Atmos. Chem. Phys. Discuss., 14, 6261–6310, 2014
www.atmos-chem-phys-discuss.net/14/6261/2014/
doi:10.5194/acpd-14-6261-2014


Atmospheric Chemistry and Physics Discussions
Open Access



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

**J. L. Schnell[1], C. D. Holmes[1], A. Jangam[1,*], and M. J. Prather[1]**

[1]Department 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

Received: 11 February 2014 – Accepted: 24 February 2014 – Published: 10 March 2014

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

Published by Copernicus Publications on behalf of the European Geosciences Union.

## 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, multi-day 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 (CTM) we find there is skill in hindcasting these extreme episodes, and thus identify a new diagnostic using global chemistry-climate models (CCM) 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 $O_3$ 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).





ACPD
14, 6261–6310, 2014

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 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 multi-day, large-scale extreme ozone episodes found in the US and Europe today.

Even at their best typical resolution ($\sim 1^{\circ} \approx 100$ 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; Rasmussen et al., 2012), although in one case the bias is negligible (Mao et al., 2013). These 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; Murazaki 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-averaged 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 $O_3$ from monitoring stations provide the basis for testing models, but measurements at individual stations are generally not representative of





ACPD
14, 6261–6310, 2014

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 $O_3$ 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 averages on a regular grid. As part of this analysis we generate an objective measure, the quality of prediction, 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 (EMEP and AirBase) for both clean and polluted periods. This assessment uses a full decade of observations (2000–2009) from three networks (EPA over the US).

In Sect. 3, we compare the maximum daily 8 h average (MDA8) grid-averaged observations over the US and Europe with the 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 %ile) or the 20 worst days in 2 years when comparing the observations to the UCI CTM. We then identify the structure of the multi-day, 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 level 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 (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 $O_3$ over the US and EU

For our observations of surface $O_3$ we use ten years (2000–2009) of hourly surface $O_3$ measurements from air quality networks in the United States and Europe (see Table 1 for summary of datasets). For the US we primarily use the Environmental Protection Agency's (EPA) 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 Europe we combine the European Monitoring and Evaluation Programme (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 datasets. The AirBase dataset 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 the EMEP and AirBase are reported as $\mu g\,m^{-3}$ and are converted to parts per billion (ppb = $10^{-6}$ mol mol$^{-1}$ = µmol mol$^{-1}$) using a temperature of 20 °C; mass essentially concentrations are multiplied by 0.5 ppb $\mu g^{-1}\,m^3$.

From these datasets we calculate the maximum daily 8 h average $O_3$ 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 (Chan and Wu, 2005; Bell et al., 2006) and climate studies, (e.g., Tagiris et al., 2007). We calculate the MDA8 by beginning the 8 h averaging period at 24:00 LT and calculating 17 8 h averages for each day, picking the maximum of those 17 (i.e. the averaging only considers windows that fully reside within one day). Thus the maximum can occur during different





ACPD
14, 6261–6310, 2014

8 h 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 $O_3$ concentrations are shown in Fig. 1.

### 2.1 Choosing a method for interpolating grid-cell averages

We develop an interpolation scheme that provides grid-cell averaged values of surface $O_3$ 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 $O_3$ 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. 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 $O_3$ 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





ACPD
14, 6261–6310, 2014

highly clustered data more like a single site (Wackernagel, 2003). Since many observation sites in the US and EU datasets 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 datasets 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 from of IDW that incorporates a declustering scheme without the need to model the underlying covariance function.

From $O_3$ observations $Z_k$ at sites $\boldsymbol{x}_k$, we interpolate the $O_3$ mole fraction at an unobserved location $x$ as a weighted sum of the observations:

$$Z(x) = \sum_{k=1}^{K} w_k \cdot Z_k \Big/ \sum_{k=1}^{K} w_k. \tag{1}$$

where $k$ is the number of observations sites and weights $w_k$ are defined as follows. In standard inverse-distance weighting $w_k = |x - \boldsymbol{x}_k|^{-\beta}$ and $\beta$ is 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 $|x - \boldsymbol{x}_k|$ exceeds a threshold $L$ to avoid meaninglessly small contributions from distant sites. We choose $L = 500$ km based on the typical scale of synoptic meteorology that influences surface $O_3$ 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





ACPD
14, 6261–6310, 2014

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 $|x - \boldsymbol{x}_k| < D$ are given equal weight to avoid singularities in the interpolation. Taken together, the weights in Eq. (1) are

$$w_k = \begin{cases} D^{-\beta}/M_k & \text{if } |x - \boldsymbol{x}_k| < D \\ |x - \boldsymbol{x}_k|^{-\beta}/M_k & \text{if } D \le |x - \boldsymbol{x}_k| \le L \\ 0 & \text{if } |x - \boldsymbol{x}_k| > L \end{cases} \tag{2}$$

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 (Huang and Hsu, 2004). Falke (1999) incorporates a temporal component by reducing the weights of highly variable (mostly urban) sites using the variance of the sites. We do not include this since we assume urban sites are representative of the true processes controlling surface $O_3$. In addition, the weights of these sites are often already significantly reduced by the declustering scheme.

We optimize the interpolation parameters using the leave-$k$-out cross-validation scheme (Cressie, 1993). This involves removing $k = 10\,\%$ of observation sites and predicting their values using the remaining observations and IDW interpolation defined above, recording the root mean square error (RMSE) of the predicted sites. This is done for 365 randomly selected sample days from 2000–2009 with different randomly selected $k$ sites for each day. The primary optimization is for $\beta$, keeping $D = 25$ km and $L = 500$ km fixed. All tested $\beta$ values use the same days and prediction sites. Where there are many nearby sites, the RMS error is at a minimum of about 6 ppb (see Fig. 2 and discussion of quality of prediction below) and does not change much for the range of $2.5 < \beta < 3.5$. The use of large $\beta$ values can lead to sharp gradients near sites, and since we seek an average concentration over a grid cell, we select the lower value of the shallow minimum, $\beta = 2.5$. Subsequently we look at the error for a range of $D$ and





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

$L$ values, and find it relatively insensitive (< 10 % change from the mean) over reasonable values ($D$ = 10 km, 25 km, 50 km; $L$ = 250 km, 500 km) and $\beta = 2.5$ (see Table S1). Thus we retain our original estimates for $D$ and $L$.

To obtain grid cell average values, we use the IDW procedure above to determine the ozone values at 25 equally spaced points in latitude and longitude within each cell and then use trapezoidal integration over the area, similar to block Kriging (Cressie, 1993). The 4 corner points are each 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 $\boldsymbol{x}_i$ 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 $X$, $(\overline{Z}(X))$ is then the weighted sum of ozone at points $\boldsymbol{x}_i$, $Z_i$:

$$\overline{Z}(X) = \sum_{i=1}^{25} w_i^* \cdot Z_i \Big/ \sum_{i=1}^{25} w_i^* \tag{4}$$

We do not report $(\overline{Z}(X))$ for grid boxes where over half of the interior points $Z(\mathbf{x}_i)$ 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 the EU we draw a similar mask but also include areas in the North Sea and in the Mediterranean





ACPD
14, 6261–6310, 2014

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 grid for the EU and US as possible, we select the value of $Q^P = 0.67$ as the cutoff for our masks. Figure 1 shows the constructed masks (grey 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 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 the EU are found in the northernmost and easternmost edges of the domain for the same reason.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

## 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 on 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° grid cell ($\sim 10^4$ km$^2$). Using the error analysis done for the sites (removing randomly 10 % of the sites), we can examine how the cell-averaged 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-averaged ozone, but is at best a lower limit. The RMSE remains small, at about 1 ppb or less, for $Q^P = 0.7$ to 100 and increases to 2 ppb for $Q^P = 0.33$. This 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 US and EU are similar, and the range of $Q^P$ is much larger than in 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° × 1°) averaged 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° × 1° cells. The $Q^P$ values for three sample cells





ACPD
14, 6261–6310, 2014

with this grid 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 and e) and combined EU (Fig. 4b, d and f) datasets for two representative percentiles, the 95th (Fig. 4a and b) and 25th (Fig. 4c and d), and their differences (95th – 25th, Fig. 4e and 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 the 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 – 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 the EU, the greatest differences are concentrated in central countries (e.g. France, Germany, northern Italy).

### 2.4 Comparison of overlapping observational $O_3$ networks

The grid-cell averaged $O_3$ 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 $Q^P$ values for each dataset 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 – EMEP and present biases for the 25th, 50th, and



Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |
ACPD
14, 6261–6310, 2014
Extreme ozone pollution episodes in a global atmospheric chemistry model
J. L. Schnell et al.

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 dataset 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 background $O_3$ so they are placed at more remote, higher altitude locations, while AirBase is selected to reflect population exposure so stations are more readily placed in the valleys where the population is greater). The bias could also reflect interpolation errors at the edge of the EMEP domain, as there are much fewer stations than in AirBase. Differences between AirBase and EMEP are much smaller in the 50th and 95th cases, with MBs of −2.7 ± 1.9 and −1.7 ± 2.2 ppb, respectively. The biases could also be due the cumulative production of $O_3$ as polluted air disperses since the EMEP 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. We present the bias (= AQS – 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 dataset 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 ∼ 5° × 5° 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





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

this difference is due to the titration of $O_3$ by $NO_x$ 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 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 cover 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 QUANTIFY inventory (Hoor et al., 2009), and do not account for daily or weekly 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.





