UNIVERSITY OF CALIFORNIA,  
IRVINE

Surface Ozone, Extreme Air Quality Episodes, and Heat Waves:  
New Applications for Chemistry-Climate Models  

DISSERTATION

submitted in partial satisfaction of the requirements  
for the degree of

DOCTOR OF PHILOSOPHY

in Earth System Science

by

Jordan Lee Schnell

Dissertation Committee:
Professor Michael J. Prather, Chair  
Professor François W. Primeau  
Assistant Professor Michael S. Pritchard

2016
DEDICATION

To my parents, Chris, Amy, and Amanda

To the Philosopher, the Physician, the Meteorologist, and the Chemist, there is perhaps no subject more attractive than that of ozone.

Cornelius Fox
Ozone and Antizone
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vii</td>
</tr>
<tr>
<td>CURRICULUM VITAE</td>
<td>viii</td>
</tr>
<tr>
<td>ABSTRACT OF THE DISSERTATION</td>
<td>ix</td>
</tr>
<tr>
<td>CHAPTER 1: Synthesis and Overview</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Appendix A Overview</td>
<td>3</td>
</tr>
<tr>
<td>Appendix B Overview</td>
<td>8</td>
</tr>
<tr>
<td>Appendix C Overview</td>
<td>11</td>
</tr>
<tr>
<td>Appendix D Overview</td>
<td>12</td>
</tr>
<tr>
<td>Discussion and future directions</td>
<td>15</td>
</tr>
<tr>
<td>CHAPTER 2: The UCI CTM surface ozone diurnal cycle</td>
<td>17</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>30</td>
</tr>
<tr>
<td>APPENDIX A: Skill in forecasting extreme ozone pollution episodes</td>
<td>37</td>
</tr>
<tr>
<td>APPENDIX B: Evaluating chemistry-climate modeling of surface ozone</td>
<td>57</td>
</tr>
<tr>
<td>APPENDIX C: Effect of climate change on surface ozone</td>
<td>74</td>
</tr>
<tr>
<td>APPENDIX D: Co-occurrence of extremes in ozone particulate matter and temperature</td>
<td>85</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Diurnal Cycles of Surface Species in the UCI CTM</td>
<td>19</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>UCI CTM Boundary Layer Height</td>
<td>27</td>
</tr>
<tr>
<td>Figure A.1</td>
<td>Location of surface ozone stations and mean MDA8</td>
<td>40</td>
</tr>
<tr>
<td>Figure A.2</td>
<td>Interpolation Error versus Quality of Prediction</td>
<td>41</td>
</tr>
<tr>
<td>Figure A.3</td>
<td>Example Day of Interpolated Surface Ozone</td>
<td>43</td>
</tr>
<tr>
<td>Figure A.4</td>
<td>Gridded Selected Percentiles of Observed Surface MDA8 Ozone</td>
<td>43</td>
</tr>
<tr>
<td>Figure A.5</td>
<td>Maps of AirBase versus EMEP of Selected Ozone Percentiles</td>
<td>44</td>
</tr>
<tr>
<td>Figure A.6</td>
<td>UCI CTM Seasonal Bias and Correlation Maps</td>
<td>45</td>
</tr>
<tr>
<td>Figure A.7</td>
<td>PDFs of Observed and UCI CTM Surface Ozone</td>
<td>46</td>
</tr>
<tr>
<td>Figure A.8</td>
<td>Time Series of MDA8 Ozone at Selected Sites</td>
<td>47</td>
</tr>
<tr>
<td>Figure A.9</td>
<td>Maps of UCI CTM Extreme Event Hindcast Skill</td>
<td>48</td>
</tr>
<tr>
<td>Figure A.10</td>
<td>Observed and UCI CTM Example AQX Episode</td>
<td>48</td>
</tr>
<tr>
<td>Figure A.11</td>
<td>Observed and UCI CTM Episode Size CCDs</td>
<td>52</td>
</tr>
<tr>
<td>Figure B.1</td>
<td>Observed and ACCMIP Diurnal and Annual Cycles</td>
<td>63</td>
</tr>
<tr>
<td>Figure B.2</td>
<td>Observed Yearly Percentiles of MDA8 Ozone</td>
<td>66</td>
</tr>
<tr>
<td>Figure B.3</td>
<td>Maps of Observed and ACCMIP Summertime Ozone Enhancement</td>
<td>66</td>
</tr>
<tr>
<td>Figure B.4</td>
<td>Observed and ACCMIP Episode Size CCDs and Enhancement</td>
<td>67</td>
</tr>
<tr>
<td>Figure C.1</td>
<td>Percentile and Seasonal Future MDA8 Ozone Changes</td>
<td>78</td>
</tr>
<tr>
<td>Figure C.2</td>
<td>Maps of ACCMIP Multimodel Mean MDA8 Ozone Changes</td>
<td>79</td>
</tr>
<tr>
<td>Figure C.3</td>
<td>ACCMIP Change in Episode Size CCDs and Enhancements</td>
<td>82</td>
</tr>
</tbody>
</table>
Figure D.1  Spatial Patterns and Time Series of Extreme Threshold  92
Figure D.2  Example Co-occurring Extreme Episode  94
Figure D.3  Maps of Co-occurring Extreme Frequency  96
Figure D.4  Spatial and Temporal Offsets of Co-occurrences  98
Figure D.5  Progression of Average Large Scale Episodes  101
Figure D.6  Co-occurring Episode Sizes and Enhancements  103
Figure D.S1  Co-occurrence Frequency versus Episode Size  107
Figure D.S2  Event Enhancement for Co-occurrences  108
Figure D.1  Spatial Patterns and Time Series of Extreme Threshold  92
Figure D.2  Example Co-occurring Extreme Episode  94
Figure D.3  Maps of Co-occurring Extreme Frequency  96
Figure D.4  Spatial and Temporal Offsets of Co-occurrences  98
Figure D.5  Progression of Average Large Scale Episodes  101
Figure D.6  Co-occurring Episode Sizes and Enhancements  103
Figure D.S1  Co-occurrence Frequency versus Episode Size  107
Figure D.S2  Event Enhancement for Co-occurrences  108
ACKNOWLEDGMENTS

I would like to express the deepest gratitude to my committee chair and advisor, Professor Michael Prather, who has first and foremost made me a better scientist. He has empowered me by allowing me to pursue research endeavors without hindrance, yet while still providing the direction when I needed help. I cannot imagine a kinder, more intelligent, and all around wonderful person to have had the pleasure of working with. Without his guidance, patience, and flexibility, the work of this dissertation would not have been possible, and I would not be the scientist I am today.

I would like to thank my committee members, Professor François Primeau and Professor Michael Pritchard, for their patience and the insight and help they have provided whenever it was needed.

I would also like to thank my parents, siblings, and loved ones for their continued love, support, and encouragement over the course of my education and their encouragement to a

I thank John Wiley and Sons for permission to include Appendix C of my dissertation, which was originally published in Geophysical Research Letters. Financial support was provided by the NSF Graduate Research Fellowship Program Grant DGE-1321846.

The text of this dissertation is a reprint of the material as it appears in Atmospheric Chemistry and Physics and Geophysical Research Letters. The co-author listed in these publication directed and supervised research which forms the basis for the dissertation. I would also like to thank the other collaborators and coauthors for these publications: Christopher Holmes, Ayush Jangam, Beatrice Josse, Vaishali Naik, Larry Horowitz, Philip Cameron-Smith, Dan Bergmann, Guang Zeng, David Plummer, Kengo Sudo, Tatsuya Nagashima, Drew Shindell, Greg Faluvegi, and Sarah Strode.
CURRICULUM VITAE

Jordan Lee Schnell

2011  B.S. in Mathematics and Environmental Science, Indiana University-Bloomington

2013  M.S. in Earth System Science, University of California, Irvine

2016  Ph.D. in Earth System Science, University of California, Irvine

PUBLICATIONS


PRESENTATIONS

Schnell, J. L., On the co-occurrence of air quality extremes and heat waves, Abstract A42F-01 presented at the AGU Fall Meeting, Dec. 17, 2015; San Francisco, CA (oral)

Schnell, J. L., Global ozone model evaluation: Lessons from ACCMIP, presented at the HTAP2 Global and Regional Model Evaluation Workshop, May 12, 2015; Boulder, CO (oral)

Schnell, J. L., A critical evaluation of present-day and future surface ozone as simulated by global chemistry-climate models in the Atmospheric Chemistry & Climate Model Intercomparison Project (ACCMIP), Abstract A51Q-05 presented at the AGU Fall Meeting, Dec. 19, 2014; San Francisco, CA (oral)

Schnell, J. L., Can global chemistry-climate models reproduce air quality extremes?, Abstract A33L-02 presented at the AGU Fall Meeting, Dec. 11, 2013; San Francisco, CA (oral).
ABSTRACT OF THE DISSERTATION

Surface ozone, extreme air quality episodes, and heat waves: New applications for chemistry-climate models

By

Jordan Lee Schnell

Doctor of Philosophy in Earth System Science

University of California, Irvine, 2016

Professor Michael J. Prather, Chair

Global change is driving chemistry, climate, and atmospheric composition to new regimes over the coming century, threatening attainment of air quality standards globally. The models used to quantify future air quality changes are often plagued with large surface ozone biases, hindering efforts to directly compare models with observations and to accurately quantify future changes. Studies of air quality extremes often rely on point-based measurements and an absolute threshold exceedance; consequently, they neither capture the large-scale, spatially coherent structures of the worst pollution episodes nor compare directly with models’ grid cell averages. This dissertation develops novel statistical approaches to commensurately compare observations and models with a specific focus on extreme pollution episodes.

The first of four studies led by the doctoral candidate develops a generalizable interpolation algorithm that converts irregularly spaced ozone measurements from surface networks in North America and Europe into maps of grid cell averaged ozone, allowing direct comparison with a global model. Air quality extreme (AQX) events are defined locally as statistical extremes of the
ozone climatology and are found to predominantly occur in clustered, coherent, multiday episodes with spatial extents of more than 1,000 km. Additionally, the University of California, Irvine Chemistry Transport Model (UCI CTM) demonstrates skill in hindcasting the observed extreme episodes, thus identifying a new diagnostic to test global chemistry-climate models. The second study evaluates the ability of the UCI CTM and a suite of models from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP) to simulate the observed, present-day surface ozone climatology over North America and Europe. The tests span temporal scales from diurnal to multi-year variability and on statistics from median geographic patterns to the timing and size of AQX episodes. We also identified and corrected an error in the UCI CTM diurnal cycle. The third study uses the ACCMIP models to quantify the effect of future climate change on surface ozone. The fourth study extends the methods to characterize the co-occurrence of surface ozone, particulate matter, and temperature extremes, providing further diagnostics for model evaluation and enabling an investigation of the multi-stressor impacts of poor air quality and heat waves.
Chapter 1

Synthesis and Overview

1. Introduction

Global change is driving chemistry, climate, and atmospheric composition to new regimes over the coming century, threatening attainment of air quality standards globally. Tropospheric ozone (O$_3$), one of the most important air pollutants, is a secondary trace gas in the atmosphere in that it is not directly emitted from any natural or anthropogenic sources but rather formed through a complex set of nonlinear chemical reactions involving nitrogen oxides (NO$_x$), CO, CH$_4$, and volatile organic compounds (VOCs) in the presence of sunlight. Given the right combination of emissions and meteorology, surface O$_3$ abundances can be elevated to extreme levels. The meteorology most favorable to extreme ozone pollution typically occurs under a slow-moving high-pressure system, which brings abundant sunshine, warm temperatures, light winds, and little to no precipitation. These conditions not only promote O$_3$ production, but also support its
accumulation along with other pollutants such as particulate matter, i.e., air quality extreme (AQX) episodes (Jacob and Winner, 2009). In addition, these stagnant conditions also favor the development of heat waves (Black et al., 2004), which also pose a great threat to human life (Rooney et al., 1998; Semenza et al., 1996; Robine et al., 2008). A prime example is the August 2003 heat wave and pollution episode over Europe, which resulted in over 35,000 deaths and has been recognized as a prototype of a potential future climate (Vautard et al., 2007). This conclusion arises from the expectation that stagnation events (Horton et al., 2014), heat waves (Beniston, 2004; Meehl and Tebaldi, 2004; Stott et al., 2004; Cowan et al., 2004), and air pollution episodes (Mickley et al., 2004; Tagaris et al., 2007; Wu et al., 2008; Gao et al., 2013; Rieder et al., 2015) to become more frequent, longer-lasting, and more intense under projected climate warming.

It is imperative that the models used for projections of future changes are evaluated against present-day observations if their simulated changes are to be trusted. In addition, the models must be compared to observations in a consistent and commensurate framework such that the observational dataset is comparable to what a global model predicts, i.e., the average O$_3$ abundance in a model grid cell. Furthermore, it is necessary to develop metrics that are not confounded by systematic model bias or future shifts in meteorological or chemical regimes.

In this dissertation, a surface O$_3$ dataset is first constructed in an effort to directly compare to a hindcast model simulation (Schnell et al., 2014; Appendix A). Extreme surface O$_3$ pollution events are defined in such a way that even a highly biased model demonstrates skill in reproducing them, providing new diagnostics to test global chemistry-climate models. These
metrics are then used to evaluate present-day simulations of surface O₃ from a suite of state-of-
the-art global models over regions where observations are widely available, validating there use
to examine changes under a future climate scenario (Schnell et al., 2015; Appendix B). The
effect of climate change on surface O₃ is then quantified using a subset of these models (Schnell
et al., 2016; Appendix B). The interpolation scheme and developed methods are then used to
extend the analysis to construct a climatology of co-occurrence and overlap of extreme episodes
of ozone, particulate matter and temperature (Appendix D).

The majority of the work of this dissertation is contained within Appendices A–D. Appendix A
describes the creation of the surface ozone dataset, the definition of extreme events, and the
statistical methods to characterize the nature of their connectedness in space and time. Appendix
B applies this dataset and metrics to a suite of global chemistry models. Appendix C uses a
subset of theses models to examine the impact of future climate change on surface ozone.
Appendix D describes the extension to extremes in particulate matter and temperature.
Appendices A-C have been previously published and as such are provided in their entirety.
Appendix D has been submitted for publication and is provided in the format required for
submission. Chapter 2 contains a short modeling exercise related to a discovery in Appendix B.

1.1 Appendix A: Skill in forecasting extreme ozone pollution episodes with a
global atmospheric chemistry model

The initial motivation of Appendix A was to develop a surface O₃ dataset that could be used to
evaluate surface O₃ simulated by the UCI CTM. Our materials included a two year (2005–2006)
hindcast of MDA8 O₃ simulated at 1° x 1° by the UCI CTM and observed hourly surface O₃ from monitoring networks in the US and Europe. Model-measurement comparisons are often performed by comparing the surface O₃ abundance in a model grid cell to a station measurement located within the grid cell, or depending on the model resolution, a “representative” station in the case of multiple measurements located within the grid cell. Picking a representative station measurement is somewhat subjective, but common methods include using the average or median value of all measurements within a grid cell, or using a station that is considered to represent the relatively cleanest air (e.g., Fiore et al., 2003). These approaches do not use all of the available measurements, and it is unclear if the model should even be able to reproduce the observed value it is being compared to. Furthermore, comparisons are unable to be made in model grid cells without observed measurements. Since the surface O₃ abundance predicted by a model is not a single point but rather the average abundance over the grid cell, we set out to create an observed surface O₃ product matches what a model predicts and one that is continuous in space.

Developing a spatially continuous surface O₃ product requires using some form of interpolation. We explored many different types ranging in complexity from simple deterministic methods, such as nearest or natural neighbor, to more complex, stochastic methods such as Kriging. Kriging and inverse-distance-weighted (IDW) are the two most common types of interpolation methods used for surface O₃, with generally only modest differences found between them (Rojas-Avellaneda and Silvan-Cardenas, 2006). Both methods use a weighted linear combination of the measurements with the weights determined by some function of the distance between the observed and desired prediction point. The advantage of IDW is its implementation simplicity whereas Kriging methods require estimating a covariance function, which can easily
be modeled incorrectly especially at short separation distances (Diem, 2003). An attractive feature of Kriging, however, is that it reduces the weights of heavily clustered sites by effectively treating them as if they were one (Wackernagel, 2003). Upon testing multiple interpolation methods, we decided to use a hybrid of IDW and Kriging. Initially the weights are determined using simple IDW, but we also include a declustering mechanism similar to Kriging in an effort to limit the influence of heavily clustered stations, those usually found in polluted urban areas. Additionally, we predict the surface O₃ abundance at 25 points equally spaced in latitude and longitude within each 1° x 1° grid cell and then average those points an average abundance over the grid cell. Furthermore, we develop a quality of prediction ($Q^p$, Eq. 5 in Appendix A) defined as the effective number of independent stations at a distance of 100 km (roughly the size of a 1° cell) that went into the interpolation. The values of $Q^p$ are an informal estimate of the error of the interpolation and are used to limit the interpolation to only locations where a reliable prediction can be made. The parameters of the interpolation were optimized for our desired 1° x 1° product using a cross-validation approach. This error analysis revealed that even sites within a 100 km radius can vary by more than 5 ppb, further demonstrating the difficulty of trying to pick a single representative station (Figure A.2). The complete formulation of the interpolation can be found in Appendix A (Eq. 1-5). We applied our interpolation to ten years (2000-2009) of station derived MDA8 surface ozone observed over the US and Europe.

We then compared our surface O₃ product to the UCI CTM hindcast of 2005–2006. Overall, we found the UCI CTM to be biased extremely high, by as much as 70 ppb in the most polluted regions during the summertime (see Fig. A.6c–d); however, it showed greater skill in reproducing summertime day-to-day variability, with correlations greater than 0.8 for much of
the US and Europe (Fig. A.6g–h). We then tested the skill of the UCI CTM in simulating exact-
day matching of the most extreme levels of pollution. Our intention was to define extremes in the
observations and UCI CTM equivalently, but also taking into account the large systematic biases
in the model.

Studies of extreme surface O₃ typically investigate when, where, and how often they occur, the
mechanisms that produce them, and how they will change in response to shifts in emissions and
climate. Extremes are most often defined as an exceedance of an absolute threshold, such as the
US EPA’s National Ambient Air Quality standard of 75 ppb for MDA8 O₃. However, an
absolute extreme definition may not ever be exceeded in some locations, while other locations
may exceed it during most days of the summer. Additionally, because the bias found in the UCI
CTM is not uncommon among even the most advanced models (e.g., Nolte et al., 2008), and
combined with the intention of employing this method to models other the UCI CTM, it is
difficult to pick an absolute definition of an extreme that applies to every location in both
observations and models. Furthermore, due to ongoing and expected future reductions in O₃
precursor emissions, an absolute threshold may not apply under future scenarios. With these
factors in mind, we defined extremes locally an exceedance of a climatological percentile. The
benefit of this definition is that it highlights the times at each grid cell when pollution levels are
highest, regardless of the absolute abundance, enabling commensurate model-measurement
comparisons and an investigation of the underlying drivers causing the extremes.

Extreme events are defined as the 10 times N worst days (i.e., highest MDA8) in an N-year
period (i.e., the ~97.3 percentile). Using this definition, we find that despite the UCI CTM’s
large systematic biases, it is able to exactly match about 25% and 33% of the events over the US and Europe, respectively, and greater than 60% in some locations. The daily maps of the observed events revealed that the extremes occur in heavily clustered episodes that persist for multiple days. Previous studies (e.g., Logan, 1989) have demonstrated that extreme ozone events occur in large-scale episodes that are driven by meteorology; however, the extent of their connectedness had not been fully realized due to incomplete observational spatial coverage and the reliance on an absolute extreme definition. The individual extreme events are clustered into episodes using an agglomerative hierarchal clustering algorithm. Ideally, this method groups together events that occur under a stagnant high-pressure system over several days. The attributes of extremes defined in this climatologic sense provided an opportunity to develop a climatology of the episodes (e.g., areal extent, duration, intensity) that can be used to evaluate present-day and compare with future simulation of free-running global chemistry-climate models (i.e., not limited to hindcast simulations). The size of an episode is calculated as the time integrated areal coverage of the events within the episode (units = 10^4 km^2 days, where 10^4 km^2 is roughly the size of a 1° x 1° grid cell).

Overall, we find that the extremes predominantly occur in coherent, multiday episodes with spatial extents >1000 km and that the UCI CTM showed skill in hindcasting them (see Fig. A.10). The work of Appendix A identifies new diagnostic tests using global chemistry-climate models to examine changes in pollution episodes under future climate warming.
1.2 Appendix B: Use of North American and European air quality networks to evaluate global chemistry-climate modeling of surface ozone

Confidence in modeled projections of future air quality is based fundamentally on our ability to accurately simulate the present-day observed climatology of surface O₃ and particulate matter over NA and EU where dense, long-term, reliable measurements are available. It is imperative that the models used for these projections are extensively evaluated in their ability to simulate the present-day O₃ climatology if their projections under future climate and emissions scenarios are to be trusted.

The models in the Atmospheric Chemistry & Climate Model Intercomparison Project (ACCMIP) used in the most recent IPCC assessment represent one of the most sophisticated attempts to date to simulate present-day and future surface O₃. The work of Appendix B uses the surface O₃ dataset and the climatology of AQX events and episodes developed in Appendix A to evaluate the present-day O₃ climatology of the ACCMIP models. A hindcast simulation was also performed with the UCI CTM over the entire time period of the observations (2000–2009) to highlight the differences and similarities of hindcast and free-running models and the statistics used to evaluate them.

The ACCMIP consists of 16 global chemistry models, but we limited our evaluation to the eight that archived hourly surface O₃ abundances. The work of Appendix A focused on only the MDA8 and as such did not evaluate higher frequency statistics such as the diurnal cycle. In Appendix A, the MDA8 was first derived at each individual station and then interpolated onto
the 1° x 1° grid. Since hourly abundances were available for the UCI CTM and the ACCMIP models, in Appendix B the observed hourly abundances were interpolated and the MDA8 was subsequently derived at each grid cell. In addition, the observational spatial coverage over NA was increased by incorporating an additional ~300 stations from the US EPA’s Clean Air Status and Trends Network (CASTNet) and Canada’s National Air Pollution Surveillance Program (NAPS).

Several new metrics were developed in Appendix B in an effort to objectively evaluate the models while avoiding issues associated with systematic model bias. For example, the shapes of the diurnal and annual cycles were approximated as a cosine curves, which were used to define the hour and month of maximum phase ($h$ and $m$, respectively) and the daily and annual peak-to-peak amplitude ($H$ and $M$, respectively). The summary the model-measurement comparisons of these cycles is shown in Fig. B.1. One major finding of Appendix B was that the UCI CTM had serious problems with simulating the summertime diurnal cycle of surface O3. Whereas the observations (and most of the ACCMIP models) showed that the maximum O3 abundances occurred sometime around 15:00 local time, the UCI CTM showed this maximum as occurring around 12:00. Additionally, the peak-to-peak amplitude of the UCI CTM’s diurnal cycle ($H$) was overestimated by about 25 ppb, a clear outlier as compared to the ACCMIP models. Resolving the issue with the UCI CTM was beyond the scope of Appendix B and was left for a subsequent analysis, which can be found in Chapter 2.

A metric of summertime enhancement was also developed, defined as the difference between the 87th and 30th percentiles ($E_{JJA}$, see Fig. B.3). The 87th percentile was chosen to represent the
median percentile value of the 92 days of summer months (June-July-August). The 30th percentile was chosen to represent a baseline O₃ level – roughly the lowest levels or spring-fall days. An additional reason for defining the baseline level as the 30th percentile was that it showed little year-to-year variability such that it was not greatly influenced by bad pollution years (Fig. B.2). In this sense, the severity of summertime pollution can be evaluated on a yearly basis, which can be used to distinguish good vs. bad pollution years. As such, the 30th percentile baseline definition was also used to describe the intensity of AQX events and episodes, defined as the absolute value of the event minus the local 30th percentile. An important discovering using this metric was that the intensity of a pollution episode increases linearly (in log) with the size of the extreme episodes (see Fig. B.4c–d). Although most models underestimated the absolute value of the episode enhancement by about 5–10 ppb across all episode sizes, they correctly reproduced the slope of the enhancement-size relationship, demonstrating that the feature was not a simply an expected result of our statistical sorting.

The statistics and subsequently identified faults in the individual models put into perspective which features the models were unable to reproduce and thus need to be approached with caution when evaluating projected future changes. Overall, the work of Appendix B demonstrated the ACCMIP models were able to reproduce the salient features of the present-day O₃ climatology and therefore provided confidence in their ability to project future changes.
1.3 Appendix C: Effect of climate change on surface ozone over North America, Europe, and East Asia

A growing body of evidence suggests that pollution episodes will become more frequent, longer lasting, and more intense under future climate change (e.g., Mickley et al., 2004; Leibensperger et al., 2008; Wu et al., 2008). These studies, however, did not evaluate changes in the pollution episodes using methods comparable to those developed in Appendices A–B. Instead, they typically rely on statistical downscaling of proxy data related to the pollution extremes, i.e., global or regional model projections of relevant meteorological data. Unfortunately, however, these correlations are derived under present-day conditions and may not apply under a future climate scenario as the correlations often reflect common underlying drivers and thus do not represent the net effect of photochemistry, meteorology, and land-atmosphere interactions and feedbacks on O₃ (Fiore et al., 2015). Other methods quantified changes in terms of a return time of an exceedance of an absolute threshold (e.g., 75 ppb), however, it is not clear what this means for a biased model and it limits an analysis to only the most polluted locations.

Our approach as described in Appendix C exploits the statistics developed in Appendices A–B, which were designed to be independent of model bias, to evaluate future changes in O₃ using the actual O₃ abundances simulated by a subset of the ACCMIP models evaluated in Appendix B. We focused on a sensitivity experiment designed to isolate the effects of climate change from emission changes on atmospheric composition. Climate change in this experiment follows the IPCC AR5 high-CO₂ scenario (RCP8.5) with year-2000 to year-2100 surface temperature increases of ~4°C, while precursor emissions for O₃ and aerosols are fixed at present day levels.
Of the eight models evaluated in Appendix B, four performed this sensitivity experiment and archived hourly surface O₃. One of these models allowed emissions of biogenic volatile organic compounds (BVOCs) to respond to changes in climate, providing an important assessment of how this possible feedback could influence future surface O₃.

Overall, we found that climate change causes increases in mean summertime and high-percentile ozone in polluted regions, while it causes decreases in clean environments (see Fig. C.1). Even with constant biogenic emissions, climate-driven surface O₃ changes are largest at high percentiles, albeit they are even larger in the model with climate-sensitive BVOCs – especially in regions with large present-day BVOC emissions. Using the cosine curve approximation of the O₃ annual cycle, we showed that climate change shifts the timing of the seasonal surface O₃ peak to earlier in the year and increases the amplitude of the annual cycle (Fig. C.2p-r). In most regions and models, climate change increases the size of AQX episodes and cause them to contain higher O₃ abundances (i.e., their enhancement relative to the 30th percentile, see Fig. C.3g-l).

1.4 Appendix D: Co-occurrence of extremes in ozone, particulate matter, and temperature

The work of Appendices A–C was focused entirely on surface O₃ and episodes of its extremes. The meteorological conditions that favor the development of O₃ pollution episodes also consistently result in enhanced abundances of particulate matter. Indeed, stagnation events are the only meteorological variable found to consistently result in increased levels of both
pollutants (Jacob and Winner, 2009). These stagnation events are also conducive to the development of heat waves.

The connection between surface O₃ and temperature is highly documented in the literature, as temperature is considered to be the single most important variable affecting surface O₃ (Bloomer et al., 2009). Studies on the relationship between temperature and particulate matter are less numerous due to either lack of available data or because the direct effect of temperature on particulate matter differs depending on the chemical makeup of the aerosol. These extreme types (ozone, particulate matter, and temperature) each have their own individual effects on human health, however, there is ample evidence that suggests that when they occur in coincidence, their health effects are nonlinearly amplified beyond the sum of their individual effects (Dear et al., 2005; Stafoggia et al., 2008; Basu, 2009; Ren et al., 2008; Li et al., 2014; Willers et al., 2016).

