consider all realizations to be equally likely and compare them to each other and to the long-term trend. However, if we were only able to simulate 5 years, we would not be able to compare to the long-term trend, and so be unable to completely quantify the likelihood of error in the context of the long-term behavior.

Formatted: Indent: First line: 0 cm

Formatted: Font: Bold

Formatted: Indent: First line: 1.27 cm

3 Results

Here we examine the spatial and temporal behavior of MOZ_2000, MOZ_2050, and MOZ_2100 and compare MOZ_2000 to present-day CASTNET observations. We introduce the moving temporal averaging windows, explore possible thresholds of acceptable error or signal strength, and examine the influence of expanding spatial averaging regions. Finally, we combine these temporal and spatial averaging techniques into a single framework.

3.1 Spatial and Temporal Comparisons

Figure 2 plots the averaged spatial distribution of the ~~daily~~ maximum daily ~~-~~8-hour average ozone ~~average~~ (~~DM8H~~MDA8 O_3) for summertime (JJA) days for 1990-2015 for the present-day ~~MOZART~~CAM-chem simulation, MOZ_2000 (Figure 2a) and ~~for~~ compares to the year 2000 for CASTNET data (Figure 2b). Some of the averaging strategies we present can average away the high ozone behavior this MDA8 O_3 metric is intended to quantify, but it is such a well-reported metric that focusing our analysis on it allows for ready comparisons to other studies. The well-known high ozone bias in the Eastern US (e.g. Lamarque et al., 2012; Travis et al., 2016; Barnes et al., 2016) is apparent, but otherwise the spatial variability over the entire Continental US is well captured. While we do examine the magnitude of surface ozone in this paper, most of our analysis is focused on the variability around the mean value (the anomaly), and as we show below, the CASTNET observations and CESM results are largely consistent in their representation of ozone variability (Figure 2, Table 1). The standard deviation of the simulated ~~DM8H~~MDA8 O_3 is large over the Eastern US and the Pacific Coast, with peak values of ± 25 ppbv over the highly populated Atlantic Coast (Figure 2c). The variability (defined as the standard deviation divided by the mean, expressed as a percentage) is lowest over the Western US (~ 15%), only slightly higher over the Eastern US (up to 25%), and highest (up to 50%) over the coastal regions (Figure 2d). We include this relative standard deviation metric since the CAM-chem biases make it difficult to compare standard deviations directly. The future climate simulations, MOZ_2050 and MOZ_2100 (Figure 2e and 2f, respectively), although run with different parametric settings than MOZ_2000 (see Section 2), simulate a similar spatial distribution of surface ozone, although under the warmer simulated climate of 2050 and 2100. These future climate simulations have a similar spatial pattern to the present-day simulation (Figure 2a), with high ozone levels in the Eastern US that increases from 2050 to 2100 (see Garcia-Menendez et al. (2017) for more details).

Figure 3 compares boxplots over the four telescoping regions (Figure 1) for MOZ_2000, the CASTNET data, the detrended CASTNET data, and for the single year 2000 for the CASTNET data (Figures 3a-d), and Table 1 summarizes relevant statistics. In order to compare CASTNET ozone to the simulated ozone, which do not have a trend over time, we detrend the CASTNET data in order to remove the impact of any temporal trends (e.g. NO_x emissions reductions) on ozone. The Northeastern US ozone bias is apparent at the smaller spatial scales (Figures 3c,d) and is less apparent when averaging over larger regions (Figures 3a,b). Figure 3e compares the year-to-year boxplots of the JJA ~~DM8H~~MDA8 O_3 for the MOZ_2000 and the detrended CASTNET data, and demonstrates the variability both in the median and spread of the ozone values in both the modeled and simulated data. While the MOZ_2000 ozone is generally higher than the CASTNET data, there are years in which the CASTNET data has higher ozone extremes. The red box plot in Figure 3e, which corresponds to the red box plot in Figure 3b, indicates that the year 2000 was an anomalously low year for observed ozone, although not the lowest.

While all the CESM CAM-chem simulations have high ozone biases in the Northeastern US (Figures 2 and 3, Table 1), their capability to simulate ozone variability is consistent with the available observations (for present day) and for expectations of ozone variability changes in the future climate (for MOZ_2050 and MOZ_2100). ~~Here we examine the variability defined as the standard deviation divided by the mean (expressed as a percent), instead of the standard deviation alone, in order to account for the model biases in the magnitude of the simulated ozone.~~ It is clear that variability increases when the size of the averaging region decreases, a fact that is well noted in the literature, as in Hawkins and Sutton (2012) for climate variables and Barnes et al. (2016) for ozone. As can be seen in in Table 1, the CASTNET variability increases as the spatial scale decreases (10%, 13%, 16%, and 20% for our telescoping regions from continental to a single Northeastern U.S. grid box), and MOZ_2000 largely captures this trend, albeit with lower overall variability (5%, 10%, 15%, and 15%). This increase in ozone variability with decreasing spatial scale is maintained in the future climate simulations (6%, 10%, 16%, and 21% for MOZ_2050 and 7%, 12%, 17%, and 20% for MOZ_2100). Table S1 contains statistics for the other telescoping regions.

3.2 Variability, Averaging Windows, and Thresholds

As we aim to quantify the potential tradeoffs that result from a particular choice of temporal and spatial scales on the assessment of ozone variability within the US, we represent the spatial scale by applying the telescoping regions (see Figure 1 and Section 2.3) and we represent the temporal scale through the use of moving averaging windows (see Section 2.4). ~~that range from 1 day up to the full 26 years for the CESM data (1990-2015), the full 24 years for the detrended CASTNET data (1991 – 2014), and the 30 years available from the future scenarios of Garcia-Menendez et al. (2017). Each averaging window, therefore, can be considered to be a "sample" of possible realizations of meteorology. For instance, a selection of an averaging window of 1 year has 26 possible slices within the 1990 – 2015 MOZ_2000 data, while a selection of an averaging window of 10 years has 17 possible slices within the CESM data (N = # years – length of window +1). In this study, we consider all realizations to be equally likely and compare them to each other and to the long-term trend. However, if we were only able to~~

~~simulate 5 years, we would not be able to compare to the long-term trend, and so be unable to completely quantify the likelihood of error in the context of the long-term behavior.~~ We frame much of the following analysis from the perspective of limited simulation length in order to approximate the question that decision-makers and modelers face when constrained by limited computational capabilities or available data: what i~~'~~s the likelihood that a particular estimate (of both the mean and the variability) is not a true representation of the true mean and variability, but rather a product of the underlying variability at the particular choice of spatial and temporal scale?

Figure 4 presents this likelihood by plotting all possible estimates of ~~DM8H~~MDA8 O_3 (as anomalies from the long-term mean) over all possible selections of averaging window (from 1 day up to the complete time series) for our telescoping regions. The semi-cyclical and highly auto-correlated nature of surface ozone is apparent at all spatial scales, with alternating cycles of anomalously high and low ozone. The temporal impact of anomalous ozone events is indicated by the vertical and right-leaning diagonal striations, which show that anomalous ozone events can impact estimates of ozone values within averaging windows up to 15 or 20 years. Figure 4 demonstrates how small-scale anomalously high or low ozone values (that come only from meteorological variability) can impact temporal averages of 5, 10, or even 20 years. For instance, a selected 5-year averaging window within the MOZ_2000 simulation averaged over the Northeastern US could be 2.5 ppbv higher or lower than the 25-year mean value of 74 ppbv, a ~~difference~~ potential error of 7%. Horizontal lines in Figure 4 mark the length of averaging windows that are needed to ensure that ozone ~~variability~~ anomaly for any selection of averaging window does not exceed a given threshold (5, 1, and 0.5 ppbv for solid, dashed, and dotted lines respectively). This potential error ~~difference~~ is larger within smaller regions and at the shorter selections of the averaging window. While the high and low ozone anomalies differ in time between CASTNET, MOZ_2000, MOZ_2050, and MOZ_2100 in Figure 4, the impact of spatial and temporal averaging is consistent.

We also quantify this variability in Supplemental Figures S1 and S2, which plots the likelihood (as a percentage) that a particular selection of spatial (rows) and temporal (x-axis) scale estimates ozone values that exceed a particular threshold (colored lines) away from the true mean value. For instance, if we are interested in characterizing ozone behavior (e.g. estimating a trend, or the mean value) in the Northeastern US, but were limited to a 5-year simulation, there is more than a 50% likelihood that the simulated ozone is 1 ppbv away from the 26-year mean, and an 80% likelihood that the discrepancy is greater than 0.5 ppbv. However, these data indicate that there is a virtual certainty that the estimate will be within 2.5 ppbv of the true mean value. We should note that, at the grid-cell level and within a 10-year period, the surface ozone variability can exceed 1 ppbv but is unlikely to exceed 2.5 ppbv (Figure 4), and that a 20-year trend is very likely to be able to identify significant ozone signals among the impact of meteorological variability on atmospheric chemistry. Our results also align with the results from Garcia-Menendez et al. (2017), which recommended that simulations need to be at least 15 years long to identify anthropogenically-forced ozone signals on the order of 1 ppbv.

Figures 4 and Supplemental Figures S1 and S2 compare the CASTNET observations to the three CESM CAM-chem simulations, and while there are minor differences, there are broad features that are consistent. First, using longer temporal averaging windows reduces the influence of small-scale ozone variability at all spatial scales, and depending on the acceptable threshold, one can select a temporal scale that effectively reduces the likelihood of exceeding that threshold to zero. Second, larger spatial scales also reduce this likelihood of exceeding a given threshold, but not as effectively as longer temporal scales. Finally, the impact of both temporal and spatial averaging on ozone variability is largely consistent for the CASTNET observations and for all three CESM CAM-chem simulations.

3.3 Selection of Temporal Averaging Scales

Figure 5 extends this analysis to examine the spatial heterogeneity of this likelihood of the meteorological variability causing ozone anomalies exceeding particular thresholds at the grid cell level. Here we plot four thresholds (0.5, 1, 2.5, and 5 ppbv) and four averaging windows (1, 5, 10, and 20 years) for the MOZ_2000 simulation. Ozone variability is highest in the Eastern US. At the grid-cell level, there are two strategies for filtering out the noise associated with natural meteorological variability (and thus enhancing signal detection capabilities): either average over longer periods, or acknowledge the level of noise and increase the threshold. For these data, it is virtually certain that any 20-year average will be within 5 ppbv of a full 25-year mean value (which itself may not be an accurate representation of a longer simulation), and virtually certain that any 1-year average will be at least 0.5 ppbv away from the mean.

~~Figure 6 and~~ Supplemental Figure S3 ~~compare~~ extends the analysis of Figure 5 by comparing the MOZ_2000, MOZ_2050, and MOZ_2100 simulations ~~by selecting one column~~ ~~(~~across the four thresholds for the 5-year averaging window~~)~~. Figure 6 similarly ~~and~~compares ~~one row (~~the 1 ppbv ozone threshold across the four averaging windows~~)~~ ~~from Figure 6 for~~for MOZ_2000~~ to equivalent plots for~~, MOZ_2050, and MOZ_2100. Interpreting Figures ~~7~~ 6 and Supplemental Figure S3 give largely consistent interpretations than the analysis above (Figure 5). Namely, that at the grid-scale level, increasing the temporal averaging window (Figure 6) or increasing the acceptable ozone threshold (Supplemental Figure S3) are effective at reducing the impact of the meteorological variability on estimates of the ozone signal. Shorter windows (or smaller thresholds) are needed in the Western US (where variability is smaller, see Figure 2d) than in the Eastern US (where variability is larger)~~,~~ ~~and grid-cells~~as well as over coastal and highly populated regions~~tend to need longer windows (or higher thresholds)~~. Finally, the 1 ppbv threshold and the 5-year averaging window plots (in either Figure 5 or Supplemental Figure S3) indicate that the spatial distribution and location of the peak variability may shift into the future, although this may be due to parametric differences between MOZ_2000, MOZ_2050, and MOZ_2100. Future simulations will be needed to check this shift in peak ozone variability.

3.4 Selection of Spatial Averaging Scales

We examine the impact of increasing the spatial averaging region (Figure 7) at four different temporal averaging windows (1, 5, 10, and 20 years) and for the smallest ozone threshold from the previous section (0.5 ppbv). It is evident that at all temporal averaging windows, expanding the number of surrounding grid cells that are averaged together consistently decreases the likelihood of exceeding the 0.5 ppbv threshold, although these reductions are relatively small at the 1-year window, especially over the Eastern U.S. While increasing the spatial averaging from a single grid-cell up to include the surrounding 81 grid cells (bottom row in Figure 7) manages to essentially smooth away much of the spatial heterogeneity in surface ozone (by moving down any column in Figure 7), it does not eliminate the likelihood of exceeding the 0.5 ppbv threshold over much of the Eastern U.S. For instance, even at a 20-year averaging window, and by averaging together the surrounding 81 grid-cells over locations in the Eastern U.S., there is still a 20-70% likelihood of exceeding the 0.5 ppbv threshold due to the small-scale impact of the meteorological variability on atmospheric chemistry.