ACPD
14, 6261–6310, 2014

### 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). Summer is typically the highest-percentile $O_3$ days and winter the lowest-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 $O_3$ (MB = CTM – OBS, Fig. 6a and b) is $+19 \pm 6$ ppb (standard deviation across the grid cells) for 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 and f, shows relatively good model hindcasting with average MCC of $0.47 \pm 0.13$ for 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 $O_3$.

The summer domain average MB (Fig. 6c and d) is larger than in winter: $+30 \pm 14$ ppb for US; $+29 \pm 8$ 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 $O_3$ in the model, possibly a consequence of $NO_x$ plumes being spread over the 100 km model grid or other non-linear interactions involving hydrocarbons and $NO_x$ (Lin et al., 2008b; Pusede and Cohen, 2012; Rasmussen et al., 2012). Supporting this hypothesis, the model's summertime bias for US has a similar pattern as our measure of pollution buildup (95th – 25th percentile, Fig. 4e, the two maps have a correlation coefficient, $r = 0.66$). For the EU, this conclusion is less obvious (Fig. 4f, $r = 0.20$).





ACPD
14, 6261–6310, 2014

In terms of MCC, the verisimilitude of the model hindcast of daily summertime pollution is quite good (Fig. 6g and h) because in this case the variability is driven synoptically by buildup of regional pollution: MCC = 0.60 ± 0.16 for US and 0.55 ± 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.

### 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) $O_3$, and the highs show the efficiency of $O_3$ 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 two years 2005–2006 from both observations and the CTM hindcast. The PDFs for winter (DJF) and summer (JJA) months are shown in Fig. 7. The observations, sorted into percentile bins (0–5 %, 5–10 % . . . ) 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-$O_3$ 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. From Fig. 7 we conclude (correctly) that during summer the CTM has a uniform bias of +30 ppb over the full range about the median (−15 ppb to +20 ppb), but that during winter it has serious errors beyond the median bias of +17 ppb probably related to the baseline tropospheric $O_3$.





ACPD
14, 6261–6310, 2014

If we had done this as a climatology comparison, we would have completely reversed this diagnosis. We show maps of model bias as calculated using independent and concurrent sampling and their difference at five representative percentiles (5th, 25th, 50th, 75th, and 95th) for the US and EU in Figs. S3 and S4, respectively. Biases at the 5th percentile calculated using independent sampling are $7 \pm 3$ ppb ($5 \pm 2$ ppb) less than concurrent sampling for the US (EU), however for increasing percentiles the trend reverses, with biases for independent sampling at the 95th percentile $9 \pm 5$ ppb ($8 \pm 4$ ppb) greater than concurrent sampling for the US (EU). We conclude that $O_3$ PDFs simply cannot be used in comparing observations with climate models.

## 4 Identifying and characterizing extreme events

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 occurring often during peak pollution episodes, but there is skill in hindcasting pollution variability. These biases hinder the ability to predict AQX 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 %ile and frequently occur a few days before local maxima in New York and Virginia, due to west-to-east motion of weather systems. If an absolute threshold, such as 75 ppb, then the peak values



Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

ACPD

14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

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 $\sim 97.3$ 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) highlights 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 $O_3$ 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





ACPD
14, 6261–6310, 2014

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 most 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 the 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 US, 37.5 % ± 14 % for EU. While CTM skill at individual grid cells in the US shows no distinct pattern, that in the 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 the EU. The east-west gradient in the EU, as well as the lack of pattern in the US, can partly be understood from the relationship between skill and $Q^P$. Low CTM skill is caused by model errors as well as errors in observations and interpolation. As shown in Fig. S5, the CTM skill is largest in grid cells with large $Q^P$ and small interpolation errors.

### 4.3 Organized episodes of AQX events

The AQX events often occur as clustered, multi-day episodes with spatial extents of more than 1000 km (note that event is a single identified AQX and an episode is a grouping of AQX event). 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 × $10^4$ $km^2$-days, and also the largest in the CTM hindcast at 1700 × $10^4$ $km^2$-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 datasets 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





ACPD
14, 6261–6310, 2014

(e.g. areal extent, duration, intensity, seasonal cycles, etc.) that can be used as metrics to test global chemistry climate models' (GCCMs) 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 hierarchal 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 one (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: Eq. (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 one 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 ($km^2$) 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 a multi-peaked or bimodal distribution. In our first algorithm these bimodal episodes were counted as a larger, single episode, but human





ACPD
14, 6261–6310, 2014

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 dangling episodes. For each episode identified with the primary algorithm, we calculate the area of the events shared with the previous day. If the ratio of the shared area divided by the area of that day is less than 0.10, we truncate the episode at the previous day and start a new episode on the current day. We do not apply this secondary algorithm to the first two or last two days of an episode, to provide flexibility for formation and dissipation. In addition, this detaching can occur more than once as we follow the evolution of an episode.

## 5 Developing climatologies

The grid-cell averaged statistics for MDA8 developed here provide a climatology of surface $O_3$ 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 $O_3$ abundance. Here we seek to develop climate records for surface $O_3$ 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 $O_3$ 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, multi-day episodes is investigated with clustering algorithms (Sect. 5.2), and we develop climate statistics of the scale of these episodes as a new dataset to evaluate CCMs (Sect. 5.3) and opening a novel test of whether climate change alters theses extreme episodes.





ACPD
14, 6261–6310, 2014

## 5.1 Weekly and annual cycles

The well-known weekly and annual cycles (Bruntz et al., 1974) in MDA8 $O_3$ concentrations are summarized for our decadal datasets 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 Supplement 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 an annual cycles using only statistic from years 2005–2006, is in the Supplement (Table S2) 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 identified in June, while in the 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 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), 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 datasets. The largest values of AQX events, the 84th, and 50th percentiles 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 to see significant evidence of a weekly cycle in the CTM, as there





ACPD
14, 6261–6310, 2014

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 numbers of events. Although seemingly trivial, this result provides us with assurance that the CTM is accurately representing the mechanisms responsible for the ozone episodes' formation and not just representing general interannual cycles.

In Table 3, the AQX frequencies for each year clearly show the extraordinary 2003 and 2006 summer heat waves in Europe, as well as a declining number of events throughout the decade (more evident for the US than the EU), associated with reductions in criteria pollutants like $NO_2$ (see www.epa.gov/airtrends/nitrogen.html and www.epa.gov/airtrends/ozone.html; Hudman et al., 2009). We also show the annual mean summertime (June, July, August) MDA8 concentrations from our interpolated product and the raw station data, both of which show excellent agreement with the annual AQX values.

### 5.2 Size distribution of extreme episodes

We define the size of an episode as the integral of AQX area over time ($km^2$-days). The area of a low-latitude grid cell in the US is about $10^4$ $km^2$, while that in EU northern latitudes is about $0.6 \times 10^4$ $km^2$. From size we can estimate two additional metrics – mean daily areal extent ($km^2$) and duration (days) of the episode. Since we only want the effective duration (i.e. the time frame that includes the majority of the episode), we do not take the total duration from first to last day. Instead, we define the duration of the peak episode to be two times the weighted standard deviation of the time indices, where the weight for each time index is the areal extent of the episode on that day. This method reduces the effect of the tails of the episode (early and late days with few events) providing a more robust measure of the duration of extreme pollution. The mean daily areal extent is simply the total size divided by the duration. Finally, we define the mean episode size, $\overline{S}$, over a given time frame (e.g. individual years, full decade)





ACPD
14, 6261–6310, 2014

as the weighted geometric mean of AQX episodes:

$$\overline{S} = \exp\left(\sum_{i=1}^{n}(S_i \cdot \ln S_i) \Big/ \sum_{i=1}^{n} S_i\right) \tag{6}$$

where $n$ is the number of episodes and $S_i$ is the size of the episode. Eq. (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, multi-day clusters that we define as AQX episodes. The complementary cumulative distribution function (CCDF = 1 – cumulative distribution function) of the percentage of the total areal extent 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 $100 \times 10^4$ km$^2$-days and about 31 % in episodes greater than $1000 \times 10^4$ km$^2$-days; for the CTM, the corresponding fractions are 66 % greater than $100 \times 10^4$ km$^2$-days and 37 % greater than $1000 \times 10^4$ km$^2$-days (Fig. 11a). For years 2005–2006 and gridded EU observations the fractions are 84 % and 67 %, respectively; and for the CTM, the fractions are 73 % and 42 %, respectively (Fig. 11b). In the 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 year 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 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, consistent with reductions in precursor emissions. For the EU, this pattern is less apparent; however the standout features are the CCDFs for 2003





ACPD
14, 6261–6310, 2014

and 2006, which have much larger episodes than other years. The annual number of AQX events and $\overline{S}$ values support this conclusion, as seen 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 $2^\circ \times 2^\circ$ dataset (typical of CCM resolution) using simple means of the MDA8 concentrations from the $1^\circ \times 1^\circ$ observational dataset. AQX events and episodes are defined as before (note that the clustering cutoff distance is essentially $2^\circ$ = 1 day). The resulting episode size CCDFs are shown in Fig. 11 (red) and are extremely similar to the $1^\circ \times 1^\circ$ case. This is encouraging for CCM comparisons. From our $1^\circ \times 1^\circ$ CTM simulation (black) we find too many small episodes, but the correct likelihood for the larger episodes that comprise about 50 % of all AQX events. This test does not use the hindcast, exact-day matching and thus should be a robust climate statistic that can test CCMs in the CMIP5 archive.