The need to study the coincidence of temperature and air pollution extremes was brought to the forefront of the climate change community following the extreme pollution episode and heat wave over Europe in 2003 that resulted in thousands of deaths. The 2003 European heat wave is sometimes considered a prototype for that expected under future climate warming, and thus it is imperative to characterize the present-day nature of their co-occurrence to construct a benchmark upon which to gauge future changes. Using the objective interpolation algorithm and the statistical definition of extreme events and their connectedness, Appendix D is the first study to investigate the co-occurrence of the three health extremes in a consistent framework. The climatology of statistics of overlap and lag and space and time are developed with the intention of providing a basis for impact studies on human health and agriculture, chemistry-climate model evaluation, and process oriented analysis of their underlying mechanisms.
The observational dataset of surface O₃ in Appendix D was extended in time to cover years 1999–2013. A commensurate dataset of gridded particulate matter abundances was developed using the interpolation method developed in Appendix A and station measurements of fine particulate matter (PM₂.₅) from monitoring networks in the Eastern US and Canada. Reanalysis data of maximum daily temperature and the meteorological variables needed to define an air stagnation index (daily average surface and 500 mb wind speeds and daily cumulative precipitation) were also used.

With the longer 15-year dataset, the trends in O₃ and particulate matter due to precursor emission reductions are more apparent than those in Appendices A-C (see Fig. D.1e) and required consideration when choosing an extreme definition. Instead of defining the events as the 150 worst days over the entire 15-year period, they are defined as the 50 worst days in three equal 5-year periods. This approach is a compromise between minimizing the effect of the trends and maintaining as much information on interannual variability as possible. The analysis was limited to the extended summer season (April–September) since the majority of O₃ and temperature extremes occur during the warm season. This definition equates to a ~94.5 percentile threshold, which despite being lower than that previous used (i.e., ~97.3 in Appendices A–C), is roughly equivalent since most O₃ extremes occur during April–September.

Overall, like that previously found for surface O₃, extremes in PM₂.₅ and temperature most often occur in connected, coherent, multi-day structures on spatial scales >1000 km. Over our Eastern North American domain, roughly a third of any two of the extreme types co-occur on average,
with many locations in the Northeast US having co-occurrence frequencies of greater than 50% (see Fig. D.3). The statistics of overlap and offsets in space and time were found to contradict simple mechanistic arguments in some locations, providing motivation for process-oriented future studies. Another major finding is that also like O₃, the largest, longest lasting episodes of extreme PM₂.₅ and temperature contain pollution levels and temperatures elevated substantially beyond the 95th percentile extreme definition threshold, i.e., by +7 ppb for O₃, +6 μg m⁻³ for PM₂.₅, and +1.7 °C for daily maximum temperature. The findings of Appendix D demonstrate the need to evaluate these extremes in a multi-stressor framework, especially given the indication of their synergistic health impacts and the likelihood they will worsen under future climate change.

1.5 Discussion and future directions

This dissertation represents a major step forward in the analysis of extreme pollution episodes and their coincident occurrence with heat waves. By creating an objective interpolation algorithm that produces a comparable quantity to that of what a global model predicts, model-measurement comparison studies can be greatly improved in their commensurability and thus their consistency. The development of a definition of extremes in terms of a historical percentile exceedance and the clustering algorithm that quantifies their spatiotemporal connectedness provides new diagnostics to evaluate both present-day and future simulations of extreme air pollution in global chemistry models. Additionally, by showing even a highly biased model can accurately hindcast extreme pollution episodes, there is the possibility for the development of early public health warning systems by forcing chemistry transport models like the UCI CTM
with forecast meteorology. The realization that the largest episodes consistently contain the
highest levels of pollution demonstrates the importance of accurately quantifying potential future
changes in the worst pollution episodes. Furthermore, this dissertation demonstrates the clear
statistical overlap and coincidence of extreme levels of ozone, particulate matter, and
temperature, and thus provides a basis for impact studies on human health and vegetation as well
as additional diagnostic tests for the models that are employed for quantifying their potential
future changes. Future work will harness the methods and metrics developed here and work
towards a process-level understanding of the mechanisms that generate these extreme episodes.
Ultimately, the work presented here can be exploited to improve model simulations of extreme
air quality episodes and heat waves such that the posterity of human health can be preserved.
Chapter 2

The UCI CTM surface ozone diurnal cycle

2.1 Introduction

The diurnal cycle of surface ozone (O₃) is driven by sunlight, meteorology (e.g., temperature and boundary layer mixing), surface deposition, and the daily cycle of precursor emissions. Accurately reproducing the diurnal cycle in a global chemistry model is thus predicated on its ability to simulate these high frequency processes.

The work of Schnell et al. (2015; Appendix B) revealed that the University of California, Irvine Chemical Transport Model (UCI CTM) had a major problem with simulating the observed summertime (JJA) surface ozone diurnal cycle over the relatively polluted regions of Eastern North America (ENA) and Southern Europe (SEU). Aside from being biased high during all hours of the day and all months of the year, the most aberrant feature of the diurnal cycle was its
shape, namely a too early and too rapid morning $O_3$ increase resulting in an inaccurate timing of when the highest $O_3$ abundances occur.

The goal of this chapter is to diagnose and remedy the mistiming and bias in the UCI CTM diurnal cycle using a newly implemented aerosol scheme. The results of subsequent sensitivity simulations (with and without aerosols) revealed the diurnal cycle problem might actually be caused by a timing shift in the driving meteorological fields. As such, we perform an additional simulation that attempts to account for the offset.

### 2.2 Diurnal cycle in the surface observations

The diurnal cycle of the June-July-August, 2005 gridded surface $O_3$ observations over a section of ENA and SEU (100 $1^\circ \times 1^\circ$ grid cells) is shown in Fig. 2.1a-b, respectively (black). *Schnell et al.* (2015; Appendix B) quantified the shape of the diurnal cycle by fitting a cosine curve to the average hourly values from the grid cells in each region, from which the hour of maximum phase ($h$) and the peak-to-peak amplitude ($H$) were derived. The appeal of this approach for model-measurement comparison is that it provides objective and continuous measure of the diurnal cycle, avoiding subjective and ambiguous results in the case of flat or multiple maxima. The triangles in Fig. 2.1a-b are plotted at $(x, y) = (h, H)$.

The lowest observed $O_3$ abundances (~20 ppb) in ENA and SEU occur at around 06:00 local time (LT), following nighttime $O_3$ loss through surface deposition and NO titration and prior to the initiation of photochemistry at sunrise. Surface $O_3$ abundances increase throughout the day
Figure 2.1. Diurnal cycles of hourly abundances (ppb) of surface (a-b) O₃, (c-d) NO, (e-f) NO₂, (g-h) N₂O₅, and (i-j) HNO₃ averaged over June-August, 2005 for the observations (black), the UCI CTM simulations without aerosols: T42 (cyan) and T159 (SAD_N, green), with aerosols (SAD_Y, blue), and with aerosols and adjusted meteorological fields (SAD_Y_3, red) over ENA (left column) and SEU (right column). Triangles in (a-b) show the observations’ and UCI CTM’s cosine fit derived values of the hour of maximum phase $h$ and peak-to-peak amplitude $H$ plotted at $(x, y) = (h, H)$. 
and reach a maximum (~55 ppb) at around 15:00 LT; thus $h$ and $H$ are roughly 15 hours and 35 ppb, respectively. The diurnal cycles of the T42 (~2.8°) UCI CTM simulation in which the problem with the diurnal cycle was first found (Schnell et al., 2015; Appendix B) (cyan) begins the morning rise about an hour earlier than the observations. Following, surface O3 abundances rapidly increase to a maximum (~130 ppb) by around 10:00 LT with an hour of maximum phase $h$ of 12:00 LT, a full three hours earlier than the observations.

Based on the mechanisms responsible for the shape of the diurnal cycle, Schnell et al. (2015; Appendix B) suggested that the cause for the UCI CTM’s rapid early morning O3 production was due to excess NOx in the early morning boundary layer, which was not sufficiently removed over the night. The nighttime loss of NOx occurs through a number of processes including titration of O3, surface deposition, and uptake by aerosols. The version of the UCI CTM in Schnell et al. (2015; Appendix B) did not include aerosol chemistry, and so as a first step, we have implemented a climatology for aerosol surface area densities, included the aerosols in the photolysis calculation, and added three heterogeneous reactions. For this work, we use a higher resolution of the UCI CTM (T159, ~1.1°) in order to eliminate the possibility that the problem was caused by dilution of point-source precursor emissions over a model grid cell.

2.3 An aerosol climatology for photolysis and heterogeneous chemistry

The UCI CTM has never implemented a prognostic aerosol package, i.e., one that would take the emissions of aerosols and their precursors (e.g., SO2, NH3, black carbon, mineral dust) and predict the abundance of sulfates, nitrates, carbonaceous, and mineral dust aerosols. Although
aerosols are clearly known to impact gas-phase chemistry (the core of the UCI CTM), a good aerosol-chemistry package is as complex and difficult to implement and proof as the gas-phase chemistry for CH₄ and O₃. For a recent application and documentation of the current UCI CTM (without aerosols) see Holmes et al. (2014).

This section describes the addition of aerosol chemistry. We have implemented a monthly mean, 3-D climatology for aerosols surface area density (SAD, cm² cm⁻³) derived from the NCAR CAM simulations made for the last IPCC (Intergovernmental Panel on Climate Change) Fifth Assessment Report (AR5), see background papers by Tie et al. (2005), Lamarque et al. (2012), and Tilmes et al. (2015). We thank S. Tilmes for support and for supplying the file aero_1.9x2.5_L26_2000-2009.nc, which contains the monthly mean, 3-D aerosol abundances (kg kg⁻¹) for 14 aerosol bins (CB1, CB2, DST01, DST02, DST03, DST04, OC1, OC2, SO4, SOA, SSLT01, SSLT02, SSLT03, SSLT04) on the native grid of the CAM used (1.9° x 2.5° x 26 levels). A full set of aerosol types and size distributions were supplied and the physical properties of the aerosol components (Tie et al., 2005; Lamarque et al., 2012) along with the mean temperatures in each level were used to integrate the SAD and calculate the density a cm² of aerosols per cm³ of air. In this case the surface area used is 4πr² for gas-aerosol collisions and not the cross-sectional area πr² used in optical cross sections. From the aero_1.9x2.5_L26_2000-2009.nc data set we created a single monthly 3-D climatology of total aerosol SAD (cm² cm⁻³), regridded to a 1° x 1° geographic grid and to standard 1-km levels in pressure altitude from 0-1 km to 23-24 km. The new data set is named SAD_12m_1x1.nc.
Heterogeneous chemical reactions between gas and condensed phase material in the atmosphere are assumed to be only three major gas-aerosol reactions (R1-R3) affecting NOx and HOx levels.

\[(R1) \quad \text{N}_2\text{O}_5 + \text{Aerosol} \rightarrow 2\text{HNO}_3 + \text{Aerosol}\]

\[(R2) \quad \text{NO}_3 + \text{Aerosol} \rightarrow \text{HNO}_3 + \text{Aerosol}\]

\[(R3) \quad \text{HO}_2 + \text{Aerosol} \rightarrow \text{Aerosol}\]

The rate of the reaction is greatly simplified and calculated only from a sticking probability \(\alpha\) (i.e., the number of collisions with the aerosol that result in a reaction). The formula used in the chemical package is:

\[(\text{Eq. 1}) \quad R_X = \alpha_X \times \sqrt{29/\text{MolWt}_X} \times 675 \times \sqrt{T} \times [X] \times [\text{Aerosol SAD}] \quad (# \text{ cm}^{-3} \text{ s}^{-1})\]

where the sticking probability is taken from JPL-2015 \((\text{Burkholder et al.}, 2015)\) and the molecular weight of \(X\) is used to scale the mean velocity. The primary impact of these reactions is to reduce NOx and HOx levels, especially in polluted (high aerosol) air masses. These reactions should reduce the early morning NOx abundances by removing the nighttime NOx reservoir species \(\text{N}_2\text{O}_5\) and \(\text{NO}_3\) over the course of the night. Reactions R1 and R2 are largely limited to nighttime chemistry as \(\text{N}_2\text{O}_5\) and \(\text{NO}_3\) undergo rapid photolysis during the day.

The other major impact of aerosols on chemistry is to scatter and absorb sunlight, thereby reducing photolysis rates (J-values) in the lower troposphere. In fact, Fast-J was designed to calculate the impact of clouds and aerosols on J-values \((\text{Bian et al.}, 2003; \text{Prather, 2015})\).
Because the aerosol climatology is for surface area and we can reasonably assume that most of these aerosols are optically active, the choice of aerosol density and effective radius have only a small impact on the optical properties. For example, if we were given aerosol mass (kg cm\(^{-3}\)) then the aerosol optical depth would scale inversely with density and effective radius (\(R_{\text{eff}}\)). Here the choices of these two parameters along with the refractive index affect the scattering phase function and single scattering albedo. Based on reported values for single scattering albedo (\(\omega = 0.90 \text{–} 0.94\)) and composition, we chose to split the SAD into equal parts of mineral dust (\(R_{\text{eff}} = 0.15 \mu\text{m}, \omega = 0.94\)) and a fossil fuel aerosol with some back carbon (\(R_{\text{eff}} = 0.14 \mu\text{m}, \omega = 0.90\)). For both aerosols the ratio of optical cross section to SAD is about 0.5.

Overall, in this version of the UCI CTM, the effect of including aerosols in heterogeneous chemistry and photolysis is to reduce both production and loss of tropospheric O\(_3\) globally by about 14% each, and to reduce the loss of CH\(_4\) by 17%. Photolysis and chemical rate changes contribute almost equally to these reductions.

### 2.4 Effect of including aerosols on the diurnal cycles of O\(_3\) and NO\(_x\)

The effect of the added aerosol scheme on the diurnal cycle of O\(_3\) and related NO\(_x\) species are summarized in Fig. 2.1 (green = without aerosols = SAD\(_N\), blue = with aerosols = SAD\(_Y\)). As expected, the rapid early morning rise in surface O\(_3\) found in the T42 simulation persists in T159 simulation without aerosols (SAD\(_N\)), and thus we conclude that resolution is not the cause for the aberrant diurnal cycle.
For the simulation that includes SAD (SAD_Y), we find that overall it has only a minimal effect on the surface O₃ diurnal cycle in both regions. Surface O₃ abundances are decreased by only about 3 to 4 ppb during daylight hours and there is essentially no change in timing or the slope of the early morning rise. For reactive nitrogen species, there is either little discernible change (NO) or changes are almost solely limited to nighttime hours (i.e., between 18:00 and 06:00 LT; vertical dashed lines in Fig. 2.1). For NO₂, the SAD_Y simulation has lower abundances throughout the night, with a deficit of about 1 ppb by 06:00 LT. As expected, decreases are also found for N₂O₅ during nighttime hours (0.1 ppb). The decreases in N₂O₅ are largely a result of R1 (and less so for R2), which also cause increased HNO₃ abundances at night.

Table 1 provides the average summertime and annual (only 2006 to allow for ample spin up time) tropospheric (surface to ~200 mb) burden of the species in Fig. 2.1 as well as CH₄ and CO over ENA, SEU, and the globe. These large-scale results generally show what is expected from the added reactions and the surface diurnal cycles in Figure 2.1. Modest decreases (~5%) are found for O₃, HO₂, NO, and NO₂, while much larger decreases (60%) are found for N₂O₅. Changes in HNO₃ are ±1% depending on the time frame (summer or annual) and the region considered. Based solely on R1 and R2, one might expect HNO₃ to increase; however, although HNO₃ increases at night, it decreases during the day (Fig. 1i-j) due to less available NO₂ for the reaction NO₂ + OH \rightarrow HNO₃.
2.4 Boundary layer height and meteorological fields

Unfortunately, the inclusion of SAD did not remedy the problem of the O3 diurnal cycle in the UCI CTM. We then evaluated the diurnal cycle of the height of the boundary layer (BL) on the suspicion it was too shallow during the early morning, causing O3 and its precursors to rapidly accumulate in the lowest model layers. The UCI CTM is forced by 3-hour meteorological fields from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecast System (IFS). The fields are not interpolated in time so the height of the BL in a model grid cell is the same for three consecutive hours. A typical diurnal cycle of the BL height should be very shallow over the night until it rapidly rises the early morning until a maximum is reached sometime during mid-afternoon.
Fig. 2.2 shows the diurnal cycle of the BL height (m) over ENA (blue) and SEU (green) averaged over June-July-August, 2005. As expected, the nighttime BL is very shallow (50 m); however, it remains shallow until around 09:00 LT in SEU and 10:00 LT in ENA. This one-hour difference reflects the timing of sunrise between the two regions relative to the 3-hour blocks of meteorological data. The timing of the dramatic increase in BL height coincides with the peak of the UCI CTM’s O3 diurnal cycle (Fig 2.1a-b, cyan, green, and blue). Once the height of the BL increases, the O3 produced and trapped at the surface is mixed vertically and the abundances drop over the remainder of the day. The height of the BL should increase shortly after sunrise (~07:00 LT), much earlier than that found. While this BL cycle is clearly off by about 3 hours (results are the same for Australia and the Sahara, not shown), we have not identified the cause and thus correction for the meteorological data.

As a simple fix, we performed a third sensitivity simulation that shifts the time used in the solar flux calculations behind by three hours (SAD_Y_3). This has no effect on the diurnal cycle of the BL height, but instead causes sunrise and thus photochemistry to begin three hours later so as to coincide with the lifting of the nocturnal BL. While this is neither an elegant or permanent fix, it should provide a simple test to see if the timing offset of the BL is responsible for the dramatic early morning rise.

The results for the SAD_Y_3 simulation are shown with red lines in Fig. 2.1 (red). Over ENA, the early morning rise begins as expected three hours later (08:00 LT), now ~2 hours later the observations. The derived phase $h$ is also three hours later than the previous UCI CTM simulations and is now offset forward compared to observations by about +2 hours. The peak-to-peak amplitude $H$ has also decreased by about 15 ppb, but it is still biased high by about 25 ppb.
Better agreement is found with the observations over SEU, where the UCI CTM misses the timing of \( h \) by +1 hour and the magnitude of \( H \) by about 15 ppb. Table 1 also provides the average summertime and annual tropospheric burdens for the SAD_Y_3 simulation. As expected, the burdens in the SAD_Y_3 simulation are comparable to the SAD_Y simulation, but with small increases in O₃ and small decreases in NO over both ENA and SEU.

These results clearly demonstrate that the root of the problem lies with the timing of the meteorological fields. For the most part, we have corrected the UCI CTM to more closely match the observed cycle in surface O₃ and to fit within the range of the other ACCMIP models. However, this simulation does not address the root of the problem and we are in contact with the U. Oslo scientists who generate these meteorological data and hope to resolve the issue.

**Figure 2.2.** Diurnal cycle of hourly boundary layer height (meters) in the UCI CTM averaged over June-July-August, 2005 for ENA (blue) and SEU (green).
2.4 Status and path forward to publication

Previous work (*Schnell et al.*, 2015; Appendix B) revealed that the UCI CTM had major problems with reproducing the observed phase and amplitude of the summertime diurnal cycle of surface O$_3$ in the polluted regions of Eastern North America and Southern Europe. The reason for the fault was hypothesized to be an excess of NO$_x$ in the early morning boundary layer, which would lead to an overestimation of O$_3$ production as compared to observations. To decrease the early morning NO$_x$ abundances, aerosol chemistry was implemented within the model; however, it had only a minimal effect on the resultant surface O$_3$ diurnal cycle. Upon further investigation, the underlying cause of the fault was found to be with the meteorology the model is forced by, namely a timing mismatch that caused the nocturnal boundary layer to persist until late morning (~10:00 LT) that led to rapid accumulation of O$_3$ in the lowest models layers. As a simple, albeit neither sophisticated nor permanent fix, the time used in the solar flux calculations was adjusted to account for the timing mismatch. This simulation brought the surface O$_3$ diurnal cycle of the UCI CTM much closer to reality. While the source of the error has been identified, it is not yet clear what steps need to be taken to remedy it, but is the subject of an ongoing investigation.

The work of this chapter has thus far only diagnosed a major fault with the UCI CTM; and although it represents a major improvement within the model, this chapter currently lacks substantive new science required for subsequent publication. However, another issue that has not yet been addressed is the diurnal cycle of ozone precursor emissions, namely NO$_x$ from transportation and biogenic volatile organic compounds such as isoprene. Currently, emissions in
the UCI CTM are provided on a monthly basis such that each time step within that month has the same fraction of the total emissions. In reality, however, traffic NOx emissions have a diurnal (and weekly) cycle that peaks in the morning and afternoon and is at a minimum at night.

It is unclear how a diurnal emission cycle would affect the UCI CTM’s O3 diurnal cycle, but it should be included for the most realistic simulation. Indeed, including a diurnal emission cycle could potentially make the diurnal cycle problem in the UCI CTM worse in terms of both phase and magnitude. For example, if morning transportation NOx emissions were increased, the rise in O3 may occur earlier, much like that of the simulations without the adjusted meteorological fields. It could also increase the overall bias by emitting more NOx during the day when it could produce more O3 and emitting less NOx at night for titration. On the other hand, it could reduce the bias in extremely polluted regions by enhancing daytime titration. The path to publication is envisioned as incorporating these diurnal emission cycles within the UCI CTM to quantify their effects on the O3 diurnal cycle and a recommendation for how modeling groups might implement them in the future.
References


Rieder, H. E., A. M. Fiore, L. W. Horowitz, and V. Naik (2015), Projecting policy-relevant metrics for high summertime ozone pollution events over the eastern United States due to


Appendix A

Skill in forecasting extreme ozone pollution episodes with a global atmospheric chemistry model

Reprint of:

Skill in forecasting extreme ozone pollution episodes with a global atmospheric chemistry model

J. L. Schnell1, C. D. Holmes1, A. Jangam1,*, and M. J. Prather1

1Department of Earth System Science, University of California, Irvine, CA 92697-3100, USA
*now at: Department of Integrated Environmental Science, Bethune-Cookman University, Daytona Beach, FL, USA

Correspondence to: J. L. Schnell (jschnell@uci.edu)

Received: 11 February 2014 – Published in Atmos. Chem. Phys. Discuss.: 10 March 2014
Revised: 9 June 2014 – Accepted: 24 June 2014 – Published: 4 August 2014

Abstract. From the ensemble of stations that monitor surface air quality over the United States and Europe, we identify extreme ozone pollution events and find that they occur predominantly in clustered, multiday episodes with spatial extents of more than 1000 km. Such scales are amenable to forecasting with current global atmospheric chemistry models. We develop an objective mapping algorithm that uses the heterogeneous observations of the individual surface sites to calculate surface ozone averaged over 1° by 1° grid cells, matching the resolution of a global model. Air quality extreme (AQX) events are identified locally as statistical extremes of the ozone climatology and not as air quality exceedances. With the University of California, Irvine chemistry-transport model (UCI CTM) we find there is skill in hindcasting these extreme episodes, and thus identify a new diagnostic using global chemistry–climate models (CCMs) to identify changes in the characteristics of extreme pollution episodes in a warming climate.

1 Introduction

Links between climate change, global atmospheric chemistry, and air pollution are noted in early climate–chemistry studies and have come to the forefront recently (e.g., Jacob et al., 1993; Johnson et al., 1999; Prather et al., 2001; Jacob and Winner, 2009; HTAP, 2010; Fiore et al., 2012; Kirtman et al., 2013). Some studies indicate that climate change may increase the intensity, duration, or frequency of ozone (O3) pollution episodes (Mickley et al., 2004; Leibensberger et al., 2008; Jacob and Winner, 2009). Future changes in air quality are undoubtedly driven foremost by changes in local emissions, and then by distant emissions, land-use change, and climate change (e.g., Steiner et al., 2006; Meleux et al., 2007; Tao et al., 2007; Lin et al., 2008a; Wu et al., 2008; Zhang et al., 2008; Doherty et al., 2009; Carlton et al., 2010; HTAP, 2010; Steiner et al., 2010; Tai et al., 2010; Hoyle et al., 2011; Lei et al., 2012; Wild et al., 2012; Stocker et al., 2013).

With climate change, several factors may affect local pollution: changing meteorological conditions, shifting background atmospheric composition, and chemistry–climate interactions that control the efficacy or residence time of pollutants. All of these factors may alter the efficiency of local emissions in generating pollution events (Weaver et al., 2009) and need systematic evaluation. Thus, global chemistry–climate models (CCMs) are a necessary component in projecting future air quality on a continental scale (Lamarque et al., 2012; Kirtman et al., 2013). Here, we provide an approach that can evaluate CCMs in terms of their ability to match this new observed climatology of ozone pollution, one that specifically examines how climate change might alter the meteorological conditions that create the multiday, large-scale extreme ozone episodes found in the US and Europe (EU) today (e.g., Barnes and Fiore, 2013).

Even at their best typical resolution (~1° ≈ 100 km), current global chemistry models are known to have high biases in their production of global tropospheric ozone from pollution (Wild and Prather, 2006). This high bias in production extends to surface ozone on a continental scale (e.g., Nolte et al., 2008; Appel et al., 2012; Lamarque et al., 2012; Ras mussen et al., 2012), although in one case the bias is negligible (Mao et al., 2013). These chemistry-transport models (CTMs) or CCMs also have serious limitations in modeling peak ozone levels (Dawson et al., 2008). The use of such
global models for air quality projections is seen as being limited until such errors are accurately diagnosed and corrected (Fiore et al., 2009; Murzakazi and Hess, 2006; Reidmiller et al., 2009). There is a need for observation-based tests of the ability of atmospheric chemistry models to simulate pollution episodes over the time- and space scales possible in a global model. In this study, we develop such diagnostics, specifically a grid-average climatology of daily surface ozone concentrations, with a focus on CTMs that should be able to simulate past events (hindcasts) using a meteorology representative of the time of the observations (e.g., ERA-Interim or GEOS MERRA). The goal is to characterize statistical errors and systematic biases in the hindcast and to provide clear metrics that can document improvements in the model.

Observations of surface \( \text{O}_3 \) from monitoring stations provide the basis for testing models, but measurements at individual stations are generally not representative of model grid cells (Valari and Menut, 2008; Dennis et al., 2010). This problem is referred to as “incommensurability” or “change of support” (Gelfand et al., 2001; Swall and Foley, 2009) and prevents ready quantitative assessment of model errors. If station observations are used to generate an observed ozone product that is directly comparable to what a model predicts, viz. the average \( \text{O}_3 \) concentration in a grid cell, then geographic patterns and statistics of the pollution episodes can be readily and commensurably tested. In Sect. 2, we present our new algorithm for mapping the individual station data onto cell averages on a regular grid. As part of this analysis we generate an objective measure, the quality of prediction \( (Q^\text{P}) \), for the mapping of each cell (i.e., how many independent points were used and how far away they are). This grid-cell product has the added advantage of allowing direct and commensurate comparison of independent sets of overlapping but not collocated observing sites, and we examine the biases between the two European ozone networks (European Monitoring and Evaluation Programme (EMEP) and AirBase) for both clean and polluted periods. This assessment uses a full decade of observations (2000–2009) from three networks (Environmental Protection Agency (EPA) over the US).

In Sect. 3, we compare the maximum daily 8h average (MDA8) grid-average observations over the US and Europe with the University of California, Irvine chemistry-transport model (UCI CTM)-simulated values for years 2005–2006. The model errors are diagnosed in terms of location, time of year, and pollution level by comparing different percentiles at each grid cell while maintaining exact-day matches (concurrent sampling) over the 2 years. Simple comparison of high- and low-end statistics of the ozone distribution is found to be misleading. In Sect. 4 we define extreme pollution events for each grid cell in a climatological sense, as the 100 worst days (i.e., highest MDA8 concentrations) in a decade (~97.3 percentile) or the 20 worst days in 2 years when comparing the observations to the UCI CTM. We then identify the structure of the multiday, continental-scale pollution episodes that make up most of these events. The CTM’s ability to match these extreme episodes is shown to have considerable skill, which degrades as the quality of prediction of the cell decreases and as random noise is added to the observations. In Sect. 5, we develop statistics of the extreme events from a decade of observations that can be used without hindcasting to compare with free-running chemistry–climate models. Using clustering algorithms, we define the size in space and time of the episodes and the fraction of all events that occur within large clusters. In Sect. 6 we conclude and discuss how to use the current climate archive Coupled Model Intercomparison Project Phase 5/Atmospheric Chemistry and Climate Model Intercomparison Project (CMIP5/ACCMIP), or to design the next-generation chemistry–climate simulations, to assess climate-driven changes in extreme ozone pollution episodes.