3.5 Combination of Spatial and Averaging Scales

We now examine the combined impact of temporal and spatial averaging on reducing the influence of small-scale ozone variability in order to enhance ozone signal detection capabilities. Table S2 summarizes our analysis by dividing the likelihood of the ozone variability estimates exceeding selected thresholds away from the long-term mean into four categories: (1) the length of the averaging window over which ozone is averaged (columns); (2) the magnitude of the ozone threshold of interest (rows); (3) the observed (CASTNET) and modeled (MOZ_2000, MOZ_2050, and MOZ_2100) ozone data (sub-columns); and (4) the size of the spatial extent over which ozone is averaged (sub-rows). A graphical representation consistent with the data presented in Table S2 is plotted in Figure 8 for the Continental US average and for three grid cells that represent various cases. In each plot in Figure 8, by moving along columns from left to right, we can see the influence of increasing the size of the temporal averaging window, and by moving along rows (from the bottom to the top), we can see the influence of increasing the spatial averaging scale. By taking in the entire plot as a whole, we can get a feel for the combined influence of both temporal and spatial averaging. Supplemental Figure S4 contains a plot for each grid cell in the Continental US.

On average within the Continental US, both temporal and spatial averaging are effective at reducing the calculated ~~DM8H~~MDA8 O_3 anomaly, although temporal averaging is more effective (Figure 8a). There are many grid cells in the Eastern and Western US coasts (Figure 8b, Supplemental Figure S4), where both spatial and temporal averaging are effective, but their combined usage is especially effective. There are also many grid cells where temporal averaging is effective, but spatial averaging is barely effective, or not effective at all (Figure 8c and Supplemental Figure S4). Finally, there are some grid cells, particularly in the Central US (Figure 8d and Supplemental Figure S4), where spatial averaging over smaller regions is effective, but spatial averaging of larger regions actually increases the calculated ~~DM8H~~MDA8 O_3 anomaly by including surrounding grid cells that have higher variability.

4 Discussion

We now return to the original three research questions posed in Section 1. First, what is the magnitude of ozone variability due to meteorology alone at the smallest scale, and what is the impact of increasing the scale of temporal and spatial averaging? In both observed and modeled ~~DM8H~~MDA8 O_3 surface data, the small-scale variability driven solely by the meteorological variability impact on atmospheric chemistry (expressed as the standard deviation as a percentage of the mean) can exceed 20% (Table 1, Figure 2d). The chemical variability examined here is the result of fluctuations in meteorology, which itself results from larger-scale climatological drivers. While variability in emissions also influences atmospheric chemistry, our analysis has removed the influence of emissions variability and isolated the variability due to meteorology. A more comprehensive analysis of chemical variability will need to account for both meteorological and emission variability, which is complicated by temporal trends in both the emissions of ozone precursor species and the climate.

There is high temporal and spatial heterogeneity of surface ozone variability (Figure 2d), with the lowest values found in the Western US (< 10%), higher values found in the Eastern US (up to 20%), and the highest values over coastal or heavily populated regions (up to 30%). Averaging over longer temporal scales (by increasing the averaging window) and over larger spatial scales (by expanding the averaging region) can reduce the magnitude of the calculated variability, with temporal averaging proving to be more effective than spatial averaging in most cases (Figure 8). In this study, we performed simple spatial averaging, but there are other methodologies for smoothing two-dimensional signals (e.g. Räisänen et al., 2011; Pogson and Smith, 2015) that could potentially increase signal detection capabilities.

Second, are there combinations of temporal and spatial averaging that maximize the filtration of calculated ozone variability, and thus maximize the potential for signal detection? Figure 8 (and Supplemental Figure S4) demonstrate clearly that there are cases in which the combined usage of temporal and spatial averaging can reduce the calculated variability better than either strategy alone (see Figure 8b), although there are many regions within the Eastern US in which spatial averaging has little to no impact on reducing the calculated variability (Figure 8c) or even results in an increase in the calculated variability (Figure 8d). There are no such cases (see Supplemental Figure S4) in which expanding the temporal averaging scale increases the calculated ozone variability. This could potentially enable region-specific averaging strategies that help decision-makers identify and meet regional air quality objectives.

Third, are these results dependent on the particular parameterizations of the CESM CAM-chem model, and are they consistent with the available CASTNET observations? The three CESM CAM-chem simulations exhibited consistent representations of ozone variability, consistent with our understanding of future changes to the climate (and meteorology) and the resulting impact on atmospheric chemistry (Table 1, Figures 4, S1, and S2). Compared to the CASTNET observations (which we detrended to remove the influence of changing precursor emissions), the present-day simulation (MOZ_2000) exhibited a high ozone bias in the Eastern US~~ (which is also evident in the future simulations, MOZ_2050 and MOZ_2100)~~, while the representation of the ozone variability is comparable (Table 1).

Finally, how may these strategies be applied to other datasets, be they chemical, meteorological, or climatological? Much of this analysis could be applied to any dataset that has spatial and temporal coverage, as long as some set of acceptable thresholds is provided. While our time step in this analysis is daily (given the ~~DM8H~~MDA8 O_3 metric), and applied only to summertime (JJA) days, any time step (i.e. hourly, monthly, annual, decadal) could be utilized as long as cyclical trends (e.g. diurnal or seasonal cycles) are removed. Indeed, the sliding-scale presentation in Figure 8 and Supplemental Figure S4 can specifically be utilized to identify particular spatial and temporal scales that are sufficient to identify signals at particular thresholds and to identify particular geographic regions that are best suited to identify a given signal. For example, Sofen et al. (2016) identified regions across the globe where additional observations would be particularly suited to improve our understanding of surface ozone behavior, and our analysis could potentially be used to identify particular temporal and spatial averaging scales that could further maximize the capability for trend detection. In particular, Sofen et al. (2016) noted that the peak in the power spectrum of the El Niño-Southern Oscillation (ENSO) on surface ozone is at the 3.8 year time scale, and that within some regions within the US, the amplitude of the ENSO influence on surface ozone approached 0.5 ppbv (and up to 1.1 ppbv globally). Our analysis shows that there are no grid cells within the Continental US where a 0.5 ppbv signal can be identified at the 5-year (or shorter) temporal averaging scale (Supplemental Figure S4), but that there are many regions – especially within the Western US – in which even a modest amount of spatial averaging can identify surface ozone signals below the 1 ppbv level with a 5-year or shorter averaging window. The type of sliding-scale analysis – in which spatial and temporal averaging are utilized individually and in combination – as presented in Figure 8 and Supplemental Figure S4 could readily be applied to a wide range of atmospheric (and other) topics to aid in the capability to identify signals that exist both in space and in time. In particular, low-frequency oscillations (e.g. ENSO, and others) and other forms of internally or externally forced trends (e.g. anthropogenic and natural changes in emissions) are readily adaptable to this type of analysis, which could address signals pertaining to precipitation, biogenic emissions, boundary layer variables, cloud properties, and many others.

Finally, we did not quantify statistical significance (as in Lewandowski et al., 2015) as our goals were to understand the general nature of ozone variability at all scales and for all signal strengths. Statistical significance testing (and other statistical techniques) can certainly provide additional information as to the strengths of ozone signals within the underlying variability, and can be used to extend these results in a case-by-case manner, but we leave this testing to future studies that can focus on particular air quality objectives at particular temporal and spatial scales. Furthermore, future research examining the impact of spatial and temporal averaging using regional-scale models, models with different resolutions, and the inclusion of urban observations could provide additional insight into understanding chemical variability and averaging techniques.

5 Conclusions

We quantified the impact of spatial and temporal averaging at different scales – both individually and combined – on estimates of summertime surface ozone variability and the resulting likelihood of over-confidence in estimates of chemical signals over the United States using CASTNET observations and the CESM CAM-chem model. We simulate three multi-decadal time periods, each with constant surface emissions, and find that this analysis is consistent across our simulated time periods, and that our results are not sensitive to particular configurations and parametric choices within the CESM CAM-chem (i.e. emissions, meteorology, and climate). We also provide a conceptual framework for gaining understanding of the influence of spatial and temporal averaging that may be adapted to a wide range of atmospheric and surface phenomena, provided sufficient spatial and temporal coverage. Here we focus on summertime surface ozone, a highly variable (in both space and time) atmospheric constituent with severe human health impacts and implications for planetary climate, which is the focus of many local, regional, and national policies. However, the resultant magnitude of these changes and ~~trends~~ signals are small compared to the magnitude of the day-to-day ozone variability, and detecting these changes and ~~trends~~ signals can be challenging. In particular, it would be impractical to delay interpreting observations for 10 – 15 years, or alternatively to expand the spatial averaging such that small-scale features are smoothed away. Nonetheless, it is unwise to over-interpret trends and signals based on observations from a limited spatial area and over a short temporal period. Our analysis and conceptual framework presented here cannot solve this tension, but it does demonstrate some strategies which can allow for a selection of spatial and temporal averaging scales, and a consideration of the error threshold, that can aid in this signal detection on a case-by-case basis. Taking into account the complex interactions involving trends and variability between emissions, chemistry, meteorology, and climatology necessitates a variety of strategies. This work quantifies the impact of spatial and temporal averaging in signal detection, which can be used in conjunction with ensembles of simulations, statistical techniques, and other strategies to further out understanding of the chemical variability in our atmosphere.

In order to quantify the impact of spatial and temporal averaging on summertime ozone variability, we start by selecting four telescoping spatial regions (the Continental US, the Eastern US, the Northeastern US, and a single grid cell within the Northeastern US) and examine all possible choices for averaging windows (ranging from daily to multi-decadal windows), although we focused primarily on averaging windows of 1, 5, 10, and 20 years. We find that – consistent with previous studies – summertime ozone variability is largest at the smallest scales, and is frequently on the order of ±10 – 20 ppbv, or which is roughly 15-20% of the mean ozone signal. In order to minimize the chemical noise that results from meteorological variability – and thus enhance the signal – we find averaging windows of 10-15 years (and sometimes longer at the smaller spatial scales) combined with modest (nearest-neighbor) spatial averaging substantially improve the capability for ~~trend~~ signal detection. We recognize that achieving a 10 – 15 year temporal averaging window is difficult, but this recommendation is consistent with recent literature (e.g. Barnes et al., 2015; Garcia-Menendez et al., 2017). For studies where 10 – 15 years of averaging is impractical, we recommend that some spatial and temporal context is provided that demonstrates that the signals being examined are robust and not the result of internal variability or noise. We also recognize that our analysis is just one strategy for enhancing signal detection capabilities, and will ideally be used alongside others, such as perturbed initial condition ensembles, running simulations with either internal or forced meteorology, and examining a region or time period with different models or parameterizations.

We show that the largest summertime ozone variability is found in the Eastern US (Figure 5, Figure S4), and subsequently there are many regions within the Eastern US where even a 20-year averaging window has a non-negligible likelihood of estimating ozone variability that is dependent (with possible error in the 1 – 3 ppbv range) on the particular years selected. In addition, over much of the Eastern US, simulations of 5-years or shorter have a substantial likelihood (40 – 90%, Figures S1 and S2) of reflecting the influence of meteorological variability on chemistry rather than the mean state of surface ozone, with the possibility of 5 – 10 ppbv error (Figure S4). While we have detrended the CASTNET observations to compare to the constant year-2000 cycled emissions in the simulations, the CASTNET time series inherently includes the compounded variability of both meteorological and emission sources. Future studies will need to expand this analysis to include trends and variability in the emissions, as well as in the meteorology.

Finally, we demonstrate a conceptual framework that allows for a “sliding-scale” view of surface ozone variability, in which both temporal and spatial averaging is examined at every grid cell within the Continental US. We show that the magnitude of estimates of ozone variability can be reduced with both temporal and spatial averaging, although temporal averaging tends to be more effective. While there are many regions in which both temporal and spatial averaging used in conjunction substantially reduce the estimate of ozone variability, there are some regions where spatial averaging is ineffective, or even counter-effective. In contrast, this is not the case for temporal averaging, which consistently reduces the magnitude of estimated ozone variability. Our analysis could be combined with other studies (e.g. Sofen et al., 2016) to guide observational and modeling strategies and identify regions and scales at which particular signals are most likely to be identified.

Code Availability

CESM CAM-Chem code is available through the National Center for Atmospheric Research / University Corporation for Atmospheric Research (NCAR/UCAR) website (http://www.cesm.ucar.edu/models/cesm1.2/), and this project made no code modifications from the released model version.

Data Availability

The raw model output is archived on the NCAR servers, and processed data is archived at https://dspace.mit.edu/handle/1721.1/114467.~~will be made available upon publication on Massachusetts Institute of Technology servers.~~

Supplemental Link

Author Contribution

BBS ran the present-day simulation, analyzed the data, and wrote the manuscript. EM ran the future climate simulations, while FGM ran the future atmospheric chemistry simulations and made the data available to BBS. NS, RP, EM, ST, and LE guided and reviewed the scientific modeling and analysis process. All authors ~~, and~~ provided feedback throughout the project and development of the manuscript.

Competing Interests

The authors declare that they have no conflict of interest.

Acknowledgements

This model development work was supported by the U.S. Department of Energy (DOE) Grant DE-FG02-94ER61937 to the MIT Joint Program on the Science and Policy of Global Change. Computational resources for this project were provided by DOE and a consortium of other government, industry, and foundation sponsors of the Joint Program. For a complete list of sponsors, see: http://globalchange.mit.edu. Additional computing resources were provided by the Climate Simulation Laboratory at NCAR's Computational and Information Systems Laboratory (CISL), sponsored by the National Science Foundation and other agencies. The National Center for Atmospheric Research is funded by the National Science Foundation. The authors would also like to thank Daniel Rothenberg for efficient processing of the ozone files.

References

Angélil, O., Stone, D., Perkins-Kirkpatrick, S., Alexander, L.V., Wehner, M., Shiogama, H., Wolski, P., Ciavarella, A., and Christidis, N.: On the nonlinearity of spatial scales in extreme weather attribution statements, Clim. Dyn., 2017.

Barnes, E. A., Fiore, A. M., and Horowitz, L. W.: Detection of trends in surface ozone in the presence of climate variability, J. Geophys. Res. Atmos., 121, 6112–6129, 2016.

Brown-Steiner, B., Hess, P. G., and Lin, M. Y.: On the capabilities and limitations of GCCM simulations of summertime regional air quality: A diagnostic analysis of ozone and temperature simulations in the US using CESM CAM-chem, Atmos. Environ., 101, 134–148, 2015.

Brown-Steiner, B., Selin, N. E., Prinn, R., Tilmes, S., Emmons, L., Lamarque, J.-F., and Cameron-Smith, P.: Evaluating Simplified Chemical Mechanisms within CESM Version 1.2 CAM-chem (CAM4): MOZART-4 vs. Reduced Hydrocarbon vs. Super-Fast Chemistry, Geosci. Model Dev. Discuss., in review, 2018.

Camalier, L., Cox, W., and Dolwick, P.: The effects of meteorology on ozone in urban areaas and their use in assessing ozone trends, Atmos. Environ., 41, 7127-7137, 2007.

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, 2012.

de Elía, R., Biner, S., and Frigon, A.: Interannual variability and expected regional climate change over North America, Clim. Dyn., 41, 1245, 2013.

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: the role of internal variability, Clim. Dyn., 38, 527, 2012.

Diem, J. E., and Comrie, A. C.: Predictive mapping of air pollution involving sparse spatial observations, Environmental Pollution, 119, 1, 99–117, 2002.

Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S. L., and Kloster, S.: Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4), Geosci. Model Dev., 3, 43-67, 2010.

Fiore, A. M., Oberman, J. T., Lin, M. Y., Zhang, L., Clifton, O. E., Jacob, D. J., Naik, V., Horowitz, L. W., Pinto, J. P., and Milly, G. P.: Estimating North American background ozone in U.S. surface air with two independent global models: Variability, uncertainties, and recommendations, Atmos. Environ., 96, 284–300, 2014.

Fiore, A. M., Jacob, D. J., Liu, H., Yantosca, R. M., Fairlie, T. D., and Li, Q.: Variability in surface ozone background over the United States: Implications for air quality policy, J. Geophys. Res. Atmos., 108, D24, 1787, 2003.

Garcia-Menendez, F., Saari, R. K., Monier, E., and Selin, N. E.: U.S. Air Quality and Health Benefits from Avoided Climate Change under Greenhouse Gas Mitigation, Environ. Sci. Technol., 49, 7580–7588, 2015.

Garcia-Menendez, F., Monier, E., and Selin, N. E.: The role of natural variability in projections of climate change impacts on U.S. ozone pollution, Geophys. Res. Lett., 44, 2911–2921, 2017.

Giorgi, F., and Bi, X.: Time of emergence (TOE) of GHG-forced precipitation change hot-spots, Geophys. Res. Lett., 36, L06709, 2009.

Guenther, A. B., Jiang, X., Heald, C. L., Sakulyanontvittaya, T., Duhl, T., Emmons, L. K., and Wang, X.: The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): An extended and updated framework for modeling biogenic emissions, Geosci. Model Dev., 5, 1471–1492, 2012.

Hawkins, E., and Sutton, R.: Time of emergence of climate signals, Geophys. Res. Lett., 39, L01702, 2012.

Jacob, D. J., and Winner, D. A.: (2009). Effect of climate change on air quality, Atmos. Environ, 43, 51–63, 2009.

Jhun, I., Coull, B. A., Schwartz, J., Hubbell, B., and Koutrakis, P.: The impact of weather changes on air quality and health in the United States in 1994–2012, Environ. Res. Lett., 10, 084009, 2015.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, Bull. Amer. Meteor. Soc., 96, 1333-1349, 2015.

King, A. D., Donat, M. G., Fischer, E. M., Hawkins, E., Alexander, L. V, Karoly, D. J., Dittus, A. J., Lweis, S. C., and Perkins, S. E.: The timing of anthropogenic emergence in simulated climate extremes, Environ. Res. Lett., 10, 094015, 2015.

Knote, C., Tuccella, P., Curci, G., Emmons, L., Orlando, J. J., Madronich, S., Baró, R., Joménez-Guerrero, P., Luecken, D., Hogrefe, C., Forkel, R., Werhahn, J., Hirtl, M., Pérez, J. L., José, R. S., Giordano, L., Brunner, D., Yahya, K., and Zhang, Y.: Influence of the choice of gas-phase mechanism on predictions of key gaseous pollutants during the AQMEII phase-2 intercomparison, Atmos. Environ., 115, 553–568, 2015.

Knutti, R.: The end of model democracy?, Clim. Change, 102, 395–404, 2010.

Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J., Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., Naik, V., Riahi, K., and Van Vuuren, D. P.: Historical (1850-2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: Methodology and application, Atmos. Chem. Phys., 10, 7017–7039, 2010.

Lamarque, J.-F., Dentener, F., McConnell, J., Ro, C.-U., Shaw, M., Vet, R., Bergmann, D., Cameron-Smith, P., Dalsoren, S., Doherty, R., Faluvegi, G., Ghan, S. J., Josse, B., MacKenzie, I. A., Plummer, D., Shindell, D. T., Skeie, R. B., Stevenson, D. S., Strode, S., Zeng, G., Curran, M., Dahl-Jensen, D., Das, S., Fritzsche, D., and Nolan, M.: Multi-model mean nitrogen and sulfur deposition from the atmospheric chemistry and climate model intercomparison project (ACCMIP): Evaluation of historical and projected future changes, Atmos. Chem. Phys., 13, 7997–8018, 2013.

Lamarque, J.-F., Emmons, L. K., Hess, P. G., Kinnison, D. E., Tilmes, S., Vitt, F., Heald, C. L., Holland, E. A., Lauritzen, P. H., Neu, J., Orlando, J. J., Rasch, P. J., and Tyndall, G. K.: CAM-chem: Description and evaluation of interactive atmospheric chemistry in the Community Earth System Model, Geosci. Model Dev., 5, 369–411, 2012.

Lawrence, M. G., Hov, Ø., Beekmann, M., Brandt, J., Elbern, H., Eskes, H., Feichter, H., and Takigawa, M.: The chemical weather, Environ. Chem, 2, 6–8, 2005.

Lewandowsky, S., Risbey, J. S., and Oreskes, N.: On the definition and identifiability of the alleged “hiatus” in global warming, Sci. Rep., 5, 16784, 2015.

Lin, M., Horowitz, L. W., Oltmans, S. J., Fiore, A. M., and Fan, S.: Tropospheric ozone trends at Mauna Loa Observatory tied to decadal climate variability, Nat. Geosci., 7, 136–143, 2014.

McComiskey, A., and Feingold, G.: The scale problem in quantifying aerosol indirect effects, Atmos. Chem. Phys., 12, 1031–1049, 2012.

Medhaug, I., Stolpe, M. B., Fischer, E. M., and Knutti, R.: Reconciling controversies about the ‘global warming hiatus,’ Nature, 545, 41–47, 2017.

Möller, D. Chemistry of the Climate System, pp. 331-334, Walter de Gruyter GmbH and Co., KG, Berlin/New York, 2010.

Monier, E., Scott, J. R., Sokolov, A. P., Forest, C. E., and Schlosser, C. A.: An integrated assessment modeling framework for uncertainty studies in global and regional climate change: The MIT IGSM-CAM (version 1.0), Geosci. Mod. Dev., 6, 2063–2085, 2013.

Monier, E., Gao, X., Scott, J. R., Sokolov, A. P., and Schlosser, C. A.: A framework for modeling uncertainty in regional climate change, Clim. Change, 131, 51–66, 2015.

Monier, E., Xu, L., and Snyder, R.: Uncertainty in future agro-climate projections in the United States and benefits of greenhouse gas mitigation, Environ. Res. Lett., 11, 055001, 2016.

Paltsev, S., Reilly, J. M., Jacoby, H. D., Eckaus, R. S., McFarland, J. R., Sarofim, M. C., Asadoorian, M. O., and Babiker, M. H.: The MIT emissions prediction and policy analysis (EPPA) model: Version 4, Rep. 125, MIT Joint Program on the Sci. and Policy of Global Change, 2005.

Pogson, M., and Smith, P.: Effect of spatial data resolution on uncertainty, Environ. Model. Softw., 63, 87–96, 2015.

Pyle, J. A., and Zavody, A. M.: The modelling problems associated with spatial averaging, Q. J. R. Meteorol. Soc., 116, 753–766, 1990.

Räisänen, J., and Ylhäisi, J. S.: How much should climate model output be smoothed in space?, J. Climate, 24, 867–880, 2011.

Rasmussen, D. J., Fiore, A. M., Naik, V., Horowitz, L. W., McGinnis, S. J., and Schultz, M. G.: Surface ozone-temperature relationships in the eastern US: A monthly climatology for evaluating chemistry-climate models, Atmos. Environ., 47, 142–153, 2012.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, R., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G-K, Bloom, S., Chen, J., Collins, D., Conaty, A., da

Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M., and Woollen, J.: MERRA: NASA's Modern-Era Retrospective analysis for Research and Applications, J. Climate, 24, 3624–3648, 2011.

Roberts, C. D., Palmer, M. D., McNeall, D., and Collins, M.: Quantifying the likelihood of a continued hiatus in global warming, Nat. Clim. Change, 5, 337–342, 2015.

Schnell, J. L., and Prather, M. J.: Co-occurrence of extremes in surface ozone, particulate matter, and temperature over eastern North America, Proc. Natl. Acad. Sci. U.S.A., 114, 11, 2854-2859, 2017.

Schurer, A. P., Hegerl, G. C., Mann, M. E., Tett, S. F. B., and Phipps, S. J.: Separating forced from chaotic climate variability over the past millennium, J. Climate, 26, 6954–6973, 2013.

Searle, K. R., Chipperfield, M. P., Bekki, S., and Pyle, J. A.: The impact of spatial averaging on calculated polar ozone loss: 2. Theoretical analysis, J. Geophys. Res, 103, D19, 25409–25416, 1998.

Sofen, E. D., Bowdalo, D., and Evans, M. J.: How to most effectively expand the global surface ozone observing network, Atmos. Chem. Phys., 16, 1445–1457, 2016.

Stevenson, D. S., Young, P. J., Naik, V., Lamarque, J. F., Shindell, D. T., Voulgarakis, A., Skeie, R. B., Dalsøren, S. B., Myhre, G., Berntsen, T. K., Folberth, G. A., Rumbold, S. T., Collins, W. J., MacKenzie, I. A., Doherty, R. M., Zeng, G., van Noije, T. P. C., Strunk, A., Bergmann, D., Cameron-Smith, P., Plummer, D. A., Strode, S. A., Horowitz, L., Lee, Y. H., Szopa, S., Sudo, K., Nagashima, T., Josse, B., Cionni, I., Righi, M., Eyring, V., Conley, A., Bowman, K. W., and Wild, O.: Tropospheric ozone changes, radiative forcing and attribution to emissions in the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos. Chem. Phys., 13, 3063–3085, 2013.

Strode, S. A., and Pawson, S.: Detection of carbon monoxide trends in the presence of interannual variability, J. Geophys. Res. Atmos., 118, 12257-12273, 2013.

Sun, J., Fu, J. S., Drake, J., Lamarque, J.-F., Tilmes, S., and Vitt, F.: Improvement of the prediction of surface ozone concentration over conterminous U.S. by a computationally efficient second-order Rosenbrock solver in CAM4-Chem, J. Adv. Model Earth. Sy., 9, 482–500, 2017.

Tilmes, S., Lamarque, J.-F., Emmons, L. K., Conley, A., Schultz, M. G., Saunois, M., Thouret, V., Thompson, A. M., Oltmans, S. J., Johnson, B., and Tarasick, D.: Technical Note: Ozonesonde climatology between 1995 and 2011 : description, evaluation and applications, Atmos. Chem. Phys., 12, 7475-7497, 2012.

Tilmes, S., Lamarque, J. F., Emmons, L. K., Kinnison, D. E., Ma, P. L., Liu, X., Ghan, S., Bardeen, C., Arnold, S., Deeter, M., Vitt, F., Ryerson, T., Elkins, J. W., Moore, F., Spackman, J. R., and Val Martin, M.: Description and evaluation of tropospheric chemistry and aerosols in the Community Earth System Model (CESM1.2), Geosci. Model Dev., 8, 1395–1426, 2015.

Tilmes, S., Lamarque, J. F., Emmons, L. K., Kinnison, D. E., Marsh, D., Garcia, R. R., Smith, A. K., Neely, R. R., Conley, A., Vitt, F., Val Martin, M., Tanimoto, h., Simpson, I., Blake, D. R., and Blake, N.: Representation of the Community Earth System Model (CESM1) CAM4-

chem within the Chemistry-Climate Model Initiative (CCMI), Geosci. Model Dev., 9, 1853–1890, 2016.

Travis, K. R., Jacob, D. J., Fisher, J. A., Kim, P. S., Marais, E. A., Zhu, L., Yu, K., Miller, C. C., Yantosca, R. M., Sulprizio, M. P., Thompson, A. M., Wennberg, P. O., Crounse, J. D., St. Clair, J. M., Cohen, R. C., Laughner, J. L., Dibb, J. E., Hall, S. R., Ullmann, K., Wolfe, G. M., Pollack, I. B., Peischl, J., Neuman, J. A., and Zhou, X.: Why do models overestimate surface ozone in the Southeast United States?, Atmos. Chem. Phys., 16, 13561–13577, 2016.

US EPA: National Ambient Air Quality Standards for Ozone: Final Rule. Fed. Regist. 80 (206), 65292-65468. 2015.

US EPA: CASTNET 2014 Annual Report Prepared by Environmental Engineering and Measurement Services, Inc. for the U.S. Environmental Protection Agency, 2016.

Weatherhead, E. C., Stevermer, A. J., and Schwartz, B. E., Detecting environmental changes and trends, Physics and Chemistry of the Earth, 27, 399-403, 2002.

Wild, O., and Prather, M. J.: Global tropospheric ozone modeling: Quantifying errors due to grid resolution, J. Geophys. Res., 111, D11305, 2006.

Young, P. J., Archibald, A. T., Bowman, K. W., Lamarque, J.-F., Naik, V., Stevenson, D. S., Tilmes, S., Voulgarakis, A., Wild, O., Bergmann, D., Cameron-Smith, P., Cionni, I., Collins, W. J., Dalsøren, S. B., Doherty, R. M., Eyring, V., Faluvegi, G., Horowitz, L. W., Josse, B., Lee, Y. H., MacKenzie, I. A., Nagashima, T., Plummer, D. A., Righi, M., Rumbold, S. T., Skeie, R. B., Shindell, D. T., Strode, S. A., Sudo, K., Szopa, S., and Zeng, G.: Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos. Chem. Phys., 13, 2063–2090, 2013.

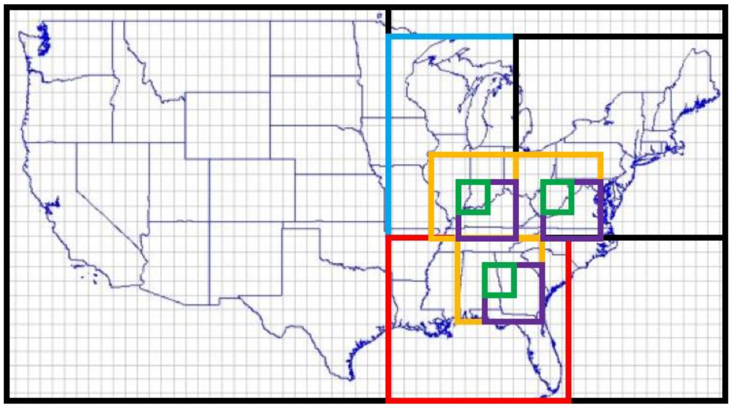
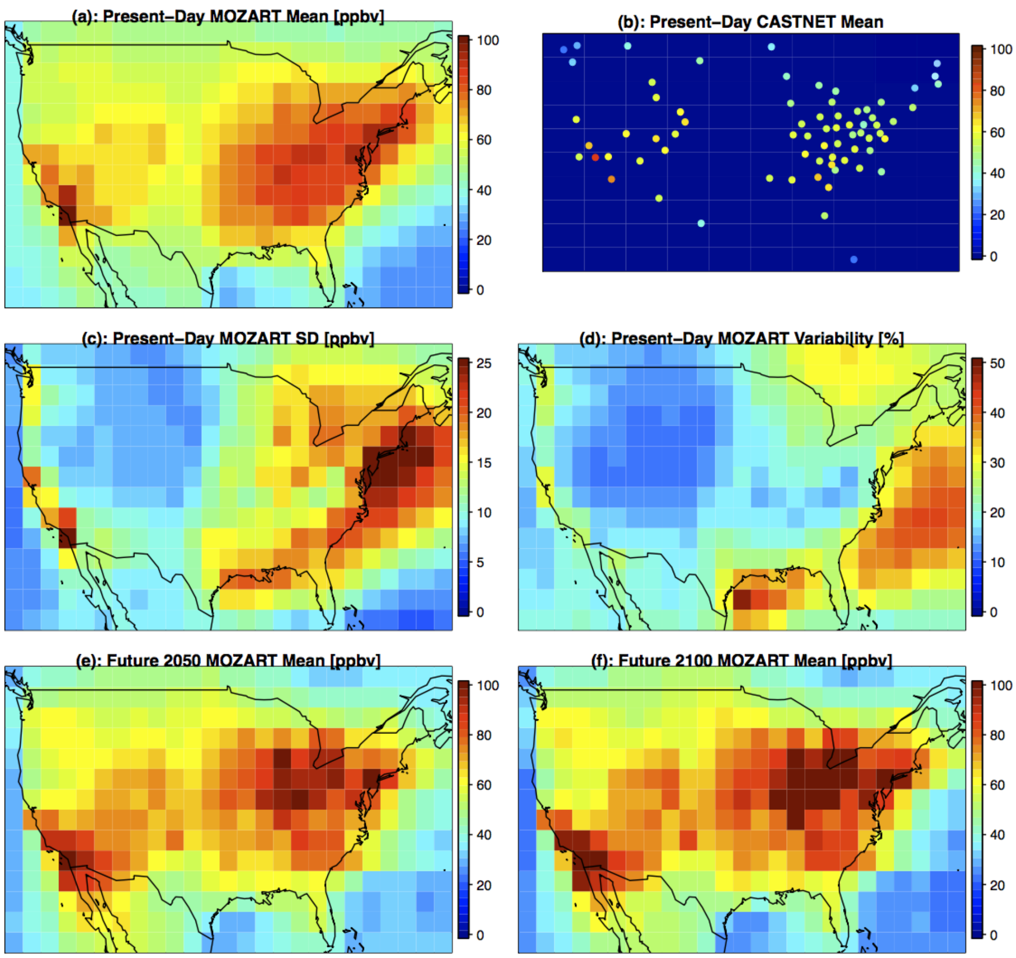


Figure 1: Telescoping Spatial Regions included in this study. The largest scale we consider is the Continental US (outer border). We focus on the Eastern US, by subdividing into three subregions: the Midwest (blue), Northeast (black), and Southeast (red). Within each subregion we telescope into a 3x3 grid cell (yellow), 2x2 grid cell (purple), and a 1x1 grid cell (green). In the paper, we only show a subset of these telescoping regions, and we include the rest in the Supplemental Material.



Formatted: Font: (Default) Times New Roman

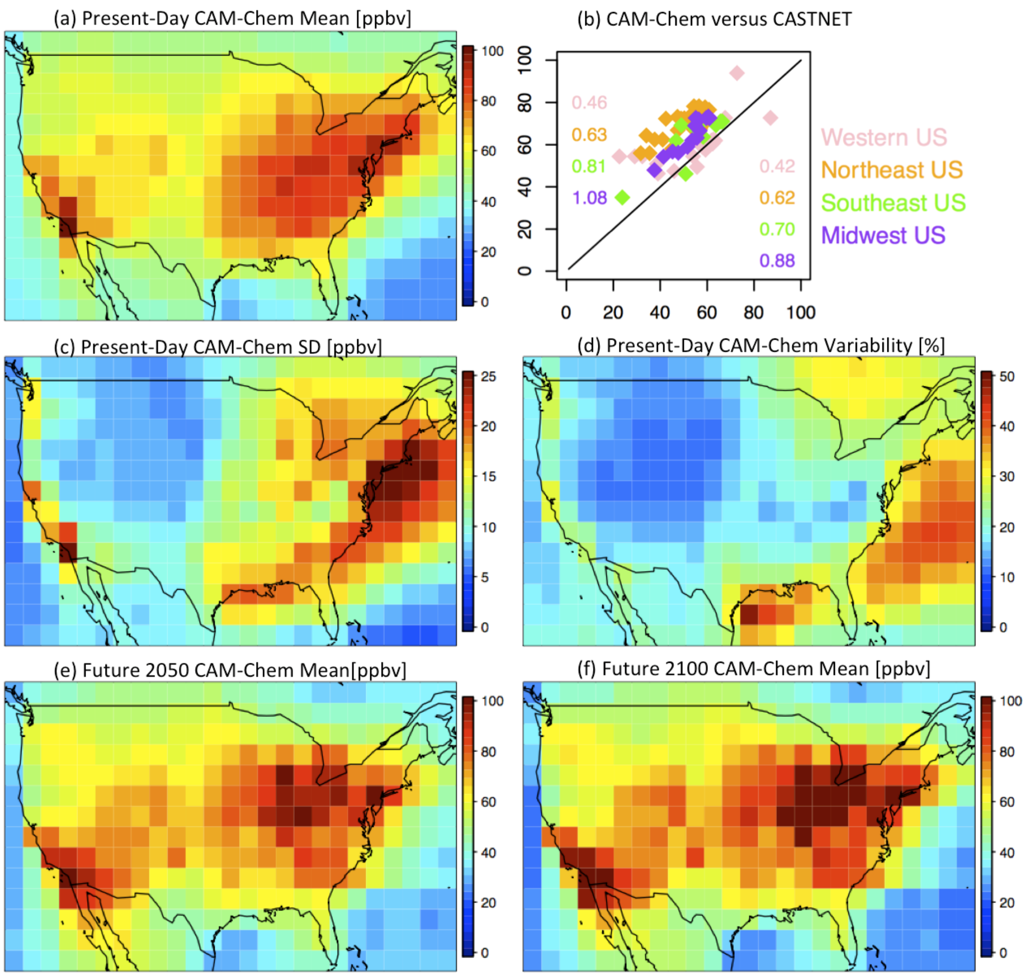
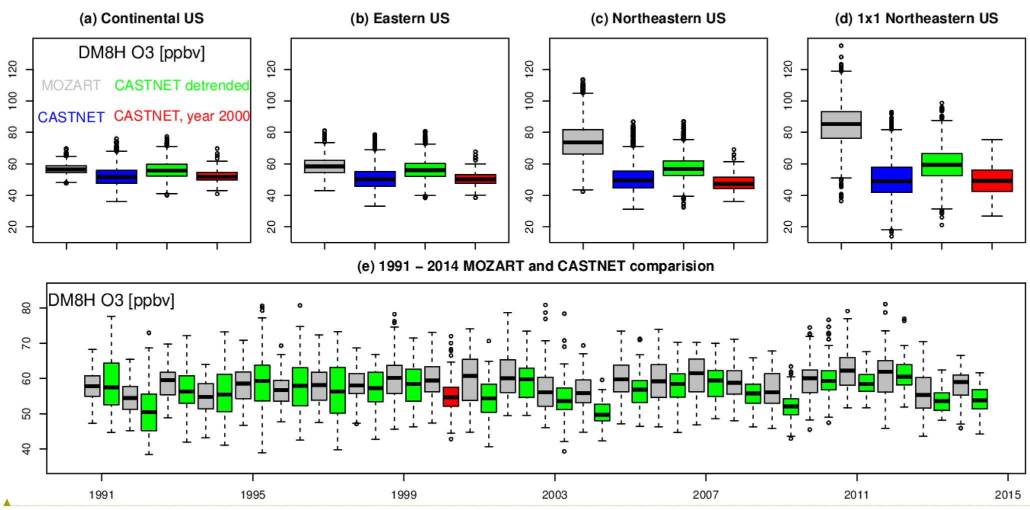


Figure 2: Continental US surface maps of (a) present-day ~~MOZART~~CAM-chem mean ~~DM8H~~MDA8 O_3; (b) CAM-Chem (y-axis) ~~present-day CASTNET mean DM8H O_3~~comparison to CASTNET observations (x-axis) for the year 2000 (see Brown-Steiner et al. (in review) for additional comparisons); (c) present-day ~~MOZART~~CAM- chem standard deviation of MDA8 O_3; (d) present-day ~~MOZART~~CAM- chem variability (standard deviation divided by mean, as a percent); (e) future ~~MOZART~~CAM- chem year 2050 mean ~~DM8H~~MDA8 O_3; and (f) future ~~MOZART~~CAM- chem year-2100 mean ~~DM8H~~MDA8 O_3. All model results are averaged over every JJA day in the time series, while the CASTNET results are only for the year 2000. The numbers in Figure 2b are slopes (left) and R^2 values (right).

Formatted: Superscript



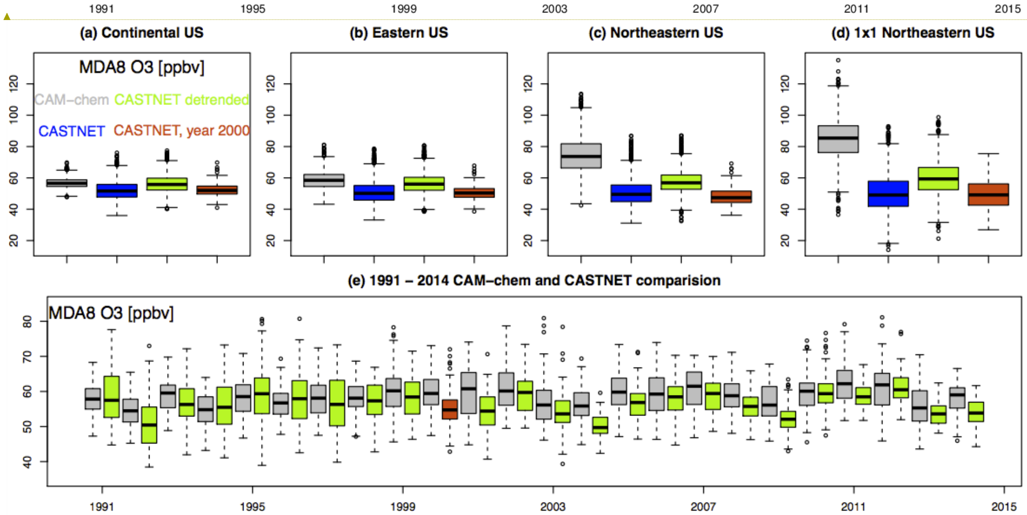


Figure 3: (a-d): Boxplots for surface ~~DM8H~~MDA8 O_3 for every summertime (JJA) day from 1991 – 2014 averaged over the Continental US, the Eastern US, the Northeastern US, and a single grid cell in the Northeastern US from ~~CESM~~ CAM-chem (grey), CASTNET observations (blue), detrended CASTNET observations centered at the year 2000 (green), and since the CAM-chem simulations have cycled year-2000 emissions and boundary conditions, the ~~detrended~~ CASTNET values for the year 2000 only (red). (e) Comparison of the yearly JJA ~~DM8H~~MDA8 O3 estimates averaged over the Eastern US for ~~MOZART~~CAM-chem (grey) and the detrended CASTNET (green) from 1991 – 2014. The single red boxplot coincides with the red boxplot in (b). The units are in ppbv, and for each boxplot the box contains the Inter Quartile Range (IQR), the horizontal line within the box is the median, and the whiskers extend out to the farthest point which is within 1.5 times the IQR with circles indicating any outliers.

Formatted: Font: (Default) Times New Roman, 10 pt

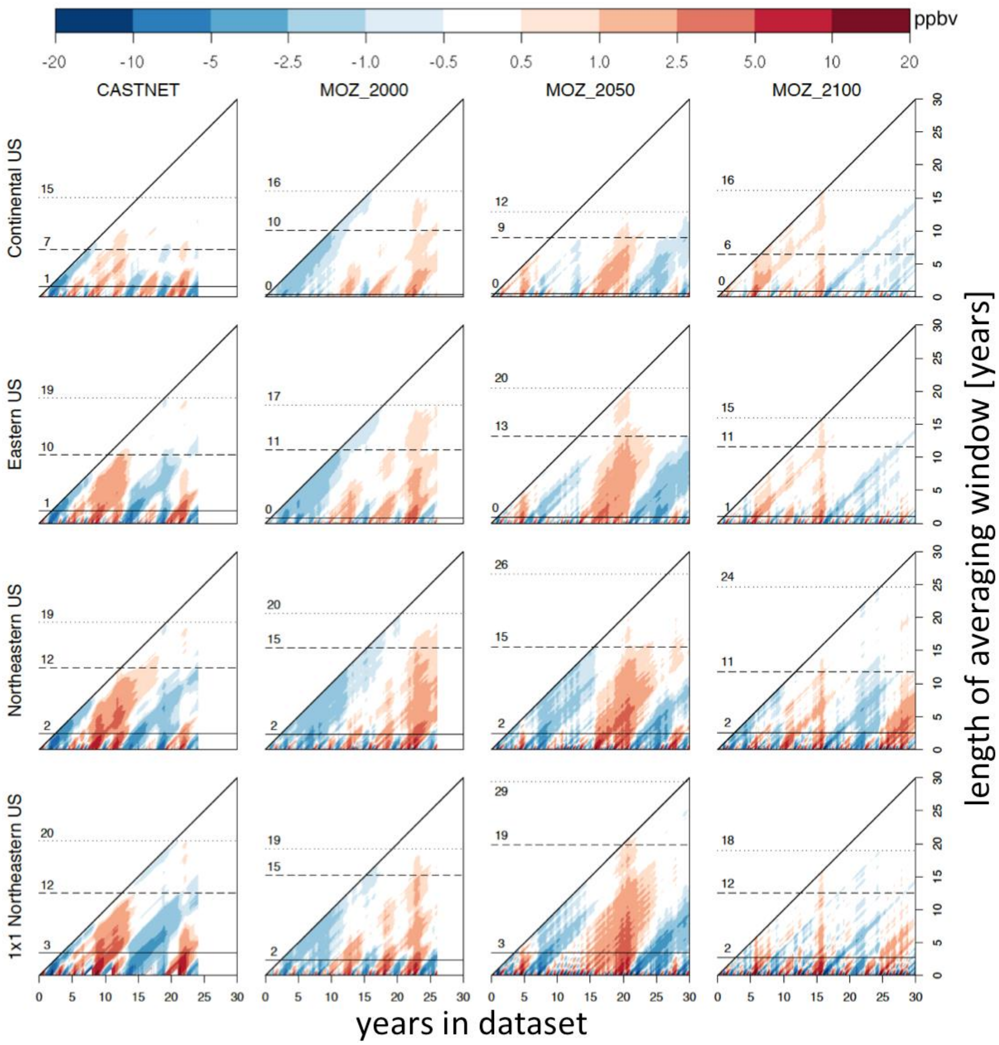
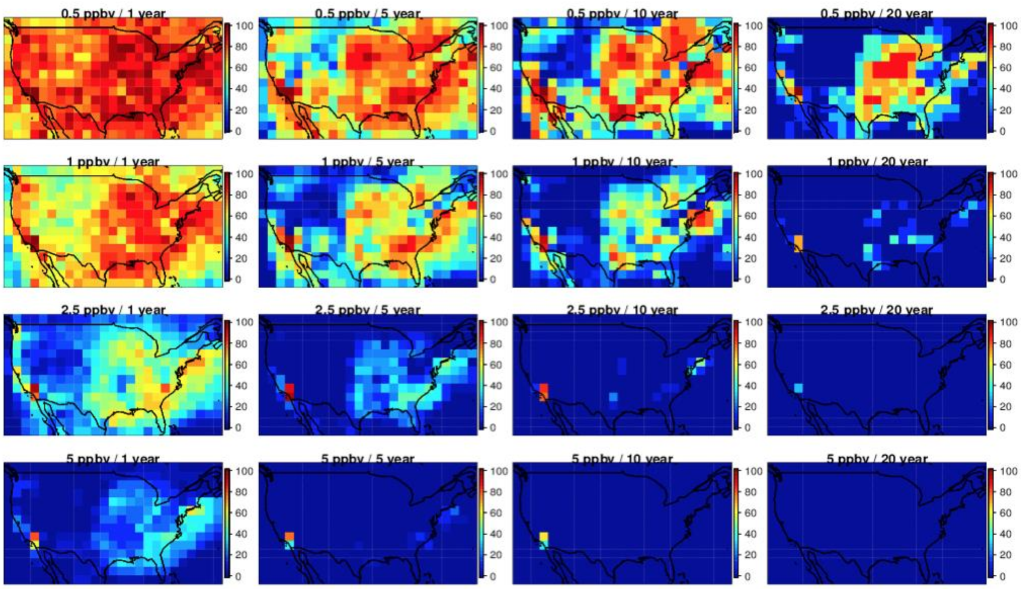
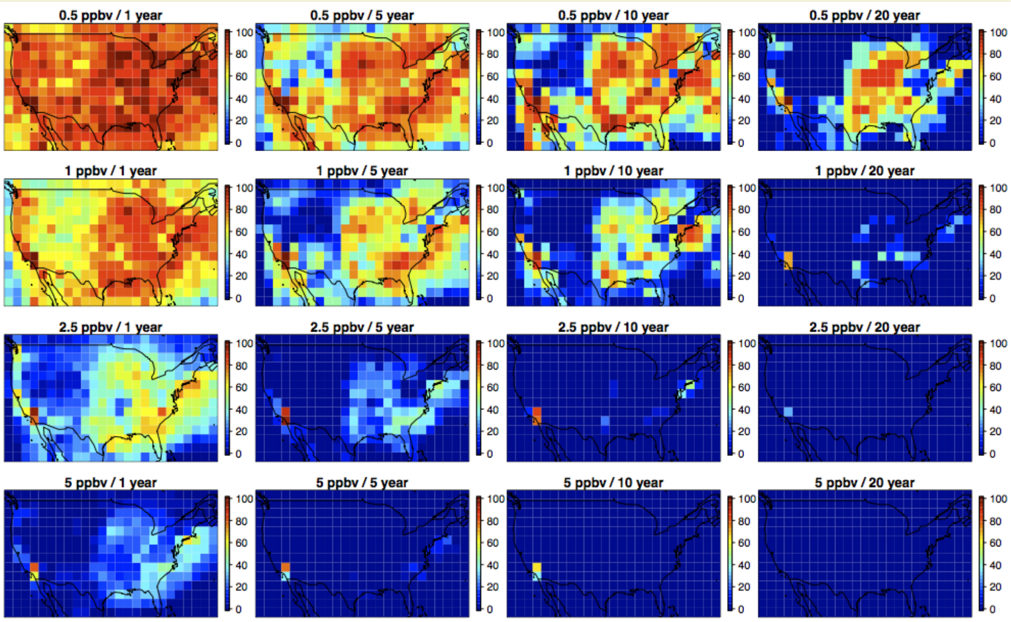


Figure 4: ~~A representation of the variability~~Comparisons of the variability represented by ~~of~~ the summertime ~~DM8H~~MDA8 O_3 anomaly (from the long-term summertime mean) for the four datasets in this study (CASTNET, MOZ_2000, MOZ_2050, MOZ_2100, columns) averaged over the four telescoping regions (CUS, EUS, NEUS, NEUS 1x1, rows). In each panel, the horizontal axis is the number of years in the dataset (24 years (1991-2014) for CASTNET, 26 years (1990-2015) for MOZ_2000, and 30 years (2036-2065 and 2086-2115) for MOZ_2050 and MOZ_2100), and the vertical axis represents the length of the averaging window (ranging from 1 day (bottom row) up to the entire time series (top pixel)). Each pixel represents the estimate of the ozone anomaly for a given averaging window (vertical axis) ending at a given time (horizontal axis). Horizontal lines indicate the length of averaging window required to guarantee that the variability drops below thresholds of 5 ppbv (solid), 1 ppbv (dashed), and 0.5 ppbv (dotted).

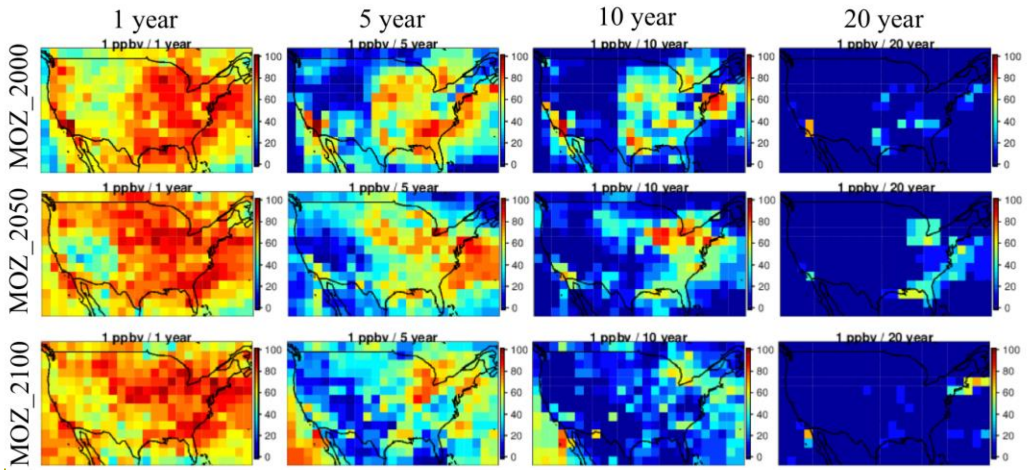


Formatted: Font: (Default) Times New Roman, 10.5 pt, Bold



Formatted: Font: (Default) Times New Roman, 10.5 pt, Bold

Figure 5: Spatial Plots over the Continental US plotting the likelihood (%) that an estimate of ozone exceeds a given threshold due to meteorological variability (rows) at the grid-cell level when using different lengths of averaging windows (columns) for the present-day CESM simulation (MOZ_2000).



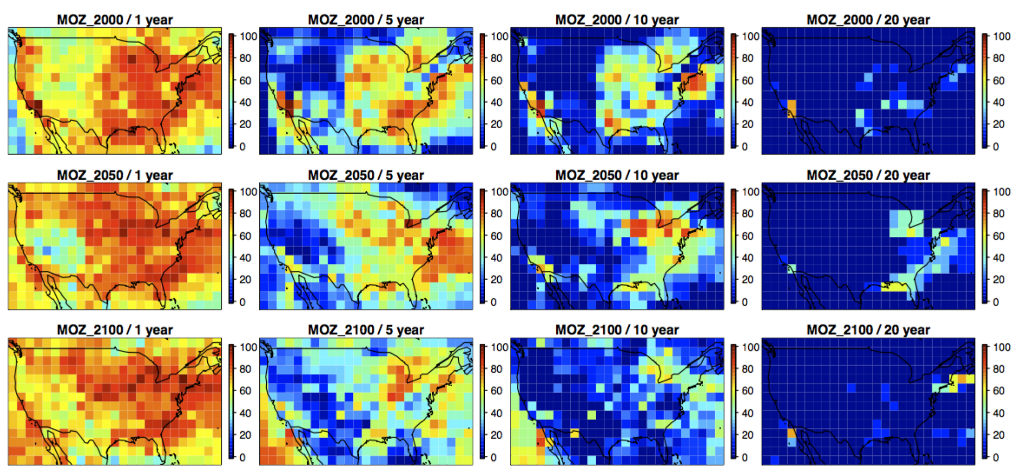
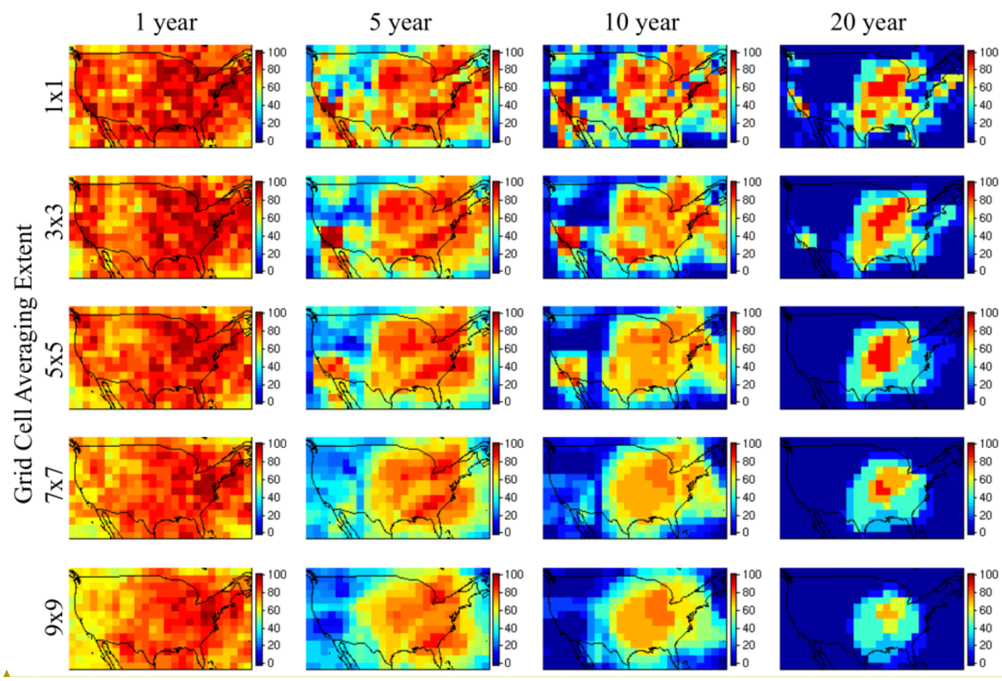


Figure 6: As in Figure 5, but only the second row (1 ppbv threshold), for present-day CAM-chem (MOZ_2000), future CAM-chem 2050 (MOZ_2050), and future CAM-chem 2100 (MOZ_2100).

Formatted: Font: (Default) Times New Roman

Formatted: Font: (Default) Times New Roman



Formatted: Font: (Default) Times New Roman, 10.5 pt, Bold

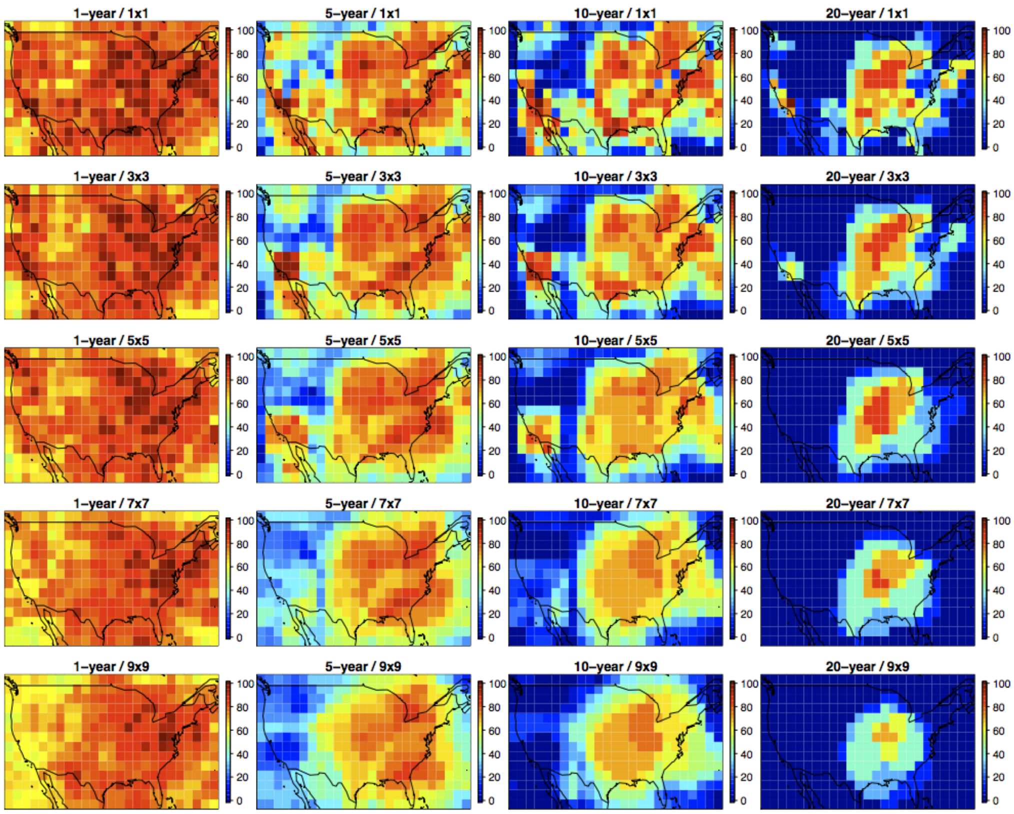
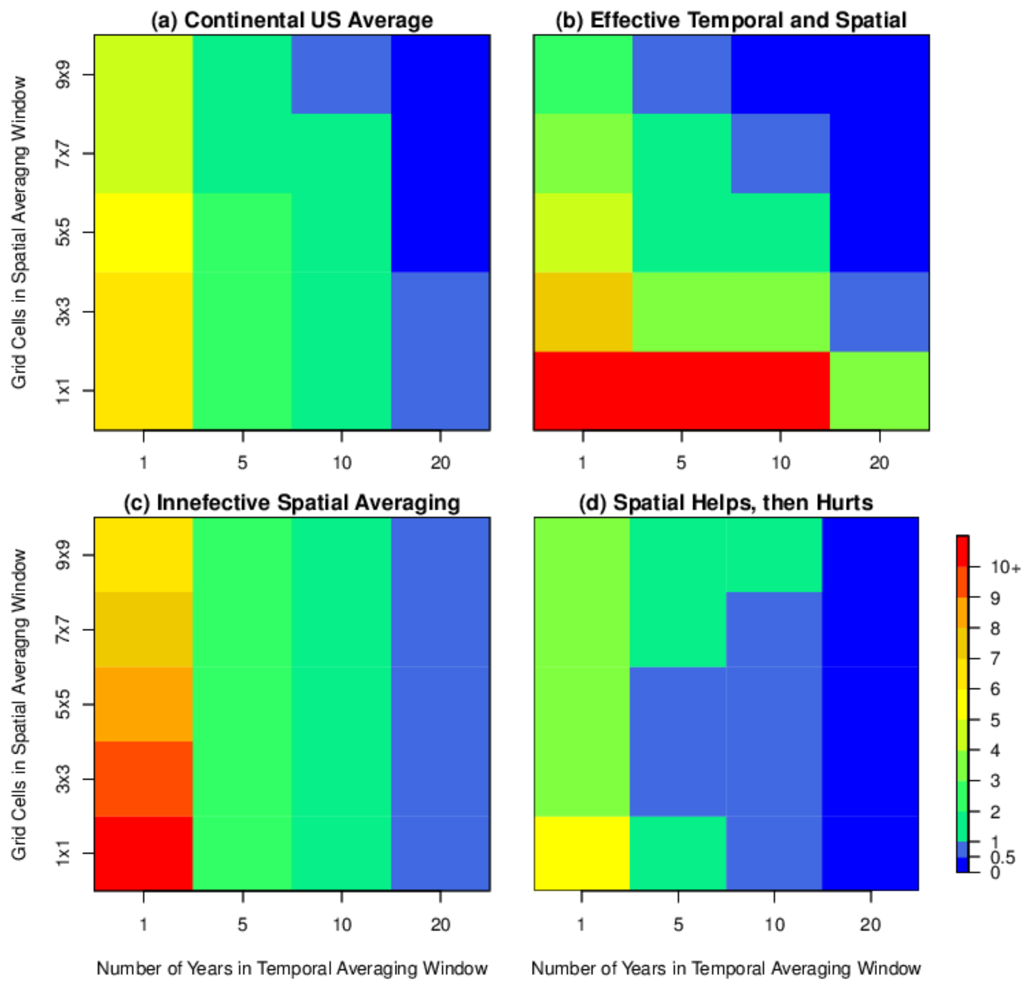


Figure 7: Combined impact of temporal and spatial averaging on reducing ozone variability on the likelihood (%) of exceeding the 0.5 ppbv threshold (as in Figures 5, 6, and Supplemental Figure S3) for the present-day MOZ_2000 simulation. The top row is the same as in Figure 6, while the lower rows have averaged the values within a 3x3, 5x5, 7x7, and 9x9 grid box surrounding each individual grid cell.

Formatted: Font: (Default) Times New Roman, 10.5 pt, Bold



Formatted: Font: (Default) Times New Roman

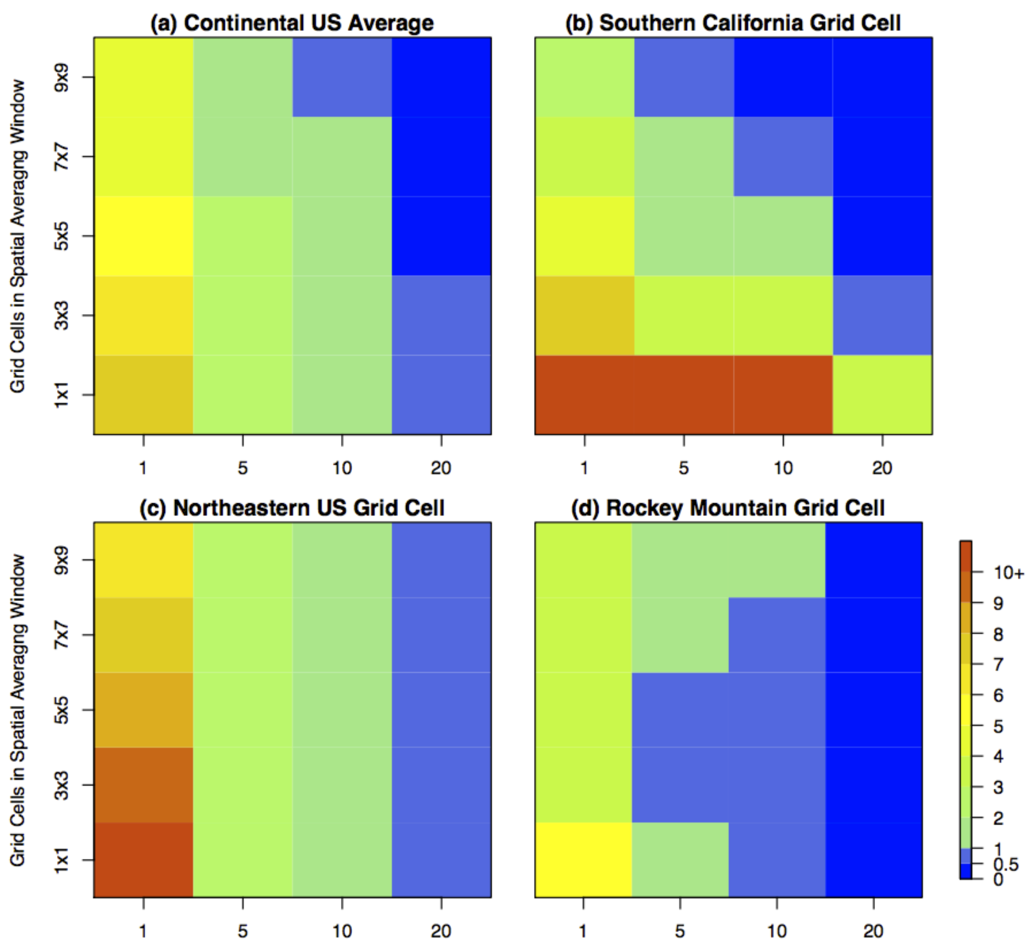


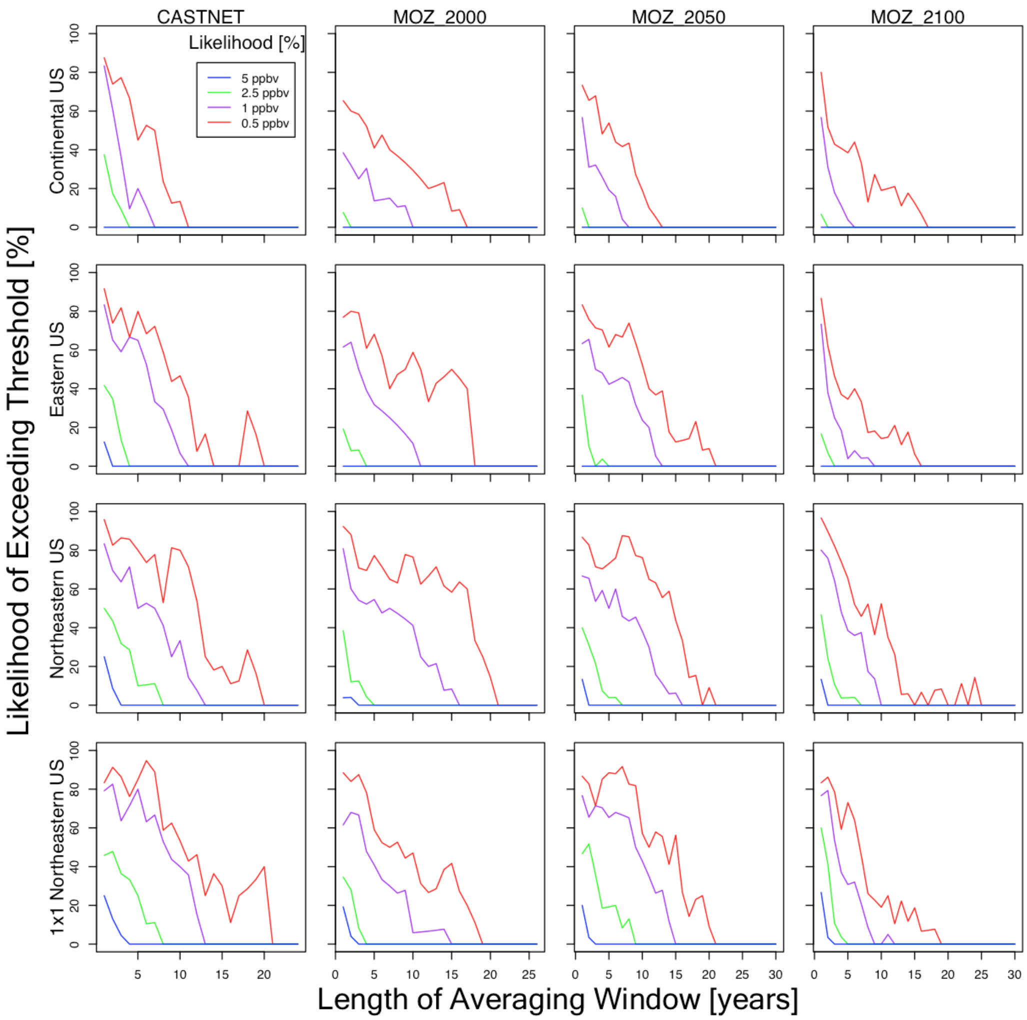
Figure 8: The maximum potential calculated ~~DM8H~~MDA8 O_3 anomaly [ppbv] from the long-term mean for (a) the Continental US average and three individual grid cells taken from (b) Southern California, demonstrating effective temporal and spatial averaging, (c) the Northeast, where spatial averaging is ineffective, and (d) the Rocky Mountains, where spatial averaging initially reduces the anomaly, but then increases the anomaly as surrounding regions get included in the spatial average. ~~demonstrating the impact of temporal and spatial averaging, with t~~The number of years included in the temporal averaging window increase~~ing~~ along the x-axis and the number of grid cells included in the spatial averaging window increase~~ing~~ along the y-axis. A full map of the Continental US can be found in the Supplemental Material (Figure S4). Note that the color scale is non-linear, and the color transitions are selected to match the thresholds established throughout this paper.

Formatted: Font: (Default) Times New Roman

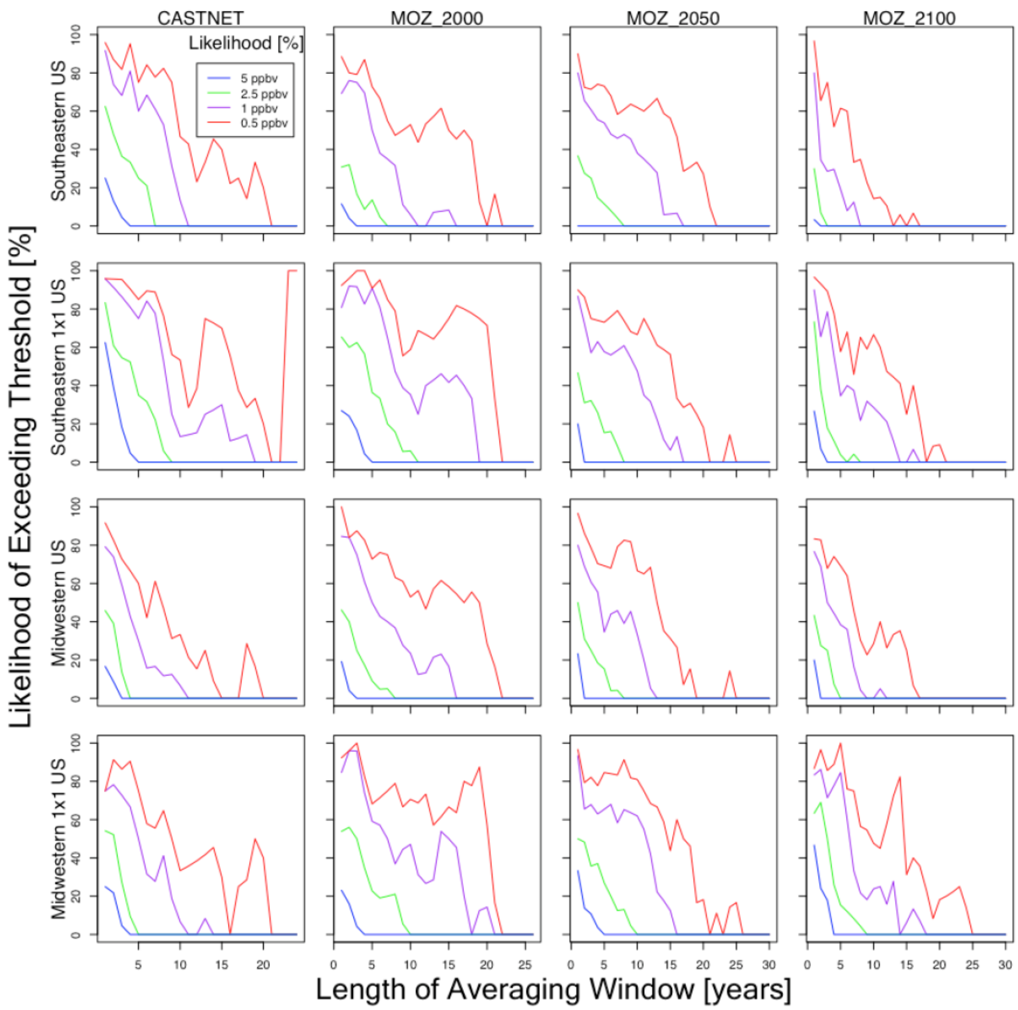
			CASTNET	MOZ_2000	MOZ_2050	MOZ_2100
Continental US	**Mean**	**ppbv**	52.4	56.7	56.8	57.4
	Standard Deviation	**ppbv**	5.04	3.08	3.54	3.73
	Variability	**%**	10%	5%	6%	7%
	Bias	**ppbv**		4.31		
Eastern US	**Mean**	**ppbv**	50.7	58.6	55.5	56.5
	Standard Deviation	**ppbv**	5.78	5.77	5.80	6.50
	Variability	**%**	11%	10%	10%	12%
	Bias	**ppbv**		7.91		
Northeastern US	**Mean**	**ppbv**	48.3	74.4	68.4	73.0
	Standard Deviation	**ppbv**	6.89	11.4	11.1	12.7
	Variability	**%**	14%	15%	16%	17%
	Bias	**ppbv**		26.1		
1x1 Northestern US	**Mean**	**ppbv**	49.6	84.9	81.1	85.1
	Standard Deviation	**ppbv**	10.2	12.8	16.7	17.3
	Variability	**%**	21%	15%	21%	20%
	Bias	**ppbv**		35.3		

Table 1: Statistical Summary of the CASTNET observations and the three CAM-chem simulations for different spatial averaging regions within the US. Variability is defined as the standard deviation divided by the mean value (in percent). Biases are only included for the present-day CAM-chem simulation compared to the CASTNET data. Similar tables for the other regions in this study are included in the Supplemental Material.