### 5.3 Developing climate statistics of AQX episodes

The episode size distributions in Fig. S6 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 $10 \times 10^4$ $km^2$-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 the 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 events or large episodes per year. If we wish to ascertain





ACPD
14, 6261–6310, 2014

whether individual years have differences in their pollution episodes in terms of areal extent or duration, then the events need to be re-normalized (i.e., 10 worst days per year). In the 100 per decade case, those years with more events will more likely to have bigger episodes, with all else being equal. This can easily be seen by the CCDFs in Fig. S6 and the $\overline{S}$ values in Table 3. Even when each year is forced to have the same number of events, the CCDF's for each of the years are not similar (see Fig. S7). Using these re-normalized 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 the 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, year 2009 (good) was found to be statistically different ($p < 0.05$) from year 2005 (bad); at the 90 % level, 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 the 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 $\overline{S}_{\text{year}}$ 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 $\overline{S}$ are derived using the same number of events. These tests, among others to be further developed, provide us with a measure of the interannual variability of meteorologically driven AQX episodes and thus allow us to test different decades from the ACCMIP climate simulations to detect a shift in such episodes that falls outside the expected variations.





ACPD
14, 6261–6310, 2014

## 6 Conclusions

Changes in air quality over the coming years of the 21st century will be determined by combined effects of anthropogenic emissions, land-use change that alters natural emissions and pollutant removal, global changes in climate that affect background atmospheric composition, and meteorological regimes that favor either good or poor air quality on regional scales. Chemistry-climate models that account for all of these processes have been used to project surface air quality changes through 2100, but their performance requires further evaluation. The most common metrics for evaluating CTMs and CCMs against observations are measures of central tendency, such as bias in the median and mean. This alone is an inadequate test of model performance, because exposure to extreme high $O_3$ causes human and crop injury and predicting these events is an important use of atmospheric chemistry models. Therefore, past studies have also evaluated models with respect to $O_3$ threshold exceedances, probability distributions, and recurrence times. In this work, we develop new datasets and metrics for evaluating simulated ozone extremes and the ozone climatology of the last 15 years. Several of these metrics test the spatial extent and duration of regional ozone extremes for the first time.

Using 10 years of surface ozone observations over the United States and Europe, we have developed a new interpolation technique for calculating gridded ozone concentrations. We apply the technique to daily MDA8 observations at hundreds of stations in the US and thousands in Europe and derive a $1^\circ \times 1^\circ$ product that has similar resolution to current CTMs. We have assessed errors in this interpolation method with resampling and noise addition techniques; these assessed errors in the gridded product are related to a $Q^{\mathrm{P}}$ metric that is much simpler to calculate. The gridded products have also been used to compare overlapping networks in the US (CASTNET and AQS) and in the EU (EMEP and AirBase) and found only small differences between them.

From the daily gridded daily MDA8 product, we calculate traditional measures of the ozone distribution – mean, median, PDFs – in each grid cell. Comparing the UCI CTM against these statistics reveals positive regional biases that are similar to other CTMs and CCMs. Model performance, as judged by the PDFs, depends on whether the model PDF is sampled concurrently with observations, like a hindcast, or independently, like a climate model. We also examine extreme $O_3$ events and episodes in both the observations and models in ways that are insensitive to model biases. Extreme events are identified relative to the local $O_3$ PDF as the 10 worst days of each year (> 97.3 %ile) and then classified into larger regional episodes based on hierarchical clustering. Despite significant model biases, we find the UCI CTM has skill in identifying observed extreme events and episodes. The model performance is best in regions with high $Q^P$, indicating that undersampling errors in observations contribute to model-observation disagreement, as well as model error.

Our goal of providing observational validation of the air quality simulated by the chemistry-climate models centers 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).

The approach outlined here also has clear applications to extreme events in satellite observations. The re-gridding procedure allows for somewhat sparse measurements and 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 geographic patterns in other quantities (e.g. aerosol optical depth or tropospheric ozone column).

Uncertainties and unresolved issues remain. Although $Q^P$ provides a measure of the quality of the cell-averaged data, it still lacks a formal uncertainty estimate. The decade analyzed here (2000–2009) has an apparent trend in $O_3$ concentrations, presumably driven by reductions in precursor emissions, and this non-stationary pattern needs to be corrected for (Turner et al., 2013). The stationarity of emissions will be important in defining the observed climatology to compare with the CMIP5/ACCMIP results.





ACPD
14, 6261–6310, 2014

**Supplementary material related to this article is available online at http://www.atmos-chem-phys-discuss.net/14/6261/2014/ acpd-14-6261-2014-supplement.zip.**

*Acknowledgements.* This research was supported by the NASA Modeling, Analysis, and Prediction Program (NNX09AJ47G, NASA NNX13AL12G), the Office of Science (BER) of the US Department of Energy (DE-SC0007021), and the Kavli chair in Earth System Science. JLS was supported by the National Science Foundation's (NSF) Graduate Research Fellowship (DGE-1321846). AJ was supported by the NSF Research Experience for Undergraduates (REU) – Summer Fellowships in Biogeochemistry and Climate Change at UC Irvine (AGS-1005042). 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 datasets used in this study.

## References

Anderson, T. W. and Darling, D. A.: Asymptotic theory of certain goodness of fit criteria based on stochastic processes, Ann. Math. Stat., 23, 193–212, doi:10.1214/aoms/1177729437, 1952.

Appel, K. W., Chemel, C., Roselle, S. J., Francis, X. V., Hu, R. M., Sokhi, R. S., Rao, S. T., and Galmarini, S.: Examination of the Community Multiscale Air Quality (CMAQ) model performance over the North American and European domains, Atmos. Environ., 53, 142–155, doi:10.1016/j.atmosenv.2011.11.016, 2012.

Bell, M. L., Peng, R. D., and Dominici, F.: The exposure-response curve for ozone and risk of mortality and the adequacy of current ozone regulations, Environ. Health Persp., 114, 532–536, doi:10.1289/ehp.8816, 2006.

Bratchell, N.: Cluster analysis, Chemometr. Intell. Lab., 6, 105–125, 1989.

Bruntz, S. M., Cleveland, W. S., Graedel, T. E., Kleiner, B., and Warner, J. L.: Ozone concentrations in New Jersey and New York: statistical association with related variables, Science, 186, 257–259, doi:10.1126/science.186.4160.257, 1974.

Carlton, A. G., Pinder, R. W., Bhave, P. V., and Pouliot, G. A.: To what extent can biogenic SOA be controlled?, Environ. Sci. Technol., 44, 3376–3380, doi:10.1021/es903506b, 2010.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

Chang, J. C. and Hanna, S. R.: Air quality model performance evaluation, Meteorol. Atmos. Phys., 87, 167–196, doi:10.1007/s00703-003-0070-7, 2004.
Cleveland, W. S., Graedel, T. E., Kleiner, B., and Warner, J. L.: Sunday and workday variations in photochemical air-pollutants in New Jersey and New York, Science, 186, 4168, doi:10.1126/science.186.4168.1037, 1974.
Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer, Verlag, London, 208 pp., 2001.
Cooley, D., Nychka, D., and Naveau, P.: Bayesian spatial modeling of extreme precipitation return levels, J. Am. Stat. Assoc., 102, 824–840, doi:10.1198/016214506000000780, 2007.
Cressie, N. A. C.: Statistics for Spatial Data, J. Wiley, New York, 1993.
Dawson, J. P., Racherla, P. N., Lynn, B. H., Adams, P. J., and Pandis, S. N.: Simulating present-day and future air quality as climate changes: model evaluation, Atmos. Environ., 42, 4551–4566, doi:10.1016/j.atmosenv.2008.01.058, 2008.
Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S. T., Scheffe, R., Schere, K., Steyn, D., and Venkatram, A.: A framework for evaluating regional-scale numerical photochemical modeling systems, Environ. Fluid Mech., 10, 471–489, doi:10.1007/s10652-009-9163-2, 2010.
Diem, J.: A critical examination of ozone mapping from a spatial-scale perspective, Environ. Pollut., 125, 369–383, doi:10.1016/S0269-7491(03)00110-6, 2003.
Doherty, R. M., Heal, M. R., Wilkinson, P., Pattenden, S., Vieno, M., Armstrong, B., Atkinson, R., Chalabi, Z., Kovats, S., Milojevic, A., and Stevenson, D. S.: Current and future climate- and air pollution-mediated impacts on human health, Environ. Health, 8, doi:10.1186/1476-069x-8-s1-s8, 2009.
Engmann, S. and Cousineau, D.: Comparing distributions: the two-sample Anderson–Darling test as an alternative to the Kolmogorov–Smirnoff test, J. Appl. Quant. Meth., 6, 1–17, 2011.
Falke, S. R.: Mapping air quality: Spatial estimation of pollutant concentrations from point monitoring data, D.Sc. Thesis, Washington University, St. Louis, Missouri, 1999.
Fiore, A., Jacob, D., Bey, I., Yantosca, R., Field, B., Fusco, A., and Wilkinson, J.: Background ozone over the United States in summer: origin, trend, and contribution to pollution episodes, J. Geophys. Res., 107, 4275, doi:10.1029/2001JD000982, 2002.
Fiore, A. M., Dentener, F. J., Wild, O., Cuvelier, C., Schultz, M. G., Hess, P., Textor, C., Schulz, M., Doherty, R. M., Horowitz, L. W., MacKenzie, I. A., Sanderson, M. G., Shindell, D. T., Stevenson, D. S., Szopa, S., Van Dingenen, R., Zeng, G., Atherton, C.,