2 Observations of surface \( \text{O}_3 \) over the US and EU

For our observations of surface \( \text{O}_3 \) we use 10 years (2000–2009) of hourly surface \( \text{O}_3 \) measurements from air quality networks in the United States and Europe (see Table 1 for summary of data sets). For the US we primarily use the EPA’s Air Quality System (AQS). The EPA’s Clean Air Status and Trends Network (CASTNET) is used for independent evaluation as described in Sect. 2.3. For EU we combine EMEP (Hjellbrekke et al., 2013) and the European Environment Agency’s AirBase network except in Sect. 2.4, where we compare these two independent but overlapping data sets. The AirBase data set includes information on the zoning type of the stations (e.g., rural, suburban, urban, traffic), and we choose to use all but the traffic stations for the most complete and representative data, a decision corroborated by Pirovano et al. (2012). The hourly measurements from EMEP and AirBase are reported as \( \text{µg m}^{-3} \) and are converted to parts per billion (ppb = \( 10^{-9} \) mol mol\(^{-1} \) = nmol mol\(^{-1} \)) using a temperature of 20°C; essentially mass concentrations are multiplied by 0.5 ppb ug\(^{-1} \) m\(^3\).

From these data sets we calculate the maximum daily 8h average \( \text{O}_3 \) concentration (MDA8), which is the primary air quality standard for the US (www.epa.gov/air/criteria.html) and is commonly used in human and agricultural health studies (Chen and Wu, 2005; Bell et al., 2006) and climate studies (e.g., Tagirits et al., 2007). We calculate the MDA8 by beginning the 8h averaging period at 24:00 LT and calculating 17 8h averages for each day, picking the maximum of those 17 (i.e., the averaging only considers windows that fully reside within 1 day). Thus the maximum can occur during different 8h intervals at adjacent sites or on consecutive days at the same station, although afternoon and early-evening maxima are most common (Bruntz et al., 1974). The location of the stations and their 10-year mean MDA8 surface \( \text{O}_3 \) concentrations are shown in Fig. 1.
Table 1. Observational data sets.

<table>
<thead>
<tr>
<th>Surface ozone network</th>
<th>Period</th>
<th>No. stations</th>
<th>URL or reference</th>
</tr>
</thead>
</table>

*CASTNET stations are used only as a validation data set and are not included in the interpolation over the US.

2.1 Choosing a method for interpolating grid-cell averages

We develop an interpolation scheme that provides grid-cell-average values of surface O₃ over the US and EU domains, essential to compare observations to a gridded model. Our goal is to use all representative station data, recognizing the heterogeneity of surface O₃ that must be averaged over to compare with gridded model simulations. The most commonly used technique used to compare observations with a gridded model is to simply average all observing sites within the grid cells to be compared (e.g., Fiore et al., 2002). This results in an incomplete domain as well as the calculated averages disproportionately representing urban stations, especially in areas where exceedances are likely to occur. Fiore et al. (2003) accounts for the clustering of urban stations by first averaging the station observations on a finer grid (0.5° x 0.5°) and then averages those cells to match the coarser model grid. In any case, Diem (2003) notes that almost all ozone-mapping methods have major problems and that this is neither a simple nor a solved task. The task here is very different from that of interpolating spatial extremes to infer regions of O₃ exceedance (e.g., Cooley et al., 2007; Padoan et al., 2010).

Inverse distance weighting (IDW) and ordinary Kriging are the most common interpolation techniques, with generally small or modest differences found between the two (Rojas-Avellaneda and Silvan-Cardenas, 2006). Both produce estimates at unmeasured points using a weighted linear combination of the values at neighboring sites, determined by some function of the separation between the unmeasured point and observation sites. The difference is that the weights in Kriging are formulated to minimize the variance in the estimated values (error) using a predefined model of the spatial covariance of the data, while the weights in IDW are determined without specific need for the covariance function.

Kriging is often favored as it provides prediction error estimates and incorporates a declustering mechanism designed to account for data redundancy, effectively treating highly clustered data more like a single site (Wackernagel, 2003). Since many observation sites in the US and EU data sets are located in close proximity to one another, some form of...
declustering is desired in our interpolation. Isaaks and Srivastava (1989) note that, when the effect of data clustering is accounted for in IDW, the advantages of using Kriging are slight. In addition, the covariance function required for Kriging can easily be modeled incorrectly, especially at short separation distances (Diem, 2003), when many sites are close in geographic space but their reported values differ by a large amount, as in the case of air pollution. Many of the geographically clustered sites in our data sets are located in urban areas associated with high variability, so the covariance function could easily be incorrectly modeled at short separation distances. Consequently, the Kriging weights given to these clustered stations would not necessarily provide the desired declustering. For this reason, we use a modified form of IDW that incorporates a declustering scheme without the need to model the underlying covariance function.

From O₃ observations Z_k at sites x_k, we interpolate the O₃ mole fraction at an unobserved location x as a weighted sum of the observations

\[ Z(x) = \sum_{k=1}^{K} w_k \cdot Z_k \]

(1)

where K is the number of observations sites and weights w_k are defined as follows. In standard inverse-distance weighting \( w_k = |x - x_k|^{-\beta} \), with \( \beta \) typically in the range \( 1 \leq \beta \leq 4 \). We optimize \( \beta \) as described below after adjusting the weights for distant and clustered observations. Weights are set to zero when \( |x - x_k| \) exceeds a threshold \( L \) to avoid meaningless small contributions from distant sites. We choose \( L = 500 \) km based on the typical scale of synoptic meteorology that influences surface O₃ and test other choices below.

We also reduce the weights of clustered stations, which tend to lie in urban areas, to avoid excessive influence of the cluster on surrounding rural regions and to avoid the shielding effect whereby an observation site screens all those that are located immediately behind it (Falke, 1999). The weight of each observation site is reduced by a factor \( M_k \) that is the number of other observation sites located within a distance \( D \) of site \( k \). We choose \( D = 25 \) km as a typical size scale for urban areas and test other choices below. Furthermore, all observation sites within the region \( |x - x_k| < D \) are given equal weight to avoid singularities in the interpolation. Taken together, the weights in Eq. (1) are

\[ w_k = \begin{cases} 
D^{-\beta} / M_k & \text{if } |x - x_k| < D \\
|x - x_k|^{-\beta} / M_k & \text{if } D \leq |x - x_k| \leq L \\
0 & \text{if } |x - x_k| > L 
\end{cases} \]

(2)

If the sum of the weights for point \( x \) from sites \( k \) is zero, a null value is given to that point. Our interpolation algorithm calculates values at points for a single day using only measurements from that day. Implementation of spatiotemporal interpolation is complex, with no specific implementation well agreed upon for applications to air quality data

Figure 2. RMSE (ppb) for the mean value of each 10th percentile of interpolated sites and grid cells, sorted by \( Q^D \).
shared with 4 grid cells, and the 12-edge points shared with 2 cells. The trapezoidal integration weights account for latitudinal variation of the points. Thus the weight $w^*_i$ of each point $x_i$ for $i = 1:25$ in the grid cell $X$ is

$$w^*_i = T_i \cos \theta_i,$$  \tag{3}

where $\theta_i$ is the latitude and $T_i$ is the trapezoidal integration weight, which takes values of 0.25 for corner points, 0.5 for edge points, and 1.0 for the interior points: The calculation of the average ozone value at the grid cell $X$, $\langle \bar{Z}(X) \rangle$, is then the weighted sum of ozone at points $x_i$, $Z_i$:

$$\bar{Z}(X) = \sum_{i=1}^{25} w^*_i \cdot Z_i,$$  \tag{4}

We do not report $\langle \bar{Z}(X) \rangle$ for grid boxes where over half of the interior points $Z(x_i)$ are zero.

### 2.2 Quality of prediction and the interpolation mask

The interpolation procedure should be limited to the region being modeled and where a reliable prediction can be made. We begin with a desired mask of $1^\circ \times 1^\circ$ cells and then check if the interpolation is adequate. For the US, we use the landmass of the contiguous states (CONUS) and include ocean cells adjacent to CONUS. For EU we draw a similar mask but also include areas in the North Sea and in the Mediterranean Sea west of Italy. We then calculate a measure of the quality of prediction, $Q^P$, for the points within this desired mask to determine the final grid masks for the US and EU. We define $Q^P$ as the effective number of independent stations at a distance of 100 km that went into the interpolation.

$$Q^P = 100^\beta \sum_{k=1}^{K} w_k,$$  \tag{5}

Thus, for $\beta = 2.5$, one station at 50 km or less distance counts as 5.7 stations, and one at 200 km counts as 0.18 stations. Grid-cell-average $Q^P$ values are calculated in the same manner as the average O$_3$ in Eq. (4). The observing sites do not always provide continuous daily data for the decade 2000–2009, and thus the numbers of sites that go into the daily interpolation of each grid cell may vary. In order to keep the masking consistent over the period, it is based on the location of all observing sites, effectively the largest possible $Q^P$ values over the time period. The declustering weighting for each site, $M_k$, is recomputed on a daily basis.

The $Q^P$ values reflect the ability of the observing network to predict O$_3$: the highest (lowest) $Q^P$ values have the smallest (largest) RMSE (Fig. 2). Using this relationship and with the intent of providing as nearly contiguous a grid for EU and the US as possible, we select the value of $Q^P = 0.67$ as the cutoff for our masks. Figure 1 shows the constructed masks (gray boxes) for the EPA (Fig. 1a) and combined EU (Fig. 1b). When comparing the EU observations with the UCI CTM, we truncate the mask northward of 65° N. Note that the mask over the US excludes parts of Montana that are too distant from sites. Figure S1 in the Supplement shows the logarithm of $Q^P$ values for all of the retained grid cells for the US and EU. The lowest $Q^P$ values for our US mask (apart from the coasts) are found from west-central Texas and north, due to the low density of observing sites in this area. The lowest values in EU are found in the northernmost and easternmost edges of the domain for the same reason.

### 2.3 Interpolation error

The error of our interpolation method can be objectively measured for the individual sites as described in Sect. 2.1. The average RMSE for the sites can be plotted as a function of our estimate of the quality of the interpolation ($Q^P$) as shown in Fig. 2. For large values of $Q^P$ the RMSE levels off at about 6 ppb. This is a measure of the small-scale, nearest-neighbor variability in ozone that is simply not resolved by our interpolation. Our analysis does show that the RMSE begins to increase when $Q^P$ falls below about 30 (effective number of independent sites at a distance of 100 km). Note that the lowest $Q^P$ value for the US is about 3, because the sites tend to be located near one another. Thus $Q^P$ is a measure of error in interpolation.

Deriving an error for the interpolated grid-cell-average values is more difficult since we have no objective measure of the cell-averaged ozone values. Clearly the minimum RMSE of 6 ppb for individual sites is an exaggeration of the error when averaging over a $1^\circ$ grid cell ($\sim 10^4 \text{ km}^2$). Using the error analysis done for the sites (removing randomly 10% of the sites), we can examine how the cell-average values change relative to standard result using the full set of sites. The RMSE for this case is also plotted in Fig. 2. It provides a measure of the error in the cell-average ozone, but is at best a lower bound. The RMSE remains small, at about 1 ppb or less, for $Q^P = 0.7$ to 100 and increases to 2 ppb for $Q^P = 0.33$. It is encouraging that relative error estimates can be made and that our cutoff of $Q^P = 0.67$ is a good choice. Note that this approach does not inform us about extrapolation error arising from, e.g., gradients near the coasts. Results for both the US and EU are similar, and the range of $Q^P$ is much larger than the site-error analysis because we are trying to interpolate cells that are distant from sites.

With the daily MDA8 O$_3$ values interpolated, we can begin to analyze the results for each domain. Figure 3 shows a sample day of grid-cell ($1^\circ \times 1^\circ$) average MDA8 O$_3$ values based on the observing sites in the northeastern US. Note the variegated nature of O$_3$ at individual sites within some $1^\circ \times 1^\circ$ cells. The $Q^P$ values for three sample cells are noted in the figure caption. Cell A has a large number of independent sites in surrounding cells; hence the $Q^P$ is very high despite only a few stations within the cell. Cell B has lower quality because the stations are more distant and located mostly in one
direction. This is even more pronounced for cell C on the edge of the domain.

Figure 4 shows the gridded, masked MDA8 ozone concentrations for both the US (Fig. 4a, c, e) and combined EU (Fig. 4b, d, f) data sets for two representative percentiles, the 95th (Fig. 4a–b) and 25th (Fig. 4c–d), and their differences (95th minus 25th, Fig. 4e–f). The percentiles here are calculated with respect to years 2005 and 2006, since these are to be compared with the CTM hindcast. The highest 95th-percentile values (~70 ppb) occur in California and then in a broad swath from Texas to New England. For EU they lie mostly around the Mediterranean. The lowest 95th percentiles occur in the northern latitudes for both the US and EU. The 25th percentile represents clean air, typically in winter, and here the largest concentrations (~40 ppb) in the US occur over the Rocky Mountains and the plains to the east, while for the EU ozone concentrations greater than 30 ppb are found only at the southern extent of the mask. Note that Greece and southern Italy stand out as maximal in both percentiles. The difference, 95th minus 25th percentile, is a measure of the pollution buildup, and it tends to follow the regions of largest emissions. California, the Midwest, and the eastern seaboard have the greatest differences in the US (>40 ppb), while in EU the greatest differences are concentrated in central countries (e.g., France, Germany, northern Italy).

2.4 Comparison of overlapping observational O3 networks

The grid-cell-average O3 MDA8 product developed here provides a ready comparison of the two independent but overlapping networks, for which individual adjacent stations are not available. For the comparison, we calculate QP values for each data set and apply a mask using a cutoff of 0.33 rather than 0.67 in order to examine a larger area. We define the bias as AirBase minus EMEP and present biases for the 25th, 50th, and 95th percentiles calculated with respect to years 2000–2009 (Fig. 5). Note that these comparisons are not exact-day matches, and hence each percentile may correspond to a different day. The AirBase data set is mostly biased low over all three percentiles, with greatest differences (below -10 ppb) for the 25th percentile in Alpine regions. In this case the area-weighted mean bias (MB) is -3.9 ± 3.1 ppb. After investigating the average altitude of stations for each network, we found this bias is possibly reflecting preferential station placement, as the mean altitude bias in the region of northern Italy and southern France is about -540 m (i.e., EMEP stations are chosen to reflect
independent sampling. This mask includes only grid cells with a $O_3$ by NOx emissions and then the cumulative production of $O_3$ as polluted air disperses. While AQS sites are generally in or near populated areas, AirBase sites are located in rural areas while AirBase sites are generally in or near populated areas.

We also present the difference between the interpolation using only AQS data compared to using only CASTNET data in Fig. S2 in the Supplement. We present the bias (= AQS minus CASTNET) for the 25th, 50th, and 95th percentiles calculated using independent sampling with respect to years 2000–2009. For the comparison, we calculate $Q_P$ values for each data set and apply a mask using a cutoff of 0.10 rather than 0.67 to examine a larger area. In addition, this value of $Q_P$ corresponds to having one station at a distance of 250 km (i.e., the station is representative of a $\sim 5^\circ \times 5^\circ$ grid cell). This figure shows that the AQS interpolation is systematically lower than the CASTNET one for almost all locations and percentiles, particularly over California and from the central plains east to New York City. The bias is least for the most polluted times (95th percentile). Similar to the EMEP–AirBase comparison, CASTNET sites are located in rural areas while AQS sites are generally in or near populated areas, and thus we believe this difference is due to the titration of $O_3$ by NOx emissions and then the cumulative production of $O_3$ as polluted air disperses.

Overall, these comparisons show excellent agreement across the networks, particularly in the high-$O_3$ events. Further comparisons of the AirBase and EMEP networks and the AQS and CASTNET networks could use a smaller mask with higher-quality score and focus on exact-day matches (concurrent sampling) as we do with the CTM hindcasts below.

3 UCI CTM simulation of years 2005–2006

We use the gridded daily $O_3$ observations described above to evaluate the UCI CTM. This model is a tropospheric CTM driven by meteorology from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecast System. The model is configured as described by Tang and Prather (2010, 2012a, b). Simulations are 1$^\circ \times 1^\circ$ resolution with 40 vertical layers, which is amongst the highest resolution for current global chemistry models, and covers 2005–2006, which is the duration of the high-resolution meteorological fields. The lowest model layer is about 80 m thick, and we use that layer-mean value as the surface $O_3$ concentration. MDA8 values are calculated from hourly simulated mole fractions in the same way as the observations.

As noted above, the MDA8 most often occurs during the afternoon, which coincides with periods of a deep convective boundary layer and avoids problems with the poorly modeled nighttime boundary layer (Lin et al., 2008b; Lin and McElroy, 2010). The present model configuration was designed for studies of stratosphere–troposphere exchange, rather than for surface air quality analysis. As a result, emissions are specified monthly, based on the Quantifying the Climate Impact of Global and European Transport Systems (QUANTIFY) inventory (Hoor et al., 2009), and do not account for daily, weekly, or monthly cycles. Because the surface $O_3$ simulation has not been optimized, the CTM performance described below may be similar to chemistry–climate models that are used for present to future scenarios.

3.1 Evaluating the central tendency of $O_3$ in models

Many global chemistry models, including the UCI CTM, predict surface $O_3$ concentrations that are higher than observations (Dawson et al., 2008; Nolte et al., 2008; Zanis et al., 2011; Appel et al., 2012; Lamarque et al., 2012; Rasmussen et al., 2012). The CTM grid-cell $O_3$ averaged over years 2005–2006 is larger than observed everywhere for both US and EU, in both summer and winter (see Fig. 6; Table S2 in...
Summer typically has the days of highest O₃ percentile. The pattern gives a level of detail that helps us identify possible sources of model error.

The winter domain model bias of the average O₃ (MB = CTM minus observation (OBS), Fig. 6a–b) is $19 \pm 6$ ppb (standard deviation across the grid cells) for the US and $18 \pm 5$ ppb for EU. The high-latitude background air (northern EU, upper Midwest US) has only a small bias ($5–15$ ppb), but air coming in from the mid-latitude oceans (east and west coast US, southern EU) has a higher bias ($20–30$ ppb) and extends beyond just polluted regions. The winter domain model correlation coefficient (MCC) derived from the daily time series of MDA8, shown in Fig. 6e–f, shows relatively good model hindcasting with an average MCC of $0.47 \pm 0.13$ for the US and $0.61 \pm 0.10$ for EU. MCC is greatest for the most part where $Q^p$ is large and lowest in coastal areas. For wintertime, most of the variability is driven synoptically by large-scale gradients in background O₃.

The summer domain average MB (Fig. 6c–d) is larger than in winter: $+30 \pm 14$ ppb for the US and $+20 \pm 8$ ppb for EU. Here the largest biases are often in polluted regions, like the Los Angeles basin and the Chicago-to-New York corridor, and the easternmost part of the EU domain. This pattern indicates exaggerated photochemical production of O₃ in the model, possibly a consequence of NOₓ plumes being spread over the $100$ km model grid or other nonlinear interactions involving hydrocarbons and NOₓ (Lin et al., 2008b; Pusede and Cohen, 2012; Rasmussen et al., 2012). Supporting this hypothesis, the model’s summertime bias for the US has a similar pattern to our measure of pollution buildup (95th minus 25th percentile, Fig. 4c, the two maps have a correlation coefficient, $r = 0.66$). For EU, this conclusion is less obvious (Fig. 4f, $r = 0.20$). In terms of MCC, the verisimilitude of the model hindcast of daily summertime pollution is quite good (Fig. 6g–h) because in this case the variability is driven synoptically by buildup of regional pollution: MCC $= 0.60 \pm 0.16$ for the US and $0.55 \pm 0.19$ for EU. In addition, the bias for each month of the year at three representative percentiles (84th, 50th, and 16th) can be derived from Table S2 in the Supplement.

### 3.2 Developing objective measures of model biases

While evaluation of the central tendency of a model provides an important test and can be used to identify bias in either hindcasts or climate simulations, it is the distribution of extremes, both high and low, that we want our climate models to simulate accurately. The lows tell us about baseline (clean-air) O₃, and the highs show the efficiency of O₃ production from the local emissions. Here we examine the distribution of MDA8, combining the daily gridded US and EU values for a season over the 2 years 2005–2006 from both observations and the CTM hindcast. The probability distribution functions (PDFs) for winter (DJF) and summer (JJA) months are shown in Fig. 7. The observations, sorted into percentile bins ($0–5\%$, $5–10\%$, etc.) calculated separately for each grid cell and plotted relative to the median, are shown in red; the CTM values, sorted independently of the observations, are in blue; and the CTM values sorted according the observed percentiles (concurrent sampling) are in green. For concurrent sampling, the CTM values are averaged for exact-day matches for each day and location of the observations that fall in that percentile bin. In a perfect model, the green and red curves would match, meaning that the CTM predicts changes relative to the median at the right time and place. The blue curve treats the CTM effectively like a climate simulation and does not try to locate the high-O₃ periods over the correct cells at the correct time. Because the CTM hindcast has errors, the sorting by observed percentiles will always result in a shallower curve, which may not even be monotonic.
To determine if air quality extreme (AQX) events involving high O$_3$ concentrations are changing with climate, we must be able to characterize those AQX events observed today and demonstrate that global chemistry models can reproduce them. As demonstrated for the UCI CTM above, surface O$_3$ concentrations in global chemistry models are often biased high, with higher biases often occurring during peak pollution episodes, but there is skill in hindcasting pollution variability. These biases hinder the ability to predict AQX events based strictly on absolute concentrations (Dawson et al., 2008; Nolte et al., 2008; Zanis et al., 2011).

We define AQX events based on the local PDF of O$_3$ concentrations, rather than based on exceeding a concentration threshold. This enables us to identify linked extreme events whose absolute magnitudes evolve over space and time. For example, Fig. 8 shows daily MDA8 O$_3$ for June 2002 in four grid cells in the Midwest and eastern US (Chicago, IL; Cincinnati, OH; New York, NY; and rural Virginia). The time series are highly correlated across these sites, but the peak magnitudes differ across sites. In Chicago, MDA8 values above 67 ppb exceed the local 97.3 percentile and frequently occur a few days before local maxima in New York and Virginia, due to west-to-east motion of weather systems. If extremes were identified based on an absolute threshold (e.g., 75 ppb), then the peak values in Chicago might not be labeled as extremes, and their connection to extremes in the eastern US might be overlooked.

4.1 Defining individual, grid-cell level ozone pollution extremes

We define the threshold value for AQX events as a frequency (return time) based on the local climatology. This is shown in Fig. 8 by the colored arrows, which are the ~97.3 percentiles, or the 100 worst days in a decade (2000–2009) for each site. This threshold varies from 68 to 78 ppb for these four grid cells, and filled circles denote the AQX events at each site. For comparison with the UCI CTM hindcast, we take the 20 worst days in years 2005–2006. Thus, over the 2 years, both CTM and observations have 20 AQX events in each grid cell. This definition of AQX highlights times at each grid cell when O$_3$ pollution is at its highest, generally when the effect of nearby precursor emissions is exacerbated by meteorology. Indeed, Lei et al. (2012) highlight the need to explore this type of method (i.e., exceedance of historical extremes) to determine their relationship to climate change. Unfortunately, by defining AQX in terms of frequency, we are unable to test for climate change impacts in terms of the number of such events alone, and must search for a suitable diagnostic that characterizes the scale and structure of large AQX episodes (see Sect. 5).

The choice of 10 days per year (upper 2.7%) instead of 20 days per year (upper 5.4%) or another number is somewhat...
arbitrary, and such choices can have undesirable results in some cases (e.g., Coles, 2001). While the top 2.7 % of O3 MDA8 may seem extreme, most of these events occur during the summer, and hence the AQX events are essentially the upper 10 % of summer days. In general, the wider the range for defining an extreme event, the easier it will be for the model to simulate.

4.2 Skill of the CTM

We define the skill of the CTM for each grid cell as the percentage of events that match the day of the observed AQX events. With this definition a random model is expected to correctly identify 2.7 % of events. This metric does not take into account the geographic pattern or persistence of AQX, for which we apply clustering algorithms (see Sect. 4.4). Skill here is calculated over all months of both years (2005–2006), although almost all AQX events occur from May to September.

Figure 9 shows the geographic pattern of CTM skill for US and EU domains. For the US it is 24.4 ± 12 % (standard deviation across grid cells) and a min-to-max range of 0 to 65 % for the grid cells (Fig. 9a). The CTM skill was slightly better for EU: 32.2 ± 17 % (Fig. 9b). For the wider AQX threshold of 94.5th percentile, the skill increases as expected and the standard deviation is reduced: 35.6 ± 11 % for the US and 37.5 ± 14 % for EU. While CTM skill at individual grid cells in the US shows no distinct pattern, that in EU shows a strong east–west trend, with significantly higher skill to the west. These patterns of skill are evident for both threshold choices with correlations ($R^2$) between them of 0.86 for the US and 0.87 for EU. The east–west gradient in EU, as well as the lack of pattern in the US, can partly be understood from the relationship between skill and $QP$. Low CTM skill is caused by model errors as well as errors in observations and interpolation. As shown in Fig. S5 in the Supplement, the CTM skill is largest in grid cells with large $QP$ and small interpolation errors.

4.3 Organized episodes of AQX events

The AQX events often occur as clustered, multiday episodes with spatial extents of more than 1000 km (note that an event is a single identified AQX event and an episode is a grouping of AQX events). Figure 10 shows an example of one of the larger episodes of the 2005–2006 period for EU, 3–8 July 2006. The episode, although not completely shown, is one of the largest observed, with a size of $1500 \times 10^4$ km$^2$-days, and also the largest in the CTM hindcast, at $1700 \times 10^4$ km$^2$-days ($10^4$ km$^2$ is our basic areal unit since our grid resolution is 1°). The skill of the CTM on these 6 days was 75.4 %, with both data sets showing the episode’s structure and trajectory. These extreme events are connected in space–time and can be reproduced in a hindcast by a global model. These attributes provide an opportunity to develop a climatology of extreme ozone episodes (e.g., areal extent, duration, intensity, seasonal cycles) that can be used as metrics to test global chemistry climate models’ (GCMs) future climate simulations.

The size of the largest AQX episodes (defining an episode as connected events as in Fig. 10) is driven by a combination of meteorology as well as regionally connected emissions and active photochemistry. To objectively identify these episodes we use an agglomerative hierarchical cluster analysis. Ideally, the clustering algorithm will connect AQX events occurring within a large, slow-moving, stagnant, high-pressure system over several days. Locations and times of AQX events are provided to the clustering algorithm, which then groups them into clusters that we call AQX episodes. The linkage criteria that define the clusters are flexible, and
we choose AQX events to be clustered if they are within a predefined cutoff in both space and time. We use the Chebyshev (maximum coordinate difference) distance metric and the single (nearest-neighbor) linkage criterion. We prescribe a cutoff value of 1 (i.e., events are not connected at greater than 1° and 1 day ahead or behind). We recognize two obvious limitations to using this linkage method: (1) we have essentially considered time as another dimension in space (i.e., 1° = 1 day), and (2) geographic distance between two grid cells varies with latitude and is not accounted for in the clustering. We consider the former to be of no consequence since a time separation cutoff of less than 1 day is not possible using daily MDA8 values to identify AQX events. Also, a larger cutoff value would be unfavorable since events could be statistically linked even if they occurred at the same grid cell and were separated by a full day. We avoid problems associated with latitudinal variations by developing statistical measures that are independent of resolution (see Sect. 5.2).