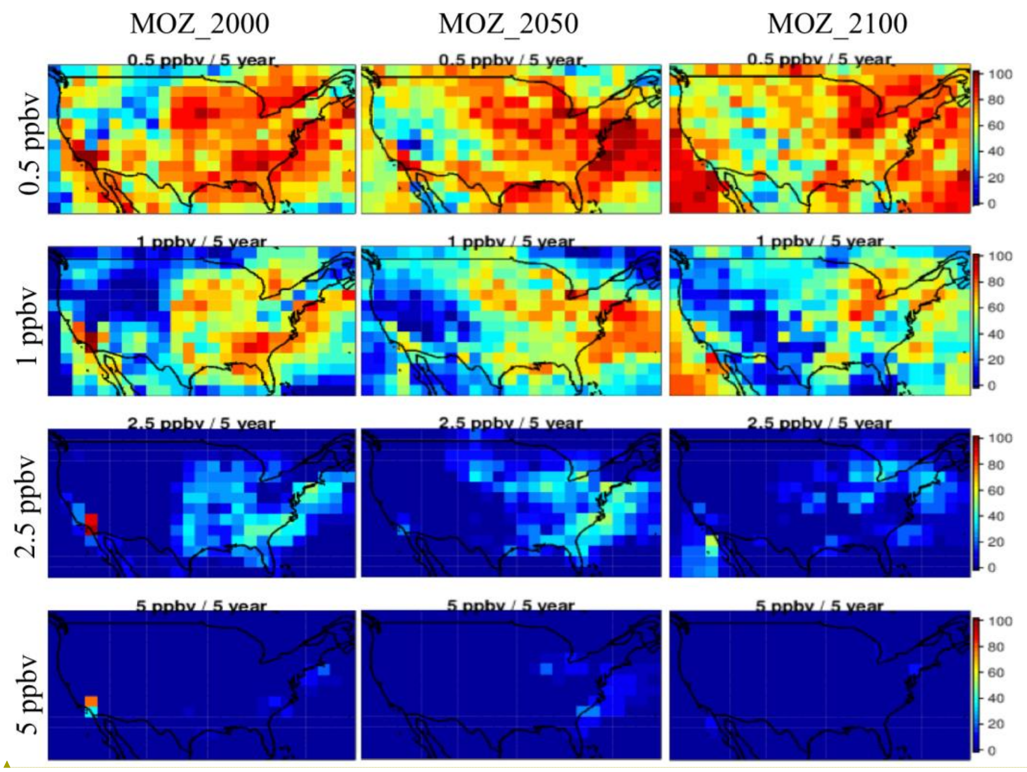
Supplemental Material



Supplemental Figure S1: The likelihood (percent, vertical axis) that an estimation of the mean MDA8 O_3 value for a given length of temporal averaging window (years, horizontal axis) is farther away from the long-term mean value than a given threshold: 5 ppbv (blue), 2.5 ppbv (purple), 1 ppbv (green), and 0.5 ppbv (blue). Individual columns represent the four datasets used in this study: CASTNET, present-day MOZART (MOZ_2000), and the two future MOZART simulations (MOZ_2050, MOZ_2100). Individual rows are spatially averaging over the telescoping regions seen in Figure 1.



Supplemental Figure S2 (as in Figure S1, but for the Southeastern and Midwestern US)



Formatted: Font: (Default) Times New Roman

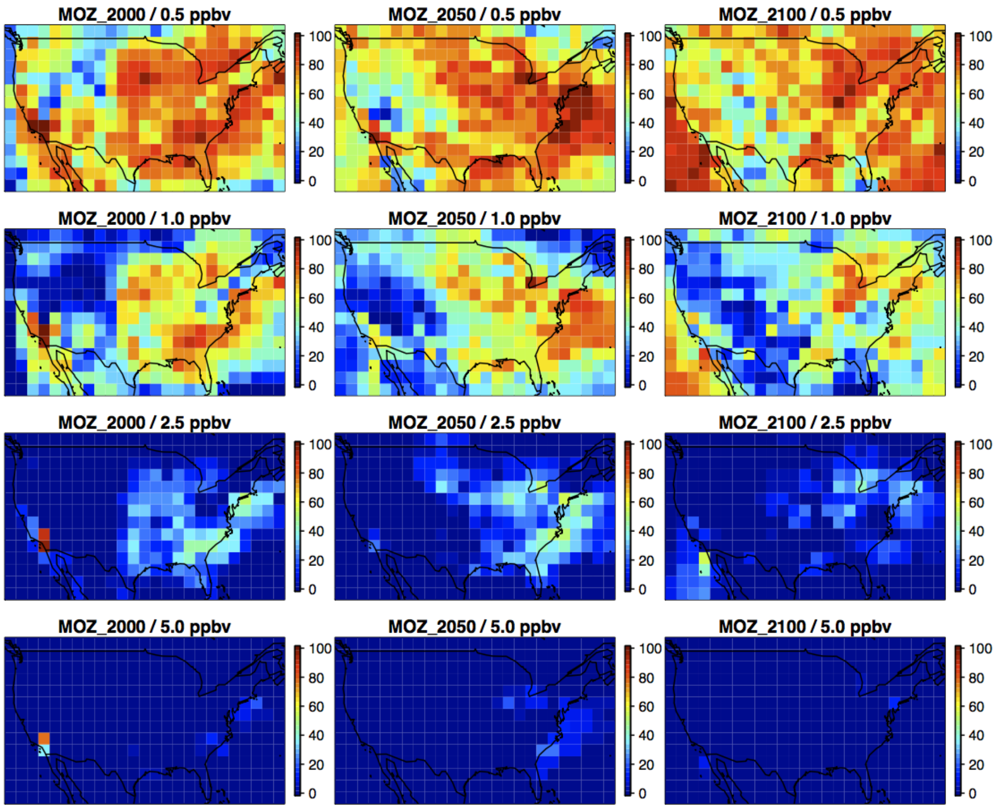
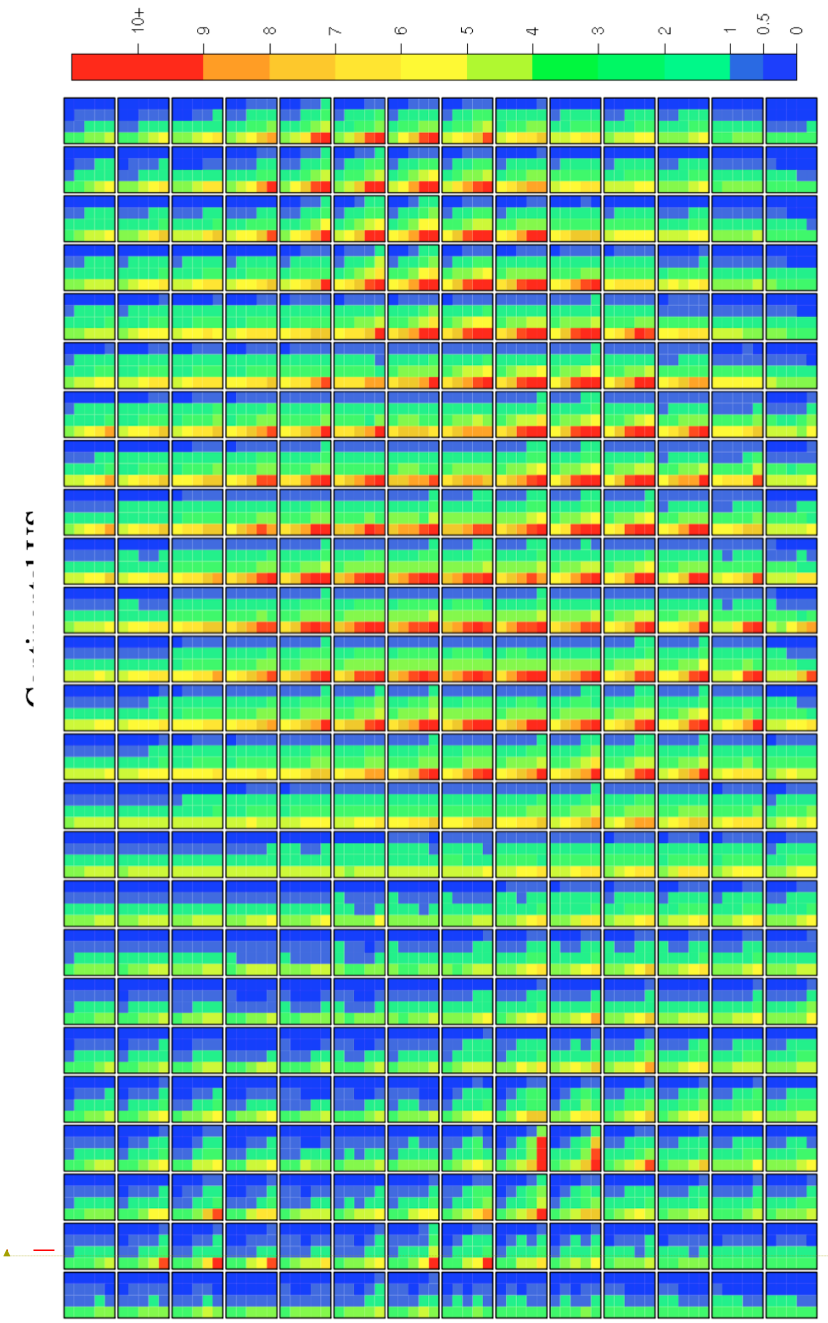


Figure S3: As in Figure 5, but only the second column ~~,~~ (5-year averaging window), for present-day CAM-chem (MOZ_2000), future CAM-chem 2050 (MOZ_2050), and future CAM-chem 2100 (MOZ_2100).~~for present-day MOZART, future MOZART 2050, and future MOZART 2100.~~

Formatted: Font: (Default) Times New Roman



Supplemental Figure S4: As in Figure 8, plotting maximum potential calculated DM8H O_3 anomaly [ppbv] from the long-term (1990 – 2014) mean, but for every grid cell in the Continental US. Covers the same Continental US extent as in Figures 5, 6, 7, and S3.

Formatted: Font: (Default) Times New Roman, 10.5 pt, Bold

			CASTNET	MOZ_2000	MOZ_2050	MOZ_2100
Southeastern US	**Mean**	**ppbv**	52.8	60.2	56.2	56.5
	Standard Deviation	**ppbv**	7.39	9.22	8.14	8.56
	Variability	**%**	14%	15%	14%	15%
	Bias	**ppbv**		7.39		
1x1 Southeastern US	**Mean**	**ppbv**	-	83.4	78.0	82.4
	Standard Deviation	**ppbv**	-	19.1	18.9	20.8
	Variability	**%**	-	23%	24%	25%
	Bias	**ppbv**		-		
Midwestern US	**Mean**	**ppbv**	53.4	71.0	74.4	79.0
	Standard Deviation	**ppbv**	6.81	11.0	12.1	13.5
	Variability	**%**	13%	16%	16%	17%
	Bias	**ppbv**		17.6		
1x1 Midwestern US	**Mean**	**ppbv**	-	84.6	98.2	109
	Standard Deviation	**ppbv**	-	16.2	22.9	26.1
	Variability	**%**	-	19%	23%	24%
	Bias	**ppbv**		-		

Supplemental Table S1: Statistical Summary of MDA8 O_3 for the CASTNET observations and the three CAM-chem simulations for the Southeastern and Midwestern US. Variability is defined as the standard deviation divided by the mean value (in percent). Biases are only included for the present-day CAM-chem simulation compared to the CASTNET data. Note that there were no CASTNET sites at the 1x1 grid cell regions.

Length of Averaging Window

Ozone Threshold	Region	5 years				10 years				15 years			
		CNT_2000	MOZ_2000	FGM_2050	FGM_2100	CNT_2000	MOZ_2000	FGM_2050	FGM_2100	CNT_2000	MOZ_2000	FGM_2050	FGM_2100
2.5 ppbv	CUS	0	0	0	0	0	0	0	0	0	0	0	0
2.5 ppbv	EUS	0	0	0	0	0	0	0	0	0	0	0	0
2.5 ppbv	NEUS	0	0	3.8	3.8	0	0	0	0	0	0	0	0
2.5 ppbv	NE1x1	10	0	19	0	0	0	0	0	0	0	0	0
1.0 ppbv	CUS	20	14	19	3.8	0	0	0	0	0	0	0	0
1.0 ppbv	EUS	25	32	42	3.8	0	12	24	0	0	0	0	0
1.0 ppbv	NEUS	30	55	50	38	0	41	38	0	0	8.3	6.3	0
1.0 ppbv	NE1x1	55	41	65	31	13	5.9	43	0	10	0	0	0
0.5 ppbv	CUS	35	41	54	38	6.7	29	19	19	0	8.3	0	13
0.5 ppbv	EUS	65	68	62	35	13	59	52	14	0	50	13	6
0.5 ppbv	NEUS	65	77	73	65	33	76	76	52	0	58	44	0
0.5 ppbv	NE1x1	70	59	88	73	67	47	57	19	30	42	56	19

% Likelihood

% Likelihood
0
1-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90

Table S2: Summary of the likelihood (%) of the ozone variability exceeding a given threshold (2.5, 1.0, 0.5 ppbv, rows) away from the long-term mean for a given an averaging window length (5, 10, 15 years, columns). We excluded the 5 ppbv threshold and the 1-year and 20-year averaging windows as they have very high or very low likelihoods, respectfully. Within each block, the Percentage Likelihood is further subdivided into the telescoping regions (CUS, EUS, NEUS, NE1x1, sub-rows) and the MDA8 O3 dataset (CASTNET, MOZ_2000, MOZ_2050, MOZ_2100, sub-columns).