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

Bergmann, D., Bey, I., Carmichael, G., Collins, W. J., Duncan, B. N., Faluvegi, G., Folberth, G., Gauss, M., Gong, S., Hauglustaine, D., Holloway, T., Isaksen, I. S. A., Jacob, D. J., Jonson, J. E., Kaminski, J. W., Keating, T. J., Lupu, A., Marmer, E., Montanaro, V., Park, R. J., Pitari, G., Pringle, K. J., Pyle, J. A., Schroeder, S., Vivanco, M. G., Wind, P., Wojcik, G., Wu, S., and Zuber, A.: Multimodel estimates of intercontinental source–receptor relationships for ozone pollution, J. Geophys. Res., 114, D04301, doi:10.1029/2008jd010816, 2009.

Fiore, A. M., Naik, V., Spracklen, D. V., Steiner, A., Unger, N., Prather, M., Bergmann, D., Cameron-Smith, P. J., Cionni, I., Collins, W. J., Dalsoren, S., Eyring, V., Folberth, G. A., Ginoux, P., Horowitz, L. W., Josse, B., Lamarque, J. F., MacKenzie, I. A., Nagashima, T., O'Connor, F. M., Righi, M., Rumbold, S. T., Shindell, D. T., Skeie, R. B., Sudo, K., Szopa, S., Takemura, T., and Zeng, G.: Global air quality and climate, Chem. Soc. Rev., 41, 6663–6683, doi:10.1039/c2cs35095e, 2012.

Gelfand, A. E., Zhu, L., and Carlin, B. P.: On the change of support problem for spatio-temporal data, Biostatistics, 2, 31–45, doi:10.1093/biostatistics/2.1.31, 2001.

Hjellbrekke, A.-G., Solberg, S., and Fjæraa, A. M.: Ozone measurements 2011, EMEP/CCC-Report 3/2013, 0–7726, Tech. Rep., Norwegian Institute for Air Research, Norway, available at: http://www.nilu.no/projects/CCC/reports/cccr3-2013.pdf (last access: 25 July 2013), 2013.

Hoor, P., Borken-Kleefeld, J., Caro, D., Dessens, O., Endresen, O., Gauss, M., Grewe, V., Hauglustaine, D., Isaksen, I. S. A., Jöckel, P., Lelieveld, J., Myhre, G., Meijer, E., Olivie, D., Prather, M., Schnadt Poberaj, C., Shine, K. P., Staehelin, J., Tang, Q., van Aardenne, J., van Velthoven, P., and Sausen, R.: The impact of traffic emissions on atmospheric ozone and OH: results from QUANTIFY, Atmos. Chem. Phys., 9, 3113–3136, doi:10.5194/acp-9-3113-2009, 2009.

Hoyle, C. R., Boy, M., Donahue, N. M., Fry, J. L., Glasius, M., Guenther, A., Hallar, A. G., Huff Hartz, K., Petters, M. D., Petäjä, T., Rosenoern, T., and Sullivan, A. P.: A review of the anthropogenic influence on biogenic secondary organic aerosol, Atmos. Chem. Phys., 11, 321–343, doi:10.5194/acp-11-321-2011, 2011.

HTAP: Hemispheric transport of air pollution 2010, Part A: Ozone and particulate matter, United Nations, Geneva, Switzerland, 278, 2010.

Huang, H. and Hsu, N.: Modeling transport effects on ground-level ozone using a non-stationary space-time model, Environmetrics, 15, 251–268, doi:10.1002/env.639, 2004.

Hudman, R. C., Murray, L. T., Jacob, D. J., Turquety, S., Wu, S., Millet, D. B., Avery, M., Goldstein, A. H., and Holloway, J.: North American influence on tropospheric ozone and the effects of recent emission reductions: constraints from ICARTT observations, J. Geophys. Res., 114, D07302, doi:10.1029/2008jd010126, 2009.

Isaaks, E. H. and Srivastava, R. M.: An Introduction to Applied Geostatistics, Oxford University Press, New York, New York, 1989.

Jacob, D. J. and Winner, D. A.: Effect of climate change on air quality, Atmos. Environ., 43, 51–63, doi:10.1016/j.atmosenv.2008.09.051, 2009.

Jacob, D. J., Logan, J. A., Gardner, G. M., Yevich, R. M., Spivakovsky, C. M., Wofsy, S. C., Sillman, S., and Prather, M. J.: Factors regulating ozone over the United States and its export to the global atmosphere, J. Geophys. Res., 98, 14817–14826, doi:10.1029/98jd01224, 1993.

Johnson, C. E., Collins, W. J., Stevenson, D. S., and Derwent, R. G.: Relative roles of climate and emissions changes on future tropospheric oxidant concentrations, J. Geophys. Res., 104, 18631–18645, doi:10.1029/1999jd900204, 1999.

Karl, T. R.: Day of week variations of photochemical pollutants in St. Louis area, Atmos. Environ., 12, 1657–1667, doi:10.1016/0004-6981(78)90314-1, 1978.

Kirtman, B., Power, S., Adedoyin, A. J., Boer, G., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A., Kimoto, M., Meehl, G., Prather, M., Sarr, A., Schaer, C., Sutton, R., Oldenborgh, G. J. v., Vecchi, G., and Wang, H.-J.: Near-term climate change: projections and Predictability, in: Climate Change 2013: The Physical Science Basis, capter 11, IPCC WGI Contribution to the Fifth Assessment Report, 2013.

Lamarque, J.-F., Emmons, L. K., Hess, P. G., Kinnison, D. E., Tilmes, S., Vitt, F., Heald, C. L., Holland, E. A., Lauritzen, P. H., Neu, J., Orlando, J. J., Rasch, P. J., and Tyndall, G. K.: CAM-chem: description and evaluation of interactive atmospheric chemistry in the Community Earth System Model, Geosci. Model Dev., 5, 369–411, doi:10.5194/gmd-5-369-2012, 2012.

Leadbetter, M.: On high-level exceedance modeling and tail inference, J. Stat. Plan. Infer., 45, 247–260, doi:10.1016/0378-3758(94)00075-1, 1995.

Lei, H., Wuebbles, D. J., and Liang, X. Z.: Projected risk of high ozone episodes in 2050, Atmos. Environ., 59, 567–577, doi:10.1016/j.atmosenv.2012.05.051, 2012.

Leibensperger, E. M., Mickley, L. J., and Jacob, D. J.: Sensitivity of US air quality to mid-latitude cyclone frequency and implications of 1980–2006 climate change, Atmos. Chem. Phys., 8, 7075–7086, doi:10.5194/acp-8-7075-2008, 2008.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

Lin, J. T., Wuebbles, D. J., and Liang, X. Z.: Effects of intercontinental transport on surface ozone over the United States: present and future assessment with a global model, Geophys. Res. Lett., 35, L02805, doi:10.1029/2007gl031415, 2008a.
Lin, J. T., Youn, D., Liang, X. Z., and Wuebbles, D. J.: Global model simulation of summertime US ozone diurnal cycle and its sensitivity to PBL mixing, spatial resolution, and emissions, Atmos. Environ., 42, 8470–8483, doi:10.1016/j.atmosenv.2008.08.012, 2008b.
Lin, J. T. and McElroy, M. B.: Impacts of boundary layer mixing on pollutant vertical profiles in the lower troposphere: implications to satellite remote sensing, Atmos. Environ., 44, 1726–1739, doi:10.1016/j.atmosenv.2010.02.009, 2010.
Mao, J., Paulot, F., Jacob, D. J., Cohen, R. C., Crounse, J. D., Wennberg, P. O., Keller, C. A., Hudman, R. C., Barkley, M. P., and Horowitz, L. W.: Ozone and organic nitrates over the eastern United States: sensitivity to isoprene chemistry, J. Geophys. Res., 118, 11256–11268, doi:10.1002/jgrd.50817, 2013.
Meleux, F., Solmon, F., and Giorgi, F.: Increase in summer European ozone amounts due to climate change, Atmos. Environ., 41, 7577–7587, doi:10.1016/j.atmosenv.2007.05.048, 2007.
Mickley, L., Jacob, D., Field, B., and Rind, D.: Effects of future climate change on regional air pollution episodes in the United States, Geophys. Res. Lett., 31, L24103, doi:10.1029/2004GL021216, 2004.
Murazaki, K. and Hess, P.: How does climate change contribute to surface ozone change over the United States?, J. Geophys. Res., 111, D05301, doi:10.1029/2005JD005873, 2006.
Nolte, C., Gilliland, A., Hogrefe, C., and Mickley, L.: Linking global to regional models to assess future climate impacts on surface ozone levels in the United States, J. Geophys. Res., 113, D14307, doi:10.1029/2007JD008497, 2008.
Padoan, S., Ribatet, M., and Sisson, S.: Likelihood-based inference for max-stable processes, J. Am. Stat. Assoc., 105, 263–277, doi:10.1198/jasa.2009.tm08577, 2010.
Pierce, T., Hogrefe, C., Rao, S. T., Porter, P. S., and Ku, J. Y.: Dynamic evaluation of a regional air quality model: assessing the emissions-induced weekly ozone cycle, Atmos. Environ., 44, 3583–3596, doi:10.1016/j.atmosenv.2010.05.046, 2010.
Pirovano, G., Balzarini, A., Bessagnet, B., Emery, C., Kallos, G., Meleux, F., Mitsakou, C., Nopmongcol, U., Riva, G. M., and Yarwood, G.: Investigating impacts of chemistry and transport model formulation on model performance at European scale, Atmos. Environ., 53, 93–109, doi:10.1016/j.atmosenv.2011.12.052, 2012.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