Since we want to characterize AQX episodes by their size, effectively a measure of their areal extent (km²) and duration (days), the robustness of the clustering algorithm, particularly the linkage across days, needs to be examined. Most episodes showed a progression of area vs. time that resembled a normal distribution. Occasionally episodes resemble

Figure 9. Skill of the CTM (i.e., percentage of events identified in the observations that were correctly reproduced in the CTM) at each grid cell for the (a) US and (b) EU for years 2005–2006. Domain mean skill and 1σ are shown for each plot.

Figure 10. Six days (3–8 August 2006) of a large AQX episode in EU. Left column is the observations and right column is the CTM.
a multi-peaked or bimodal distribution. In our first algorithm these bimodal episodes were counted as a larger, single episode, but human discernment identifies them as two different episodes adjoined by only a small number of events. Our revised algorithm defines a cutoff in order to separate these extreme episodes. For each episode identified with the primary algorithm, we calculate the area of the events for each day and the area of events that are shared with the previous day (i.e., the same grid cell on 2 consecutive days). If the ratio of the shared area divided by the total area of that previous day is less than 0.10, we truncate the episode at the previous day. We do not apply this secondary algorithm to the first 2 or last 2 days of an episode, to provide flexibility for formation and dissipation. In addition, this detaching can occur more than once as we change alters these extreme episodes.

5 Developing climatologies

The grid-cell-average statistics for MDA8 developed here provide a climatology of surface O₃ that can be used to test and evaluate CCMs. This approach holds promise given that one global CTM has skill in hindcasting specific years and events in spite of some large systematic errors in surface O₃ abundance. Here we seek to develop climate records for surface O₃ over the US and EU that can be used to improve both CTMs and CCMs and to develop confidence in CCM projections of changing air quality in a warming climate. First, we develop statistics for the basic cycles of O₃ over a week, a season, and a year, using a decade of observations (Sect. 5.1). These statistics present a useful climatology for testing the means and perhaps standard deviations (see Chang and Hanna (2004) for more examples), but extreme high- and low-probability events are not so useful as a climatology (Sect. 3.2). The characterization of AQX events as large-scale, multiday episodes is investigated with clustering algorithms (Sect. 5.2), and we develop climate statistics on the scale of these episodes as a new data set to evaluate CCMs (Sect. 5.3) and opening a novel test of whether climate change alters these extreme episodes.

5.1 Weekly and annual cycles

The well-known weekly and annual cycles (Bruntz et al., 1974) in MDA8 O₃ concentrations are summarized for our decadal data sets in Table 2, where we combine typical measures (16th, 50th, 86th percentiles in ppb) with AQX frequencies (based on 100 per decade). Higher percentiles are of interest, but then the geographic patterns need to be examined. The table gives an average over the entire domain (US or EU), and the results for each grid cell or region can be derived from the supplementary data, but are not shown here. The day-of-the-week and month-of-the-year statistics include a decade of observations (years 2000–2009). The direct comparison with the CTM, for weekly and annual cycles using only statistic from years 2005–2006, is in the supplementary material (Table S2 in the Supplement) and shows excellent agreement, except for the weekly cycle, an expected result (see below).

For Table 2, the annual cycle of the number of AQX events in the US follows a normal distribution with most events in spite of some large systematic errors in surface O₃ concentration (ppb) corresponding to the 84th, 50th, and 16th percentiles for each month of the year and day of the week for the week of the 2000–2009 observations in the US and EU. The 84th- and 16th-percentile values are given relative to the 50th percentile. Correlation coefficients ($R^2$) are defined with respect to the number of AQX events per month of the year or day of the week.

Table 2. Domain mean number of air quality extreme events (AQX) defined for the grid-cell interpolated MDA8 O₃ series and the MDA8 O₃ concentration (ppb) corresponding to the 84th, 50th, and 16th percentiles for each month of the year and day of the week for the 2000–2009 observations in the US and EU. The 84th- and 16th-percentile values are given relative to the 50th percentile. Correlation coefficients ($R^2$) are defined with respect to the number of AQX events per month of the year or day of the week.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>US AQX</td>
<td>9.1</td>
<td>8.3</td>
<td>9.9</td>
<td>10.2</td>
<td>10.6</td>
<td>10.9</td>
<td>11.0</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O₃ 84%</td>
<td>+13.4</td>
<td>+13.8</td>
<td>+14.4</td>
<td>+14.7</td>
<td>+14.5</td>
<td>+14.2</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O₃ 50%</td>
<td>39.4</td>
<td>39.5</td>
<td>39.4</td>
<td>39.5</td>
<td>39.5</td>
<td>40.1</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O₃ 16%</td>
<td>-11.6</td>
<td>-11.5</td>
<td>-12.0</td>
<td>-12.2</td>
<td>-12.2</td>
<td>-11.9</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU AQX</td>
<td>9.9</td>
<td>8.7</td>
<td>9.2</td>
<td>9.9</td>
<td>10.9</td>
<td>10.5</td>
<td>11.0</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O₃ 84%</td>
<td>+12.2</td>
<td>+12.6</td>
<td>+12.8</td>
<td>+13.3</td>
<td>+13.5</td>
<td>+13.5</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O₃ 50%</td>
<td>35.8</td>
<td>34.6</td>
<td>34.5</td>
<td>34.6</td>
<td>34.4</td>
<td>35.3</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O₃ 16%</td>
<td>-10.5</td>
<td>-10.8</td>
<td>-11.3</td>
<td>-11.4</td>
<td>-11.5</td>
<td>-11.3</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Atmos. Chem. Phys., 14, 7721–7739, 2014
events identified in June, while in EU the cycle is slightly weighted towards spring months. Similar patterns are seen in the 84th- and 50th-percentile values, while the highest values in the 16th percentile are slightly weighted towards the spring. These MDA8 values corresponding to these percentiles show excellent agreement with the monthly AQX frequencies. For the 2005–2006 case (Table S2 in the Supplement), July dominates in the EU observations due to the 2006 summer having 14 out of 20 of the events, while in the CTM June had the most, with 2006 having slightly less events than the observations at 12 out of 20 events.

The weekly cycle is also evident in both observational data sets. The largest values of AQX events, the 84th percentile, and the 50th percentile, generally occur at the end of the week (Friday, Saturday, Sunday), a phenomenon termed “the weekend effect” with lower values in the beginning of the week (Cleveland et al., 1974; Karl, 1978; Tonse et al., 2008; Pierce et al., 2010). For the 16th percentile, the trend is less obvious. The 84th-percentile values show excellent agreement with the day-of-week AQX frequencies. As expected, we did not see significant evidence of a weekly cycle in the CTM, as there is not a parameterization for the day of the week within the model. The mean skill of the CTM was generally higher for months and days that had higher combined values in the 84th- and 50th-percentile values, while the highest values in the 16th percentile generally were 84 and 67 %, respectively; for the CTM, the corresponding fractions are 66% greater than 100 \times 10^4 \text{km}^2\text{-days} and about 31 % in episodes greater than 1000 \times 10^4 \text{km}^2\text{-days}; for the CTM, the corresponding fractions are 66% greater than 100 \times 10^4 \text{km}^2\text{-days} and 37 % greater than 1000 \times 10^4 \text{km}^2\text{-days} (Fig. 11a). For years 2005–2006 and gridded EU observations, about 74 % of all events occurred in episodes greater than 100 \times 10^4 \text{km}^2\text{-days} and about 31 % in episodes greater than 1000 \times 10^4 \text{km}^2\text{-days}; for the CTM, the corresponding fractions are 66% greater than 100 \times 10^4 \text{km}^2\text{-days} and 37 % greater than 1000 \times 10^4 \text{km}^2\text{-days} (Fig. 11a).

The sensitivity of these diagnostics to grid resolution needs to be determined as we have differing resolution across CCMs and the climatology is a useful model diagnostic only if it is robust across different model resolutions. We create a 2° \times 2° data set (typical of CCM resolution) using simple means of the MDA8 concentrations from the 1° \times 1° observational data set. AQX events and episodes are defined as

\[
\bar{S} = \exp \left( \frac{\sum_{i=1}^{n} (S_i \cdot \ln S_i)}{\sum_{i=1}^{n} S_i} \right),
\]

where \(n\) is the number of episodes and \(S_i\) is the size of the episode. Equation (6) was chosen over the simple arithmetic mean to reduce the influence of the numerous small episodes while giving more weight to larger episodes.

The majority of AQX events are grouped into large-area, multiday clusters that we define as AQX episodes. The complementary cumulative distribution function (CCDF = 1 minus cumulative distribution function) of the percentage of the total area of all events as a function of episode size is shown in Fig. 11. For years 2005–2006 and gridded US observations, about 74 % of all events occurred in episodes greater than 100 \times 10^4 \text{km}^2\text{-days} and about 31 % in episodes greater than 1000 \times 10^4 \text{km}^2\text{-days}; for the CTM, the corresponding fractions are 66% greater than 100 \times 10^4 \text{km}^2\text{-days} and 37 % greater than 1000 \times 10^4 \text{km}^2\text{-days} (Fig. 11a). For years 2005–2006 and gridded EU observations the fractions are 84 and 67 %, respectively; for the CTM, the fractions are 73 and 42 %, respectively (Fig. 11b). In EU, the events are clustered into larger-size episodes.

Figure 11 also shows that the decadal climatology (years 2000–2009) of episode sizes (green) is quite different from the 2 yr climatology (blue) that overlaps with the CTM hindcast. Thus, interannual variability is an important factor that must be considered, but interannual variability is also an important diagnostic that provides a key test for the CCMs as well as a metric that can help assess the significance of changes between two different decades. This is especially evident when each year’s individual CCDF is examined (see Fig. S6 in the Supplement). In addition to climate variability in AQX episodes, there is the problem of stationarity in the observations due primarily to continuing mitigation of emissions. For the US, a clear pattern of decreasing episode sizes for successive years in the decade can be seen, consistent with reductions in precursor emissions. For EU, this pattern is less apparent; however the standout features are the CCDFs for 2003 and 2006, which have much larger episodes than other years. The annual number of AQX events and (\(\bar{S}\)) values support this conclusion, as seen in Table 3.

The sensitivity of these diagnostics to grid resolution needs to be determined as we have differing resolution across CCMs and the climatology is a useful model diagnostic only if it is robust across different model resolutions. We create a 2° \times 2° data set (typical of CCM resolution) using simple means of the MDA8 concentrations from the 1° \times 1° observational data set. AQX events and episodes are defined as.

www.atmos-chem-phys.net/14/7721/2014/
Table 3. Climatology of O₃ air quality and extreme episodes (AQX) observations over the US and EU (2000–2009). Each grid cell has AQX events defined as the 100 worst days per decade, except for AQXyr, which is normalized to have 10 events per year. The mean AQX size (\(\bar{S}\)) for the 10 events-per-year case is computed from Eq. (6) after the clustering algorithm that couples nearest neighbors and successive days, with units of 10⁴ km squared days (km²d), where 10⁴ km² is about a 1⁰ × 1⁰ grid cell. Average summertime (JJA) MDA8 O₃ (ppb) from the grid-interpolated data (grid) is area weighted, but the station average (station) is raw with all stations equally weighted. The mean (\(\mu\)) and standard deviation (\(\sigma\)) of the annual values over the decade are given. Correlation coefficients (\(R^2\)) are defined with respect to the number of AQX events per year. Using the stations’ redundancy weightings derived here gives a slightly greater \(R^2\), but still less than that for the gridded O₃.

<table>
<thead>
<tr>
<th>Unit</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>(\mu \pm \sigma)</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>AQX events</td>
<td>No.</td>
<td>13.5</td>
<td>11.5</td>
<td>16.5</td>
<td>15.0</td>
<td>4.6</td>
<td>11.2</td>
<td>13.3</td>
<td>8.1</td>
<td>4.6</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>((\bar{S}))</td>
<td>10⁴ km²d</td>
<td>618</td>
<td>373</td>
<td>1239</td>
<td>581</td>
<td>82</td>
<td>435</td>
<td>515</td>
<td>186</td>
<td>70</td>
<td>32 413±363</td>
</tr>
<tr>
<td>AQXyr</td>
<td>events</td>
<td>No.</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0±0.0</td>
</tr>
<tr>
<td></td>
<td>((\bar{S}))</td>
<td>10⁴ km²d</td>
<td>264</td>
<td>295</td>
<td>337</td>
<td>276</td>
<td>217</td>
<td>329</td>
<td>222</td>
<td>232</td>
<td>208</td>
<td>199 256±50</td>
</tr>
<tr>
<td>O₃ (grid)</td>
<td>ppb</td>
<td>49.3</td>
<td>49.4</td>
<td>51.4</td>
<td>50.1</td>
<td>45.5</td>
<td>48.8</td>
<td>50.7</td>
<td>47.5</td>
<td>46.2</td>
<td>43.7</td>
<td>48.3±2.4</td>
</tr>
<tr>
<td>O₃ (station)</td>
<td>ppb</td>
<td>51.3</td>
<td>52.1</td>
<td>55.0</td>
<td>51.0</td>
<td>46.9</td>
<td>50.8</td>
<td>52.0</td>
<td>50.1</td>
<td>48.8</td>
<td>45.0</td>
<td>50.3±2.8</td>
</tr>
<tr>
<td>EU</td>
<td>AQX events</td>
<td>No.</td>
<td>7.4</td>
<td>8.3</td>
<td>11.0</td>
<td>19.9</td>
<td>10.0</td>
<td>8.2</td>
<td>16.5</td>
<td>6.0</td>
<td>8.3</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>((\bar{S}))</td>
<td>10⁴ km²d</td>
<td>280</td>
<td>502</td>
<td>187</td>
<td>793</td>
<td>415</td>
<td>287</td>
<td>2528</td>
<td>210</td>
<td>240</td>
<td>140 558±718</td>
</tr>
<tr>
<td>AQXyr</td>
<td>events</td>
<td>No.</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0±0.0</td>
</tr>
<tr>
<td></td>
<td>((\bar{S}))</td>
<td>10⁴ km²d</td>
<td>388</td>
<td>419</td>
<td>237</td>
<td>446</td>
<td>404</td>
<td>319</td>
<td>1149</td>
<td>437</td>
<td>305</td>
<td>367 447±255</td>
</tr>
<tr>
<td>O₃ (grid)</td>
<td>ppb</td>
<td>40.1</td>
<td>41.7</td>
<td>44.3</td>
<td>47.3</td>
<td>42.7</td>
<td>41.4</td>
<td>45.2</td>
<td>41.2</td>
<td>41.5</td>
<td>40.0</td>
<td>43.4±2.3</td>
</tr>
<tr>
<td>O₃ (station)</td>
<td>ppb</td>
<td>43.5</td>
<td>46.6</td>
<td>45.7</td>
<td>54.9</td>
<td>45</td>
<td>45.1</td>
<td>49.5</td>
<td>43.5</td>
<td>44.1</td>
<td>44.6</td>
<td>46.2±3.5</td>
</tr>
</tbody>
</table>

5.3 Developing climate statistics of AQX episodes

The episode size distributions in Fig. S6 in the Supplement show clear differences across the years; however we need an objective measure of these differences. The Anderson–Darling (AD) test (Anderson and Darling, 1952) compares two CDFs (equivalently CCDFs) and gives a confidence level that they occur from the same underlying and unknown distribution (the AD null hypothesis). The AD test is non-parametric, distribution free, does not require normality, and it is more sensitive to differences in the tails of the distribution than the widely used Kolmogorov–Smirnoff test (Engmann and Cousineau, 2011). We compare the distributions in Fig. 11 for episodes larger than 10 × 10⁴ km²-days (10 to 16 connected grid cells) since we are mostly interested in the largest episodes and, further, more than 90% of the events are in episodes of size greater than this. For the US, the CTM hindcast was found to be statistically different (\(p<0.05\)) from the observations, while for EU both distributions are the same (\(p<0.05\)).

By defining AQX events as the 100 worst days per decade, we can quantify interannual variability in the number of events or large episodes per year. If we wish to ascertain whether individual years have differences in their pollution episodes in terms of areal extent or duration, then the events need to be renormalized (i.e., 10 worst days per year). In the 100-per-decade case, those years with more events will more likely have bigger episodes, with all else being equal. This can easily be seen by the CCDFs in Fig. S6 in the Supplement and the (\(\bar{S}\)) values in Table 3. Even when each year is forced to have the same number of events, the CCDFs for each of the years are not similar (see Fig. S7 in the Supplement). Using these renormalized AQX episode size distributions, we test if we can statistically identify “good” and “bad” years (based on row one of Table 3) by comparing the individual years to one another. The AD test shows that, in EU, year 2006 (a relatively bad year) was statistically different from several years (2000, 2001, 2002, 2004, 2005) at the 95% confidence level and 2009 at the 90% level. For the US, the year 2009 (good) was found to be statistically different (\(p<0.05\)) from the year 2005 (bad); at the 90% level, the year 2005 was also found to be different from years 2000 and 2003. The tests can also be performed on the distributions of areal extent. For example, the year 2006 in EU was once again found to be statistically different (\(p<0.05\)) than the years listed above for the distributions of areal extent. At the 90% level, it was different from all years except 2007. Finally, the mean episode size (Table 3, denoted \(\bar{S}\)yr for the 10-per-year case) also varies from year to year and shows a strong agreement with the annual number of AQX events in the 100-per-decade case. This agreement provides strong evidence that the severity of a given year is largely dependent on its meteorology, since all years’ values of \(\bar{S}\)yr are derived using the same number of
Figure 11. Complementary cumulative distribution function of the percentage of the total areal extent of all individual AQX events as a function of AQX episode size (10^4 km^2-days) they are clustered into for the (a) US and (b) EU. Results are shown for the 2 yr observations at 1° and 2°, the CTM at 1°, and the 10 yr observations at 1°. Note: only latitudes <65° N were used for the 10 yr EU OBS.

6 Conclusions

In evaluating a future scenario for air quality, one can identify four major contributing factors: (1) global emissions that alter atmospheric composition and thence baseline levels (lowest percentiles) of near-surface O_3 and particulate matter (PM); (2) global changes in climate that also alter these baselines (e.g., temperature, water vapor, convection, lightning, biogenic emissions); (3) climate-driven changes in the meteorological regimes over polluted regions that lead to AQX episodes; and (4) changes in the efficacy of local emissions to generate pollution within a governance region (e.g., air quality management district, an EU country). While these factors are all part of a coupled system, an integrated model that combines all would be almost impossible to verify. Thus an assessment approach would be to evaluate each of them separately using observations and an ensemble of models (e.g., HTAP, 2010; Kirtman et al., 2013). This paper focuses on factor (3), providing clear measures of bias and skill in global chemistry models run in hindcast mode, and developing climatologies that can be used to test climate models and to detect a climatic shift in AQX episodes.

The approach developed here establishes a reliable method for gridding the air quality station observations so that direct comparison with global atmospheric chemistry models can be made. We then examine climatologies of surface ozone (percentiles, seasonality, probability distributions, AQX episodes) based on the observations and use them to test a chemistry-transport model (UCI CTM) run in hindcast mode, attempting to simulate each day’s MDA8 O_3 concentrations for the years 2005–2006. Surprisingly, we find that the often-used test of the probability distribution of MDA8 O_3 values over a region gives different results when testing a hindcast model than when treating the identical model simulation as climate statistics. Nevertheless, comparing the gridded observations directly with the hindcast MDA8 O_3 values clearly defines model deficiencies in terms of biases, baseline values (lowest percentiles at ocean boundaries), seasonality, and the ability to predict the relative increase in O_3 during high-pollution events. When used to test a chemistry–climate model, more caution is needed.

AQX events are defined here in terms of the return time of such events for each cell (i.e., as in climate extremes) rather than as an absolute O_3 threshold. Such definition clearly identifies large-scale pollution episodes associated with stagnant meteorological regimes. The AQX events (10 worst days per year = 97.3 percentile) contain a disproportionately large fraction of the excess MDA8 O_3. We test the ability the UCI CTM to hindcast the 1000 km, multiday giant AQX episodes that include most of the individual, cell-based AQX events. Although we have no formal error estimate of the gridding procedure, we feel our quality of prediction (Q^P) provides a similar quantity, as shown with both the observations themselves (Fig. 2) and with the ability of the UCI CTM to hindcast AQX events (Fig. S4 in the Supplement).
We also tested our interpolation algorithm by applying random noise to the raw station data and then recalculating the cell-average values. This analysis, although not shown, revealed the CTM’s skill did not significantly degrade until the amplitude reached ±10 ppb.

Our goal of providing observational validation of the air quality simulated by the chemistry–climate models is centered on the size and duration of AQX episodes and their interannual variability. This is a bias-free test as shown with the UCI CTM, and should be able to identify when more bad years occur in a decade under a future climate, independent of global changes in baseline levels of pollutants. Our statistics will be used to test the chemistry–climate models used in the recent IPCC assessment (CMIP5/ACCMIP).

One advantage of the approach here is that it can be readily applied to satellite observations. The regridding allows for somewhat sparse measurements resulting from day-to-day cloud obscuration to be filled to a regular grid with a measure of the quality of the prediction ($Q^P$). Our definition of AQX events takes into account natural gradients in aerosol optical depth or tropospheric ozone column.

Uncertainties and unresolved issues remain. Although $Q^P$ provides a measure of the cell-averaged data, it still lacks a formal uncertainty estimate. The decade analyzed here (2000–2009) has an apparent trend in O3 concentrations driven at least in part by reductions in precursor emissions (Turner et al., 2013). For climate statistics, this non-stationary pattern needs to be recognized and if possible corrected for. One could remove a linear trend from the station observations prior to their use in the interpolation or calculate a fit to the O3 precursor emissions over the decade and adjust the data year by year. In terms of AQX events, one could define them on a year-by-year basis and look at size only; however the absolute interannual variability over a decade remains a very important test of the models.

The Supplement related to this article is available online at doi:10.5194/acp-14-7721-2014-supplement.

Acknowledgements. This research was supported by an NSF Graduate Fellowship to J. L. Schnell; the NSF REU (1005042) support of A. Jangam; the NASA Modelling, Analysis, and Prediction Program (NNX13AL12G); the Office of Science (BER) of the US Department of Energy (DE-SC0007021); and the Kavli Chair in Earth System Science. We are grateful to the US Environmental Protection Agency’s (EPA) Air Quality System (AQS) and Clean Air Status and Trends Network (CASTNET), the European Monitoring and Evaluation Programme (EMEP), and the European Environment Agency’s (EEA) air quality database (AirBase) for providing the data sets used in this study.

Edited by: A. Pozzer

References


Atmos. Chem. Phys., 14, 7721–7739, 2014


Appendix B

Use of North American and European air quality networks to evaluate global-chemistry-climate modeling of surface ozone

Reprint of:

Use of North American and European air quality networks to evaluate global chemistry–climate modeling of surface ozone

J. L. Schnell$^1$, M. J. Prather$^1$, B. Josse$^2$, V. Naik$^3$, L. W. Horowitz$^4$, P. Cameron-Smith$^5$, D. Bergmann$^6$, G. Zeng$^6$, D. A. Plummer$^7$, K. Sudo$^{8,9}$, T. Nagashima$^{10}$, D. T. Shindell$^{11}$, G. Faluvegi$^{12}$, and S. A. Strode$^{13,14}$

$^1$Department of Earth System Science, University of California, Irvine, CA, USA
$^2$GAME/CNRM, Météo-France, CNRS – Centre National de Recherches Météorologiques, Toulouse, France
$^3$UCAR/NOAA Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Princeton, NJ, USA
$^4$Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Princeton, NJ, USA
$^5$Lawrence Livermore National Laboratory, Livermore, CA, USA
$^6$National Institute of Water and Atmospheric Research, Lauder, New Zealand
$^7$Canadian Centre for Climate Modeling and Analysis, Environment Canada, Victoria, British Columbia, Canada
$^8$Department of Earth and Environmental Science, Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan
$^9$Department of Environmental Geochemical Cycle Research, Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan
$^{10}$Center for Regional Environmental Research, National Institute for Environmental Studies, Tsukuba, Japan
$^{11}$Nicholas School of the Environment, Duke University, Durham, NC, USA
$^{12}$NASA Goddard Institute for Space Studies and Columbia Earth Institute, Columbia University, New York, NY, USA
$^{13}$NASA Goddard Space Flight Center, Greenbelt, MD, USA
$^{14}$Universities Space Research Association, Columbia, MD, USA

Correspondence to: J. L. Schnell (jschnell@uci.edu)

Received: 16 February 2015 – Published in Atmos. Chem. Phys. Discuss.: 16 April 2015
Revised: 9 September 2015 – Accepted: 9 September 2015 – Published: 25 September 2015

Abstract. We test the current generation of global chemistry–climate models in their ability to simulate observed, present-day surface ozone. Models are evaluated against hourly surface ozone from 4217 stations in North America and Europe that are averaged over $1^\circ \times 1^\circ$ grid cells, allowing commensurate model–measurement comparison. Models are generally biased high during all hours of the day and in all regions. Most models simulate the shape of regional summertime diurnal and annual cycles well, correctly matching the timing of hourly ($\sim 15:00$ local time (LT)) and monthly (mid-June) peak surface ozone abundance. The amplitude of these cycles is less successfully matched. The observed summertime diurnal range ($\sim 25$ ppb) is underestimated in all regions by about 7 ppb, and the observed seasonal range ($\sim 21$ ppb) is underestimated by about 5 ppb except in the most polluted regions, where it is overestimated by about 5 ppb. The models generally match the pattern of the observed summertime ozone enhancement, but they overestimate its magnitude in most regions. Most models capture the observed distribution of extreme episode sizes, correctly showing that about 80 % of individual extreme events occur in large-scale, multi-day episodes of more than 100 grid cells. The models also match the observed linear relationship between episode size and a measure of episode intensity, which shows increases in ozone abundance by up to 6 ppb for larger-sized episodes. We conclude that the skill of the models evaluated here provides confidence in their projections of future surface ozone.
1 Introduction

We test simulated present-day surface ozone in global chemistry–climate models on temporal scales from diurnal to multi-year variability and on statistics from median geographic patterns to the timing and size of extreme air quality episodes. The tests use gridded hourly surface ozone abundances based on a decade of observations from 4217 air quality monitoring sites in North America (NA) and Europe (EU). Chemistry-climate models provide a valuable means for projecting future air quality in a changing climate (Kirtman et al., 2013), but recent assessments have lacked commensurate observational comparisons to establish their credibility in reproducing current cycles of surface ozone over polluted regions (Young et al., 2013). Model–measurement comparisons to date have identified model faults, yet they have often been limited to monthly statistics, biased to picking clean-air sites over limited parts of the continents (Fiore et al., 2009; Doherty et al., 2013), and avoided evaluating diurnal cycles and the patterns of major pollution episodes (Schnell et al., 2014, henceforth S2014).

The factors driving future surface ozone (O₃) changes include (1) local-to-regional emissions, (2) global-scale emissions of air pollution transported across continents and oceans, (3) global emissions and physical climate change that alters the hemispheric-scale abundances of tropospheric O₃, and (4) climatic shifts in the meteorology that creates the worst pollution episodes. Factors (1), (2), and (3) have been studied extensively with global chemical transport models (CTMs) and chemistry–climate models (CCMs), and there is some agreement on model projections given an emissions scenario (e.g., Prather et al., 2003; Reimann et al., 2009; HTAP 2010; Wild et al., 2012; Doherty et al., 2013; Young et al., 2013). The importance of (4), however, lies in the recognition that air quality extremes (AQX), the worst pollution episodes in a decade, are triggered by meteorological conditions. Air quality absolute exceedances are known to occur in multi-day, spatially extensive episodes over the USA (Logan, 1989; Seinfeld et al., 1991), but it was not until the regular gridding of all station data over North America and Europe and the statistical definition of extremes in S2014 that the extent, coherence, and decadal variability of the episodes became clear. If climate change increases the duration and/or extent of the worst decadal AQX episodes, then the overall health impact of poor air quality may be worse than expected based on precursor emission changes alone (Fiore et al., 2012). A warming climate appears to increase the number of stagnation days (Horton et al., 2014) and may decrease the frequency of ventilating midlatitude cyclones (e.g., Mckeeley et al., 2004), but it is unclear how these meteorological indices relate to surface O₃ or particulate matter, especially with respect to the worst AQX episodes as identified in S2014.