Prather, M., Ehhalt, D., Dentener, F., Derwent, R., Dlugokencky, E. J., Holland, E., Isaksen, I., Katima, J., Kirchhoff, V., and Matson, P.: Atmospheric chemistry and greenhouse gases, Chapter 4 in IPCC Third Assessment Report, Cambridge U. Press, New York, 2001.

Pusede, S. E. and Cohen, R. C.: On the observed response of ozone to $NO_x$ and VOC reactivity reductions in San Joaquin Valley California 1995–present, Atmos. Chem. Phys., 12, 8323–8339, doi:10.5194/acp-12-8323-2012, 2012.

Rasmussen, D. J., Fiore, A. M., Naik, V., Horowitz, L. W., McGinnis, S. J., and Schultz, M. G.: Surface ozone–temperature relationships in the eastern US: a monthly climatology for evaluating chemistry-climate models, Atmos. Environ., 47, 142–153, doi:10.1016/j.atmosenv.2011.11.021, 2012.

Reidmiller, D. R., Fiore, A. M., Jaffe, D. A., Bergmann, D., Cuvelier, C., Dentener, F. J., Duncan, B. N., Folberth, G., Gauss, M., Gong, S., Hess, P., Jonson, J. E., Keating, T., Lupu, A., Marmer, E., Park, R., Schultz, M. G., Shindell, D. T., Szopa, S., Vivanco, M. G., Wild, O., and Zuber, A.: The influence of foreign vs. North American emissions on surface ozone in the US, Atmos. Chem. Phys., 9, 5027–5042, doi:10.5194/acp-9-5027-2009, 2009.

Rojas-Avellaneda, D. and Silvan-Cardenas, J.: Performance of geostatistical interpolation methods for modeling sampled data with non-stationary mean, Stoch. Env. Res. Risk A, 20, 455–467, doi:10.1007/s00477-006-0038-5, 2006.

Steiner, A. L., Tonse, S., Cohen, R. C., Goldstein, A. H., and Harley, R. A.: Influence of future climate and emissions on regional air quality in California, J. Geophys. Res., 111, D18303, doi:10.1029/2005jd006935, 2006.

Steiner, A. L., Davis, A. J., Sillman, S., Owen, R. C., Michalak, A. M., and Fiore, A. M.: Observed suppression of ozone formation at extremely high temperatures due to chemical and biophysical feedbacks, P. Natl. Acad. Sci. USA, 107, 19685–19690, doi:10.1073/pnas.1008336107, 2010.

Stocker, T., Qin, D., and Platner, G.: Climate Change 2013: The Physical Science Basis, Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Summary for Policymakers (IPCC, 2013), Cambridge, United Kingdom and New York, NY, USA, 2013.

Swall, J. L. and Foley, K. M.: The impact of spatial correlation and incommensurability on model evaluation, Atmos. Environ., 43, 1204–1217, doi:10.1016/j.atmosenv.2008.10.057, 2009.

Tagaris, E., Manomaiphiboon, K., Liao, K., Leung, L., Woo, J., He, S., Amar, P., and Russell, A.: Impacts of global climate change and emissions on regional ozone and fine par-





ACPD
14, 6261–6310, 2014

ticulate matter concentrations over the United States, J. Geophys. Res., 112, D14312, doi:10.1029/2006JD008262, 2007.
Tai, A. P. K., Mickley, L. J., and Jacob, D. J.: Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: implications for the sensitivity of $PM_{2.5}$ to climate change, Atmos. Environ., 44, 3976–3984, doi:10.1016/j.atmosenv.2010.06.060, 2010.
Tang, Q. and Prather, M. J.: Correlating tropospheric column ozone with tropopause folds: the Aura-OMI satellite data, Atmos. Chem. Phys., 10, 9681–9688, doi:10.5194/acp-10-9681-2010, 2010.
Tang, Q. and Prather, M. J.: Five blind men and the elephant: what can the NASA Aura ozone measurements tell us about stratosphere-troposphere exchange?, Atmos. Chem. Phys., 12, 2357–2380, doi:10.5194/acp-12-2357-2012, 2012a.
Tang, Q. and Prather, M. J.: Tropospheric column ozone: matching individual profiles from Aura OMI and TES with a chemistry-transport model, Atmos. Chem. Phys., 12, 10441–10452, doi:10.5194/acp-12-10441-2012, 2012b.
Tao, Z. N., Williams, A., Huang, H. C., Caughey, M., and Liang, X. Z.: Sensitivity of US surface ozone to future emissions and climate changes, Geophys. Res. Lett., 34, L08811, doi:10.1029/2007gl029455, 2007.
Tonse, S. R., Brown, N. J., Harley, R. A., and Jinc, L.: A process-analysis based study of the ozone weekend effect, Atmos. Environ., 42, doi:10.1016/j.atmosenv.2008.05.061, 2008.
Turner, A. J., Fiore, A. M., Horowitz, L. W., and Bauer, M.: Summertime cyclones over the Great Lakes Storm Track from 1860–2100: variability, trends, and association with ozone pollution, Atmos. Chem. Phys., 13, 565–578, doi:10.5194/acp-13-565-2013, 2013.
Valari, M. and Menut, L.: Does an Increase in Air Quality Models' Resolution Bring Surface Ozone Concentrations Closer to Reality?, J. Atmos. Ocean Tech., 25, 1955–1968, doi:10.1175/2008jtecha1123.1, 2008.
Wackernagel, H.: Multivariate Geostatistics: an Introduction with Applications, 3rd Edn., Springer, Berlin, 387 pp., 2003.
Weaver, C., Liang, X., Zhu, J., Adams, P., Amar, P., Avise, J., Caughey, M., Chen, J., Cohen, R., Cooter, E., Dawson, J., Gilliam, R., Gilliland, A., Goldstein, A., Grambsch, A., Grano, D., Guenther, A., Gustafson, W., Harley, R., He, S., Hemming, B., Hogrefe, C., Huang, H., Hunt, S., Jacob, D., Kinney, P., Kunkel, K., Lamarque, J., Lamb, B., Larkin, N., Leung, L., Liao, K., Lin, J., Lynn, B., Manomaiphiboon, K., Mass, C., McKenzie, D., Mickley, L.,





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

O'Neill, S., Nolte, C., Pandis, S., Racherla, P., Rosenzweig, C., Russell, A., Salathe, E., Steiner, A., Tagaris, E., Tao, Z., Tonse, S., Wiedinmyer, C., Williams, A., Winner, D., Woo, J., Wu, S., and Wuebbles, D.: A preliminary synthesis of modeled climate change impacts on us regional ozone concentrations, B. Am. Meteorol. Soc., 90, 1843–1863, doi:10.1175/2009BAMS2568.1, 2009.

Wild, O. and Prather, M. J.: Global tropospheric ozone modeling: quantifying errors due to grid resolution, J. Geophys. Res., 111, D11305, doi:10.1029/2005jd006605, 2006.

Wild, O., Fiore, A. M., Shindell, D. T., Doherty, R. M., Collins, W. J., Dentener, F. J., Schultz, M. G., Gong, S., MacKenzie, I. A., Zeng, G., Hess, P., Duncan, B. N., Bergmann, D. J., Szopa, S., Jonson, J. E., Keating, T. J., and Zuber, A.: Modelling future changes in surface ozone: a parameterized approach, Atmos. Chem. Phys., 12, 2037–2054, doi:10.5194/acp-12-2037-2012, 2012.

Wu, S. L., Mickley, L. J., Jacob, D. J., Rind, D., and Streets, D. G.: Effects of 2000–2050 changes in climate and emissions on global tropospheric ozone and the policy-relevant background surface ozone in the United States, J. Geophys. Res., 113, D18312, doi:10.1029/2007jd009639, 2008.

Zanis, P., Katragkou, E., Tegoulias, I., Poupkou, A., Melas, D., Huszar, P., and Giorgi, F.: Evaluation of near surface ozone in air quality simulations forced by a regional climate model over Europe for the period 1991–2000, Atmos. Environ., 45, 6489–6500, doi:10.1016/j.atmosenv.2011.09.001, 2011.

Zhang, Y., Hu, X. M., Leung, L. R., and Gustafson, W. I.: Impacts of regional climate change on biogenic emissions and air quality, J. Geophys. Res., 113, D18310, doi:10.1029/2008jd009965, 2008.





ACPD
14, 6261–6310, 2014

**Table 1.** Observational datasets.

| Surface ozone network | Period | # stations | URL or reference |
|---|---|---|---|
| US EPA Air Quality System (AQS) | 2000–2009 | 1608 | http://www.epa.gov/ttn/airs/aqsdatamart |
| US EPA Clean Air Status and Trends Network (CASTNET)* | 2000–2009 | 92 | http://epa.gov/castnet/javaweb/index.html |
| European Monitoring and Evaluation Programme (EMEP) | 2000–2009 | 162 | Hjellbrekke et al. (2010) |
| European Environment Agency's air quality database (AirBase) | 2000–2009 | 2123 | http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-7#tab-european-data |

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





ACPD
14, 6261–6310, 2014

**Table 2.** Domain mean number of air quality extreme events (AQX) defined for the grid-cell interpolated MDA8 $O_3$ series and the MDA8 $O_3$ 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.

| | | Unit | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| US | AQX | # | 0.0 | 0.2 | 0.8 | 10.6 | 17.8 | 30.2 | 25.0 | 22.9 | 11.0 | 1.2 | 0.1 | 0.0 | 1.00 |
| | $O_3$ 84 % | ppb | +6.7 | +7.1 | +7.6 | +8.8 | +10.2 | +12.8 | +11.7 | +12.0 | +12.3 | +11.3 | +8.0 | +6.7 | 0.77 |
| | $O_3$ 50 % | ppb | 30.3 | 36.1 | 42.6 | 48.2 | 48.9 | 48.6 | 49.2 | 48.0 | 42.9 | 35.1 | 30.7 | 28.6 | 0.71 |
| | $O_3$ 16 % | ppb | −8.0 | −7.9 | −7.9 | −9.0 | −10.4 | −13.1 | −13.9 | −14.2 | −12.6 | −9.8 | −8.3 | −7.8 | 0.48 |
| EU | AQX | # | 0.0 | 0.1 | 3.4 | 22.4 | 24.5 | 21.1 | 25.6 | 19.0 | 3.7 | 0.1 | 0.0 | 0.0 | 1.00 |
| | $O_3$ 84 % | ppb | +7.2 | +6.5 | +6.6 | +8.2 | +9.6 | +12.8 | +15.7 | +15.1 | +12.9 | +7.5 | +7.8 | +8.1 | 0.84 |
| | $O_3$ 50 % | ppb | 26.9 | 32.3 | 39.6 | 44.6 | 44.7 | 42.7 | 40.4 | 38.9 | 33.5 | 28.1 | 24.8 | 23.8 | 0.75 |
| | $O_3$ 16 % | ppb | −7.8 | −9.9 | −7.0 | −6.7 | −6.9 | −8.7 | −10.2 | −10.3 | −7.5 | −7.1 | −8.1 | −8.6 | 0.66 |
| | | | Sun | Mon | Tue | Wed | Thu | Fri | Sat | $R^2$ | | | | | |
| US | AQX | # | 9.1 | 8.3 | 9.9 | 10.2 | 10.6 | 10.9 | 11.0 | 1.00 | | | | | |
| | $O_3$ 84 % | ppb | +13.4 | +13.8 | +14.4 | +14.7 | +14.5 | +14.5 | +14.2 | 0.77 | | | | | |
| | $O_3$ 50 % | ppb | 40.4 | 39.5 | 39.4 | 39.5 | 39.5 | 39.5 | 40.1 | 0.00 | | | | | |
| | $O_3$ 16 % | ppb | −11.6 | −11.5 | −12.0 | −12.2 | −12.2 | −12.2 | −11.9 | 0.19 | | | | | |
| EU | AQX | # | 9.9 | 8.7 | 9.2 | 9.9 | 10.9 | 10.5 | 11.0 | 1.00 | | | | | |
| | $O_3$ 84 % | ppb | +12.2 | +12.6 | +12.8 | +13.3 | +13.5 | +13.5 | +13.0 | 0.87 | | | | | |
| | $O_3$ 50 % | ppb | 35.8 | 34.6 | 34.5 | 34.5 | 34.6 | 34.4 | 35.3 | 0.03 | | | | | |
| | $O_3$ 16 % | ppb | −10.5 | −10.8 | −11.3 | −11.4 | −11.5 | −11.3 | −11.0 | 0.00 | | | | | |





ACPD
14, 6261–6310, 2014

**Table 3.** Climatology of $O_3$ 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 $AQX_{year}$, which is normalized to have 10 events per year. The mean AQX size $\overline{S}$ ($\overline{S}_{year}$ 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 $10^4$ km squared days (km$^2$ d), where $10^4$ km$^2$ is about a $1^\circ \times 1^\circ$ grid cell. Average summertime (JJA) MDA8 $O_3$ (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 $O_3$.

| | | Unit | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | $\mu \pm \sigma$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| US | AQX events | # | 13.5 | 11.5 | 16.5 | 15.0 | 4.6 | 11.2 | 13.3 | 8.1 | 4.6 | 1.7 | $10.0 \pm 5.0$ | 1.00 |
| | $\overline{S}$ | $10^4$ km$^2$ d | 618 | 373 | 1239 | 581 | 82 | 435 | 515 | 186 | 70 | 32 | $413 \pm 363$ | 0.78 |
| | $AQX_{year}$ events | # | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | $10.0 \pm 0.0$ | 0.00 |
| | $\overline{S}_{year}$ | $10^4$ km$^2$ d | 264 | 295 | 337 | 276 | 217 | 329 | 222 | 232 | 208 | 199 | $256 \pm 50$ | 0.55 |
| | $O_3$ (grid) | ppb | 49.3 | 49.4 | 51.4 | 50.1 | 45.5 | 48.8 | 50.7 | 47.5 | 46.2 | 43.7 | $48.3 \pm 2.4$ | 0.96 |
| | $O_3$ (station) | ppb | 51.3 | 52.1 | 55.0 | 51.0 | 46.9 | 50.8 | 52.0 | 50.1 | 48.8 | 45.0 | $50.3 \pm 2.8$ | 0.85 |
| EU | AQX events | # | 7.4 | 8.3 | 11.0 | 19.9 | 10.0 | 8.2 | 16.5 | 6.0 | 8.3 | 4.4 | $10.0 \pm 4.8$ | 1.00 |
| | $\overline{S}$ | $10^4$ km$^2$ d | 280 | 502 | 187 | 793 | 415 | 287 | 2528 | 210 | 240 | 140 | $558 \pm 718$ | 0.43 |
| | $AQX_{year}$ events | # | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | $10.0 \pm 0.0$ | 0.00 |
| | $\overline{S}_{year}$ | $10^4$ km$^2$ d | 388 | 419 | 237 | 446 | 404 | 319 | 1149 | 437 | 305 | 367 | $447 \pm 255$ | 0.25 |
| | $O_3$ (grid) | ppb | 40.1 | 41.7 | 44.3 | 47.3 | 42.7 | 41.4 | 45.2 | 41.2 | 41.5 | 40.0 | $43.4 \pm 2.3$ | 0.94 |
| | $O_3$ (station) | ppb | 43.5 | 46.6 | 45.7 | 54.9 | 45 | 45.1 | 49.5 | 43.5 | 44.1 | 44.6 | $46.2 \pm 3.5$ | 0.85 |





ACPD
14, 6261–6310, 2014

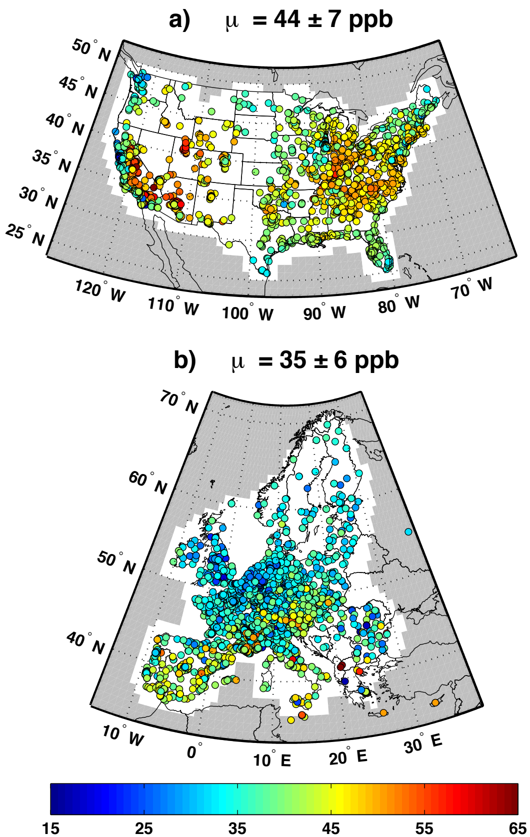


**Fig. 1.** Location of surface $O_3$ monitoring stations and their 10 year (2000–2009) mean MDA8 (ppb) for (**a**) US (EPA AQS) and (**b**) EU (combined EMEP and AirBase). The mask for interpolating the $1^\circ \times 1^\circ$ grid cells is also shown with light gray indicating cells with $Q^P < 0.67$ used here (see text).