The models in the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP; Lamarque et al., 2013) were used in the recent assessment of the Intergovernmental Panel on Climate Change (IPCC; Kirtman et al., 2013) and represent the most advanced attempt to simulate global surface O₃ in a future climate. However, in order to place any confidence in their projections, their ability to simulate the observed, present-day surface O₃ climatology must be evaluated. In this paper we present the first such model–measurement comparisons, specifically addressing (4) by applying the methodologies from S2014 to the current generation of CCMs in an effort to quantify their ability to simulate the decadal statistics of the AQX episodes. Due to the complexity and nonlinearity of the underlying processes, accurately simulating surface O₃ over both clean and polluted environments is a formidable task for global models with resolutions of 100 km at best. For example, it has been shown that choices in the parameterization of surface deposition can shift modeled surface O₃ levels by 10 ppb or more (Val Martin et al., 2014). Moreover, there are new, phenologically based land-surface models for interactions between atmospheric chemistry and the biosphere (Bueker et al., 2012) that have yet to be fully implemented in global models. In any case, both recent and future land-use change is expected to impact surface O₃ abundances (Ganzeveld et al., 2010). Thus, we recognize that this model–measurement comparison is just one of the first steps in evaluating global model simulations of surface O₃ pollution. A summary of the observational and model data sets as well as a brief overview of the methods developed in S2014, and used here, is presented in Sect. 2. Model–measurement comparisons are presented in Sect. 3 with concluding remarks and further discussion in Sect. 4.

2 Data and methods

2.1 Observations of surface O₃

We use 10 years (2000–2009) of hourly surface O₃ measurements from air quality networks in NA and EU. Following S2014, in NA we use 1633 stations from the US Environmental Protection Agency’s (EPA) Air Quality System (AQS) and also increase the spatial coverage in NA by including 92 stations from the US EPA’s Clean Air Status and Trends Network (CASTNet) and 207 stations from Environment Canada’s National Air Pollution Surveillance Program (NAPS). The data sets used for EU remain the same as S2014: 2123 stations from the European Environment Agency’s air quality database (AirBase) and 162 stations from the European Monitoring and Evaluation Programme (EMEP; Hjellbrekke et al., 2013). Table 1 provides a summary of the observational data sets.

A major advance by S2014 was the generation of average surface O₃ abundance in a grid cell from observational products, one that could be directly compared to gridded model output. The station measurements are used to gen-
Table 1. Observational data sets (2000 to 2009).

<table>
<thead>
<tr>
<th>Domain</th>
<th>Surface ozone network</th>
<th>No. stations</th>
<th>URL or reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America (NA)</td>
<td>US EPA Air Quality System (AQS)</td>
<td>1633</td>
<td><a href="http://www.epa.gov/ttn/airs/aqsdamatart">http://www.epa.gov/ttn/airs/aqsdamatart</a></td>
</tr>
<tr>
<td>Europe (EU)</td>
<td>European Monitoring and Evaluation Programme (EMEP)</td>
<td>162</td>
<td>Hjellbrekke et al. (2013)</td>
</tr>
</tbody>
</table>

Table 2. Model summary.

<table>
<thead>
<tr>
<th>(Abbreviation) model</th>
<th>Modeling center</th>
<th>Member*</th>
<th>Resolution (lat. × lon)</th>
<th>No. years</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) MOCAGE</td>
<td>MeteoFrance</td>
<td>r2i1p1, v2</td>
<td>2° × 2°</td>
<td>4</td>
<td>Josse et al. (2004)</td>
</tr>
<tr>
<td>(B) GFDL-AM3</td>
<td>GFDL</td>
<td>r1i1p1, v2</td>
<td>2° × 2.5°</td>
<td>10</td>
<td>Donner et al. (2011)</td>
</tr>
<tr>
<td>(C) CESM-CAM-SF</td>
<td>LLNL-NCAR</td>
<td>r1i1p1, v4</td>
<td>~1.9° × 2.5°</td>
<td>10</td>
<td>Naik et al. (2013)</td>
</tr>
<tr>
<td>(D) UM-CAM</td>
<td>NIWA</td>
<td>r1i1p1, v2</td>
<td>2.5° × 3.75°</td>
<td>10</td>
<td>Cameron-Smith et al. (2006)</td>
</tr>
<tr>
<td>(E) CMAM</td>
<td>CCCma</td>
<td>r1i1p1, v2</td>
<td>~3.7° × 3.75°</td>
<td>10</td>
<td>Lamarque et al. (2013)</td>
</tr>
<tr>
<td>(F) MIROC-CHEM</td>
<td>JAMSTEC-NIU-NIES</td>
<td>r1i1p1, v2</td>
<td>~2.8° × 2.8125°</td>
<td>10</td>
<td>Sciocca et al. (2008)</td>
</tr>
<tr>
<td>(G) GISS-E2-R</td>
<td>GISS</td>
<td>r1i1p3, v1</td>
<td>2° × 2.5°</td>
<td>5</td>
<td>Watanabe et al. (2011)</td>
</tr>
<tr>
<td>(H) GEOSCCM</td>
<td>NASA-GSFC</td>
<td>r1i1p1, v1</td>
<td>2° × 2.5°</td>
<td>10</td>
<td>Koch et al. (2006)</td>
</tr>
<tr>
<td>(I) UCI CTM</td>
<td>UCI</td>
<td>–</td>
<td>~2.8° × 2.8125°</td>
<td>10</td>
<td>Shindell et al. (2013)</td>
</tr>
</tbody>
</table>

* The format $r < N > i < M > p < L > vX$ distinguishes among closely related simulations by a single model where the set of integers ($N, M, L, X$) formatted as shown (e.g., r2i1p1, v2) define each model simulation’s realization number ($N$), initialization method ($M$), perturbed physics version ($L$), and version of publication-level data set ($X$).
one CGCM) with archived hourly surface O₃, incorporating the years from each model most closely aligned with observations. Most models provide 10 years of data, starting in either model year 2000 or 2001. In any case, all ACCMIP simulations are climatologically representative of the average 2000s with respect to meteorology and emissions. Table 2 provides a brief summary and the references of the models used in this study. Detailed descriptions of the ACCMIP models can be found in Lamarque et al. (2013) and references therein.

We also include a hindcast simulation over the same period as the observations from the University of California Irvine Chemical Transport Model (UCI CTM) performed at T42L60 resolution (Holmes et al., 2013) to both compare our model with the current generation models and to highlight differences between model simulations using free-running and hindcast meteorological conditions. The UCI CTM had many updates since the 1° × 1° × L40 version (Tang and Prather, 2010) used by S2014, but calculates similar, not unexpectedly high-biased patterns of surface O₃.

For commensurate comparison of the models and measurements, we regird the modeled hourly O₃ abundances (typically at 2 to 3° resolution) to the same 1° × 1° cells as the observations using first-order conservative mapping (i.e., proportion of overlapping grid-cell areas). Modeled hourly abundances are adjusted by 1 h per 15° longitude to be consistent with the local time of the observations. Our two major domains are NA bounded by 25–49° N and 125–67° W and EU bounded by 36–71° N and 11° W–34° E. A further masking drops coastal grid cells for which the quality of prediction index, \(P < 2/3\) (the number of independent stations at an effective distance of 100 km used to calculate the grid-cell values), see S2014 and Fig. S2 in the Supplement. Table S1 in the Supplement provides the latitudes and longitudes used in the final masking for both domains. Because of their differing chemical regimes, some of our analyses split the NA domain into western (WNA) and eastern (ENA) regions at 96° W, and EU into southern (SEU) and northern (NEU) regions at 53° N.

### 2.3 Air quality extremes

We define AQX events on a daily basis using local (i.e., grid-cell) climatologies to identify the 10 times N worst days (i.e., percentile; e.g., the 100 worst days in a decade). The space–time connectedness of the AQX events into episodes is defined using a hierarchical clustering algorithm described in S2014. Because AQX episodes span across the regions, statistics for these analyses are done only on the two major domains NA and EU. The total size of an AQX episode (\(S\), units = km² days) is calculated by integrating the areal extent of an episode (km²) through time (days). For a given set of episodes, the mean size \(\overline{S}\) is calculated as a weighted geometric mean, with the weights equal to the AQX episode sizes (Eq. 6 in S2014). Because the lower native resolutions of the models typically map onto four to eight contiguous 1° × 1° grid cells, the modeled episode sizes have artificial minima; however, S2014 demonstrated that this has little effect on the resultant episode size distributions.

### 3 Results

#### 3.1 Diurnal cycles

We test the models’ abilities to reproduce the observed shape (i.e., phase and amplitude) of the diurnal cycle, averaged over summer (JJA) and winter (DJF) months. For each of the four regions, average hourly values (local solar time) are calculated as the area-weighted mean of all grid cells’ O₃ abundances. We calculate the phase (\(h\), hour of peak O₃ abundance, with \(h = 0.0\) corresponding to 00:00 LT) and peak-to-peak amplitude (\(H\), ppb difference from minimum to maximum) of the diurnal cycle using a cosine fit with a period of 24 h. Although the diurnal cycle could be more accurately represented by a higher-order fit, this simple method provides objective and continuous measures of \(h\) and \(H\) for each data set, avoiding subjective, ambiguous results in cases of flat and/or multiple maxima.

Figure 1a–h show the diurnal cycle of the observations and models averaged over JJA (top row) and DJF (second row) in WNA, ENA, SEU, and NEU (columns from left to right). A triangle for each data set is plotted as \((r, y) = (h, H)\). The large number of data points (~10⁶ × 24 h per model) provides a smooth and robust estimate of each data set’s diurnal cycle. The color scheme and model abbreviations in the legend of Fig. 1 are common to all similar figures and text throughout. The Taylor diagrams (Taylor, 2001) in Fig. S3a–h in the Supplement show an alternate, commonly used summary of the results in terms of the correlation coefficient \((R)\), the normalized standard deviation (NSD), and centered root-mean-square difference. Figures 1 and S3 in the Supplement show very similar quantities (e.g., model–measurement discrepancies in \(h\) and \(H\) roughly correspond to \(R\) and NSD, respectively); however, we consider the representation in Fig. 1 to be more useful. The panels of Fig. S3 in the Supplement correspond to panels in Fig. 1 in terms of region and variable. Summary statistics on diurnal cycles, annual cycles, and AQX events for ENA are presented in Table 3, with all regions and additional statistics provided in Tables S2–S4 in the Supplement.

The shape of the diurnal cycle of O₃ is driven primarily by sunlight, meteorology (e.g., temperature and variations in boundary layer mixing), surface deposition, and the daily cycle of precursor emissions. The hour of the maximum phase \(h\) occurs when these factors align, usually in midafternoon. Indeed, for seven of eight region-seasons in Fig. 1a–h, the observed value of \(h\) ranges from 14.8 to 15.5 h. For DJF in NEU, where photochemical O₃ formation is negligible, there...
and S2 in the Supplement) for the ACCMIP models is typical of all eight region-seasons (up to 28 ppb), although some models (e.g., C and E in JJA, E in DJF) show little or no mean bias, even though they underestimate $H$ in JJA by about 25 % like all ACCMIP models.

The underestimation of the summertime diurnal amplitude $H$ by most ACCMIP models suggests that they either underestimate net daytime production or have too little nighttime loss of $O_3$ or its precursors through either in situ chemical loss or dry deposition. From the derivative of the diurnal cycles in Fig. 1a–d, there are two periods of model–observation discrepancy: in the early morning ($\sim 06:00$ LT) models underestimate the observed slope and in the early evening ($\sim 19:00$ LT) they overestimate it. The models generally match the observed slope to within $\pm 1\%$ during midday and throughout the night. Thus the model error is to underestimate net $O_3$ production in the early morning and overestimate it in early evening, which may be caused by the lack of a diurnal emission cycle in these global models. The mismatch of the slope in the early morning, during which the boundary layer grows rapidly, may be caused by the models underestimating entrainment of free troposphere air. We find no clear evidence that modeling errors in the nocturnal planetary boundary layer (Lin et al., 2008) or missing near-surface processes affect the diurnal cycle on a regional average.

Underestimated daytime production could result from limited representation of VOC chemistry, since discrepancies are largest in summer when VOCs play a larger role. Indeed, model A, which simulates the most chemical species of VOC chemistry, since discrepancies are largest in summer when VOCs play a larger role. Indeed, model A, which simulates the most chemical species of

### Table 3. Example summary statistics for the observations (OBS), the ACCMIP models (A–H), and the UCI CTM (I) for eastern North American (ENA) summer (JJA) and winter (DJF) diurnal cycles, annual cycle of MDA8, annual cycle of AQX events, and North American (NA) AQX episodes (100 AQX events per decade case).

<table>
<thead>
<tr>
<th>Data</th>
<th>Metric, description (unit)</th>
<th>OBS</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJA diurnal cycle</td>
<td>$h$, maximum phase (hour)</td>
<td>15.0</td>
<td>17.0</td>
<td>16.1</td>
<td>16.5</td>
<td>15.5</td>
<td>15.8</td>
<td>15.2</td>
<td>15.7</td>
<td>16.0</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td>$H$, peak-to-peak amplitude (ppb)</td>
<td>29.1</td>
<td>28.3</td>
<td>28.4</td>
<td>21.8</td>
<td>22.7</td>
<td>21.8</td>
<td>22.6</td>
<td>12.1</td>
<td>18.5</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td>MB, mean bias (ppb)</td>
<td>–</td>
<td>19.0</td>
<td>24.4</td>
<td>1.1</td>
<td>12.2</td>
<td>3.5</td>
<td>17.9</td>
<td>21.1</td>
<td>12.9</td>
<td>37.0</td>
</tr>
<tr>
<td>DJF diurnal cycle</td>
<td>$h$, maximum phase (hour)</td>
<td>15.1</td>
<td>18.0</td>
<td>16.7</td>
<td>15.7</td>
<td>15.3</td>
<td>14.0</td>
<td>15.9</td>
<td>14.8</td>
<td>16.3</td>
<td>16.1</td>
</tr>
<tr>
<td></td>
<td>$H$, peak-to-peak amplitude (ppb)</td>
<td>9.1</td>
<td>6.7</td>
<td>7.5</td>
<td>11.3</td>
<td>7.8</td>
<td>5.8</td>
<td>6.9</td>
<td>2.4</td>
<td>10.6</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td>MB, mean bias (ppb)</td>
<td>–</td>
<td>10.2</td>
<td>15.2</td>
<td>9.8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>4.0</td>
<td>30.1</td>
<td>5.5</td>
</tr>
<tr>
<td>MDA8 annual cycle</td>
<td>$m$, maximum phase (month)</td>
<td>5.3</td>
<td>5.8</td>
<td>6.0</td>
<td>3.7</td>
<td>5.8</td>
<td>5.7</td>
<td>6.1</td>
<td>6.0</td>
<td>6.2</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>$M$, peak-to-peak amplitude (ppb)</td>
<td>20.7</td>
<td>29.8</td>
<td>29.1</td>
<td>12.8</td>
<td>32.7</td>
<td>25.9</td>
<td>31.5</td>
<td>3.5</td>
<td>20.3</td>
<td>64.6</td>
</tr>
<tr>
<td></td>
<td>MB, mean bias (ppb)</td>
<td>–</td>
<td>16.9</td>
<td>16.6</td>
<td>6.8</td>
<td>4.2</td>
<td>–</td>
<td>8.1</td>
<td>20.1</td>
<td>8.0</td>
<td>24.8</td>
</tr>
<tr>
<td></td>
<td>$T_{\text{MDA8, 87th-30th percentile}}$ (ppb)</td>
<td>22.8</td>
<td>33.0</td>
<td>27.5</td>
<td>19.4</td>
<td>27.0</td>
<td>22.4</td>
<td>28.3</td>
<td>19.1</td>
<td>21.9</td>
<td>56.0</td>
</tr>
<tr>
<td></td>
<td>$R_{\text{MDA8,JJA, spatial correlation of $E_{\text{MDA8}}$ maps}}$</td>
<td>1.00</td>
<td>0.70</td>
<td>0.81</td>
<td>0.52</td>
<td>0.69</td>
<td>0.69</td>
<td>0.34</td>
<td>0.27</td>
<td>0.69</td>
<td>0.71</td>
</tr>
<tr>
<td>AQX event annual cycle</td>
<td>$m_{\text{AQX, maximum phase (month)}}$</td>
<td>5.5</td>
<td>6.2</td>
<td>6.8</td>
<td>3.2</td>
<td>6.2</td>
<td>6.4</td>
<td>6.6</td>
<td>6.8</td>
<td>7.7</td>
<td>6.6</td>
</tr>
<tr>
<td></td>
<td>$R_{\text{MDA8,AQX, correlation of AQX and MDA8 cycles}}$</td>
<td>0.84</td>
<td>0.76</td>
<td>0.78</td>
<td>0.88</td>
<td>0.78</td>
<td>0.82</td>
<td>0.80</td>
<td>0.78</td>
<td>0.70</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>$T_{\text{AQX, AQX threshold – 30th percentile (ppb)}}$</td>
<td>0.84</td>
<td>0.76</td>
<td>0.78</td>
<td>0.88</td>
<td>0.78</td>
<td>0.82</td>
<td>0.80</td>
<td>0.78</td>
<td>0.70</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>$R_{\text{AQX,JJA, spatial correlation of $E_{\text{AQX}}$ maps}}$</td>
<td>1.30</td>
<td>0.70</td>
<td>0.78</td>
<td>0.28</td>
<td>0.63</td>
<td>0.53</td>
<td>0.44</td>
<td>0.60</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>NA AQX episodes</td>
<td>$(\overline{s})$, weighted geometric mean AQX episode size (104 km2 days)</td>
<td>415.0</td>
<td>128.0</td>
<td>229.0</td>
<td>1426.0</td>
<td>461.0</td>
<td>290.0</td>
<td>522.0</td>
<td>243.0</td>
<td>774.0</td>
<td>463.0</td>
</tr>
<tr>
<td></td>
<td>CCD100, fraction of AQX events’ areas in AQX episodes$\geq 100 \times 10^4$ km2 days (%)</td>
<td>79.0</td>
<td>56.1</td>
<td>73.7</td>
<td>92.6</td>
<td>85.3</td>
<td>76.1</td>
<td>80.3</td>
<td>73.0</td>
<td>83.0</td>
<td>80.2</td>
</tr>
<tr>
<td></td>
<td>CCD1000, fraction of AQX events’ areas in AQX episodes$\geq 1000 \times 10^4$ km2 days (%)</td>
<td>38.0</td>
<td>9.7</td>
<td>12.8</td>
<td>69.2</td>
<td>30.8</td>
<td>19.2</td>
<td>43.6</td>
<td>12.7</td>
<td>48.7</td>
<td>37.5</td>
</tr>
<tr>
<td></td>
<td>$\Delta T_S$, average increase in $T_S$ for AQX episodes of size $S$ (ppb dec$^{-1}$)</td>
<td>2.9</td>
<td>9.9</td>
<td>4.6</td>
<td>0.8</td>
<td>2.3</td>
<td>2.9</td>
<td>3.0</td>
<td>2.9</td>
<td>3.5</td>
<td>6.0</td>
</tr>
</tbody>
</table>
all the ACCMIP models in addition having the largest VOC emissions, is one of the few models to consistently overestimate $H$. In contrast, C and E are two of the better-performing models despite their comparatively simple representation of VOC chemistry (C – only isoprene, E – none). The only models to include the small and relatively uncertain fractional yield of HNO$_3$ from the reaction of HO$_2$ and NO are A, G, and H (Lamarque et al., 2013). This reduces daytime production and could partly explain why the models G and H consistently underestimate $H$ more than others, however model A overestimates $H$.

The ACCMIP models reproduce the phase of the observed diurnal cycle in both seasons despite not accounting for hourly variation in emissions. The weekly emission-driven cycles in MDA8 O$_3$ were diagnosed by S2014, but we do not apply that diagnostic here because the models did not include such variability in emissions. The lack of hourly variation of emissions may account for the overall underestimates of $H$ by the ACCMIP models, since NO emissions can be lost heterogeneously at night, less effectively than those during the morning and afternoon peaks in traffic. In addition, if the early morning peak in transport NO$_x$ emission was included, the modeled morning rise in O$_3$ would most likely...
be augmented, thus yielding larger values of $H$. The ACCMIP models use a wide range of boundary layer mixing schemes but consistently underestimate $H$. The boundary layer schemes may be responsible for these underestimates; however, Menut et al. (2013) notes that at least for one model, increasing its vertical resolution results in very small surface $O_3$ changes.

The UCI CTM’s values of $h$ and $H$ show that it drastically overestimates net daytime $O_3$ production, especially during early morning hours. Its values of $h$ are about 2.4 h earlier than observed in the highest three region-seasons, and in contrast to the ACCMIP models its $H$ values are too large by 10 s of ppb. This diagnostic identifies a serious problem with the UCI CTM diurnal cycle over polluted regions that needs to be investigated (e.g., missing heterogeneous loss of NO$_2$ at night, capped boundary layer in the morning) and which will be done after publication of this research. S2014 found that the UCI CTM accurately hindcast the summertime probability distribution of MDA8 $O_3$, the occurrence of AQX events, and the size of these episodes, albeit with high bias of about +29 ppb in JJA over both NA and EU. This new diurnal diagnostic has clearly identified model errors and pathways to improve our model as well as models like G, which gravely underpredicts the amplitude of the diurnal cycle. The tests shown here emphasize a large-scale average over different photochemical regimes in the four regions, and thus individual model developers may wish to analyze the observations for smaller regions using the data sets generated here, which are available by request from the corresponding author.

### 3.2 Annual cycle

We test the models’ abilities to reproduce the observed phase and amplitude of the annual cycle over the four regions. Average monthly values for each region are calculated as the area-weighted mean of all encompassed cells’ MDA8 $O_3$ abundance, reflecting the EPA air quality metric (www.epa.gov/air/criteria.html). Similar to the diurnal cycle, we derive the phase ($m$, month of peak $O_3$ abundance, with $m = 0.0$ corresponding to 1 January) and peak-to-peak amplitude ($M$, ppb difference from minimum to maximum) using a cosine fit assuming 12 equally spaced monthly means. Figure 1i–l show the annual cycle of the observations and models over our four regions with triangles plotted for each model and data set as $(x, y) = (m, M)$. The filled gray curve shows ±1 standard deviation of each monthly mean based on 10 years of observations. This interannual variability is quite narrow, much less than the spread across models. As for the diurnal cycle, the Taylor diagrams in Fig. S3i–l show an alternate presentation of the annual cycle results with summary statistics given in Tables 3 and S3 in the Supplement.

In northern midlatitudes, processes that drive the shape of the annual cycle are similar to those of the diurnal cycle (i.e., sunlight, temperature, and precursor emissions) but occur on continental to hemispheric scales. Dry deposition through stomatal uptake and large-scale meteorological conditions including stratosphere–troposphere exchange and the position of the jet stream (Barnes and Fiore, 2013) also play important roles. These surface observations show the same well-known cycle that has been seen in the northern hemispheric midlatitude troposphere from ozone sondes and clean-air remote sites (Logan, 1989; Fiore et al., 2009): lowest values in late fall (ND), increasing through winter (JFM) followed by a broad flat peak over spring–summer (AMJJA). The lower reactivity region NEU peaks in April and declines until January, indicating meteorologically driven increases through the winter (e.g., stratospheric influx). The observations show a phase $m$ = 5.6, 5.3, 5.5, and 4.3 month of year for WNA, ENA, SEU, and NEU, respectively; and corresponding amplitudes $M$ = 22, 21, 26, and 17 ppb. By fitting a cosine curve to each grid cell’s time series, we find that in terms of specific locations, the earliest $m$ occur in Canada, Florida, and NEU while the latest $m$ occur in California, south-central NA, and SEU (not shown). Most ACCMIP models have $m$ within ±1 month of the observations, generally earlier in NEU, later in ENA and SEU, and split in WNA. Models C and G have difficulty producing the observed seasonal cycles, and their derived phases are not meaningful.

The amplitude $M$ is controlled by both meteorology and photochemistry. For the very large regional values of $M$, it is clearly chemical, occurring in regions with large $O_3$ precursor emissions: California with ~40 ppb, the Great Lakes region with ~30 ppb, and northern Italy with ~45 ppb (not shown). The smallest values of $M$ (~15 ppb) are found in northwest and southeast NA and in NEU. The ACCMIP models generally underestimate $M$ by about 5 ppb in WNA, SEU, and NEU, while they overestimate it by about 5 ppb in ENA. The low values of $M$ for C and G suggest they are either overestimating net production of $O_3$ in winter or underestimating it in summer; however, their wintertime biases (see Fig. 1e–h, Tables 3 and S2 in the Supplement) indicate that wintertime production or representation of wintertime physical climate could be causing the low $M$ values.

The annual cycles here are constructed using the MDA8 $O_3$ derived from hourly data. Many models, including eight other ACCMIP models not analyzed here, do not report hourly surface $O_3$ but only monthly means (i.e., the average of all hours within a month). We chose MDA8 values to conform to the US EPA primary air quality standards and statistics, but if we used monthly averages then more models could be evaluated. Unfortunately, without at least daily diagnostics (e.g., daily mean or maximum value) analysis of percentile patterns and AQX events and episodes (see Sects. 3.3–3.7) are precluded. Further, we tested the difference in annual cycles diagnosed both ways and found that the bias of a model can differ and thus these two diagnostics cannot be mixed. For example, the ACCMIP ensemble mean bias for JIA using MDA8 averages is 2, 11, 11, and 8 ppb in WNA, ENA, SEU, and NEU, respectively; however, the corresponding bias using 24 h averages is consistently larger at 6, 14,
13, and 9 ppb. This result was expected since the ACCMIP model ensemble generally has the largest biases outside of MDA8 hours. These conclusions are generally true for all seasons and models, as illustrated in Fig. S4 in the Supplement, which shows the mean bias (model minus observed) of MDA8 minus 24 h average for each model, season, and region.

For the UCI model, excess production in the diurnal cycle is also evident in the annual cycle, overestimating $M$ in all regions, most in ENA ($\pm 44$ ppb) and least in NEU ($\pm 9$ ppb). In addition, the month of peak abundance is always later than observed, sometimes by more than 1 month. Not unexpectedly, the bias in $M$ using 24 h averages is significantly less than that using MDA8 (e.g., +30 vs. +44 ppb in ENA) because the largest errors occur near midday. We conclude that using 24 h averages to construct the annual cycle is basically a different, almost independent diagnostic than that constructed from the daily MDA8 $O_3$, and further it would predict different health impacts if used to project summertime surface $O_3$ in a future climate.

### 3.3 AQX events

Next, we test the models’ ability to reproduce the annual cycle of the individual AQX events, identified for each grid cell as the 100 days with the highest MDA8 in the decade (40 in 4 years for A, 50 in 5 years for G). Figure 1m–p show the annual cycle of AQX events for the observations and models over our four regions. The filled gray curve shows ± 1 standard deviation for each month based on 10 years of observations. The interannual variability is much larger than that seen in the observed MDA8 cycle with most models falling in its range in SEU and NEU but not in WNA or ENA. An alternate presentation as Taylor diagrams is shown in Fig. S3 in the Supplement, and the summary statistics are given in Tables 3 and S4 in the Supplement. The month of maximum AQX events for most models is within ± 1 month of that observed in each region ($m_{AQX}$ in Tables 3 and S4 in the Supplement). Based on S2014, we expect the annual cycle of AQX events to be highly correlated with that of MDA8, as the observations show correlations $R_{MDA8}$ (i.e., AQX vs. MDA8) of 0.81 to 0.87 for all regions. For the ACCMIP models this correlation is not as good, but they still show $R_{MDA8} > 0.70$ (Tables 3 and S4 in the Supplement). Models whose monthly MDA8 correlates well with observed MDA8 also have monthly AQX events that correlate well with observed. Nevertheless, matching the AQX events annual cycle is more difficult than matching the cycle of MDA8 (Tables 3, S3, S4, and Fig. S3 in the Supplement) because AQX events are driven by meteorological extremes which are not necessarily represented in these climatological simulations.

The UCI CTM also reproduces the annual AQX events well, and since it is a hindcast, we can extend the analysis to how well it identifies each AQX event on an exact-match basis (“model skill” by S2014). For a climatological model that exactly matches the annual cycle (i.e., matching the number of AQX events in each month) but is synoptically random in each month, a skill score of ~8% is expected; however, the UCI hindcast correctly identifies 28, 33, 33, and 21% of AQX individual cell events in WNA, ENA, SEU, and NEU, respectively.

### 3.4 Mapping $O_3$ percentiles and enhancements

We can define baseline levels of $O_3$ from observations as the statistically lowest percentiles (National Research Council, 2009). Baseline levels are independent of attribution to specific emissions or policy relevance implied by US EPA’s use of the term background. We can expect, or possibly assume, that baseline levels are not influenced by recent, locally emitted or produced pollution (HTAP, 2010). To estimate the daytime enhancement in summertime $O_3$, presumably caused by continental emissions, we first want to define a baseline level for each grid cell as a lower percentile of the daily surface $O_3$. We seek a percentile that represents the cleanest air possible over the summer season (even if it is never realized during the summer), and one that does not change across years. We use MDA8 rather than 24 h average data to prevent nighttime values from determining the baseline. We calculate percentiles for each cell on an annual basis and then derive regional area-weighted averages of the percentiles. The resulting percentiles by region (Fig. 2) show that the year-to-year variability is small below the 40th percentile, but the largest pollution years are evident at and above the 50th percentile. Thus, we select the 30th percentile as each grid cell’s baseline level, which corresponds roughly to the lower levels of spring–fall days. One might argue choosing, for example, the 10th percentile of JJA to estimate summertime enhancement; however, this assumes JJA in all models is the peak of the annual cycle and still sees clean air. We define $O_3$ enhancement ($E_X$, unit = ppb) here as the difference between the 30th percentile and any larger value, where subscripts will describe the reference value.

To estimate the summertime $O_3$ enhancement from local to continental-scale pollution, we assume that the 92 days of JJA are the highest $O_3$ values of the year, pick their median value (87th percentile), and subtract from it the spring–fall baseline (30th percentile). Maps of the summer enhancement $E_{JJA}$ (i.e., 87th minus 30th percentile) in NA and EU in observations and models are shown in Fig. 3. While $O_3$ levels for the 87th, 30th, and other percentiles vary considerably from cell to cell (see S2014), the maps of observed $E_{JJA}$ show mostly large-scale structures.

Many models (A, B, D, E, F, H, I) have similar patterns of $E_{JJA}$ over NA, with large enhancements (30 to 50 ppb) from the Mississippi through the Ohio River valley to the northeast, whereas the observations show such a pattern but with smaller enhancements (25 to 30 ppb). Model A greatly overestimates $E_{JJA}$ in the most polluted areas (e.g., California, northeast NA, south and central EU) as well as coastal ar-
Figure 2. Values of MDA8 O₃ (ppb) for years 2000 to 2009 corresponding to the 10th, 20th, . . . , 90th, 95th, and 97.3 (i.e., AQX threshold) percentiles in (a) WNA, (b) ENA, (c) SEU, and (d) NEU. The percentile for each line is shown at the beginning of the curves in each panel.

Figure 3. Summertime O₃ enhancement \( E_{JJA} \) is the difference between the 87th and 30th percentile of the gridded surface MDA8 O₃ (ppb) over (left two columns) NA and (right two columns) EU for the observations (O), ACCMIP models (A–H), and UCI CTM (I). The values of model I are scaled by 0.5 so the same color scale can be used.

The extremely large bias near the Gulf of Mexico is unique to model A, presumably resulting from natural JJA emission sources such as lighting NOₓ, wildfires, or biogenic VOCs since the area is not known for

Figure 4. (a–b) Complementary cumulative distribution (CCD) of the percentage of total areal extent of all individual AQX events (100-per-decade case) as a function of AQX episode size ($S$, $10^4$ km$^2$ days) for the observations (O), ACCMIP models (A–H), and UCI CTM (I) in (a) NA and (b) EU. Dashed vertical lines show the graphical representations of CCD$_{100}$ and CCD$_{1000}$. Mean episode size ($\bar{S}$) for each data set and domain is given in the legend as (NA/EU). (c–d) Density scatterplot of the observations enhancement of AQX episodes $\bar{E}_S$ vs. their size $S$ ($\bar{E}_S$ binned at 2.5 ppb increments from $<15$ ppb to $>55$ ppb, $S$ binned at each log decade) in (c) NA and (d) EU. The gray scale represents the relative percentage of AQX episodes in each $(x, y) = (S, \bar{E}_S)$ bin and includes percent ranges of $\leq 5\%$ (white), 5–10, 10–15, and $>15\%$ (darkest gray) where the size bins (i.e., columns) are normalized to sum to 100\%. The overlain curves show the observation’s and each model’s area-weighted mean enhancement $\bar{E}_S$ for each size bin. The values of $\bar{E}_S$ in each size bin for models A and I have been scaled by 0.5 since they are largely outside the range of the others.

large anthropogenic sources. Two models (C, G) are unusually uniform across NA (except California). Surprisingly, this sorting of the models does not hold for EU. For example, there must be some clue as to why model B greatly overestimates $\bar{E}_{\text{JJA}}$ over NA but underestimates it over EU. Such behavior from model C (uniform $\bar{E}_{\text{JJA}}$) may be expected since the tropospheric VOC chemistry is highly simplified. The uniform pattern of $\bar{E}_{\text{JJA}}$ is also somewhat evident in EU for model E, which has even simpler VOC chemistry compared to C, although this may be due to biases in the representation of physical climate rather than chemistry.

The $\bar{E}_{\text{JJA}}$ diagnostic provides an excellent geographically resolved test for CCM development. It also provides a useful measure of O$_3$ regional pollution changes in a future climate with shifting O$_3$ baselines due to hemispheric-scale changes in methane, water vapor, temperature, and stratospheric influx. Over each of our four regions, we calculate the average summertime enhancement $\bar{E}_{\text{JJA}}$ (see Tables 3 and S3 in the Supplement), expecting to find the values and model–measurement differences similar to those found in the seasonal amplitude $M$. Indeed, this is true, albeit $\bar{E}_{\text{JJA}}$ is generally smaller than $M$. In addition, the spatial pattern of the values and model–measurement differences are also consistent between $\bar{E}_{\text{JJA}}$ and $M$ (not shown).

3.5 AQX episode size

We examine the models’ ability to simulate the observed distribution of AQX episode sizes over the decade 2000–2009. Our hierarchical clustering analysis identifies connected-cell, multi-day AQX episodes of size $S$ (given here in units of $10^4$ km$^2$ days). We do not split the NA and EU domains here because episodes span across regions. Figure 4a–b show the distribution of episode sizes in the observations and each model as the complementary cumulative distribution (CCD, %), i.e., the fraction of total AQX area-day events occurring in episodes of size $S$ or larger.
For NA observations, the fraction of AQX area-weighted events that occur in episodes with $S > 100 \times 10^4 \text{ km}^2$ days (CCD$_{100}$) is 79 %; and those with $S > 1000 \times 10^4 \text{ km}^2$ days (CCD$_{1000}$) is 38 %. For EU observations, most AQX events also occur in large episodes: CCD$_{100} = 80 \%$ and CCD$_{1000} = 35 \%$. Model C is aberrant in having extremely large episodes (e.g., NA CCD$_{100} = 93 \%$), which fall mostly in the spring rather than summer months (see Fig. 1m–p). This may result from the model’s simplified chemistry or unrealistic widespread stratospheric intrusion of O$_3$. In any case, this model’s summertime high ozone events are obscured. Model A, with much more complex chemistry, however, shows significantly smaller episodes. For CCD$_{100}$, the other models (B, D–I) are close to the observed: 73–85 % for NA and 71–85 % for EU. For CCD$_{1000}$, however, this model spread diverges substantially: 13–69 % for NA and 15–70 % for EU. In general, models A, B, E, and G do not produce the larger episodes and thus their physical climate may lack the synoptically correlated persistent stagnation episodes. The UCI CTM, using observed meteorology, captures the shape of the observed CCDs extremely well compared to the free-running climate of the ACCMIP models.

Integrating over all episodes, we calculate the weighted geometric mean size (S) (see S2014). Observations have mean episode sizes (S) of 415 ($10^4 \text{ km}^2$ days) and 444 in NA and EU, respectively. Models C, D, F, H, and I are biased high in (S), while models A, B, E, and G are biased low for both NA and EU (Fig. 4a–b, Tables 3 and S4 in the Supplement).

### 3.6 Non-stationarity and possible trends

One problem with diagnosing decadal AQX size statistics is that they can be biased when more AQX events occur at one end of the decade due to a trend in O$_3$ precursor emissions. A greater density of events in one summer generally means larger episodes. A linear fit of annually derived O$_3$ percentiles calculated over years 2000–2008 (2009 was excluded due to lack of NO$_x$ and VOC emission data, see below) for each of the four regions (Fig. S5 in the Supplement) shows clearly decreasing surface O$_3$ abundances at the higher percentiles (see also Fig. 2), presumably through reductions in NO$_x$ and VOC emissions (Hudman et al., 2009; Xing et al., 2014). To test if these trends are emission-driven or artifacts of the meteorological time slice, we analyze the UCI CTM results (dashed lines, Fig. S5 in the Supplement), which are forced by observed meteorology but have constant anthropogenic pollution emissions over the time period. We also obtain total NO$_x$ and VOC emissions from version 4.2 of the Emission Database for Global Atmospheric Research (EDGAR, EC-JRC/PBL, 2009) for years 2000–2008 (2009 was unavailable at time of publication) and calculate their trends over the period. Over WNA and ENA, meteorology seems to be driving the small positive trends at lower O$_3$ percentiles (where UCI and observed trends roughly agree), but above the 60th percentile (where UCI and observed trends diverge) emissions reductions are the most likely cause. In SEU and NEU, the trends are less conclusive for either meteorology or emission based, but most EU NO$_x$ reductions occurred prior to 2000 (Xing et al., 2015). Koumoutsaris and Bey (2012) compare GEOS-Chem hindcasts with NA and EU trends at a limited number of stations from CASTNet and EMEP (~40 in each domain) and find similar trends. They also attribute the negative trends at high percentiles to reduced precursor emissions; however, they attribute the positive trends at low percentiles to changing background O$_3$ as opposed to changing meteorology posited here.

In an effort to correct the AQX decadal statistics for changes in O$_3$ precursors, we searched for correlations on a cell-by-cell basis between high-percentile MDA8 O$_3$ vs. NO$_x$ emissions on an annual basis for years 2000–2008. No simple linear relation emerged, and we could find no satisfactory way to “correct” the observations for this regionally varying, monotonic but nonlinear, decline in NO$_x$ and VOC emissions that did not corrupt the data. The post-CMIP5 plans for the Chemistry-Climate Model Initiative (CCMI) include hindcast simulations with time-dependent emissions that will allow for the simulation of the observed O$_3$ non-stationarity.

One option for analyzing extremes in a non-stationary decadal data set is to define AQX events annually on a 10-per-year basis. This approach greatly dampens the observed episode mean size and across-year standard deviation from 415 ± 307 (100 per decade) to 249 ± 67 (10 per year) in NA and from 444 ± 720 to 355 ± 48 in EU. Moreover, it gives a false positive impression of the severity of air pollution in extreme years. Thus, we maintain our primary analysis with AQX defined as 100-per-decade. In parallel with Fig. 4a–b, we show the CCDs using a 10-per-year basis for AQX in Fig. S6a–b in the Supplement.

### 3.7 Severity of pollution in largest episodes

As a measure of O$_3$ produced during AQX events/episodes, we map out the enhancement at the AQX threshold level $E_{AQX}$ (~97.3 percentile) as shown in Fig. S7 in the Supplement (parallel to Fig. 3, also relative to the local 30th percentile). We also calculate the average AQX enhancement $E_{AQX}$ over our regions (Tables 3 and S4 in the Supplement). For ENA, the ACCMIP modeled range of $E_{AQX}$ is 29–52 ppb, spanning the observed of 35 ppb (Table 3). This average result is encouraging for the ACCMIP models except that, as for $E_{10A}$ (Fig. 3), the pattern match is not as good (Tables 3, S3 and S4 in the Supplement).

Of the 100 AQX events in each cell, many will lie above the local AQX threshold value. We expect that larger, longer-duration episodes accumulate more O$_3$, and thus these super episodes might have O$_3$ enhancements (relative to the 30th percentile) well above the AQX threshold enhancement, $E_{AQX}$. For each AQX event, we calculate an enhancement...
low-O3 regions with very low 30th percentile baselines, re-
stand them to be the effect of very polluted air masses being
mostly along the coasts at the edge of the mask. We under-
small fraction of events (see Sect. 3.5), and we find them
These small episodes are uncommon, representing only a

$E_S = \text{area-weighted average enhancement}$ of O3 driven by pol-

$S$ as the MDA8 value of that AQX event minus the local

30th percentile value. For each episode of size $S$, we calcu-
late the area-weighted average enhancement $E_S$. Figure 4c–d
plot the observed density distribution of all $E_S$, quantized ev-
every 2.5 ppb for $E_S$ and every decade in $10^4 \text{km}^2$ days for $S$.
These plots show large variability in the observed $E_S$ fre-
quency (gray pixels) and yet a consistent picture of the mean
enhancements as a function of $S$ (open circles). For episode
sizes of 0.3 (i.e., 0.1 to 0.99), 3, and 30, $E_S$ is almost constant
($\sim 32 \text{ ppb for both NA and EU}$), but for sizes 300 and 3000 it
increases almost linearly per decade. We calculate this slope
$\Delta E_S$ as the average of the 30-to-300 increase (1 decade in $S$)
plus half of the 30-to-3000 increase (2 decades), getting
values of 2.9 (NA) and 1.7 (EU) ppb increase per decade of
episode size. Similar results are seen for the 10-per-year
AQX definition (Fig. S6c–d in the Supplement), with $\Delta E_S$
of 2.7 (NA) and 3.3 (EU). The slope $\Delta E_S$ is not simply an
expected result from our statistical sorting since in NA we
find that compared to the observations, model C has slope
that is a factor of about 4 smaller, while A has a slope nearly
a factor of 4 larger, and F has a negative slope.

The models generally produce the shape of $E_S$ vs. $S$, al-
though most models (except A and I, see Fig. 4 caption) under-
estimate the enhancement for all sizes. The obvious dis-
crepancies are for NA episodes, where many models predict
that the largest enhancements occur in the smallest episodes
($S = 0.3$). This anomaly does not occur for EU episodes.
These small episodes are uncommon, representing only a small
fraction of events (see Sect. 3.5), and we find them mostly along the coasts at the edge of the mask. We understand them to be the effect of very polluted air masses being advected to the neighboring ocean cells which are typically low-O3 regions with very low 30th percentile baselines, resulting in large enhancements from the highly polluted air. The observations are interpolated and not capable of following a pollution shift offshore. Thus the models are probably correct, but the method of masking and station interpolation makes this discrepancy a systematic feature. The lack of such a feature in EU can be understood by the lack of such sharp coastal gradients. Overall, most models agree with the observations, showing that the super-episodes have the largest O3 enhancements.

4 Conclusions and discussion

Confidence in modeled projections of future air quality is based fundamentally on our ability to accurately simulate the present-day observed climatology of surface O3 and particu-
late matter over NA and EU where dense, long-term, reliable measurements are available. In this work we evaluate the surface O3 climatologies from eight global models (six CCMs, one CTM, and one CGCM) that reported hourly surface O3 as part of the ACCMIP. In addition we test the UCI CTM simulation as an exact hindcast of the 2000–2009 decade of observations used here. Our tests follow the unique approach of S2014 in which over 4000 heterogeneously spaced air quality stations are used to calculate the hourly O3 averaged over $1^\circ \times 1^\circ$ grid cells that can then be compared un-
ambiguously with the modeled grid. Diagnostics include the
hourly diurnal cycle, monthly seasonal cycles, and sizes and
intensity of air quality extreme episodes. For the most part, the models are biased high during all hours of the day, all
months of the year, and in all regions.

Averaged over large regions, the ACCMIP models simu-
late the shape of the observed summertime diurnal cycle well, with the hour of maximum within $\pm 1$ h of observed
($\sim 15:00 \text{ LT}$). The observed peak-to-peak amplitude (25 to 29 ppb over the more polluted regions) is not as well matched and typically underestimated by about 7 ppb. The UCI CTM hindcast, which performed well in the S2014 tests except for a uniform high bias, clearly fails these new diurnal tests and indicates model error in the morning boundary layer chemistry. In general, the ACCMIP models simulate the observed regional annual cycle of monthly mean MDA8 O3. They match the month of maximum to within $\pm 1$ month of observed (mid-June), although two models are in error with almost no annual cycle and no clear maximum. The other models overestimate the peak-to-peak amplitude of the observed cycle by about 5 ppb (20 %) in the most polluted region (eastern North America) while underestimating it by about 5 ppb in the other three regions. Model skill in matching the annual cycle of AQX events is fair but not good. This annual cycle has much larger interannual variability than that of MDA8 O3, and many models shift the month of maximum AQX events to later in the summer than is observed.

Measures of the enhancement in surface O3 driven by pol-
lution are derived from the statistics of the decade of daily
gridded MDA8 values. For our measure of summertime en-
hancement (87th minus 30th percentile), the models gen-
ern the observed monthly mean MDA8 O3. They match the month of maximum to within $\pm 1$ month of ob-
served (mid-June), although two models are in error with al-
most no annual cycle and no clear maximum. The other models overestimate the peak-to-peak amplitude of the observed cycle by about 5 ppb (20 %) in the most polluted region (eastern North America) while underestimating it by about 5 ppb in the other three regions. Model skill in matching the annual cycle of AQX events is fair but not good. This annual cycle has much larger interannual variability than that of MDA8 O3, and many models shift the month of maximum AQX events to later in the summer than is observed.

About 80 % of the AQX events in NA and EU occur in
large, connected, multi-day episodes consisting of 100 grid
cells or more. This result is closely matched by all but two
models, with C producing much larger episodes and A much
smaller ones. It remains unclear whether such errors result
from chemical or physical processes. The observations show
that super-sized episodes of 100 cells or more have success-
vously greater O3 levels as they become bigger, with the 100-
times-larger episodes having 4–6 ppb greater O3. Most, but
not all of the models match this increase. It is likely that
larger, longer-lasting episodes allow for greater accumula-
tion of O3 from neighboring pollution sources.
4.1 What are the best air quality diagnostics for model development?

For testing and identifying the model strengths and weaknesses and improving simulations of air quality, modelers save a large number of diagnostics during the model development process. This typical model development process is far less limiting than the experiment analyzed here, which is based on the voluntary contributions of many models and many terabytes of diagnostics imposed in the ACCMIP. We would still recommend saving the diagnostic of hourly surface O₃ over a decade or more of simulation from which all of the primary diagnostics here can be readily derived and compared with the observations. To segue from the surface O₃ over NA and EU to the sondes and remote sites, a monthly averaged 3-D O₃ would probably suffice. Hourly data observed at coastal or mountain sites likely include a diurnal meteorology that is not represented in the global models, even at a resolution of 0.5° x 0.5°. Furthermore, the 24 h and MDA8 averages show different biases and should not be treated as the same diagnostic. There may be inventive ways to avoid the massive hourly data sets by storing the diurnal cycle as a monthly mean and calculating MDA8 inline or just storing the maximum daily O₃ value, which would then require similar analysis of the observations.

The open questions are what model simulations are practical and which would be most useful to identify model errors. The ACCMIP simulations forced by a decade of 2000s climate-model sea surface temperatures are useful in comparing decadal statistics, but the UCI CTM hindcast provides unique tests on the ability to simulate specific events and years. Even if the observed sea surface temperatures were used, the synoptic extreme events would not likely coincide with the observed, so a hindcast meteorology based on reanalysis for forecast fields provides an important test of the model.

The surface O₃ data here are based on an interpolation algorithm that was optimized for the 50–100 km scale averages. Thus, the supplied grid-cell averaged data could be regenerated at 0.5° resolution, but if one wants 10 km cell averages for regional models then the parameters in the current algorithm would need to be revised and re-optimized. The surface O₃ data set will be expanded to include more than 2 decades (1993–2015) and thus longer simulations would be desirable to investigate interannual variability.

4.2 What are the most important tests for these chemistry–climate models, assuming that hindcasts and detailed emission data are not being used?

Another major question is what emissions to use. With ACCMIP the choice of a single year of representative emissions for the decade was the optimal choice. The downward trending emissions in NA and EU over the 2000–2009 decade, however, created a non-stationary data set, 1993–2015, will make the comparison between models and measurements more awkward. Model developers will need to take some account of this non-stationarity, possibly as a sensitivity study using two different emissions sets representative of the early and late periods of observations, when not tracking emission changes each year.

An emissions problem not resolved here is whether the modeled diurnal cycle over heavily polluted regions in summer would be affected by imposing a more accurate diurnal and weekly cycle in emissions. This is probably beyond what can be imposed in a model intercomparison project such as the ACCMIP but should be part of the individual model development as a sensitivity assessment.

The four-region decadal average statistics here provide a fairly broad view of the models’ ability to predict the buildup of O₃ and extreme events in polluted regions. Clear examples of model error are identified. The general agreement of the diurnal cycle between models and measurements still needs to be tested with diurnal emissions. Going beyond the mean regional cycles, the ability to test models at the grid-cell level provides clear geographic coverage, identifying patterns of the discrepancy that are sometimes disturbing, as shown in Fig. 3, but not developed further in this paper. The next study of the CMIP-generated surface O₃ needs to evaluate this.

4.3 What tests provide the best confidence in model prediction of future air quality?

Accurate projections of future air quality rely on our ability to predict the changes in both baseline level and pollution buildup in response to both specified future climatic conditions and a change in local-to-global emissions. Both the baseline and the amount of O₃ produced from pollution are likely to change and need to be assessed separately. For that purpose, we find that the maps of summertime (87th percentile) and baseline (30th percentile) and their difference are one of the more important tests of a model’s simulation of the present day. The annual cycle of monthly means is also in some way a measure of the summertime enhancement but not as useful as the percentiles. One key measure of future change would be in the size and intensity of extreme episodes. The intensity needs to be assessed relative to the baseline, but the size of the episodes clearly relates to their intensity and would be independent of shifts in baseline. Thus the AQX statistics based on the daily MDA8 values here are an important model test.

The Supplement related to this article is available online at doi:10.5194/acp-15-10581-2015-supplement.
Acknowledgements. Research at UCI was supported by NASA grants NNX09AJ47G, NNX13AL12G, NNX15AE35G, and DOE award DE-SC0007021. J. L. Schnell was supported by the National Science Foundation’s Graduate Research Fellowship Program (DGE-1321846). The work of D. Bergmann and P. Cameron-Smith was funded by the US Dept. of Energy (BER), performed under the auspices of LLNL under contract DE-AC52-07NA27344, and used the supercomputing resources of NERSC under contract no. DE-AC02-05CH11231. G. Zeng acknowledges the use of New Zealand’s national HPC facilities that are provided by the NZ eScience Infrastructure and funded jointly by NeSI’s collaborator institutions and through the Ministry of Business, Innovation & Employment’s Research Infrastructure Programme. The simulations with MIROC-CHEM was supported by the Global Environment Research Fund (S-7) by the Ministry of the Environment Japan and completed with the supercomputer (NEC SX-8R) at the National Institute for Environmental Studies (NIES). We are grateful to the US Environmental Protection Agency’s (EPA) Air Quality System (AQS) and Clean Air Status and Trends Network (CASTNet), Environment Canada’s National Air Pollution Surveillance Program (NAPS), the European Monitoring and Evaluation Programme (EMEP), and the European Environment Agency’s (EEA) air quality database (AirBase) for providing the observational data sets used in this study. We are also grateful to the British Atmospheric Data Centre (BADC), which is part of the NERC National Centre for Atmospheric Science (NCAS), for collecting and archiving the ACCMIP data.

Edited by: L. Ganzeveld

References


Appendix C

Effect of climate change on surface ozone over North America, Europe, and East Asia

Reprint of:

Effect of climate change on surface ozone over North America, Europe, and East Asia

Jordan L. Schnell1, Michael J. Prather1, Beatrice Josse2, Vaishali Naik3, Larry W. Horowitz4, Guang Zeng5, Drew T. Shindell6, and Greg Faluvegi7

1Department of Earth System Science, University of California, Irvine, California, USA, 2GAME/CNRM, Météo-France, CNRS–Centre National de Recherches Météorologiques, Toulouse, France, 3UCAR/NOAA Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Princeton, New Jersey, USA, 4Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Princeton, New Jersey, USA, 5National Institute of Water and Atmospheric Research, Lauder, New Zealand, 6Nicholas School of the Environment, Duke University, Durham, North Carolina, USA, 7NASA Goddard Institute for Space Studies and Columbia Earth Institute, Columbia University, New York, New York, USA

Abstract The effect of future climate change on surface ozone over North America, Europe, and East Asia is evaluated using present-day (2000s) and future (2100s) hourly surface ozone simulated by four global models. Future climate follows RCP8.5, while methane and anthropogenic ozone precursors are fixed at year 2000 levels. Climate change shifts the seasonal surface ozone peak to earlier in the year and increases the amplitude of the annual cycle. Increases in mean summertime and high-percentile ozone are generally found in polluted environments, while decreases are found in clean environments. We propose that climate change augments the efficiency of precursor emissions to generate surface ozone in polluted regions, thus reducing precursor export to neighboring downwind locations. Even with constant biogenic emissions, climate change causes the largest ozone increases at high percentiles. In most cases, air quality extreme episodes become larger and contain higher ozone levels relative to the rest of the distribution.

1. Introduction

Future surface ozone (O₃) will be determined by multiple factors, including changes in stratosphere–troposphere exchange [e.g., Zeng and Pyle, 2003], changes in anthropogenic and natural emissions of O₃ precursors, plus climate-induced shifts in meteorology and background tropospheric chemistry [Jacob and Winner, 2009; Fiore et al., 2012, 2015]. Future scenarios of O₃ precursors adopted in the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5) [Cubasch et al., 2013; Kirtman et al., 2013; Myhre et al., 2013] project lower overall anthropogenic emissions by 2100 in most locations, but one scenario has large increases in methane (also an O₃ precursor). Future air quality is expected to improve over North America and Europe in all scenarios and improve worldwide if methane increases are avoided [e.g., Kirtman et al., 2013; Pfister et al., 2014]. These IPCC results combine climate and emission changes. Climate change in and of itself will likely reduce O₃ in unpolluted conditions [Johnson et al., 1999] but possibly increase surface O₃ in polluted regions (e.g., “climate change penalty” of Wu et al. [2008]), thus offsetting the benefits of precursor emission reductions. Here we examine the role of climate change on continental surface O₃ using global chemistry model simulations with fixed anthropogenic emissions of ozone precursors (and fixed methane abundances) as climate changes from 2000 to 2100.

Climate change will affect surface O₃ through numerous temperature-driven pathways [e.g., Jacob and Winner, 2009; Thambiran and Diab, 2010; Fiore et al., 2012]. While these pathways have a physical basis, the magnitude is usually derived from present-day correlations with observed surface O₃ and meteorological proxy variables meant to represent the pathway [e.g., Ordonez et al., 2005; Camalier et al., 2007]. Fully coupled chemistry-climate model simulations on the time-space scales relevant for air quality studies are uncommon, so future O₃ changes are often statistically downscaled using proxy data, i.e., global or regional climate model projections of relevant meteorological variables [e.g., Mahmud et al., 2008; Holloway et al., 2008]. Unfortunately, these O₃ meteorology correlations may not apply to future climate since they often reflect a common underlying driver and do not represent the net effect of photochemistry, meteorology, and land-atmosphere interactions on O₃ [Fiore et al., 2015]. For example, large O₃ increases often occur during periods with clear skies, high temperatures, and light winds (i.e., stagnation events), which allow for increased O₃ production and accumulation of nearby emissions [Logan, 1989; Hogrefe et al., 2004; Mickley et al., 2004;
Leibensperger et al., 2008]. Thus, one can define a stagnation index based on year 2000 meteorology and argue that a year 2100 increase in this index implies a proportionate increase in the worst pollution days [Horton et al., 2014]. However, a stagnation index is just that, a proxy index, and requires that the underlying processes responsible for the present-day correlations be maintained in the future.

Our approach uses hourly surface O₃ simulated by models participating in the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP) conducted in support of the IPCC AR5 [Lamarque et al., 2013; Naik et al., 2013; Stevenson et al., 2013; Voulgarakis et al., 2013; Young et al., 2013]. Other direct studies are generally either limited to a single model, restricted regions, or relied on annual and seasonal average diagnostics [e.g., Doherty et al., 2013; Clifton et al., 2014; Rieder et al., 2015]. We use hourly O₃ to identify the daily and synoptic variability and air quality extremes (AQX), enabling an investigation of climate-driven changes in pollution episodes at high temporal resolution.

2. Data and Domains

2.1. ACCMIP Models

We analyze model results from two ACCMIP experiments: acchist (a present-day climatology, henceforth CL2000) and Em2000Cl2100 (present-day O₃ precursors with year 2100 climate, henceforth CL2100). The ACCMIP models that correctly implemented the experiments and saved hourly surface O₃ are used here: MOCAGE, GFDD-AM3, UM-CAM, and GISS-E2-R (see Table S1 in the supporting information) [Lamarque et al., 2013, and references therein]. Climate change follows the IPCC AR5 high-CO₂ scenario (RCP8.5) with global mean CL2000 to CL2100 surface temperature increases of ~4°C. These two simulations isolate the effects of climate change by fixing anthropogenic emissions of O₃ and aerosol precursors and the methane abundances used for chemistry to present-day levels while forcing the physical climate (and some natural precursor emissions; see below) by RCP8.5 sea surface temperatures and sea ice distributions. Schnell et al. [2015] (hereinafter S2015) evaluated the models’ present-day surface O₃ climatologies against North American and European observations. Although generally biased high in surface O₃, the models capture the shape and amplitude of the diurnal and annual cycles, large-scale percentile patterns, and the size distribution of AQX episodes. The analysis of S2015 thus provides some confidence in the modeled future changes (see S2015 for a full description of the metrics used here).

While anthropogenic emissions in the ACCMIP were specified, natural emissions were not, so the models used different formula for natural emissions with differing responses to climate changes [Young et al., 2013]. For example, lightning NOₓ emissions are usually parameterized as a function of deep convection and hence are climate sensitive. Across our four models, these emissions ranged from 4.4 to 7.7 Tg N a⁻¹ for CL2000 and increased by 21% to 44% for CL2100 [Lamarque et al., 2013]. Climate-driven biogenic volatile organic compound (BVOC) emission changes may be responsible for O₃ changes in some regions [Lin et al., 2008]. Isoprene, the most abundant BVOC, is chemically very reactive, either increasing or decreasing near-surface O₃ under high or low NOₓ conditions, respectively. Only GISS-E2-R incorporates climate-driven isoprene emissions that increase with increasing temperature [Guenther et al., 1995], and thus, CL2100 has greater (natural) BVOC emissions than CL2000 [Young et al., 2013]. Other direct climate-sensitive feedbacks from increased CO₂ include inhibited isoprene emissions [Heald et al., 2009] or reduced stomatal uptake of O₃ [Ainsworth and Rogers, 2007], but neither were included here.

Each model’s hourly surface O₃ abundances (typically at 2°–3°) are remapped to a common 1° × 1° grid using first-order conservative mapping to facilitate a more direct intermodel comparison and maintain consistency with S2015. Maximum daily 8 h averages (MDA8) are calculated from the hourly O₃ abundances after adjusting to local solar time. All analyses shown here are performed using the MDA8.

2.2. Domains and Regions

We investigate changes over three continental-scale domains: North America (NA) bounded by 25°N–49°N and 125°W–67°W, Europe (EU) bounded by 36°N–71°N and 11°W–34°E, and East Asia (EA) bounded by 8°N–42°N and 69°E–130°E. Due to differing chemical and climatic regimes, we split each domain into two regions: NA into western (WNA) and eastern (ENA) regions at 96°W, EU into southern (SEU) and northern (NEU) regions at 33°N, and EA into southern (SEA) and northern (NEA) regions at 30°N. The split location for each domain is somewhat arbitrary but roughly divides the domains in half and coincides with natural
breaks in the seasonal chemical regime. Henceforth, the terms “domain” and “region” are used to describe the whole and split domains, respectively.

3. Results

We evaluate 21st century climate-driven changes from decadal differences ($\Delta = \text{CL2100 minus CL2000}$) in the probability distribution (PD) of MDA8 surface O$_3$ to identify the full scope of changes, from the cleanest to most polluted conditions. At each grid cell and simulated period, we compute seasonal averages as well as percentiles from the daily values. From the multiyear PDs, we calculate seasonal medians as the percentile of the median day in each season. If the seasons were alike, then all would have a percentile ranking of 50%. Regional average changes are calculated as the area-weighted average of the derived changes in all encompassed grid cells. Climate-driven $\Delta$ changes are presented in order of increasing percentile, from the low O$_3$ of boreal winter to the high O$_3$ of summer.

3.1. Changes in Incoming Boundary and Low-Percentile O$_3$

Boundary O$_3$ is defined here as the statistically cleanest air impinging on the continents, providing the baseline O$_3$ upon which pollution accumulates. For our six continental regions, boundary O$_3$ largely originates from the marine boundary layer. Modeled surface O$_3$ changes over adjacent ocean basins are presented in Figure S1 for mean winter (December-January-February, DJF) and Figure S2 for mean summer (June-July-August, JJA). These seasonal average changes should reflect the most basic hemispheric-scale, climate-driven shifts such as increased water vapor.

Over the tropical and subtropical oceans, photochemistry drives a net O$_3$ loss of $\sim$10%/d, which increases with temperature and water vapor [Johnson et al., 1999], thus enhancing loss in the CL2100 simulation. From Figure S1, all models except MOCAGE show positive $\Delta$DJF (+1 to +5 ppb) over the Pacific and Atlantic Oceans north of $\sim$20°N and a sharp transition to negative $\Delta$DJF ($\sim$1 to $\sim$8 ppb) to the south. This $\Delta$DJF edge over the Pacific (and Atlantic to a lesser extent) augments the north-south gradient of CL2000 DJF O$_3$ (not shown). All models show negative $\Delta$DJF over the Indian Ocean adjacent to SEA. We find negative $\Delta$JJA (Figure S2) over all three oceans (except GISS-E2-R over the Pacific and Atlantic). Negative $\Delta$ changes are easily attributable to enhanced destruction in a warmer, wetter marine boundary layer. A cause for positive $\Delta$DJF in the extratropics is less clear but may be due to enhanced vertical mixing with high O$_3$ in the free troposphere or possibly enhanced wintertime photochemical production (e.g., from increased lightning NOx).

Nevertheless, all models except MOCAGE show increased DJF boundary O$_3$ everywhere except SEA, and all models except GISS-E2-R show decreased JJA boundary O$_3$.

Figure 1 summarizes percentile and seasonal MDA8 O$_3$ changes for the four models over six continental regions (i.e., 24 model regions). The colored bars show each model region’s MDA8 O$_3$ $\Delta$ change averaged over 10 equal percentile bins (a uniform color means that the entire PD has shifted, whereas red at the high end and blue at the low end mean that the worst days became more polluted and the best days became cleaner). The median percentile of each season’s day is shown for CL2000 (black filled markers) and for CL2100 (color filled markers), where the fill color identifies the $\Delta$ change in the season’s average. Observed CL2000 (S2015) seasonal medians are given for NA and EU regions. The dashed lines at the 13th and 87th percentiles respectively represent the minimum and maximum possible values for a seasonal median (i.e., if all days in a season have the lowest or highest abundances in the entire period).

Modeled CL2000 seasonal percentiles agree with present-day observations over NA and EU in that fall (September-October-November, SON) or winter (DJF) typically has the lowest O$_3$ percentiles (Figure 1) [see also S2015]. Because $\Delta$DJF (and $\Delta$SON, not shown) is positive over the midlatitude oceans, we expect $\Delta$DJF and $\Delta$SON to be positive and they are 0 to +5 ppb for 16 of 24 model regions. Changes in the lowest 10% ($\Delta$5th), however, do not follow these patterns but show decreases ($\sim$1 to $\sim$4 ppb for 16 of 24 model regions). This pattern fits across models, including UM-CAM with the lowest DJF percentiles, and likely reflects increased daily variability of low-season O$_3$ in CL2100 (Table S2). For subtropical SEA, we expect $\Delta$5th to be related to $\Delta$JJA when monsoons and tropical, low-O$_3$ air occurs. Indeed, all models show both negative $\Delta$5th and $\Delta$JJA over SEA, consistent with their negative $\Delta$JJA over the Indian Ocean. These $\Delta$ changes in monsoonal surface O$_3$ could be attributed to enhanced photochemical loss, alterations to the location or intensity of the Asian monsoon [Christensen et al., 2013], or both.
The geographic patterns of the multimodel mean of $\Delta$5th and $\Delta$DJF are shown in Figures 2a–2c and 2d–2f, respectively, with individual model-domain maps in Figures S3 and S4 and model region averages in Table S2. Consistent with Figure 1, $\Delta$5th is predominately negative and $\Delta$DJF is positive, with some similarity in their geographic patterns. One standout feature in NA $\Delta$DJF is the break at 40°N with positive $\Delta$DJF to the south and negative to the north. The cause may be photochemical with warmer temperatures in the south but is more likely due to circulation changes. Clear differences between $\Delta$5th and $\Delta$DJF for NA and EU occur in areas most influenced by clean maritime air (negative $\Delta$5th and positive $\Delta$DJF), but this is less obvious for EA. Thus, in maritime regions, $\Delta$DJF is driven by CL2100 increases in maritime DJF O3, and extreme low-O3 events ($\Delta$5th) have shifted seasons, generally out of DJF and into JJA or SON (Figure S5).

Figure 1. Percentile and seasonal MDA8 surface O3 changes (CL2100 minus CL2000, ppb) for the ACCMIP models MOCAGE (A), GFDL-AM3 (B), UM-CAM (C), and GISS-E2-R (D) over (a) WNA, (b) ENA, (c) SEU, (d) NEU, (e) SEA, and (f) NEA. Colored bars show the absolute MDA8 surface O3 change averaged over 10 equal percentile bins. The median percentiles of the days in winter (DJF), spring (MAM), summer (JJA), and fall (SON) are shown for modeled CL2000 (black markers) and CL2100 (colored markers), where the CL2100 fill colors correspond to the change in each season’s average. Observed CL2000 median seasonal percentiles (S2015) are plotted on the top edge of NA and EU regions’ subplots (green markers). All changes correspond to the single color bar. Dashed lines at the 13th and 87th percentiles respectively represent the minimum and maximum possible values for a median season percentile.

The geographic patterns of the multimodel mean of $\Delta$5th and $\Delta$DJF are shown in Figures 2a–2c and 2d–2f, respectively, with individual model-domain maps in Figures S3 and S4 and model region averages in Table S2. Consistent with Figure 1, $\Delta$5th is predominately negative and $\Delta$DJF is positive, with some similarity in their geographic patterns. One standout feature in NA $\Delta$DJF is the break at 40°N with positive $\Delta$DJF to the south and negative to the north. The cause may be photochemical with warmer temperatures in the south but is more likely due to circulation changes. Clear differences between $\Delta$5th and $\Delta$DJF for NA and EU occur in areas most influenced by clean maritime air (negative $\Delta$5th and positive $\Delta$DJF), but this is less obvious for EA. Thus, in maritime regions, $\Delta$DJF is driven by CL2100 increases in maritime DJF O3, and extreme low-O3 events ($\Delta$5th) have shifted seasons, generally out of DJF and into JJA or SON (Figure S5).
The 30th percentile represents a photochemical baseline O3 season regardless of seasonal ranking differences; e.g., the lowest O3 season is not always winter but typically occurs between the 20th and 40th percentiles (Figure 1). The 30th percentile is stable, showing little year-to-year variation in observations and in present-day ACCMIP model simulations (S2015). The multimodel mean of $\Delta 30^{th}$ over each domain is

Figure 2. ACCMIP multimodel mean of MDA8 O3 changes (CL2100 minus CL2000, ppb) in the (a–c) 0–10th percentile bin ($\Delta 5^{th}$), (d–f) winter ($\Delta$DJF), (g–i) 30th percentile ($\Delta 30^{th}$), (j–k) summer ($\Delta$JJA), (m–o) 87th percentile ($\Delta 87^{th}$), and (p–r) peak-to-peak amplitude of the annual cycle ($\Delta M$) over North America (left column), Europe (middle column), and East Asia (right column). Individual model-domain maps for each quantity can be found in the supporting information.

The 30th percentile represents a photochemical baseline O3 season regardless of seasonal ranking differences; e.g., the lowest O3 season is not always winter but typically occurs between the 20th and 40th percentiles (Figure 1). The 30th percentile is stable, showing little year-to-year variation in observations and in present-day ACCMIP model simulations (S2015). The multimodel mean of $\Delta 30^{th}$ over each domain is
shown in Figures 2g–2i, with individual model-domain maps in Figure S6 and model region averages in Table S2. Consistent with $\Delta$DJF, we find positive $\Delta$30th (0 to +3 ppb) over most of NA and EU. Increases are also found over the Himalayan Plateau, likely reflecting increased exchange with the free troposphere or stratosphere [Hsu et al., 2005]. Comparing $\Delta$30th and $\Delta$5th over NA shows that lower O$_3$ prevails for at least 4 months over the desert southwest and the Gulf Coast. We associate the large negative $\Delta$30th over most of EA (−5 ppb), especially the near-ocean areas, with changes in the Asian summer monsoon since the median percentile of JJA days is within ±5% of the 30th percentile in all models for both CL2000 and CL2100 (Figure 1).

### 3.2. Changes in High Percentiles and Seasonal Cycle of O$_3$

The highest surface O$_3$ abundances typically occur when precursors, sunlight, and meteorology align. Hemispheric-scale processes that alter the basic latitudinal surface O$_3$ gradients (e.g., the summertime midlatitude jet location [Barnes and Fiore, 2013; Barnes and Polvani, 2013]) and vertical mixing also play important roles.

The seasons with the highest median percentile are summer (JJA) for ENA and SEU; spring (March-April-May, MAN) for WNA, NEU, and NEA; and either DJF or MAM for SEA (Figure 1). We calculate the timing of peak O$_3$ ($m$, days of the year) from the phase of a cosine fit to each grid cell’s monthly averaged MDA8 O$_3$. The seasonal amplitude \( M \) (ppb) is also derived from this fit. To see the representativeness of these fits, see S2015 (Figures 1–11). Individual model-domain maps of $\Delta m$ are shown in Figure S7 with model region averages in Table S3.

GFDL-AM3 and UM-CAM show $\Delta m < 0$ (i.e., peak O$_3$ shifting earlier) in all regions, consistent with an earlier study using transient simulations of the fully coupled GFDL-CM3 over the eastern U.S. [Clifton et al., 2014]. However, Clifton et al. (2014) find that the climate impact reinforces the potentially much larger impact of RCP8.5 O$_3$ precursor emission reductions, particularly with regard to the phase of the seasonal cycle. The largest negative $\Delta m$ (−15 days) are found over regions where the peak O$_3$ season is early (MAM), while the smallest are found over JJA peak regions. An earlier arrival of peak O$_3$ indicates a climate-driven O$_3$ increase in spring, a decrease in summer, or both. Among the four models, GFDL-AM3 has the largest negative $\Delta m$ and the largest positive $\Delta$MAM. Overall, 14 of 24 model regions have positive $\Delta$MAM and 17 have negative $\Delta$JA (Table S2). GISS-E2-R shows later arrivals of peak O$_3$ ($\Delta m > 0$) in NA and EU regions, possibly related to its climate-sensitive BVOCs, but it also had the poorest simulation of present-day seasonality over WNA and NEU (S2015), thought to be related to excessive wintertime stratosphere-troposphere exchange [Shindell et al., 2013]. MOCAGE also has $\Delta m > 0$ for all regions except WNA (Table S3); however, MOCAGE has $\Delta m < 0$ over most of NA, India, and western China (Figure S7).

GISS-E2-R accounts for five of the seven model regions with positive $\Delta$JA, almost certainly due to its climate-sensitive BVOCs. Indeed, its largest $\Delta$JA are in regions with large BVOC emissions: ENA (+7.1 ppb) and SEU (+9.3 ppb). Figures 2j–2l show the multimodel mean of $\Delta$JA over each domain, with individual model-domain maps in Figure S8 and model region averages in Table S2. We find negative $\Delta$JA over most of southern NA, northwest EU, and all of EA except northeast China. Negative $\Delta$JA in the near-ocean areas reflects decreases in boundary O$_3$. Only GISS-E2-R has positive $\Delta$JA in these areas because it has increases in boundary O$_3$. Positive $\Delta$JA is found over the western and northern edge of NA, the Ohio River Valley extending to the northeast, all of SEU, and northeastern China. These multimodel mean increases largely reflect GISS-E2-R; however, all models show positive $\Delta$JA over polluted regions such as northeastern NA, the Po Valley, and northeastern China.

Overall, during the peak photochemical season JJA, climate change increases O$_3$ in polluted regions and decreases O$_3$ in nearby cleaner regions, broadly consistent with previous findings [e.g., Johnson et al., 1999; Wu et al., 2008]. We hypothesize that warmer temperatures increase the efficiency of precursors to produce O$_3$ in polluted regions, consequently reducing precursor availability in neighboring, cleaner, downwind locations, where NO$_x$ is usually more efficient in producing O$_3$. On a much broader scale, Doherty et al. [2013] use one model to show that the more rapid thermal decomposition of organonitrites expected in a warming climate can lead to a few ppb increase in the annual average of surface O$_3$ over land and a corresponding decrease over the oceans. The ACCMIP models used here include all climate-driven effects on photochemical kinetics, and we find that this climate-driven shift in surface O$_3$ has a much greater magnitude (up to 10 ppb).
is found particularly at higher percentiles of O₃ (i.e., the AQX episodes; see section 3.3), and occurs on the 200 km scale within the continents.

The negative ΔJJA in south central NA and large positive ΔJJA over the northeast may indicate a CL2100 westward extension of the Bermuda High and subsequent changes in the related Great Plains low-level jet [e.g., Eder et al., 1993; Fiore et al., 2003; Shen et al., 2015]. Some studies suggest an ~5° westward shift in this pattern by 2100 [Li et al., 2012, 2013], which would decrease O₃ in central NA through enhanced flux of low-O₃ air from the Gulf of Mexico but increase O₃ in northeast NA by extending the high-pressure system’s stagnation conditions. This pattern is seen in all four models (Figures S8a, S8b, S8g, and S8j) but at different magnitudes.

The 87th percentile was chosen in S2015 to represent the median value of a hypothetical season containing all high-O₃ days. It allows for peak O₃ to shift away from JJA and thus is a more robust measure of the photochemical O₃ season. The multimodel mean of δ87th over each domain is shown in Figures 2m–2o, with individual model-domain maps in Figure S9 and model region averages in Table S2. We find δ87th is predominantly positive and largest over the most polluted regions (e.g., California, ENA, SEU, and northeast China). This feature is largely consistent with ΔJJA patterns, but it is more pronounced and missing the JJA reductions due to CL2000-enhanced onshore flow in south central NA and the monsoon regions of EA.

The amplitude of the O₃ annual cycle increases with climate change: 17 of 24 model regions show positive ΔM (Figures 2p–2r and S10 and Table S3). Like δ87th and ΔJJA, the largest positive ΔM are found in the most polluted regions whereas the largest negative ΔM are found in southern NA and northern EU (where ΔJJA is most negative). The mostly positive ΔM may appear contradictory to general findings of negative ΔJJA and positive ΔDJF but instead indicates a shift in peak O₃ away from JJA and in the lowest O₃ away from DJF. The geographic patterns of ΔM, ΔJJA, and δ87th coincide roughly with areas where the proxy index for stagnation days changes in a future climate [Horton et al., 2014].

### 3.3. Changes in Air Quality Extreme (AQX) Episodes

Air quality extreme (AQX) events are defined for each grid cell as the 10 times N worst days (i.e., highest MDA8) in an N year period [see Schnell et al., 2014]. Because AQX events are defined as a return time, we cannot evaluate climate-driven changes in the number of events; however, we can quantify shifts in their relative intensity, seasonality, and the space-time clustering of events into episodes.

We evaluate AQX seasonality by deriving the timing of maximum phase (τAQX, days of the year) from a cosine fit to each grid cell’s monthly binned AQX events. With climate change, AQX events generally shift to earlier in the year (ΔτAQX < 0 for 16 of 24 model regions), by a greater amount than the seasonal cycle of MDA8 (Table S3). Because of its climate-driven BVOC emissions, GISS-E2-R has more AQX events in late summer and positive ΔτAQX for NA and EU, consistent with its large ΔJJA.

Neighboring AQX events connected in space and/or time are clustered into AQX episodes [Schnell et al., 2014]. Episode sizes S range from a single-cell, 1 day event (10⁻¹⁰ km² d) to multiday episodes spanning hundreds of kilometers (10⁻⁴ km² d). For each model region, we calculate (i) a complementary cumulative distribution of episodes as a function of size (CCDₜ₈₇, percent of events’ areas in episodes of size S or larger); (ii) a mean episode size ⟨S⟩, 10⁴ km² d, weighted geometric mean with weights equal to S [Schnell et al., 2014, equation 6]; (iii) an episode enhancement as a function of size (Eₜ, ppb, time-area weighted average of the magnitude of all AQX event cells above each cell’s 30th percentile O₃ value); and (iv) the average enhancement increase with episode size (ΔEₜ/ΔS, ppb/decade) for episodes S ≥ 30 × 10⁴ km² d (S2015). Climate-driven changes have different units: (i) ΔCCDₜ as absolute differences in CCDₜ (%; Figures 3a–3f); (ii) Δ⟨S⟩ as percent relative to CL2000 to account for the large intermodel and interregion range of episode sizes (Table S4); and (iii) ΔEₜ as a 2-D color-contour plot of the multimodel mean percent change in Eₜ (designated by the color bar) as a function of enhancement for each log-scale size bin in S (Figures 3g–3l). For each half-decade size bin (edges at 3, 10, 30, 100, 300, 1000, and 10⁻⁴ km² d), the sum of ΔEₜ is zero; a red color indicates where particular enhancements Eₜ are relatively more prevalent, and a blue color, where less prevalent.

Figures 3a–3f show ΔCCDₜ for each model region plotted at each half decade in S. As expected, the largest ΔCCDₜ is found for S > 300 × 10⁴ km² d where the CCDₜ curve is steepest (S2015), but the sign is often split between the models in each region. The models are similarly split for Δ⟨S⟩, with 11 of 24 model regions showing positive values (Table S4). The sign of Δ⟨S⟩ and ΔCCDₜ for the largest S generally match for each model region.
These mixed results provide little confidence in how the distribution of sizes or mean size of future AQX episodes will change. Yet for the very largest episodes there is some consensus: 17 of 24 model regions, and at least two models in each region, show increases in the average size of the largest two episodes (Table S4).

The mean value of \( ES \) averaged over each size bin is shown for each model and the multimodel mean for CL2000 (circles) and CL2100 (triangles) in Figures 3g–3l. In all regions except NEU (the least photochemically active), CL2100 episodes show greater enhancements, particularly for \( S > 100 \times 10^4 \text{ km}^2 \text{ d}^{-1} \). The color-contour plot shows the mean area-weighted enhancement \( E_S \) (ppb above the 30th percentile) of each size bin for CL2000 (solid line, open circles) and CL2100 (dashed lines, open triangle) for each ACCMIP model (colors) and the multimodel mean (black). Columns from left to right correspond to regions: WNA, ENA, SEU, NEU, SEA, and NEA.

4. Summary and Conclusions

We investigate the effect of climate change on MDA8 surface \( \text{O}_3 \) over North America, Europe, and East Asia using four global models that participated in ACCMIP and archived hourly surface ozone abundances. We use climate change projections from scenario CL2000 (decade of the 2000s) to CL2100 (first decade of the 2100s, high-CO2 RCP8.5 scenario with global mean temperature increase of \( \sim 4^\circ\)C) and fixed anthropogenic \( \text{O}_3 \) precursor emissions (and methane concentrations for chemistry) for CL2000 and CL2100. We focus on North America and Europe for which present-day observations can be used to test the models (see S2015) and on East Asia, which consists of the midlatitude northern part and the monsoonal southern part.

Overall, the diverse patterns of \( \Delta M \) among the four models, plus the large climate-driven changes in BVOCs in one model, make it difficult to present a simple plot of future \( \text{O}_3 \) (even with fixed anthropogenic emissions). We present individual model as well as multimodel mean results but feel that the model mean
plots cover up current modeling uncertainty in this simulation. Nevertheless, we find some consistencies that we believe will survive a more thorough assessment. Climate change shifts the timing of peak O$_3$ to earlier in the year and increases the amplitude of the annual cycle. Similarly, climate change spreads out and shifts the timing of air quality extreme (AQX) events to earlier in the year. There are clear seasonal differences in baseline O$_3$ levels entering the continents for CL2100, but the cause is unclear.

Increases in summertime mean and high-percentile O$_3$ are generally found in polluted environments, with decreases found in clean environments. We propose that this pattern of "the most polluted get worse while their neighbors get better" reflects an augmented efficiency of precursor emissions to generate surface O$_3$ in the polluted regions under future climate change (warmer temperatures, more water vapor, and faster chemical kinetics), thus reducing export of precursors to neighboring downwind locations. All models show climate-driven increases in summertime surface O$_3$ over the northeast U.S., the Po Valley, and northeast China. Even with constant biogenic emissions, climate change increases O$_3$ at the upper tail of the probability distribution in most models and regions. In most cases, AQX episodes become larger and contain higher O$_3$ levels relative to the rest of the distribution. Thus, the extremes become more hazardous.

Further studies of the climate-driven changes in air quality, as opposed to local emissions-driven changes, should include a wider range of models and assess the balance between changes in anthropogenic and natural emissions of O$_3$ precursors. Broader chemistry-climate model participation is needed to develop more robust findings. Including models with more ensemble members and simulation years may further reduce the uncertainty of climate change impacts on O$_3$ and related photochemistry. Nevertheless, the indication here that the most extreme air pollution episodes are more likely than not to become more extensive and more severe poses a serious challenge for our posterity in managing air quality.

Acknowledgments
Research at UCI was supported by NASA grants NNX09JA73G, NN113AL12G, and NNX15AE35G and DOE award DE-SC0007021. J.L. Schnell was supported by the National Science Foundation’s Graduate Research Fellowship Program (DGE-1321846). G. Zeng acknowledges the use of New Zealand’s national HPC facilities that are provided by the NZ eScience Infrastructure and funded jointly by NeSI’s collaborator institutions and through the Ministry of Business, Innovation and Employment’s Research Infrastructure Programme. We are grateful to the British Atmospheric Data Centre (BADC), which is part of the NERC National Centre for Atmospheric Science (NCAS), for collecting and archiving the ACCMIP data. The hourly ACCMIP data at the models’ native resolutions can be obtained at http://badc.nerc.ac.uk, while the remapped MDA8 ozone data can be obtained by contacting the corresponding author.

References
Appendix D

Co-occurrence of extremes in ozone, particulate matter, and temperature

Adapted from:

Co-occurrence of extremes in ozone, particulate matter, and temperature

Jordan L. Schnell and Michael J. Prather

Department of Earth System Science, University of California, Irvine, CA 92697-3100, USA

Corresponding author: Jordan L. Schnell, Department of Earth System Science, University of California, Irvine, CA 92697-3100, USA, +1 (812) 630-7993, jschnell@uci.edu

Keywords: extremes; air quality; ozone; particulate matter; heat waves;

Classification: PHYSICAL SCIENCES

Significance statement

Exposure to extreme temperatures and high levels of the pollutants ozone and particulate matter pose a major threat to human health. Heat waves and pollution episodes share common underlying meteorological driver and thus often coincide, which can synergistically worsen their health impacts beyond the sum of their individual effects. Furthermore, there is evidence that pollution episodes and heat waves will worsen under future climate change making it imperative to understand the nature of their co-occurrence. In this paper, using fifteen years of surface observations over the eastern USA and Canada, we show that the extremes cluster together in often overlapping large-scale episodes and that the largest episodes have the hottest temperatures and highest levels of pollution.
Abstract

Heat waves and air pollution episodes pose a serious threat to human health and may worsen under future climate change. In this paper, we use fifteen years of commensurately gridded (1° x 1°) surface observations of extended summer (April–September) surface ozone (O3), fine particulate matter (PM2.5), and maximum temperature (TX) over the Eastern US and Canada to construct a climatology of the coincidence, overlap, and lag in space and time of the extremes. Extremes of each quantity are defined climatologically at each grid cell as the 50 days with the highest values in three 5-year windows (~95th percentile). We find that any two extremes co-occur on average between 27–35% of the time and >50% of the time in the Northeast US. The extremes also show consistent offsets in space in time, which contradict simple mechanistic explanations in some locations. The extremes most often occur in large-scale, multi-day, spatially connected episodes on scales of >1,000 km episodes and are clearly driven by meteorology, namely stagnation. The largest, longest-lived, clusters have the highest incidence of co-occurrence and contain extremes well above the 95th percentile threshold, i.e., by +7 ppb for O3, +6 μg m⁻³ for PM2.5, and +1.7 °C for TX. The results demonstrate the need to evaluate these extremes as synergistic co-stressors to accurately quantify their impacts on human health.

Introduction

Statistically extreme events in pollution, weather, and climate often pose risks to human health. In this paper, using fifteen years of observations over Eastern US and Canada, we exam three health extremes: surface ozone (O3) and fine particulate matter (PM2.5 = aerosols with diameter of 2.5 μm or less) and heat waves (T). O3 and PM are the two major air pollutants threatening human health (1). Heat waves also pose a major threat to human life (2–4). The average
magnitude of local air pollution is controlled primarily by local and regional emissions of the pollutants and their precursors; but, like heat waves (5), extreme air pollution is often driven by synoptic meteorology (6–7).

These three health extremes are often associated with slow moving high-pressure systems that accumulate pollutants and heat due to the overlying meteorological conditions, viz. high temperatures, abundant solar insolation, low precipitation, and low wind speeds (7–10). Aside from the direct influence of meteorology, there exists multiple interactions and feedbacks that act further to exacerbate extreme conditions. For example, the high temperatures during a heat wave increase emission rates of biogenic volatile organic compounds (BVOCs), which augment the production of surface O₃ and secondary organic aerosols (PM₂.₅). The drought-like conditions often accompanied by heat waves can inhibit stomatal uptake of ozone (11) and through soil moisture feedbacks can both amplify the heat wave and worsen air quality (12). Energy demand for air conditioning increases during heat waves (13), which can increase pollutant abundances to due greater anthropogenic emissions.

In any case, the extremes often co-occur as a result of their shared underlying drivers, greatly increasing the risks to human health (14). The imperative to understand the co-occurrence of health extremes is driven in part by the recognition that episodes of extreme temperatures (15–18) and poor air quality (19–24) may become more frequent, longer lasting, and more intense in a warming climate, where many climate-driven feedbacks can alter air quality independent of emissions (e.g., see reviews in refs. 8, 10, 25 and ref. 26).
That combined extremes produce greater impacts or risks than those summed simply from single extremes acting alone is a prevalent concept in the climate change community (27). In that case however, the multiple stressors go beyond physical extremes – such as the health events analyzed here – to include economic, social and political events (see Fig 1.5 in ref. 28). There is evidence that combined pollution extremes and heat waves are such synergistic stressors (i.e., impact modifiers) and that combined extremes produce disproportionately greater adverse health impacts (e.g., refs. 14, 29–33).

Here we use the methodology developed in refs. 23, 34–35 to calculate regularly gridded, daily surface values for O₃, PM₂.₅ and temperature. These quantities are continuous with real units (parts per billion by mole fraction (ppb); micrograms per cubic meter (µg m⁻³); and degrees Celsius), thus allowing us to develop a probability distribution for each grid cell and define extreme events in terms of their return time or frequency (days per year). An important, relevant finding from the previous work was that the mean absolute levels of O₃ in an extreme pollution episode generally increase with its geographic extent and duration.

The use of a gridded daily air stagnation index (ASI, ref. 36) as a proxy for health extremes has been further developed by refs. 37 and 26 and we include the ASI here because stagnation describes the basic meteorological conditions that exacerbate pollution episodes and heat waves. However, this index is Boolean (true or false) and cannot be used to define extreme events in terms of frequency or magnitude (e.g., in the Southeastern US about 40% of all summer days are considered stagnant, but in the more polluted corridor from Washington DC to Maine, there are half as many; see Fig. 1D).
We begin by defining extreme events, taking into account the non-stationarity of all three health extremes. We then present a case study of an episode of co-occurring extremes and build a climatology of the statistics of overlap and lag in space and time. We conclude by attempting to understand the unique spatiotemporal pattern found for each of our health extremes. Overall, the clear identification of large, overlapping, extreme pollution episodes – quantified in terms of observations of $O_3$, $PM_{2.5}$, and $T$ – provides a baseline evaluation for any model used to project future air quality and even heat waves.

**Defining extreme events and episodes.** Previous work (24, 34–35) defined $O_3$ extremes over the 2000–2009 decade in a local climatological manner as the 100 days (~97.3 percentile over the full year) with the highest maximum daily 8-hour average (MDA8). This approach enabled us to clearly identify large-scale multi-day pollution episodes consisting of connected locally extreme, daily events, that were not so readily seen using EPA’s absolute threshold approach: i.e., neighboring regions often have large systematic differences in absolute ozone levels; and thus a 75 ppb exceedance in one cell may not occur in its neighbor even though both cells had a 97th percentile event. Here we continue this approach to defining extremes in $O_3$, $PM_{2.5}$, and daily maximum temperature ($TX$). These three quantities can be treated identically, using their local probability distributions to define an extreme event; but the stagnation index, being Boolean, cannot define statistical extremes in terms of severity.

$PM_{2.5}$ extremes can occur in both summer and winter in Eastern North America (ENA), but because the vast majority of temperature and surface $O_3$ extremes occur between late-spring and early-fall, our statistics consider only days during the extended summer season (April 1–
September 30). We adopt the 95th percentile as the local extreme threshold, because that is equivalent to our 100-days-per-decade definition if all extremes occur during this extended summer period.

Fig. 1A–C show the 95th percentile over the entire period (1999–2013) for the three observed extremes at each grid cell. The 95th percentile of MDA8 O₃ (Fig. 1A) ranges from about 45 to 80 ppb across the domain, with the highest abundances in the Ohio River Valley and along the eastern urban corridor. The 95th percentile of PM₂.₅ (Fig. 1B) ranges from about 15 to 30 µg m⁻³ with a swath of >25 µg m⁻³ almost everywhere east of the Mississippi River between 33°N and 43°N. The 95th percentile of TX (Fig. 1C) is highest in the southwest corner of the domain (>37 °C) and monotonically decreases to its minimum over the Great Lakes and the northeast (<23 °C). Fig. 1D shows stagnation day frequency, which is greatest in the south and accounts for >30% of extended summer days in almost all grid cells south of 40°N, and >40% in a swath across the southern states (LA–AL–MS–GA). Stagnation frequency is much less common in the north and drops dramatically northward across 40°N (from ~32% to ~24%). Averaged over our domain, ~26% of the days are stagnant, many more than the upper 5% used for our extreme criteria.

**Trends in air quality and temperature.** Precursor emission reductions resulted in a clear decreasing trend in the 2000–2009 gridded MDA8 O₃ time series in ref. 34, causing a greater number of extremes to be identified in the early part of the decade. Emission reductions have been similarly effective in reducing PM₂.₅. Fig. 1E shows the domain averaged, annual value of the quantities shown in Fig. 1A–D. Air quality management has reduced the highest values of
both MDA8 O₃ and 24-h average PM₂.₅, with annual 95th percentile trends from 1999–2013 of
–0.9 ppb yr⁻¹ for O₃ and –0.8 µg m⁻³ yr⁻¹ for PM₂.₅. These reductions are much greater than those
in the median, with trends for O₃ and PM₂.₅ of –0.4 ppb yr⁻¹ and –0.3 µg m⁻³ yr⁻¹, respectively
(not shown). Temperature extremes are also occurring on an overall warming background (38),
but the trend here is minimal (+0.02 °C yr⁻¹). Stagnation frequency has no obvious trend, but the
interannual variability (IAV) follows the other three quantities (see correlations r in Fig. 1E).

These trends must be accounted for when identifying extreme events. Once solution is to force
each year to have the same number of extreme events (i.e., 10), but then IAV information is lost.
Indeed, the 95th percentile values of O₃ and PM₂.₅, when calculated annually, show remarkably
similar IAV, with a correlation coefficient of 0.89 over 1999–2013 (Fig. 1E). Some factors like wildfires may drive large IAV in PM (8), but not necessarily O\textsubscript{3}. The magnitude of this correlation indicates that for the 95th percentile extremes, meteorology is the most likely factor (e.g. refs. 39–40).

To minimize contamination by background trends while preserving interannual meteorological variability, we break the 1999–2013 O\textsubscript{3} and PM\textsubscript{2.5} time series into three 5-year windows. To avoid similar bias in temperature due to any warming trends, we treat the temperature record similarly. We follow the methods of ref. 33 and define extreme events in MDA8 O\textsubscript{3} (O\textsubscript{3X}), 24-h average PM\textsubscript{2.5} (PM\textsubscript{X}), and daily maximum temperature (T\textsubscript{XX}) at each grid cell as the 50 days with the highest values in each five-year window (~95th percentile).

**An example of co-occurrences: 20–28 June 2005.** Fig. 2 shows columns (A–D) of a daily sequence (rows) of extreme maps for the multi-day pollution episode of 20–28 June 2005. The first three columns show O\textsubscript{3X} (blue), PM\textsubscript{X} (green), and T\textsubscript{XX} (red). The fourth column (Fig. 2D) shows the combination of the first three columns, identifying cells with single or co-occurring extremes. For a single occurrence the cell retains the same color as in first 3 columns. For co-occurrences the cell is colored with the combined colors: O\textsubscript{3X}+PM\textsubscript{X} = blue+green = cyan; O\textsubscript{3X}+T\textsubscript{XX} = blue+red = magenta; PM\textsubscript{X}+T\textsubscript{XX} = green+red = yellow; all three, however, are black (instead of white). Stippling denotes a grid-cell stagnation day. An animation showing maps of the identified extremes, co-occurrences, and stagnation for each day of the 15-year record can be found in the SI.
Fig. 2. Nine day (20–28 June, 2005) episode progression for (A) O3X (blue), (B) PMX (green), (C) TXX (red), and (D) their combinations. Colors in (D) correspond to the combined individual RGB triplets: i.e., O3X+PMX (cyan), O3X+TXX (magenta), PMX+TXX (yellow), and O3X+PMX+TXX (black). Stippling denotes an identified stagnation day (white stipples in (D) if the grid cell has all three extreme types).
The episode begins on 20 June (Fig. 2, top row) on the western edge of the domain with O3X and PMX. The episode grows in size and slowly propagates eastward over the next three days. O3X first extends in a thin southwest-northeast filament before eventually covering the majority of the domain on 24 June. PMX coverage first extends eastward, but also displays a clockwise movement pattern after a few days, presumably a result of the combined influence of anticyclonic circulation and westerly flow. This feature highlights the potential role of meteorological transport on the location and timing of PMX. For example, because the highest absolute PM$_{2.5}$ abundances are found in the southeast (both in this episode example and on average, see Fig. 1B), an anticyclonic circulation pattern can transport high-PM$_{2.5}$ air to relatively cleaner areas such as the Midwest and Great Lakes region. TXX are rare until 23 June (day 4), where then they occur in the northwest quadrant of the domain. The episode reaches its maximum size on 24 June (day 5): 87% of the grid cells in the domain have at least one type of extreme, 64% have at least two, and 23% have all three. Over the final three days the episode decreases in size and continues to propagate eastward. The identified ASI days show erratic overlap with the health extremes, sometimes coinciding with a health extreme almost exactly (PMX on 22 June) and sometimes being exclusive (TXX on 23–24 June). However, the rows of Fig. 4D show that in most cases at least one type of extreme occurs when the grid cell is considered stagnant (68% of all grid cells and days). While this example is one of the larger episodes, it is hardly unusual (see discussion below and the animation in the SI), and we find that stagnation and most extreme levels of O$_3$, PM$_{2.5}$, and TX occur in large-scale, multi-day, coherent structures. The individual extremes also tend to group together within their own smaller scale clusters but often overlap (i.e., co-occurrences).
Fig. 3. Co-occurrence frequency (% of 150 events) for (A) O3X+PMX, (B) O3X+TXX, (C) PMX+TXX, and (D) O3X+PMX+TXX.

Fig. 3 shows the frequency of co-occurrence over the entire 15-year period (% of 150 events) at each grid cell for the three two-event combinations (i.e., O3X+PMX, O3X+TXX, and PMX+TXX in Fig. 3A–C, respectively) and the three-event combination (Fig. 3D). Averaged over the domain, O3X+PMX is most likely (35%), followed by PMX+TXX (29%), O3X+TXX (27%), and O3X+PMX+TXX (15%). The greatest co-occurrence frequencies for all event combinations are found in the Northeast US, greater than 50% for any two co-occurring events and about 30% for all three. The lowest co-occurrence frequencies are found in Florida, the Gulf Coast, and the western edge of the domain.

The extremes may show consistent offsets in space and time due to set spatial patterns in nearby precursor emissions or meteorology and related transport. For example, temperature is a well-known driver of O₃ production (41), and thus TXX may precede O3X in grid cells with large precursor emissions. Due to prevailing westerlies, this mechanism may result in O3X displaced westward of TXX.
**Systematic offsets in space.** We test for systematic spatial offsets by calculating the frequency of co-occurrence as a function of a spatial lag averaged across all grid cells. The 1° x 1° domain extends 27° in latitude and 30° in longitude, but because the extremes are unlikely to be correlated beyond ~500–1000 km, we limit the lags to ±8° and do not extrapolate any lags beyond the masked domain.

Fig. 4A–C show the domain average co-occurrence as a function of spatial lag for the corresponding event co-occurrence types in Fig. 3A–C. Also provided is the fraction in each quadrant and the 2-D weighted centroid. Note that the 0-lag center point represents the domain average co-occurrence in Fig. 3A–C. The sign notation is as follows: for PMX w.r.t. O3X (Fig. 4A), the color contours show the likelihood (% of events) that PMX occur at the plotted latitude-by-longitude offset in 1° cells from O3X. The most likely co-occurrences are identified by the darkest red colors, and the average offset over this domain is shown with a white 'X'. For example, Fig. 4A shows that the highest coincidence of PMX+O3X (35%) is roughly centered on (latitude, longitude) = (0°, 0°) but that on average PMX occur to the southwest of O3X (–0.7°, –2.2°). For TXX+O3X (Fig. 4B) the highest co-occurrences are in the northeast quadrant with an average of (+1.3°, +0.5°), and for TXX+PMX it is similar.

These spatial offsets can be manifestations of temporal offsets because large episodes typically propagate eastward across the domain. For example, the westward displacement of PMX with respect to O3X (Fig. 4A) may be evidence that PMX tend to occur after O3X at a given grid cell (i.e., after the O3X episode has moved east).
Systematic offsets in time. We test for systematic temporal offsets by calculating the frequency of co-occurrence at each grid cell as a function of time lag, from −7 to +7 days. Fig. 4D–F show the weighted average lag (days) at each grid cell for the corresponding event co-occurrence types in Fig. 3A–C. For example, Fig. 4D (PMX w.r.t. O3X) shows positive values (~1 day) over much of the domain, which means that PMX occur about a day after O3X. We can speculate on the mechanisms for this lag, but an understanding will require a well-tested model that reproduces these observations. The PMX delay, for example, could be that PM2.5 simply takes longer to accumulate than O3 or that the different diurnal cycles and averaging windows cause a shift in PM2.5 to the following morning.
The TXX w.r.t. O3X map (Fig. 4E) shows a sharp regime shift at 40°N. To the north of 40°N, TXX occur 1–2 days after O3X; but to the south, TXX precede O3X by about 1 day (excepting parts of Florida and Georgia). To the south we find the expected T–O3 causal relationship in which O3 abundances increase following a temperature increase. To the north we may be seeing evidence of high temperatures suppressing O3. This suppression has been previously investigated in the US (42, 43) and attributed to enhanced PAN decomposition and reduced isoprene emissions, but it is not clear why this relationship changes at 40°N. The TXX w.r.t. PMX map (Fig. 4F) is essentially Fig. 4E minus Fig. 4D. TXX typically occur after PMX, except for the southwestern corner of the domain and parts of Canada.

**Average progression of large episodes.** The systematic spatial and temporal offsets between the three different extreme events suggests there may be a typical evolution of a multi-extreme episode, such as which extremes appear first and where. The highest O3 levels are associated with the largest episodes (35); hence we limit our analysis to episodes (i.e., a connected set of any of the three extremes; see Data and Methods and ref. 34) that last at least seven days and have at least 350 events of each type within the central 7 days. These episodes are constructed by connecting grid cell days with any of the three extreme types (as opposed to only connecting extremes of the same type). We truncate these large-scale episodes to the central 7 days by defining the central day as the rounded value of the event-weighted average of the time indices and then limiting the episode to at most ±3 days from the central day. Thus the average daily spatial coverage of each extreme over the 7 days is about 50 grid cells or 7° x 7° with the most events on the central day. Over the 15-year period, there are 30 such weeklong, multi-extreme
episodes (i.e., on average about 2 weeks of each year or 210 days in all); however, some episodes last longer than 7 days but are truncated as described above.

In Fig. 5, we show the 7-day sequence in maps of the average large-scale episodes in O3X, PMX, and TXX. The 7 rows show daily maps from 3 days before to 3 days after the central day of the episode. Each 1° cell in each daily map shows the fraction (%) of the 30 possible days that have an extreme event designated by the column. If the extreme events were random with respect to the large-scale episodes then the frequency would be that of the events themselves, 5%, and we see such large random areas at the beginning and end of the 7 days for all three extremes.

Three days before the central day of an episode (Fig. 5, top row), O3X are most likely in the area west of 90°W from 35°N to 45°N (orange–red = 25% = 5 times random), PMX dominate south of 40°N, while TXX are most common north of 40°N except in New England. Over the next three days (Fig. 5, rows 2–4), the fraction increases for all extreme types, with the highest fractions moving generally eastward. For the central day, O3X occur with greatest frequency (>30%) in a band 35°N–45°N across the domain; PMX, in a band 32°N–41°N; while TXX occupy the top half of the domain (36°N–48°N) and are most common overall (>40% in some grid cells). O3X disappear most quickly with little pattern left at +2 days (Fig. 5, row 6), while TXX remain frequent in the northeast at +3 days (Fig. 5, row 7). This average pattern closely resembles days 22–28 June of the sample episode in Fig. 2. It is not clear whether these observed patterns could provide a semi-empirical prediction capability; nevertheless, they do provide stringent observational tests for models used in air quality predictions. Most notably, the size of heat waves (TXX) is clearly larger (space and time) than either pollution extreme.
Fig. 5. Progression of average large-scale episodes shown as the frequency (%) of events in each grid cell over 30 weeklong episodes in 15 years. The rows show the 7-day sequence about the central day. Columns from left-to-right show the fraction for O3X, PMX, and TXX, respectively. If events were random with respect to the episodes, then the fraction would be 5%.
Size and enhancement of extreme episodes. The clustering algorithm identifies connected-cell, multi-day episodes of size $S$ (given here in units of $10^4$ km$^2$ days, roughly the size of a $1^\circ \times 1^\circ$ grid cell). Fig. 6A shows the distribution of episode sizes for each type of extreme as the complementary cumulative distribution (CCD), i.e., the fraction of events’ time integrated area occurring in episodes of size $S$ or larger (see Fig. 11 of ref. 34). Fig. 6A can be read as saying that $\sim 100\%$ of all events occur in episodes with $S > 1 \times 10^4$ km$^2$ days (by definition), $90\%$ occur with $S > 30 \times 10^4$ km$^2$ days, and that $40\%$ of O3X and PMX events occur in episodes of size $S > 500 \times 10^4$ km$^2$ days. For TXX the episodes are larger, and the larger episodes contain a larger fraction of the events: $40\%$ of TXX events occur in episodes of size $S > 900 \times 10^4$ km$^2$ days.

Average episode sizes, $\langle S \rangle$, calculated as a geometric mean with weights equal to $S$, confirm that TXX episodes are clearly larger: $\langle S_{TXX} \rangle = 478$; $\langle S_{PMX} \rangle = 282$; and $\langle S_{O3X} \rangle = 249 \times 10^4$ km$^2$ days. The annual variation in $\langle S \rangle$ (Fig. 6B) shows moderate correlation across the three health extremes, with paired correlation coefficients ranging from 0.21 to 0.74. O3X and TXX are most highly correlated, but TXX clearly shows years with large-scale heat waves that are not matched in PMX and O3X (i.e., 1999, 2006, 2007, 2012).

Co-occurring extremes at a cell-by-cell level are generally more likely for larger episodes and on days with maximum spatial coverage (Fig. S1), excepting the largest TXX episodes ($S > 300 \times 10^4$ km$^2$ days) where co-occurrence frequency decreases and likely reflects the propensity of TXX episodes to be significant larger than those of O3X and PMX.

Identifying the size of an episode is highly valuable in terms of predictive capability regarding its severity. Using the 95th percentile definition as the baseline for events in each cell, we find a
Fig. 6. Episode sizes and enhancements. (A) Complementary cumulative distribution (CCD, %) of the total areal extent of extreme events as a function of episode size $S$ ($10^4$ km$^2$-days) for O3X (blue), PMX (green), and TXX (red) episodes. Average episode sizes $\langle S \rangle$ over the 15-year period are provided in the inset. (B) Annual derived average episode size $\langle S \rangle$ for each extreme type and paired correlation coefficients ($r$). (C) Average value of the extreme events (ppb for O3X, $\mu$g m$^{-3}$ for PMX, °C for TXX) relative to the 95th percentile as a function of episode size. Also provided is the slope of the enhancement per log-decade in $S$. 
clear relationship between episode size and the absolute value of the extremes that is linear in log(S) as shown in Fig. 6C. For episodes with S = 1,000 x 10^4 km^2 days compared to those with S = 1 x 10^4 km^2 days O_3 abundances are 7 ppb higher, PM_{2.5} abundances are about 6 µg m^-3 higher, and TX is about 1.7 °C higher. Consistently across all episode sizes, the larger episodes have the highest pollution and hottest temperatures. The extremes are also on average highest when it co-occurs with one or both other extremes and lowest when it occurs by itself (Fig. S2). This effect is most pronounced for the combination of all three extremes in the far northeast, with enhancements above the 95th percentile threshold of up to +10 ppb, +9 µg m^-3, and +2 °C for O3X, PMX, and TXX, respectively.

Discussion and consequences

By combining detailed site measurements of surface O_3 and PM_{2.5} over the Eastern US and Canada with meteorological observations of surface temperatures, we create a consistently mapped climatology of air pollution and heat waves on a standard 1° x 1° grid. Daily values of O_3 (MDA8, ppb), PM_{2.5} (24-hr average, µg m^-3), and maximum temperature (°C) are derived for April–September 1999–2013 and reported in the SI. This 15-year climatology is intended for analysis of the structure and co-occurrence of extreme air pollution episodes and heat waves (this work), impact studies on human health and agriculture, and chemistry-climate model evaluation of these health extremes. Hence, all datasets are objectively interpolated in space to give average values over each grid cell (ref. 34).

For our three health effects (ozone, particles, temperature), we define extreme events climatologically at each cell based on the percentile distribution: 10 worst days of the extended
summer season, about the 95th percentile over April–September. All three data sets have some
trend over the 15 years, and thus we define the extrema separately in each of the three 5-year
segments. Extreme events tend to cluster into multi-day, spatially connected episodes with
spatial scales of order 1,000 km or more. For the largest episodes, values for ozone, particles and
temperature are well above the statistical threshold defining their extremes. Meteorology clearly
drives the extremes as the three different types of extreme episodes often coincide or are slightly
offset in space or time. The sequencing of events does not always support simple mechanistic
arguments, e.g., warmer temperatures make ozone pollution more severe, because the ozone
events precede temperature events for much of the Eastern US. There are obviously many
mechanisms driving these patterns of extremes of air pollution and temperature; and thus these
observations present the evidence for a model evaluation of cause.

Large-scale, overlapping extreme episodes pose the greatest potential health risk not only
because they coincide, but also because these episodes are found to have the highest pollution
levels and hottest temperatures. Thus, a multi-stressor approach must be taken when evaluating
impacts on human health and vegetation. Furthermore, there is evidence that heat waves and
pollution episodes will intensify under a warming climate in some regions, and thus we need to
develop climate models that have skill in reproducing large-scale pollution episodes and heat
waves.
Data and Methods

We utilize a combination of surface monitoring stations and meteorological data over a 15-year period (1999–2013) to identify climatologically extreme events in each cell of a regular 1° x 1° grid over Eastern North America (ENA). For surface ozone (O₃), we use observed hourly abundances from the US EPA’s Air Quality System (AQS; https://www.epa.gov/aqs) and Clean Air Status and Trends Network (CASTNet; https://www.epa.gov/castnet) and Environment Canada’s National Air Pollution Surveillance Program (NAPS; http://maps-cartes.ec.gc.ca/rnspa-naps/data.aspx). The AQS and NAPS networks also provide observed daily average fine particulate matter (PM₂.₅) abundances over the same time period. The hourly O₃ and daily PM₂.₅ abundances are interpolated onto a 1° x 1° grid over continental eastern North America (ENA = 96°W–66°W, 24°N–50°N) following the algorithm of ref. 34. The maximum daily 8-hour average (MDA8) of O₃ is derived at each grid cell from the interpolated hourly abundances. The disparity between the O₃ and PM₂.₅ averaging windows reflects what is typically used for regulatory purposes and health impacts. For temperature we use the European Centre for Medium range Weather Forecasting (ECMWF; http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/) reanalysis data, which provides gridded, six-hourly maximum temperatures at 2-m height. Daily maximum temperature (TX) is calculated as the maximum of each day’s four values. ECMWF data is taken from the available 0.5° x 0.5° grid, and four such cells are averaged to calculate our 1° x 1° gridded product.

As an additional analysis tool, we calculate the stagnation indices following ref. 37 using reanalysis data of 500 mb and 10 m wind speeds (2.5° remapped to 1°) from the National Centers for Environmental Prediction (NCEP/NCAR;
http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html) and daily cumulative precipitation from NASA’s Global Precipitation Climatology Project (GPCP; ref. 44). This index is Boolean (true or false) such that a grid cell day is considered stagnant when 10 m wind speed < 3.2 m s\(^{-1}\), 500 mb wind speed < 13.0 m s\(^{-1}\), and cumulative precipitation < 1 mm.

Large-scale, multi-day pollution episodes are defined from the extreme events in each cell using the clustering algorithm developed in ref. 34. This method links events if they are located within 1 day or 1° in latitude and longitude of one another, and thus episodes can be assigned a size (in km\(^2\) days) and followed throughout their synoptic development. We use 'event' to describe a single day statistical extreme at a cell and 'episode' to describe a cluster of such events.

**Supporting Information**

![Graphs showing co-occurrence frequency (%)](https://example.com/supinfo.png)

**Fig. S1.** Co-occurrence frequency (%) as a function of episode size (S, 10\(^4\) km\(^2\) days) for (A) O3X, (B) PMX, and (C) TXX episodes.
Fig. S2. Average value of (A–D) O3X (ppb), (E–H) PMX (µg m\(^{-3}\)), and (I–L) TXX (°C) above the 95th percentile when occurring by themselves (A, F, K) and when they co-occur with one (B–C, E, G, I–J) or both (D, H, L) other extreme types.
References