ACPD
14, 6261–6310, 2014

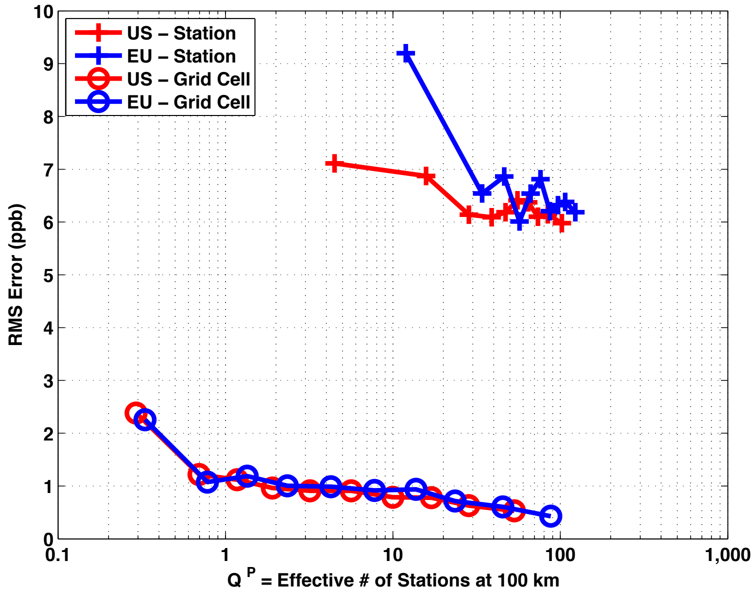


**Fig. 2.** RMS error (ppb) for the mean value of each 10th percentile of interpolated sites and grid cells, sorted by $Q^P$.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

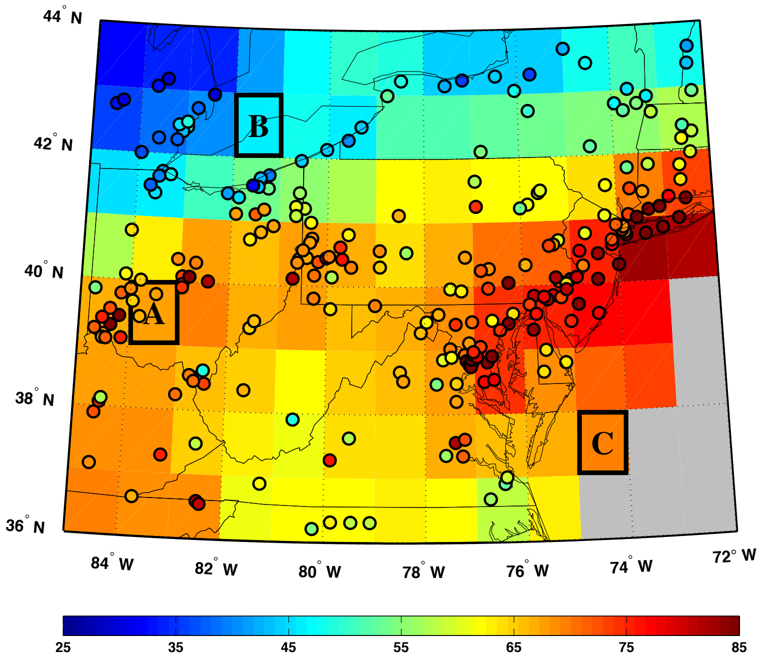


**Fig. 3.** Surface $O_3$ MDA8 (ppb) on 11 August 2005 over a section of the US and Canada. Values from the individual EPA AQS stations are overlain on the grid-cell average interpolated here (see text). Boxes marked A-C have respective $Q^P$ values of 60.1, 15.4, and 6.6. Grey cells are outside the range of interpolation (i.e. $Q^P < 0.67$).





ACPD
14, 6261–6310, 2014

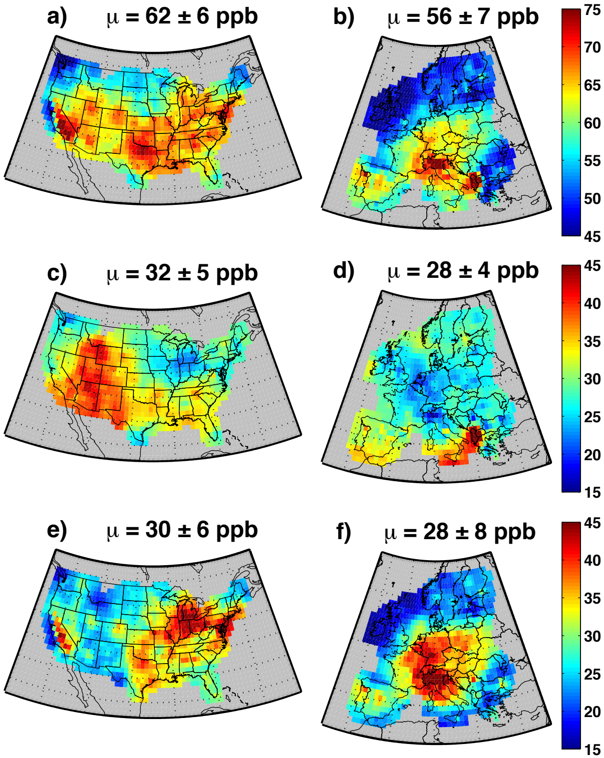


**Fig. 4.** Gridded surface MDA8 $O_3$ (ppb) corresponding to the (**a, b**) 95th percentile, (**c, d**) 25th percentile, and (**e, f**) their difference (95th – 25th) calculated with respect to years 2005–2006. Left column (**a, c, e**) shows results for the US and the right column for the EU (**b, d, f**). Note the change in color bars from (**a, b**) to (**c, d, e, f**).

ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.



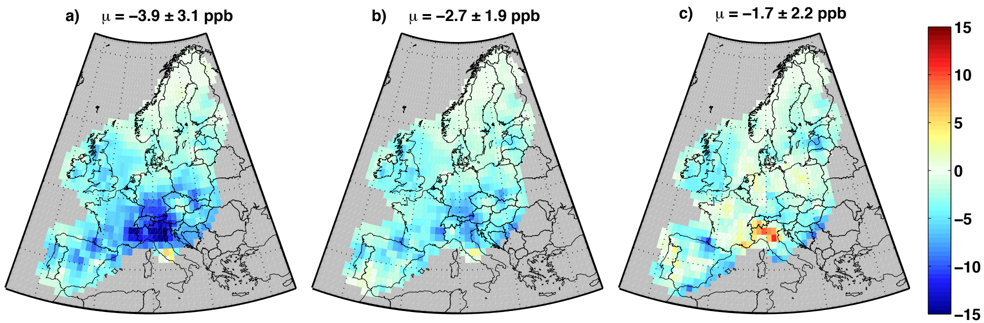


**Fig. 5.** Bias of the gridded MDA8 $O_3$ concentration (ppb) created using only AirBase stations vs. using only EMEP stations for the years 2000–2009 (bias = AirBase – EMEP). Biases are shown for the (**a**) 25th, (**b**) 50th, and (**c**) 95th percentiles and are calculated using independent sampling. This mask includes only grid cells with a $Q^P$ greater than 0.33 for both datasets. The area-weighted mean bias and 1$\sigma$ for each percentile is given with the graph. All mean biases are negative.

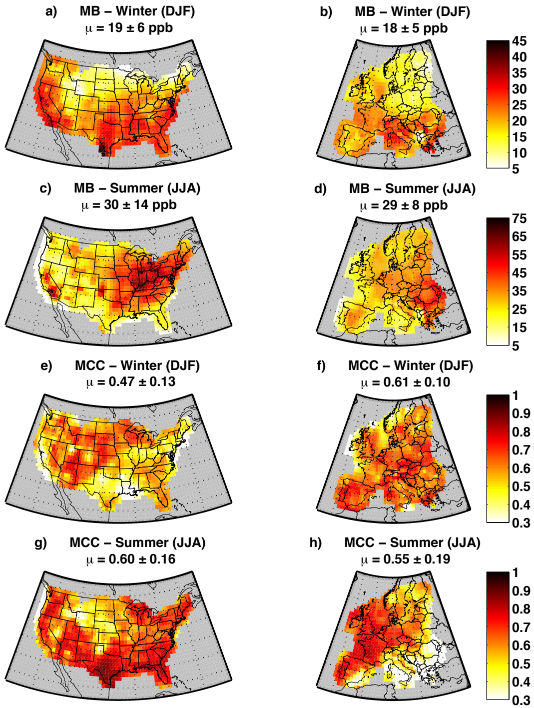


**Fig. 6.** Top two rows (**a, b, c, d**) shows the model mean bias (MB = CTM – observed) of surface MDA8 $O_3$ (ppb) calculated using independent sampling. Bottom two rows (**e, f, g, h**) shows the model correlation coefficient (MCC). Left column (**a, c, e, g**) is the US and right column (**b, d, f, g**) is EU. Both MB and MCC are calculated with respect to years 2005–2006. First and third rows (**a, b**, **e, f**) are for winter months (DJF) and second and fourth rows (**c, d, g, h**) are for summer months (JJA). The area-weighted mean and $1\sigma$ is given for each plot. Note the difference in color scales for MB in winter and summer and between MB and MCC.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

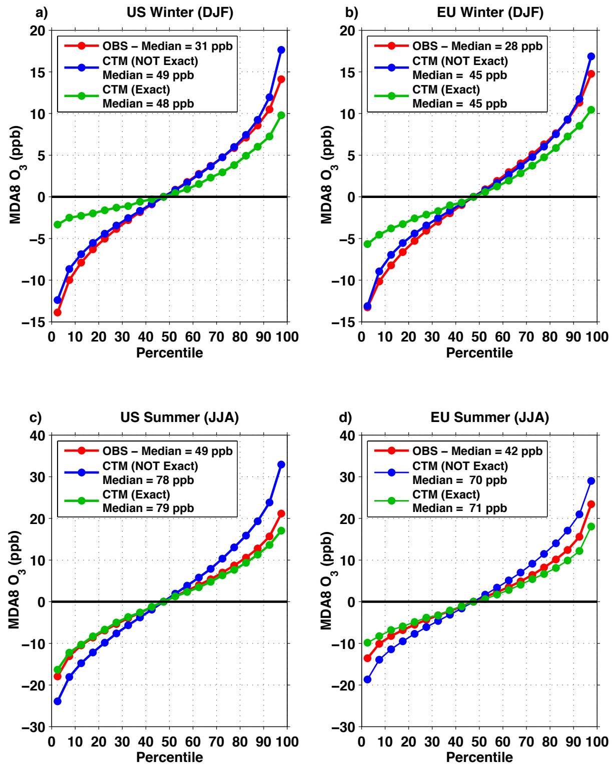


**Fig. 7.** PDFs of surface MDA8 $O_3$ (ppb) for the observations and CTM binned at every 5th percentile for years 2005–2006. PDFs of the CTM are shown for both independent (NOT Exact) and concurrent sampling (Exact). Left column (**a, c**) is US and right column (**b, d**) is EU. Top row (**a, b**) shows the PDFs for winter months (DJF), and bottom row (**c, d**) for summer months (JJA). The median of each PDF was subtracted prior to plotting and is listed in the legend.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

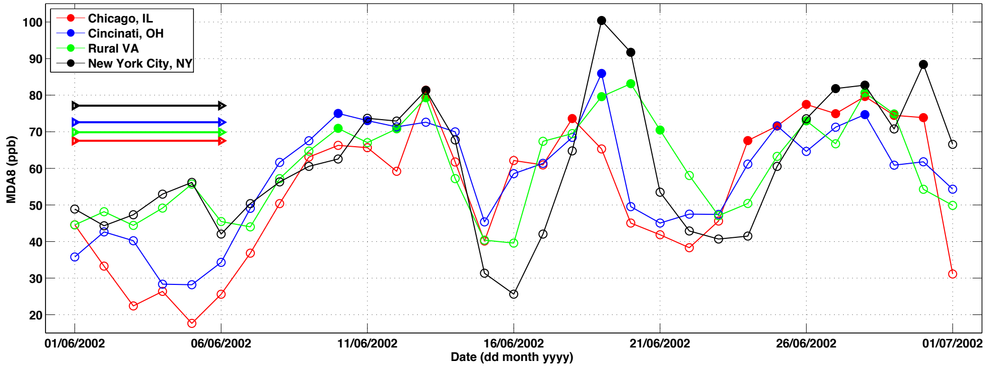


**Fig. 8.** Time series (1 July–1 August 2002) of surface MDA8 $O_3$ (ppb) for four grid cells in the US observations encompassing from west to east: Chicago, IL, Cincinnati, OH, rural area, VA, and New York City, NY. The colored arrows on the left denote the $O_3$ concentration corresponding to an AQX event (97.3 percentile) for each location, calculated with respect to years 2000–2009.





ACPD
14, 6261–6310, 2014

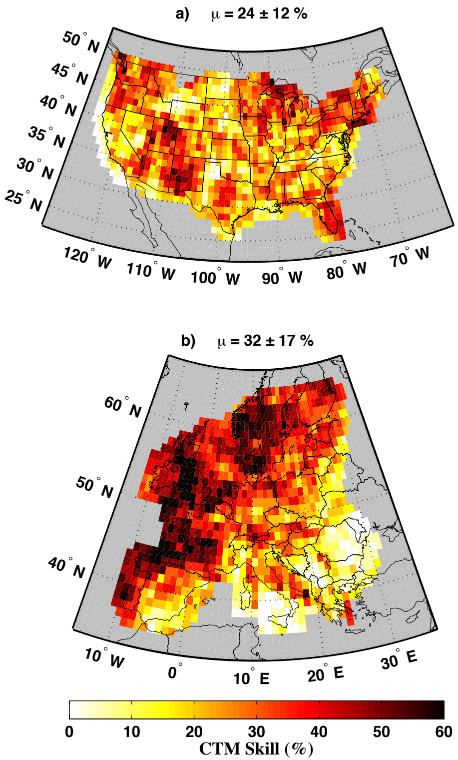


**Fig. 9.** Skill of the CTM (i.e. percentage of events identified in the observations that were correctly reproduced in the CTM) at each grid cell for the (**a**) US and (**b**) EU for years 2005–2006. Domain mean skill and 1$\sigma$ are shown for each plot.



ACPD
14, 6261–6310, 2014

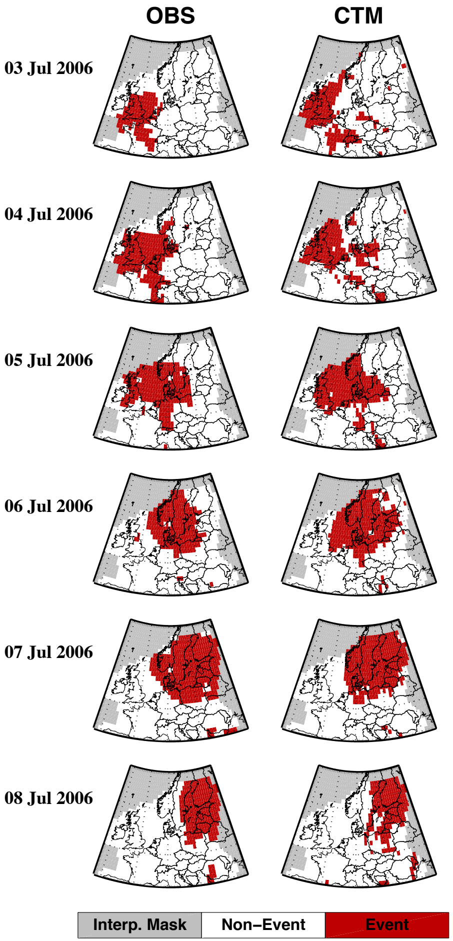


**Fig. 10.** Six days (3–8 August 2006) of a large AQX episode in the EU. Left column is the observations and right column is the CTM.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◄◄ | ►►
◄ | ►
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

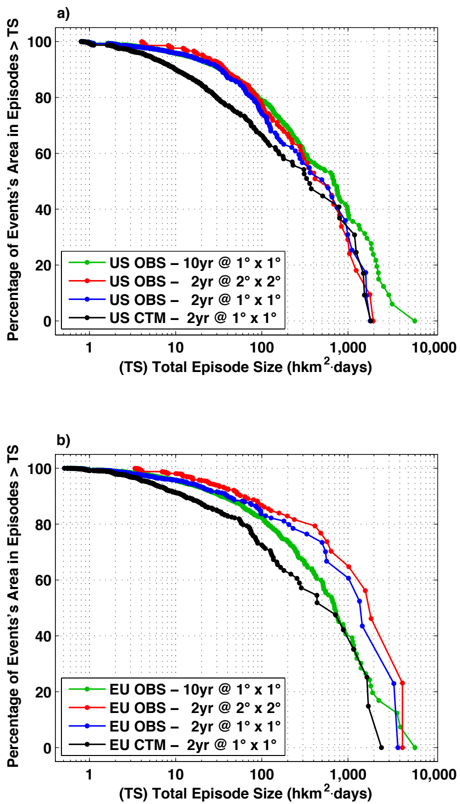


**Fig. 11.** Complementary cumulative distribution function of the percentage of the total areal extent of all individual AQX events as a function of AQX episode size ($10^4$ km$^2$-days) they are clustered into for the (**a**) US and (**b**) EU. Results are shown for the 2 year observations at 1° and 2°, the CTM at 1°, and the 10 year observations at 1°. Note: only latitudes < 65° N were used for the 10 year EU OBS.





ACPD
14, 6261–6310, 2014

Extreme ozone pollution episodes in a global atmospheric chemistry model

J. L. Schnell et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀ | ▶
◀ | ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion