AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2013JD021435

Key Points:
- CCM results disagree with measured long-term changes of O_3 concentrations
- Reduced confidence implied for estimates of radiative forcing of tropospheric O_3
- Understanding of tropospheric O_3 as it is incorporated into models is incomplete

Supporting Information:
- Readme
- Figures S1–S20

Correspondence to:
D. D. Parrish,
David.D.Parrish@noaa.gov

Citation:
Parrish, D. D., et al. (2014), Long-term changes in lower tropospheric baseline ozone concentrations: Comparing chemistry-climate models and observations at northern midlatitudes, *J. Geophys. Res. Atmos.*, *119*, 5719–5736, doi:10.1002/2013JD021435.

Received 29 DEC 2013
Accepted 17 APR 2014
Accepted article online 24 APR 2014
Published online 13 MAY 2014

Long-term changes in lower tropospheric baseline ozone concentrations: Comparing chemistry-climate models and observations at northern midlatitudes

D. D. Parrish[1], J.-F. Lamarque[2], V. Naik[3], L. Horowitz[4], D. T. Shindell[5], J. Staehelin[6], R. Derwent[7], O. R. Cooper[1,8], H. Tanimoto[9], A. Volz-Thomas[10], S. Gilge[11], H.-E. Scheel[12,13], M. Steinbacher[14], and M. Fröhlich[15]

[1]NOAA ESRL Chemical Sciences Division, Boulder, Colorado, USA, [2]National Center for Atmospheric Research, Boulder, Colorado, USA, [3]UCAR/NOAA Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA, [4]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA, [5]Goddard Institute for Space Studies, National Aeronautics and Space Agency, New York, New York, USA, [6]Institute for Atmospheric and Climate Science, ETHZ, Zürich, Switzerland, [7]rdscientific, Newbury, Berkshire, UK, [8]CIRES, University of Colorado, Boulder, Colorado, USA, [9]National Institute for Environmental Studies, Tsukuba, Ibaraki, Japan, [10]IEK-8, Forschungszentrum Juelich, Juelich, Germany, [11]Hohenpeissenberg Meteorological Observatory, German Meteorological Service, Hohenpeissenberg, Germany, [12]Karlsruhe Institute of Technology, IMK-IFU, Garmisch-Partenkirchen, Germany, [13]Deceased, [14]Swiss Federal Laboratories for Materials Science and Technology, Duebendorf, Switzerland, [15]Air Pollution Control and Climate Change Mitigation, Environment Agency, Vienna, Austria

Abstract Two recent papers have quantified long-term ozone (O_3) changes observed at northern midlatitude sites that are believed to represent baseline (here understood as representative of continental to hemispheric scales) conditions. Three chemistry-climate models (NCAR CAM-chem, GFDL-CM3, and GISS-E2-R) have calculated retrospective tropospheric O_3 concentrations as part of the Atmospheric Chemistry and Climate Model Intercomparison Project and Coupled Model Intercomparison Project Phase 5 model intercomparisons. We present an approach for quantitative comparisons of model results with measurements for seasonally averaged O_3 concentrations. There is considerable qualitative agreement between the measurements and the models, but there are also substantial and consistent quantitative disagreements. Most notably, models (1) overestimate absolute O_3 mixing ratios, on average by ~5 to 17 ppbv in the year 2000, (2) capture only ~50% of O_3 changes observed over the past five to six decades, and little of observed seasonal differences, and (3) capture ~25 to 45% of the rate of change of the long-term changes. These disagreements are significant enough to indicate that only limited confidence can be placed on estimates of present-day radiative forcing of tropospheric O_3 derived from modeled historic concentration changes and on predicted future O_3 concentrations. Evidently our understanding of tropospheric O_3, or the incorporation of chemistry and transport processes into current chemical climate models, is incomplete. Modeled O_3 trends approximately parallel estimated trends in anthropogenic emissions of NO_x, an important O_3 precursor, while measured O_3 changes increase more rapidly than these emission estimates.

1. Introduction

Chemical transport models (CTMs) and chemistry-climate models (CCMs; e.g., the models included in the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP) [*Lamarque et al.,* 2013]) are ambitious efforts to synthesize virtually our entire knowledge of atmospheric chemistry. They provide calculations of atmospheric composition through the depth of the atmosphere over the entire globe. The models can not only simulate present-day conditions but also provide reproductions of the past and predictions of the future. Thus, these models can generally address all questions relating to the concentrations of atmospheric species or the variability of those concentrations in location and time. However, these models are so complex that it is difficult to judge the confidence that should be placed on the answers provided to such questions. Quantitative comparisons of model calculations with well-characterized measurements can help to provide a basis for that judgment.

Ozone (O_3) is a molecule central to the chemistry of the troposphere, where it is primarily of secondary origin, produced through photochemical oxidation of methane (CH_4), carbon monoxide (CO), and volatile organic compounds (VOC) in the presence of nitrogen oxides (NO_x). Downward transport from the stratosphere is an

additional significant source of tropospheric O_3. Photolysis of O_3 is the primary source of OH radicals, which are the major initiator of the photochemical oxidation cycles of the troposphere [*Levy,* 1971]. Ozone itself is a radiatively active gas, so that any changes in atmospheric O_3 concentrations contribute to the radiative forcing of climate change. Ozone is also an important contributor to degraded air quality, with enhanced concentrations associated with negative impacts on human health, agricultural and forest yields, and natural ecosystems [e.g., *Royal Society*, 2008].

Observational records from the nineteenth century at a few specific surface sites indicate that tropospheric O_3 concentrations have increased significantly since preindustrial times [e.g., *Mickley et al.*, 2001, and references therein]. Many models have attempted to calculate this increase throughout the troposphere and thereby provide assessments of the resulting radiative forcing and air quality impacts. A long-standing, significant concern regarding these assessments is that compared to observations, models overestimate preindustrial O_3 concentrations [e.g., *Wang and Jacob,* 1998; *Horowitz,* 2006; *Young et al.,* 2013]. Efforts have been made to identify model improvements that would correct the model overestimates [e.g., *Mickley et al.,* 2001; *Parrella et al.*, 2012]. Questions have also been raised regarding the reliability of the early O_3 measurements [e.g., *Staehelin et al.*, 1994] so that it can plausibly be argued that the model-derived O_3 concentrations are more representative of the preindustrial atmosphere than are the observations [e.g., *Stevenson et al.*, 2013]. Resolution of this disagreement would increase confidence in our knowledge of historical increases in tropospheric O_3 concentrations and in model-based assessments of the associated radiative forcing and air quality impacts.

Our goal in this paper is twofold: first, to present a systematic approach for comparing long-term, model-calculated tropospheric O_3 concentrations with observational records from the last half of the twentieth century up to recent years and second, to provide some initial comparisons of results from three global CCMs with baseline O_3 measurements in the lower troposphere at northern midlatitudes. This time period is expected to include most of the observed O_3 increase since preindustrial times, since the increase of anthropogenic ozone precursor emissions has been larger since World War II [e.g., *Lamarque et al.*, 2010, Figure 1]. These comparisons may provide insight into the model-observation disagreement discussed above. Notably, we will be comparing some of the most recently developed models, which may give more realistic estimates of past ozone changes, with well-characterized measurements made with more modern instrumentation; the comparison will not depend upon nineteenth century measurements, which have served as benchmarks for many previous comparisons. An in-depth analysis of intermodel differences will be left for future work; both *Young et al.* [2013] and *Eyring et al.* [2013] address some differences in modeled tropospheric O_3 concentrations.

2. Models and Observations

The observations and model results compared in this paper are only briefly described here, with references given for more complete descriptions. In particular, the models and simulations included here were contributed to the Coupled Model Intercomparison Project Phase 5 (CMIP5) and are well documented; *Eyring et al.* [2013] present a table summarizing the models, and *Lamarque et al.* [2013] describe the models in detail.

For the analysis described in section 3, we sampled the model monthly mean O_3 concentrations including all times of day at the longitude, latitude, and altitude of the observation sites over the periods specified in the following sections for each model. It is likely that the model-measurement comparisons are sensitive to some degree to the model layer (altitude) chosen for comparison with the observations and to the relatively coarse horizontal resolution of the models. Since our primary focus is on the long-term O_3 changes rather than the absolute O_3 concentrations, we have not investigated this sensitivity in detail; as we will show, the long-term O_3 changes can be analyzed in a manner that is insensitive to the choice of model layer or spatial position.

The following analysis could have been improved, particularly with respect to comparison of absolute O_3 concentrations, through consideration of high time resolution model output. Such an approach would likely improve comparison of surface ozone observations with output from a coarse scale model. In a previous comparison of global CTM output with 54 European surface stations in the EMEP rural ozone network, it was found optimal to compare the model predictions of the means of the maximum daily (i.e., midafternoon) ozone mixing ratios with the comparable measurements [*Derwent et al.,* 2004]. This was because the marked diurnal cycles found at continental rural stations due to surface uptake of ozone under shallow nocturnal

Table 1. Ozone Data Sets Investigated in This Work

Monitoring Site	Latitude/Longitude	Elevation (km)	Dates
	Europe		
Arkona-Zingst, Germany	54°26′N/12°44′E	0.00	1956–2010
Mace Head, Ireland	53°10′N/9°30′W	0.02	1989–2010
Hohenpeissenberg, Germany	47°48′N/11°01′E	1.0	1971–2010
Arosa, Switzerland	46°47′N/9°41′E	1.8	1950s, 1989–2010
Zugspitze, Germany[a]	47°25′N/10°59′E	3.0	1978–2009
Sonnblick, Austria[a]	47°3′N/12°57′E	3.1	1990–2011
Jungfraujoch, Switzerland[a]	46°33′N/7°59′E	3.6	1930s, 1990–2010
	North America		
U.S. Pacific Coast MBL	38–48°N/123–124°W	0–0.24	1985–2010
Lassen NP California U.S.	40°32′N/121°35′W	1.76	1988–2010
North American FT	25–55°N/90–130°W	3.0–8.0	1984–2011
	Asia		
Japanese MBL	38–45°N/138–142°E	0–0.11	1998–2011
Mount Happo, Japan	36°17′N/137°48′W	1.85	1991–2011

[a]Zugspitze, Jungfraujoch, and Sonnblick are treated together as Alpine sites in some of the analyses.

stable layers are not simulated in global models. During the midafternoon, surface observations reflect ozone levels in a deeper atmospheric layer and provide a better comparison with model predictions. Future comparisons will benefit from consideration of monthly means of midafternoon maximum ozone mixing ratios, in addition to monthly means, to enable a more careful and meaningful assessment of model performance against observations at surface continental stations.

2.1. Observational Datasets

Ozone trends in the troposphere have been evaluated over a variety of longer and shorter time periods by different techniques and based on different measurement and analysis approaches and data sets [see *Oltmans et al.,* 1998, 2006, 2013, and references therein]. In this work we will limit our consideration to the data that were the focus of two recent analyses. *Logan et al.* [2012] present a critical analysis of changes in O_3 over Europe based upon observations from alpine surface sites, sondes, and commercial aircraft. They construct a mean time series from 1978 to 2009 using data from three alpine surface sites in central Europe—Jungfraujoch, Zugspitze, and Sonnblick—and demonstrate that this time series is generally consistent with sonde and regular aircraft data available from the same region. There is excellent agreement between data sets since 1998, although there are some substantial differences between the sondes and other data at earlier times. This alpine data set is the longest and best characterized long-term record of lower troposphere O_3 concentrations available to us. *Parrish et al.* [2012] quantify O_3 changes at 11 relatively remote northern midlatitude locations that are believed to represent baseline O_3 (here understood as representative of continental to hemispheric scales) over the past six decades. The sites were selected based upon the quality and length of their measurement records and to provide some representation for all three northern midlatitude continents. They comprise six European sites (beginning in the 1950's and before) including two of the alpine sites considered by *Logan et al.* [2012], three data sets from western North America (beginning in 1984), and two from Asia (beginning in 1991). Table 1 gives some information regarding the sites including the dates of available data. The mean monthly measurement data are derived from archived data sets as described by *Parrish et al.* [2012]. Ozone concentrations are consistently expressed as mixing ratios in units of nmol O_3/mole air, referred to as ppbv throughout the paper.

The data sets analyzed here are the same as those analyzed by *Logan et al.* [2012] and *Parrish et al.* [2012], except for the following differences. *Logan et al.* [2012] considered the available alpine data through 2009, and for each year calculated a single seasonal average O_3 concentration including all sites with data available for that year. Here we consider the additional data that have become available since the *Logan et al.* [2012] analysis (2010 and 2011 at Sonnblick and 2010 at Jungfraujoch), and we calculate all available seasonal averages for each site separately. All resulting single site seasonal averages are considered in the analysis. Additionally, *Logan et al.* [2012] excluded January–May 1982 data from Zugspitze when computing trends,

since they are apparently outliers (note the ~+15 ppbv outlier in Figure 1). We include those data (as did *Parrish et al.* [2012]). The derived long-term O_3 changes with and without this exclusion are not statistically significantly different, although inclusion of those data widens the confidence limits of the derived changes. For the North American free troposphere data set, *Parrish et al.* [2012] considered the data of *Cooper et al.* [2010], who used a particle dispersion model to filter out data with a recent, strong influence from the North American boundary layer. Those data extended through 2008. Here we use the unfiltered data set of *Cooper et al.* [2012], which includes three additional years of data. *Cooper et al.* [2010] show that filtering the data to exclude North American influence did not lead to statistically significant differences in the derived long-term changes.

Logan et al. [2012] conclude that the alpine time series they constructed is useful for evaluation of hindcast simulations of ozone and show that the temporal variability of ozone is similar on spatial scales of 500–1000 km in the lower and middle troposphere. *Parrish et al.* [2012] note that the observed baseline O_3 concentration changes exhibit a high degree of zonal uniformity. The similarity noted in both of these analyses indicates the large spatial scale of the processes affecting O_3 in the lower troposphere. This spatial similarity plus the selection of relatively remote sites for analysis indicate that these data sets are useful for comparison to calculations from global models, which at present cannot resolve regional distributions of emissions and fine topographical features.

There are other data sets that possibly could be considered, particularly those provided by ozone sondes. These measurements provide valuable information for tropospheric climatology; however, data quality concerns regarding tropospheric long-term changes, particularly for historical Brewer Mast sondes, remain [*Logan et al.*, 2012; *Schnadt Poberaj et al.*, 2009]. For this reason, we have not included these data sets in the following analysis.

2.2. Community Atmosphere Model With Chemistry

The global three-dimensional Community Atmosphere Model is expanded to include interactive chemistry (CAM-chem) [*Lamarque et al.,* 2012] to calculate distributions of gases and aerosols in the troposphere and the lower to midstratosphere, from the surface to approximately 40 km. This model shares much of its parameterizations with MOZART-4 [*Emmons et al.,* 2010]. The standard model configuration includes a horizontal resolution of 1.9° (latitude) by 2.5° (longitude) and 26 hybrid levels, with a time step of 30 min. In order to simulate the evolution of the atmospheric composition over the model vertical range, the chemical mechanism used in this study is formulated to provide an accurate representation of both tropospheric and stratospheric chemistry as initially described in *Lamarque et al.* [2008]; this mechanism includes 81 chemical species involved in 197 reactions. The emissions are as described in *Lamarque et al.* [2010] through year 2000 with later emission kept at their 2000 level. Extensive comparisons with observations (satellite, aircraft, and ground based) are discussed in *Lamarque et al.* [2012]. In addition, CAM-chem has participated in a variety of model intercomparisons, such as described in the ACCMIP special issue (http://www.atmos-chem-phys.net/special_issue296.html). The comparisons described in section 3 use the transient climate simulation results for the period 1951 to 2009 and are extensively described by *Lamarque et al.* [2010]. Because this particular configuration was run by NCAR, the model results are identified as NCAR CAM-chem.

2.3. Geophysical Fluid Dynamics Laboratory Coupled Model

The GFDL-CM3 is a coupled atmosphere–ocean-land-ice model [*Donner et al.,* 2011; *Griffies et al.*, 2011] that simulates climate physics and tropospheric and stratospheric chemistry interactively over the full model domain [*Austin et al.*, 2013; *Naik et al.,* 2013]. The standard model configuration uses a finite-volume atmospheric dynamical core on a cubed sphere with horizontal grid varying from 163 km (at the six corners of the cubed sphere) to 231 km (near the center of each face) over the globe, a resolution denoted as C48 (model results are interpolated to 2° latitude × 2.5° longitude grid). The vertical coordinate includes 48 hybrid pressure levels ranging in thickness from 70 m at the surface to 1–1.5 km in the upper troposphere to 2–3 km in most of the stratosphere with a top level at 0.01 hPa (~86 km). The CM3 model simulates atmospheric distributions of 97 chemical species interacting via 236 reactions throughout the model domain, with a time step of 30 min. Stratospheric and tropospheric chemistry are simulated seamlessly by combining the stratospheric chemistry formulation of *Austin and Wilson* [2010] and the tropospheric chemistry mechanism of *Horowitz et al.* [2003, 2007]. *Naik et al.* [2013] and *Austin et al.* [2013] provide detailed description and evaluation of tropospheric and

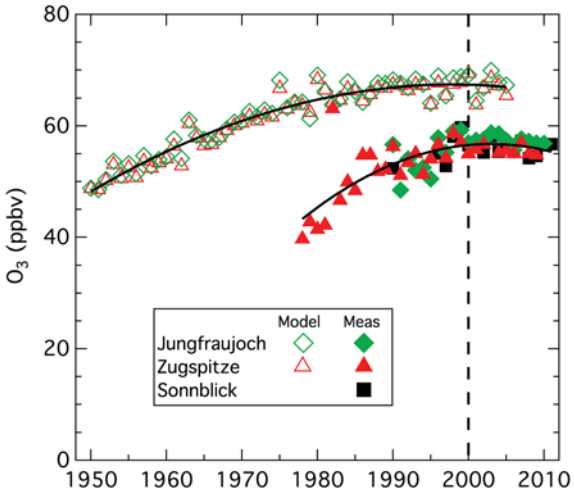

Figure 1. Seasonally averaged springtime (March, April, and May) O_3 concentrations at alpine sites in Europe. Closed and open symbols give measurements and GFDL CCM results, respectively. The solid lines give quadratic fits to respective results. The vertical dashed line indicates the year 2000 reference.

stratospheric chemistry, respectively, simulated by the model. Long timescale coupled transient simulations of CM3 have been performed for a range of experiments in support of the IPCC Fifth Assessment Report (AR5). Here, we consider results from one member of the five-member ensemble historical (1860–2005) simulations [*John et al.,* 2012; *Austin et al.,* 2013; *Eyring et al.,* 2013]. The runs were forced with time-varying, spatially distributed anthropogenic and biomass burning emissions as described in [*Lamarque et al.,* 2010] through 2000 with later emission trends following the RCP4.5 scenario [*Lamarque et al.*, 2012]. Natural emissions of O_3 precursors, except lightning NO_x, were held fixed at 2000 levels. Lightning NO_x emissions were calculated interactively as a function of subgrid convection in the model and therefore vary in time. The comparison described in section 3 is based on model results for the period 1950 to 2005.

2.4. Goddard Institute for Space Studies Climate Model

The GISS-E2-R model is a coupled atmosphere–ocean-land-ice model that simulates climate physics and chemistry interactively over the full model domain [*Shindell et al.,* 2013]. The model was run at 2° latitude by 2.5° longitude resolution, with increased effective resolution for tracers by carrying higher-order moments at each grid box. The configuration used had 40 vertical hybrid sigma layers from the surface to 0.1 hPa (80 km). ACCMIP diagnostics for GISS-E2-R were saved from the GISS-E2-R CMIP5 transient climate simulations as those included fully interactive chemistry and aerosols. Those simulations were spun up for more than 1000 years, after which an ensemble of five simulations was performed for 1850–2012. The gas phase chemistry scheme included both tropospheric and stratospheric chemistry, with 156 chemical reactions among 51 species, with a time step of 30 min. Detailed evaluation of the chemistry in this model has been documented previously [*Shindell et al.,* 2013]. Anthropogenic and biomass burning emissions are those described in *Lamarque et al.* [2010]. Natural emissions include NO_x from lightning and isoprene, both of which vary with climate, and prescribed emissions of other biogenic VOCs and NO_x from soils. The comparison described in section 3 is based on model results for the period 1931 to 2012.

3. Analysis and Results

Our primary focus is on model-measurement comparisons of long-term changes in tropospheric O_3 concentrations. Figure 1 shows the measured seasonal average O_3 for one data set (European alpine) during one season (spring) and compares these measurements to the results from one example model calculation. To effectively compare long-term O_3 changes, we analyze polynomial fits (quadratic fits shown in Figure 1) to the model results and to the measurement data in order to extract and compare the long-term changes that underlie the interannual variability. *Logan et al.* [2012] present similar quadratic fits to the data of Figure 1, and *Parrish et al.* [2012] utilized quadratic fits to all of the data sets that they examined. The functional fits utilized in this work are chosen to adequately capture decadal scale O_3 changes without influence from interannual scale variations. The coefficients of the functional fits provide a convenient means to quantitatively compare the long-term O_3 changes between models and measurements. As discussed in *Parrish et al.* [2012], the time scale will be referenced to zero in the year 2000 to facilitate interpretation of the coefficients derived from the functional fits. This fitting process is mathematically equivalent to approximating the long-term O_3 concentration evolution by the first few terms of a power series expansion with year 2000 as the origin.

Different time periods will be considered when comparing European and western North American/Asian data sets due to the different time periods covered by the measurements. European data extend over much of the post–World War II period (when the majority of the increase in total anthropogenic precursor emissions is

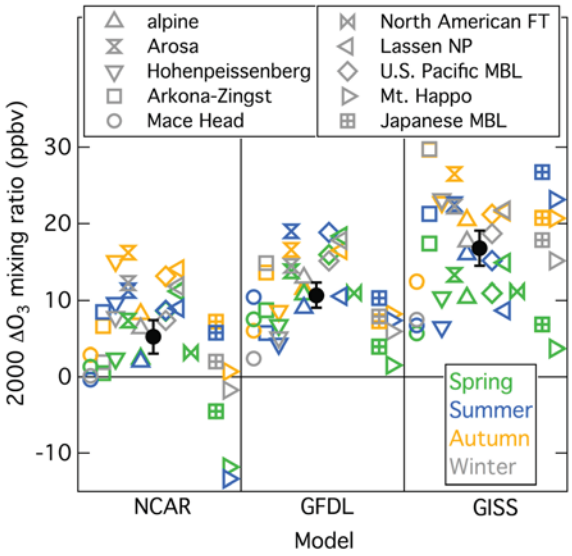

Figure 2. Difference between modeled and measured seasonally averaged year 2000 O_3 concentrations (ΔO_3) for five European (left of black symbols with error bars), three western North American (right of symbols), and two Asian data sets (far right in each panel). Data sets and seasons are indicated by different symbols and colors, respectively, as indicated in the annotations. The three panels compare results from the three indicated models. The black symbols indicate averages over all ten data sets for each of the models; the error bars give 2σ confidence limits of the averages.

believed to have occurred), while the western North American/Asian measurements began only in the mid-1980s when most of that increase had ended. Thus, the measurement record in Europe allows an analysis of the model response to the increase in total anthropogenic precursor emissions. The measurement record from western North America/Asia allows an analysis for the time period when Asian emissions have increased dramatically.

All of the comparisons presented in this work are based on seasonal averages; i.e., means of O_3 concentrations over 3 month periods including all times of day: spring (March, April, and May), summer (June, July, and August), etc. As discussed by *Parrish et al.* [2012], seasonal averages provide a good compromise between minimizing variability associated with shorter time period averages (e.g., monthly) while still providing information on seasonal dependence of long-term O_3 changes.

3.1. Comparison of Absolute O_3 Concentrations: Year 2000 Intercepts

The difference between the model and measurements in the European alpine springtime O_3 (Figure 1) varies over time. Here we select the year 2000 as a reference for all comparisons of absolute O_3 concentrations and take the intercepts of the quadratic fits with year 2000 as a measure of the absolute O_3 concentrations from both the model results and the measurements. In the example shown in Figure 1, those intercepts are 67.4 ± 0.6 ppbv for the GFDL model results and 56.6 ± 0.9 ppbv for the measurements (here and elsewhere, 95% confidence limits are indicated unless otherwise stated.) We quantify the difference in absolute O_3 concentrations between model and measurements as the difference in the respective year 2000 intercepts:

$$\Delta O_3 = O_3(\text{model})_{2000} - O_3(\text{measurement})_{2000}. \quad (1)$$

For the results in Figure 1, $\Delta O_3 = 10.8 \pm 1.1$ ppbv.

The ΔO_3 results for all 10 data sets in all four seasons for the three CCM models are summarized in Figure 2. Each symbol represents a ΔO_3 calculation from equation (1) analogous to that illustrated in Figure 1. The sites are organized in each panel with the European sites on the left, North American data sets to the right of the solid symbols, with the two Asian sites on the far right. Within each continent the sites are organized from left to right in order of increasing altitude. With the exception of some negative ozone biases for the NCAR model at the Japanese sites, all of the ΔO_3 values are positive, indicating that (as exemplified in Figure 1) the three models each overestimate O_3 concentrations in the lower free troposphere throughout northern midlatitudes.

The comparisons of the absolute O_3 concentrations in Figure 1 are expected to be sensitive to the limited horizontal and vertical resolution of the models. The models are sampled at the altitude of the station, which may well be above the ground level assumed in the model. For example, at the elevated European alpine sites the model results will be representative of the free troposphere while the surface measurements are likely affected by upslope flow of boundary layer air from lower altitudes. Since our primary focus in this paper is on the long-term changes of O_3 concentrations, we have made no systematic effort to investigate this sensitivity. Nevertheless, in Figure 2 the model-measurement differences found at the alpine sites (vertical triangles) agree well with the average differences (solid symbols with error bars) for all seasons in all three models, which suggests that errors arising from this uncertainty are not large.

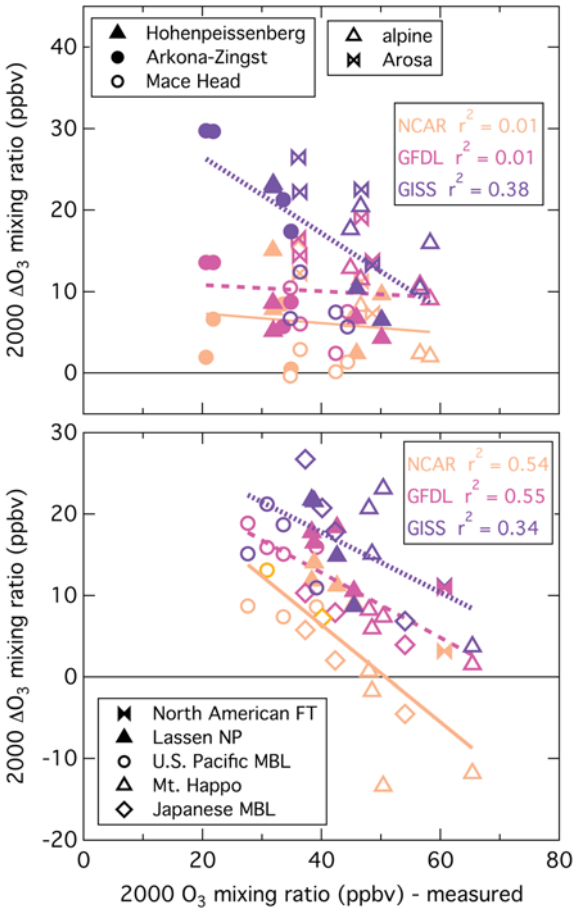

Figure 3. Dependence of ΔO_3 on magnitude of seasonal average measured O_3 concentrations for (top) European and (bottom) North American and Asian data sets. Data sets and models are indicated by different symbols and colors, respectively, as indicated in the annotations. Lines indicate results of linear regressions.

The three models differ in their overestimation of O_3 concentrations. Considering together all seasons and all sites, the NCAR, GFDL, and GISS models on average overestimate O_3 by 5.2 ± 2.2, 10.7 ± 1.6, and 16.8 ± 2.3 ppbv, respectively (where the indicated uncertainties are 95% confidence limits of the averages, assuming that each ΔO_3 determination is independent). There are smaller systematic differences in the ΔO_3 values between seasons and continents. The three models on average overestimate average O_3 in autumn (14.2 ± 2.8 ppbv) significantly more than in spring (7.4 ± 2.4 ppbv), with intermediate overestimates for the other two seasons (summer = 10.5 ± 3.2 ppbv and winter = 11.8 ± 3.0 ppbv). The average overestimate for the three North American data sets (13.7 ± 1.8 ppbv) is significantly larger than for Asia (7.2 ± 4.1 ppbv), with the European average (11.1 ± 1.9 ppbv) intermediate.

Within Europe there are, on average, no systematic differences among stations with respect to site elevation; the average biases of model results from the marine boundary layer to the alpine and free troposphere data sets are statistically not significantly different. However, Figure 3 indicates that the models generally yield a greater overestimate of the seasonal average O_3 concentrations at the stations with the lowest observed O_3 concentrations, particularly for the North American and Asian data sets. The squares of the correlation coefficients annotated in Figure 3 indicate that approximately 1% to 55% of the variability in the ΔO_3 values can be explained by the measured average O_3 concentrations. Evidently, the models have greater difficulty in reproducing smaller observed seasonal average O_3 concentrations, although certainly other factors also contribute to the model-measurement differences.

3.2. Measurement and Model Shape Factors

Parrish et al. [2012] found significant similarity in the long-term O_3 concentration changes throughout northern midlatitudes when those changes were expressed as percent changes relative to the year 2000 intercepts. We take advantage of that similarity in this analysis through normalization of all observations and model results by dividing all seasonal averages by the respective year 2000 intercepts. Figure 4 illustrates this normalization process for summertime European observations and results from the GISS model. Figures 4a and 4b show the unnormalized data and model calculations, and Figures 4c and 4d show the results after dividing the data and model results by the 2000 intercept from the respective quadratic fits. The normalized O_3 concentrations exhibit similar long-term temporal evolution at all of these relatively remote European sites in both the measurements and model results. A greater degree of scatter is apparent in the normalized measurement data, which may reflect interannual variability, local effects not captured by the model, and perhaps in some cases, measurement problems.

Least square polynomial fits to the normalized data and model results provide a means to quantitatively compare the long-term changes in the observations and model results. We refer to these polynomial fits as "shape factors," since they capture the temporal evolution of the normalized O_3 concentrations at all of the

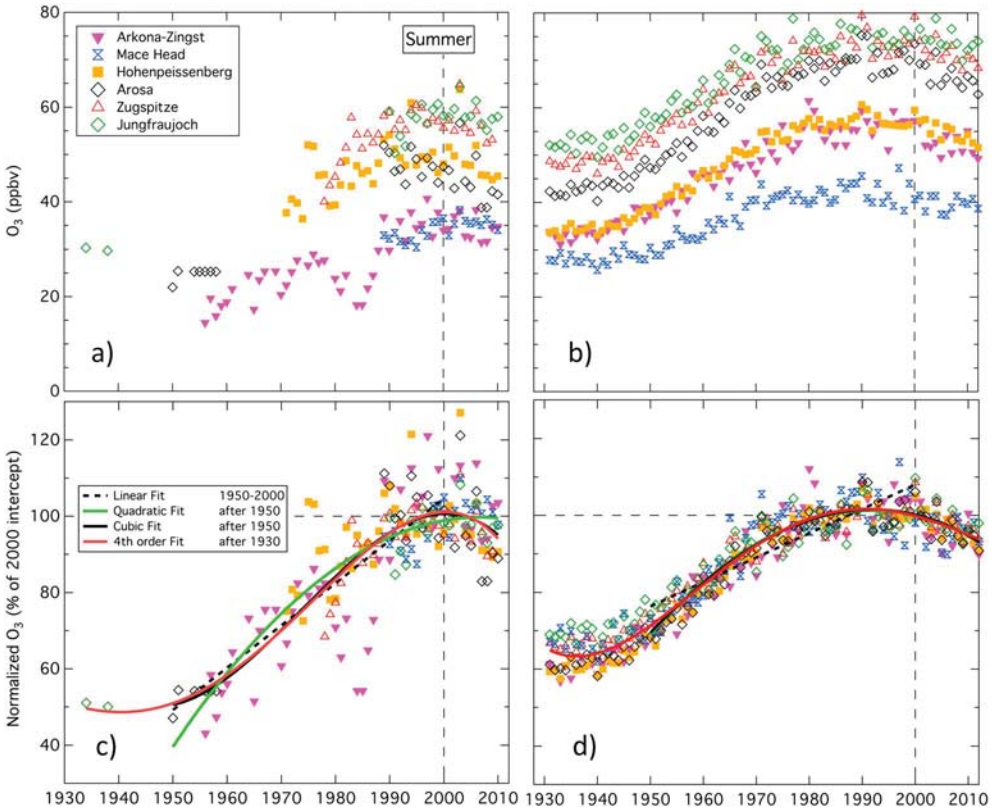

Figure 4. Seasonally averaged, summertime O_3 mixing ratios (a) measured and (b) modeled by GISS CCM at six European sites, and (c, d) those results normalized to year 2000 intercepts. The curves in Figures 4c and 4d are least square polynomial fits to normalized results from all sites; these curves include the years of data annotated in Figure 4c. In Figure 4d, it is difficult to discern the quadratic and cubic curves, as they generally lie beneath the fourth order polynomial curve. The black dashed line indicates the linear least squares fit to all data from 1950 to 2000.

European sites. Figures 4c and 4d illustrate polynomial fits of increasing order for the European sites. We find that four polynomial functions of increasing order are useful for defining the shape factors and for providing a quantitative basis for comparing model results with measurements: (1) linear fits between 1950 and 2000 (dashed black lines), (2) quadratic fits for all results after 1950 (green lines), (3) cubic fits for all results after 1950 (solid black lines), and (4) fourth-order polynomial fits to all results after 1930 (red lines).

The selection of these four polynomial fits for the following analyses are based upon both the quantity to be compared and the maximum number of terms justified by goodness-of-fit considerations, although we will allow some subjectivity in this criterion for consistency of analysis among all of the data sets. There are competing considerations in choosing the number of terms in the polynomial fits; an additional term more accurately describes the temporal evolution but simultaneously decreases the precision (i.e., increases the confidence limits) with which the coefficients of the fit can be determined. In general, the number of terms in the polynomial fit that is statistically justified increases as the length of the data record increases. In the following discussion, the rationale for the selection of the polynomial order is discussed. Supporting information Figures S1–S8 show the corresponding analyses for the measurements and the results of all three models in all four seasons. Table 2 gives the coefficients for the shape factors derived from the observations; they may be used to reproduce these measurement-based shape factors for other purposes, such as comparison with other model results.

The European shape factors from the three models are compared to those from measurements in each season in Figure 5. Two features are particularly notable. First, in all seasons the three models give similar shape factors. Second, the measurement-derived shape factors from all European sites and from only the alpine sites agree reasonably well. It should also be noted that the polynomial fits give physically unreasonable decreasing O_3 concentrations for the measurements and model results at the earliest times; these end effects illustrate the limits of fidelity of the fits at the extremes of the time series.

Table 2. Coefficients of Polynomial Fits That Define the Shape Factors Derived From the Measurements[a]

Season	a	b	c ($\times 10^{-2}$)	d ($\times 10^{-4}$)	e ($\times 10^{-6}$)	Years of Fit
			European Alpine			
Spring	99.935 ± 1.1	0.24309 ± 0.15	−4.4927 ± 1.2	—	—	1998–2011
Summer	99.641 ± 1.4	−0.12083 ± 0.19	−5.5201 ± 1.5	—	—	1998–2011
Autumn	99.297 ± 1.2	0.15778 ± 0.16	−3.6852 ± 1.2	—	—	1998–2011
Winter	99.348 ± 1.2	0.55859 ± 0.17	−4.7280 ± 1.3	—	—	1999–2011
			Europe			
Spring	100.45 ± 1.5	0.42533 ± 0.18	−4.0812 ± 1.6	−5.6916 ± 3.2	—	1951–2010
Summer	101.11 ± 2.0	0.03828 ± 0.21	−6.2334 ± 2.6	−11.3230 ± 9.5	−5.4281 ± 9.5	1934–2010
Autumn	99.168 ± 1.7	0.35747 ± 0.21	−1.5011 ± 1.8	−0.1440 ± 3.5	—	1950–2010
Winter	99.405 ± 1.8	0.81373 ± 0.22	−2.7518 ± 1.9	−3.9481 ± 3.7	—	1950–2010
			North America and Asia			
Spring	99.565 ± 1.4	0.93265 ± 0.16	−2.1775 ± 2.2	—	—	1984–2011
Summer	99.795 ± 2.8	0.84819 ± 0.27	−6.3693 ± 4.4	—	—	1988–2011
Autumn	99.897 ± 2.4	0.38730 ± 0.25	−3.1554 ± 4.1	—	—	1988–2011
Winter	99.349 ± 2.2	0.91580 ± 0.23	−2.5665 ± 3.9	—	—	1988–2011

[a]The fits are of the form $y = a + bt + ct^2 + dt^3 + et^4$ where the coefficients in the table have been divided by the factor indicated below the respective coefficient symbol. The unit of t is years.

There is substantial qualitative similarity between the models and measurements. As has been well established by many studies, O_3 concentrations over Europe have increased markedly over the past decades, and this is clearly exhibited by the model results as well as by the measurements. The earliest measurements and the results of the GISS model chosen as an example here, which extend back to 1930, both indicate a

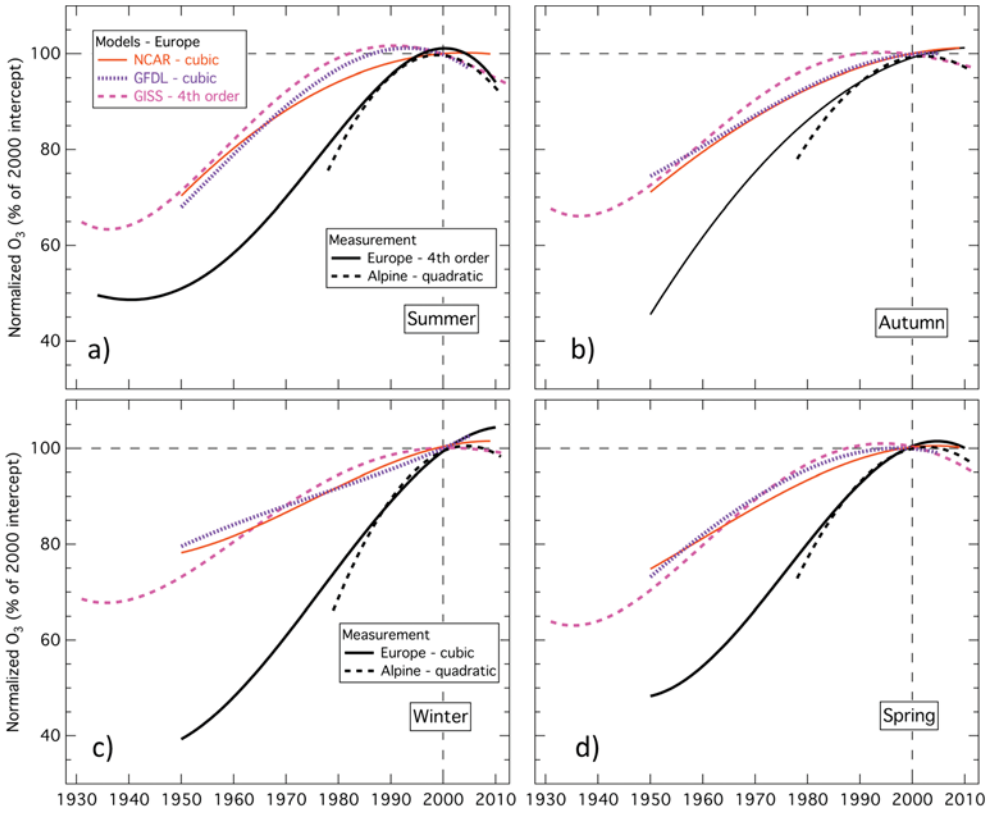

Figure 5. Comparison of shape factors for four seasons from three models with those from observations at European and Alpine sites. The models and observational data sets are identified in the annotations. (a) Functional fits are indicated except (b–d) cubic fits are shown for the Europe observations as annotated in Figure 4c.

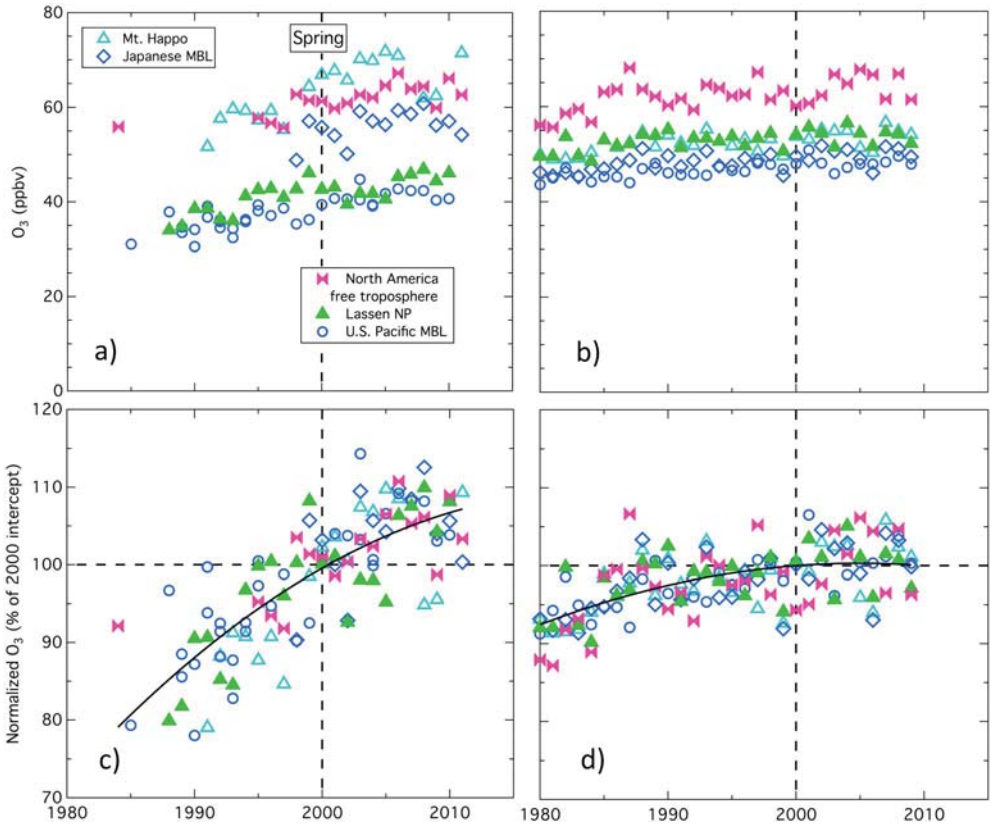

Figure 6. Long-term seasonally averaged, springtime O_3 mixing ratios (a) measured and (b) modeled by NCAR CCM for five North American and Asian data sets, and (c, d) those results normalized to year 2000 intercepts. The curves are the least square quadratic fits to normalized data from all sites in Figure 6c and normalized model results from 1965 and later years in Figure 6d.

leveling off of O_3 concentrations before about 1950. These earliest measurements are quite limited, since they represent only 7 and 5 days of measurements in two summers (1934 and 1938) [*Crutzen*, 1988], but they were made by well-developed spectroscopic and chemical instrumental techniques. Since approximately the late 1990s, there has been a slowing of the O_3 concentration increase, and at least at some sites in some seasons, the increasing trend has reversed. This has been previously noted in observations, particularly at the European alpine sites [*Logan et al.,* 2012] as shown in Figure 1 and also at other northern midlatitude sites [*Parrish et al.,* 2012, and references therein]. The models do generally capture a maximum in the O_3 concentrations and a subsequent decrease more recently. However, there are quantitative differences between model results and measurements, with the modeled changes significantly smaller than those observed. Another important difference is that the models find the largest long-term O_3 changes in the summer and the smallest in winter, while the measurements document contrasting behavior—larger changes in winter and smaller changes in summer.

A similar normalization process and extraction of shape factors is illustrated in Figure 6 for western North American and Asian data and model results. The period spanned by the measurements is shorter, so no more than quadratic fits are statistically justified in defining the measurement shape factors. Although the model calculations extend to earlier years, we choose quadratic fits to the model results beginning in 1965, a period that gives precise determination of the parameters of the quadratic fits without requiring additional polynomial terms to adequately describe the shape factors. Supporting information Figures S9–S16 show the corresponding analyses for the measurements and the results of all three models in all four seasons, and Table 2 gives the coefficients for the shape factors derived from the observations. Figure 7 compares the model and measurement shape factors for the western North American and Asian data sets. Similar long-term changes are noted, with initial increases slowing and in some cases reversing. Again, there are

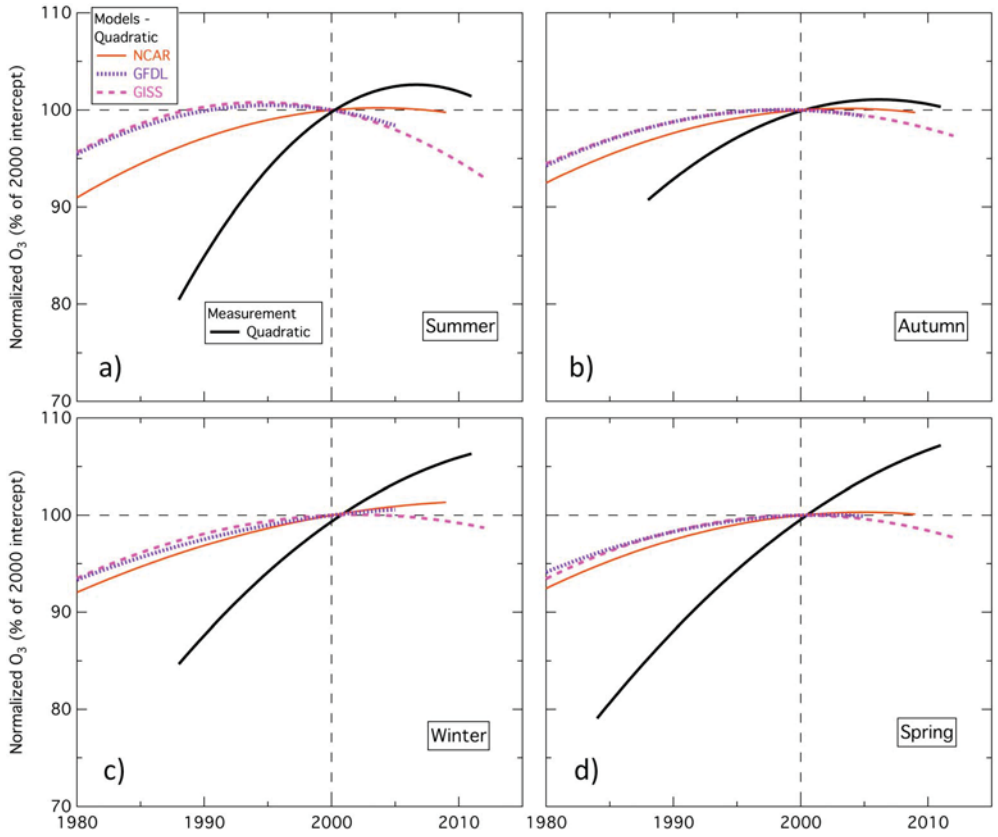

Figure 7. (a–d) Comparison of shape factors for four seasons from three models with those from observations at North American and Asian sites. The models and the observational data set are identified in the annotations, and the functional fits are all as indicated in Figure 7a. The fits to the model results include 1965 and later years.

noticeable quantitative differences between model results and measurements. The following sections will focus on quantifying the extent of agreement and identifying areas of disagreement at all of these midlatitude locations.

One consequence of the similarity of both the measured and the modeled shape factors apparent across the European sites and across the North American and Asian data sets is that exact correspondence of location is not important for comparing relative O_3 changes (i.e., shape factors) between measurement and models. This correspondence is important in comparing the absolute O_3 concentrations as in Figures 2 and 3, but for comparison of the quantitative properties of the shape factors, measurements and models need not be precisely colocated. Thus, these comparisons are not expected to be sensitive to the local environment of the measurement site, or the particular model cell or level selected for the comparisons.

In this analysis, we have chosen to consider only the three sets of shape factors included in Table 2. This choice is made for several reasons. First, there are no unambiguous, statistically significant differences between the polynomial fits to the ozone trends at different sites included in each of the three data sets, so no greater number of shape factors is statistically justified. Second, the precision of the derived coefficients defining the shape factors increases with the number of data considered, so the greater number of sites combined, the greater the precision of the derived parameters. Finally, separately treating the European alpine sites alone and all European sites together allows comparison of shape factors derived over different time periods; notably, the parameters derived from the fits to these two separate European data sets are not statistically significantly different.

3.3. Long-Term O_3 Concentration Changes

Examining the European results in Figures 4c and 4d suggests multiple approaches to quantifying the long-term changes in O_3 concentrations that have occurred over the last half of the twentieth century. Here we adapt the approach of *Parrish et al.* [2012] and fit both the measurements and model results with linear regressions for the

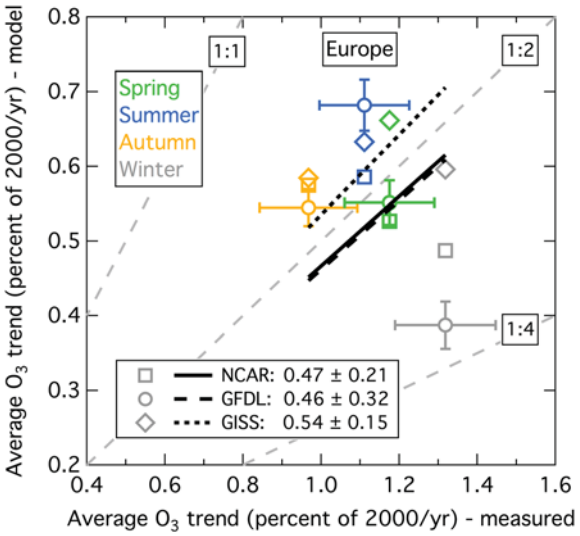

Figure 8. Comparison of modeled and measured average annual changes in seasonally averaged O_3 concentrations for the normalized European results over the 1950–2000 period. Models and seasons are indicated by different symbols and colors, respectively, as annotated. Representative error bars indicate the 95% confidence limits of the measured and GFDL modeled changes. The black lines indicate linear least squares fits (with the intercepts forced through the origin) to the results of the three different models. The identity of the lines and their slopes with 95% confidence limits are annotated. The dashed grey lines give the indicated model to measurement ratios.

years 1950–2000. The black dashed lines in Figures 4c and 4d are examples of these fits. The slopes of these lines provide the quantification of the long-term change over the 50 year period. Figure 8 compares the long-term changes at the European sites calculated by the models with those derived from the measurements. The results from the measurements are between about 1.0 (autumn) and 1.3 (winter) in units of percent of year 2000 intercept per year, which correspond to factors of 2 to nearly 3 increases from 1950 to 2000. These measurement results are consistent with those reported by *Parrish et al.* [2012].

Each of the models on average reproduces approximately one half of the O_3 increases seen in the measurements between 1950 and 2000. For all seasons and all models, the points in Figure 8 lie near or below the 1:2 model to measurement line. In general, the models do not reproduce the seasonal variation observed in the changes, with wintertime changes more greatly underestimated by the models; the result is an overall negative correlation ($r = -0.40$) between the models and measurements considering all points in Figure 8.

Two features of using linear fits to quantify long-term changes should be noted. First, although a linear fit is utilized, there is no assumption that the change over the 50 year period was actually a constant linear increase; as discussed by *Parrish et al.* [2012], the linear fit gives an excellent approximation for the average annual change even for nonlinear changes. Given the scatter of the measurements illustrated in Figure 4c, fitting a higher degree polynomial to the measurement data over the 1950–2000 period is not statistically justified, and the slope of the black dashed line provides a good approximation to the average annual change. However, the scatter in the model results is smaller, and significant deviations from a linear change are obvious (Figure 4d). Nevertheless, the linear fit again gives an excellent approximation for the average annual change; calculation of the average annual change from the higher-order polynomial fits illustrated in Figure 4d agree with the linear fit result for all models in all seasons to within 4%. Second, the units of the linear slope may be confusing. Percent per year often indicates an exponential trend, as the percent is based upon an ever-changing reference quantity. Here the reference quantity is fixed at the year 2000 intercept, so the units used here (percent of year 2000 intercept per year) do correspond to a linear increase. It is perhaps worth noting that the maximum physically reasonable slope possible is 2% of year 2000 intercept per year, since that would represent a change from zero to the year 2000 intercept over the 50 year period considered.

A different approach is used to quantify the trends in O_3 concentrations in the western North American and Asian data sets; we take the slope of the quadratic shape factor in the year 2000 (Figures 6c, 6d, and 7) as a measure of the rate of change of O_3 concentrations in that year. The year 2000 is near the center of the period covered by the available data, so the slope in that year gives a good approximation for the average annual change over the measurement record. The slope in the year 2000 is also easily obtained from the quadratic fits (Table 2 for the measurements) to the measurements and model results shown in Figure 7. Figure 9 compares the trends of the North American and Asian sites calculated by the models with those derived from the measurements. The results from the measurements are about 0.4 in autumn and 0.8 to 0.9 in other seasons (units of % of year 2000 intercept per year), which are in reasonable accord with those reported by *Parrish et al.* [2012].

None of the models accurately reproduces the increases seen in the western North American and Asian measurements. The NCAR model does indicate generally increasing O_3 concentrations in 2000 but captures

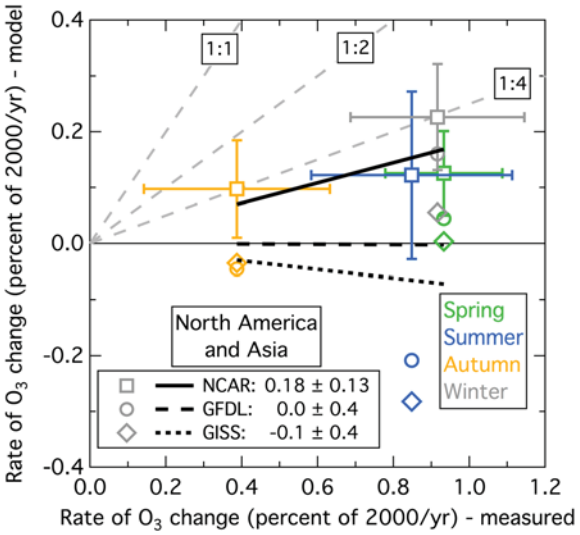

Figure 9. Comparison of modeled and measured changes in seasonally averaged O_3 concentrations for the year 2000. These changes are for the normalized North American and Asian results. Models and seasons are indicated by different symbols and colors, respectively, as annotated. Representative error bars indicate the 95% confidence limits on the measured and NCAR modeled changes. The black lines indicate linear least squares fits (with the intercepts forced through the origin) to the results of the three different models. The identity of the lines and their slopes with 95% confidence limits are annotated. The dashed grey lines give the indicated model to measurement ratios.

less than one third of the increase; the other two models suggest no significant increasing trends. Comparison of the model and measurement curves in Figure 7 indicates that the models do find increasing trends in the decades before 2000, but they are smaller and generally these model trends had slowed or reversed by the year 2000. The measurements also suggest that the rate of the increasing trends is decreasing, but where the trends have reversed, that reversal was after 2000. Consequently there is little agreement between the trends predicted by the models and those found by the measurements in the year 2000.

3.4. Rate of Change of O_3 Concentration Trends

As indicated in Figures 1, 5, and 7 measurements and models generally agree that O_3 concentrations throughout northern midlatitudes over the past decades are characterized by increasing trends that have slowed and, in some cases, have recently reversed. The coefficients of the second-order term in the polynomial fits defining the shape factors provide a quantitative comparison of the slowing of the trends between measurements and models. This coefficient quantifies the curvature of the shape factor and has units of percent of year 2000 intercept per year per year. It is physically equal to one half of the rate of change of the decrease (or increase if the coefficient were positive) in the O_3 concentration trend. If a quadratic fit is adequate to quantify the shape factor, then the curvature of the shape factor is constant; if higher-order terms are statistically significant in the polynomial fits that define the shape factors, then the curvature of the shape factor changes with time, and the coefficient of the second-order term gives the curvature in the year 2000. Figure 10 compares the coefficient of the second-order term from the model results with those from the measurements in the three regions under consideration. The models on average underestimate the curvature of the shape factors by factors of 2–4 in all seasons and all data sets, although there are specific seasons in specific data sets when the agreement is better (or worse). However, for most data sets, the models do capture a significant fraction of the seasonal variation in the shape factor curvature as is indicated by the positive correlation coefficients annotated in Figure 10.

3.5. Estimate of Year of Maximum O_3

In the preceding sections, we have quantified disagreements between models and measurements. In some cases, these disagreements are substantial, but the models still do have considerable skill in describing the tropospheric O_3 distribution and its decadal scale temporal variability. One example of this skill is illustrated in Figure 11. As discussed above, the shape factors derived from the measurements and models agree that northern midlatitude O_3 concentrations have increased and have reached, or will reach, a maximum that is, or will be, followed by decreasing concentrations. Figure 11 plots those measured and modeled maximum O_3 concentrations for the European alpine sites as a function of the year of the maximum; in many cases, those modeled maxima are extrapolations based on the simple assumption that shape factors derived from past modeled changes in ozone can predict future changes. There is certainly no guarantee that these extrapolations are reliable predictions of what the model results would show if the calculations were extended into the future, but they may provide some insight into the long-term O_3 changes calculated by the models. There is obvious quantitative disagreement between the models and measurements in Figure 11.

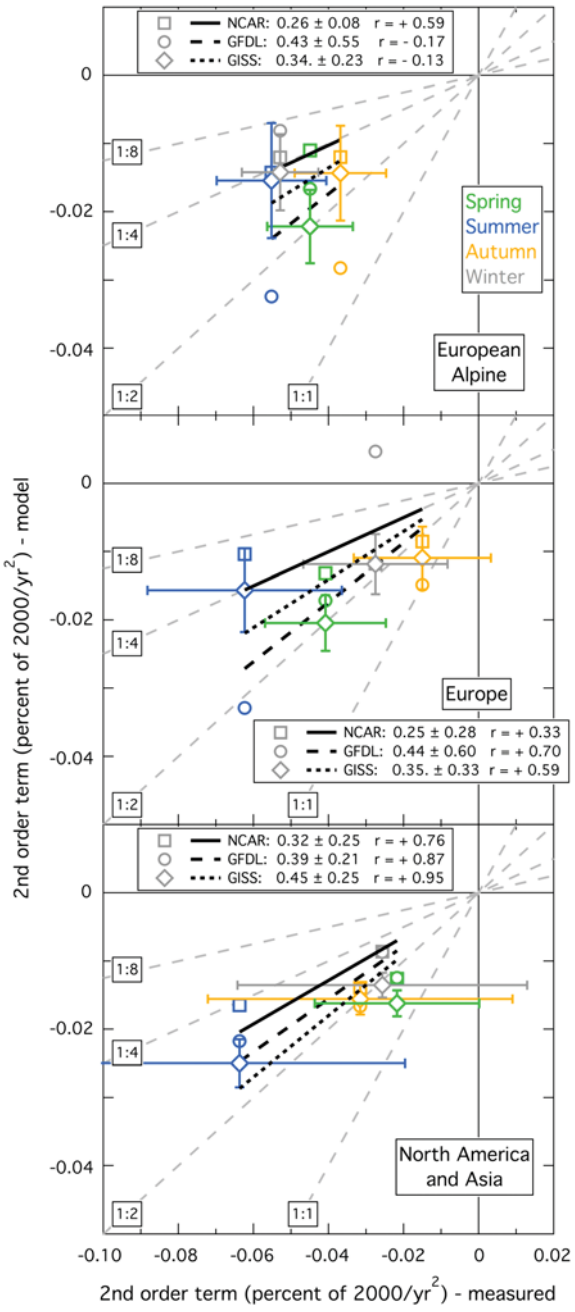

Figure 10. Comparison of modeled and measured second order coefficient in least square polynomial fits to time series of seasonally averaged O_3 concentrations. These comparisons are for the three indicated regions. Models and seasons are indicated by different symbols and colors, respectively, as annotated. Representative error bars indicate the 95% confidence limits on the measured and GISS modeled changes. The black lines indicate linear least squares fits (with the intercepts forced through the origin) to the results of the three different models. The identity of the lines and their slopes with 95% confidence limits are annotated. The dashed grey lines give the indicated model to measurement ratios.

The models consistently overestimate the maximum O_3 concentrations by 2 to 20 ppbv, and the years of the model maxima are usually late (by as much as 28 years) but in some cases are early (by as much as 7 years). However, there are significant features of qualitative agreement that may give some indication of the reasons for the quantitative disagreement.

The seasonal cycles (points connected by line segments in Figure 11) calculated by all three models have distinct similarities to that measured. In all cases, summer has the highest maximum and the earliest peak, while winter has the lowest maximum and the latest peak, with spring and autumn falling between. Considering the results of the three models, there is an apparent anticorrelation between the magnitude of the maxima and the years of those maxima. This relationship is such that overall the points for the four seasons from the three models appear to define a curve, with the highest maxima appearing in the earliest years. This behavior may be an indication of how the models respond in differing ways to changing emissions.

4. Summary and Conclusions

Similarities in long-term changes in lower tropospheric O_3 concentrations measured throughout large regions have been previously noted. *Logan et al.* [2012] examined O_3 over Europe and found that the temporal variability of O_3 in the central part of the continent is similar over spatial scales of 500–1000 km in the lower and middle troposphere and conclude that the similarity of the temporal behavior, including long-term changes, of O_3 at Zugspitze (an alpine site in the Alps) and Mace Head (a marine boundary layer site on the west coast of Ireland) demonstrates the large spatial scale of the processes affecting ozone. *Parrish et al.* [2012] found little if any evidence for statistically significant differences in average rates of increase among data sets from Europe, western North America, and eastern Asia. Here we have examined the same data sets considered by those two studies and show that the relative, long-term O_3 changes are well quantified by two sets of seasonal shape factors. One set represents the changes at all European sites (Figure 5) and the other the changes in the North American and Asian data sets (Figure 7). Additionally, as shown by the summertime shape factors in Figure 12a, there is significant similarity between these two sets of shape factors representing these large regions; the primary difference is that the

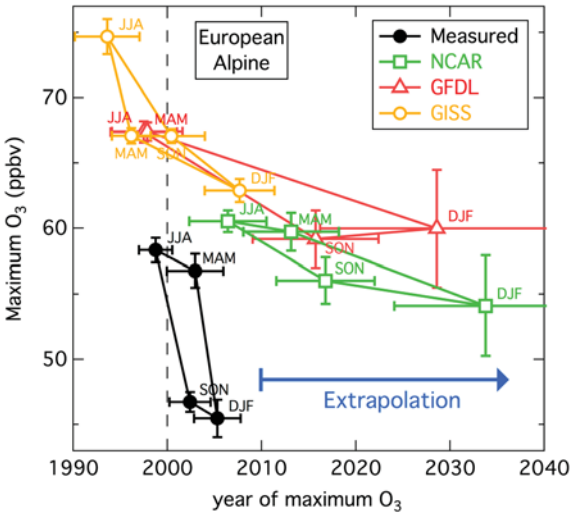

Figure 11. Relationship between the maximum seasonal average O_3 concentration and the year of that maximum calculated from the quadratic fits to the measurements (solid black symbols) and model results (colors and shapes as annotated) for the European alpine sites. The error bars give 95% confidence limits. The maxima after 2010 are extrapolated from the quadratic fits.

North American and Asian data sets approximately parallel the European changes with a 5 to 10 year delay. Such similarities are seen in all four seasons (see supporting information Figures S17–S20).

We can have high confidence in the measurement records from which these shape factors are derived for at least two reasons. First, the European shape factors separately derived from the data sets considered by *Logan et al.* [2012] and by *Parrish et al.* [2012] agree well (Figures 5 and 12a). Second, separate considerations of each of four seasons give comparable results; each seasonal data set is independent, but the results are generally consistent (Figures 5 and 7) with interpretable differences. When multiple data sets over all seasons give consistent results, the confidence in the resulting conclusions is significantly enhanced compared to a conclusion based upon a single data set.

Similarity between O_3 changes throughout northern midlatitudes may be expected given that the O_3 lifetime in the troposphere is 20–30 days or longer depending on season and altitude [*Fusco and Logan*, 2003; *Stevenson et al.,* 2006]. This lifetime is long with respect to the time scale of zonal transport, implying that O_3 is transported on at least intercontinental scales. Thus, changes in the anthropogenic O_3 budget on one continent affect O_3 inflow into downwind continents. Outside of urban areas, this inflow (i.e., baseline O_3) represents a large majority of observed O_3 concentrations. Therefore, at least for regions of zonal transport such as northern midlatitudes, a great deal of longitudinal similarity is expected in long-term O_3 changes.

The three CCMs examined here agree that two sets of seasonal shape factors accurately quantify the changes at all European sites and in the North American and Asian data sets in all seasons (Figures 4, 6, and S1–S16). These shape factors have marked similarities among the models but differ significantly from those derived from measurements (Figures 5 and 7). The models on average overestimate absolute O_3 concentrations throughout northern midlatitudes by ~5 to 16 ppbv (Figure 2). These overestimates are approximately independent of altitude and season but correlate with measured seasonal average O_3 concentration (Figure 3), with the lowest measured concentrations most difficult to match by the models. As a consequence of this correlation, measured seasonal average O_3 concentrations among all data sets vary over a larger range than those derived from models. The models capture only a fraction (~ 50%, Figure 8) of observed O_3 changes from 1950 to 2000 in Europe and little of the year 2000 instantaneous rates of change in North America and Asia (Figure 9). Finally, the rates of these long-term changes are themselves changing, i.e., generally slowing as indicated by the curvature of the shape factors, and the models capture only a fraction (~25 to 45% throughout all northern midlatitude locations, Figure 10) of this rate of slowing.

The comparisons of model calculations with long-term measurements of O_3 concentrations in the lower troposphere at northern midlatitudes find significant agreement but also areas of disagreement. These disagreements are profound enough to raise three major concerns. First, if our models cannot accurately reproduce past O_3 concentrations, what confidence can we place on their prediction of future concentrations? Second, estimates of present-day radiative forcing of tropospheric O_3 are provided by the historic O_3 concentration changes estimated by global models. Since models underestimate these changes by about a factor of approximately two, the radiative forcing may also be underestimated. The radiative forcing of O_3 is most sensitive to concentration changes in the mid and upper troposphere, while we have investigated O_3 changes in the lower troposphere. However, the lifetime of O_3 in the free troposphere is long compared to the time scale for vertical mixing so that the entire vertical O_3 profile is expected to shift

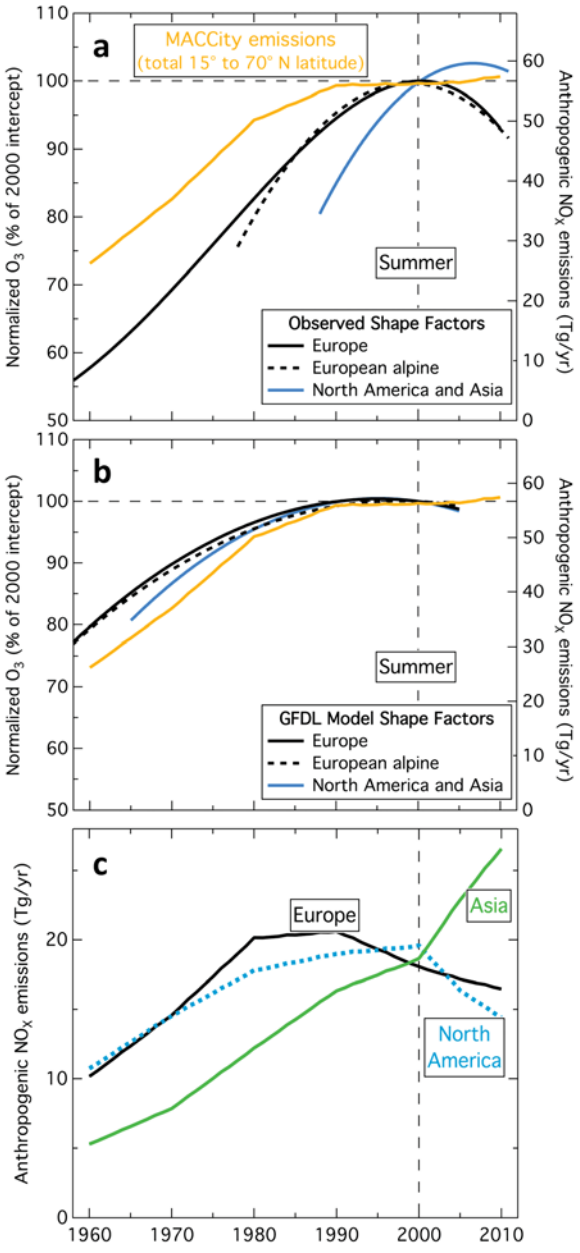

Figure 12. Comparison of (a) observed and (b) modeled long-term relative changes in O_3 concentrations with estimated changes in emissions of NO_x, an important O_3 precursor at northern midlatitudes and (c) over northern midlatitude continents. Emissions include 15 to 70°N latitude on each continent. European, Asian, and North American emissions are totaled from −30 to 60, 60 to 180, and −180 to −30°E, respectively. For comparison here, the European shape factor is renormalized to 100% in the year 2000 by dividing by the first polynomial factor (a) from Table 2. All emissions are taken from the MACCity Emission Inventory [*Granier et al.*, 2011; *Diehl et al.*, 2012; *Lamarque et al.*, 2010; *van der Werf et al.*, 2006] downloaded through the ECCAD emission portal (http://eccad.sedoo.fr/eccad_extract_interface/JSF/page_meta.jsf—accessed 27 September 2013).

regardless at which altitude the source or sink terms change. Finally, there is the question of what is missing from our understanding of the tropospheric O_3 budget, at least in the context of how that understanding is incorporated into our present generation of global atmospheric models. Efforts to modify emission inventories [e.g., *Mickley et al.*, 2001] or to include additional atmospheric chemical cycles in the model chemistry [*Parrella et al.*, 2012] have not yet been entirely satisfactory.

The common behavior of the three models in underestimating the historical increases does suggest that there is a problem common to all three models. Such a problem could possibly arise from inaccuracies in the historical emissions estimates or from a missing (or inadequately simulated) chemical or physical process in the models. It seems less likely that a smaller, more specific problem is the cause, such as bias in wet deposition of some ozone precursor or in the effect of clouds on photolysis rates, since the models implement such specific processes differently. Estimates of emissions trends over the past decades are similar between the models [*Lamarque et al.*, 2010], so inaccuracies in this area are a potential problem that should receive particular attention. There is indeed very limited information on historic emission trends and a documented lack of consistency in historic emission data [*Schultz et al.*, 2007]. *Ordóñez et al.* [2007] have argued that increases in stratospheric input of O_3 may have contributed to changes in O_3 concentrations at the alpine sites, at least from 1992 to 2004. However, *Logan et al.* [2012] conclude that changes in stratospheric input cannot explain O_3 increases over Europe in the 1980s, which indicates that changes in stratospheric input do not play a major role in the observed O_3 changes from 1950 to 2000.

Identified model-measurement disagreements may guide model improvement, but such guidance is not yet obvious. *Logan et al.* [2012] note that the observed O_3 changes "provide a serious challenge to current understanding of the processes that control tropospheric ozone, particularly the increases year-round in the 1980s and in summer in the 1990s when emissions of the key precursor, NO_x, were constant or decreasing over North America and Europe, and Chinese emissions were relatively low."

This concern is emphasized in Figure 12a, where inventories indicate that total anthropogenic NO_x emissions at northern midlatitudes increased by only about 14% from 1980 to 2010, yet summertime O_3 concentrations changed markedly, especially relative to preindustrial concentrations (estimated as equal to approximately 50% on the left ordinate in Figure 12a). There is also little correspondence between observed O_3 changes and continental scale emission changes (Figure 12c) on either the same or upwind continents. Evidently, there is no simple relationship between the anthropogenic emissions changes included in the MACCity Emission Inventory and the observed O_3 concentrations. In contrast, the model results (e.g., Figure 12b) more closely follow the global emissions. Similar comparisons for other seasons (see supporting information Figures S17–S20) show similar or greater lack of correspondence with observed O_3 concentrations and similar agreement for the results from all three models. These results suggest that models would agree much more closely with observations if the anthropogenic emissions were much lower in earlier decades than currently estimated.

It will also be valuable to carefully assess several other aspects of model performance. For one, Figure 3 indicates that the models generally more greatly overestimate seasonal average O_3 concentrations for the lowest observed O_3 concentrations. This behavior may indicate that the treatment of O_3 loss processes within models has shortcomings. In this regard, the balance between photochemical O_3 production and destruction is a sensitive function of the modeled NO_x concentration fields throughout the troposphere, which are poorly constrained by measurements, especially at the low NO_x concentrations that mark the transition from O_3 destruction to production. Second, Figure 7 shows that the decadal changes in O_3 observed over Europe are largest in winter, while the models predict the largest increases in summer. This implies that models do not capture the observed evolution of the O_3 seasonal cycle over past decades [*Parrish et al.*, 2013]; a more systematic investigation of the modeled and measured seasonal cycles may be enlightening. Third, models find long-term O_3 changes over North America and Asia are similar to those over Europe, while measured O_3 changes show some marked differences (e.g., Figures 12a and S17–S20); attempts to resolve this inconsistency may give insight into critical dependencies of model performance. Finally, in this paper we have limited our consideration to northern midlatitudes. This limitation is driven by both the location of the largest anthropogenic emissions and the longest high quality data records in this region. However, similar comparisons for other latitudes may be enlightening.

Acknowledgments

The CESM project (which includes the CAM-chem model) is supported by the National Science Foundation and the Office of Science (BER) of the U. S. Department of Energy. NCAR is operated by the University Corporation of Atmospheric Research under sponsorship of the National Science Foundation. The authors are grateful to the Umweltamt of Graubünden and J. Thudium for providing the Arosa data sets, to the German Umweltbundesamt for providing recent data from Zingst, and to P.G. Simmonds and T.G Spain for providing the Mace Head data, to A.J. Manning for sorting the Mace Head data into baseline and non-baseline observations, to NOAA Earth System Research Laboratory, Global Monitoring Division, for providing data from Trinidad Head, to U.S. National Park Service 2002 for providing Lassen NP data, and to Acid Deposition Monitoring Network in East Asia for providing the Mount Happo and Japanese MBL data. The data analyzed here are available from the sources acknowledged above or the authors themselves. D. Parrish acknowledges support from NOAA's Health of the Atmosphere and Atmospheric Chemistry and Climate Programs. H. Tanimoto acknowledges support from the Global Environment Research Fund of the Ministry of the Environment, Japan (S-7-1).

References

Austin, J., and R. J. Wilson (2010), Sensitivity of polar ozone to sea surface temperatures and halogen amounts, *J. Geophys. Res.*, *115*, D18303, doi:10.1029/2009JD013292.

Austin, J., L. W. Horowitz, M. D. Schwarzkopf, R. J. Wilson, and H. IlLevy (2013), Stratospheric ozone and temperature simulated from the preindustrial era to the present day, *J. Clim.*, *26*, 3528–3543, doi:10.1175/JCLI-D-12-00162.1.

Cooper, O. R., et al. (2010), Increasing springtime ozone mixing ratios in the free troposphere over western North America, *Nature*, *463*, 344–348, doi:10.1038/nature08708.

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

Crutzen, P. J. (1988), Tropospheric ozone: An overview, in *Tropospheric Ozone*, edited by I. S. A. Isaksen and D. Reidel, pp. 3–31, Dordrecht.

Derwent, R. G., D. S. Stevenson, W. J. Collins, and C. E. Johnson (2004), Intercontinental transport and the origins of the ozone observed at surface sites in Europe, *Atmos. Environ.*, *38*, 1891–1901.

Diehl, T., A. Heil, M. Chin, X. Pan, D. Streets, M. Schultz, and S. Kinne (2012), Anthropogenic, biomass burning, and volcanic emissions of black carbon, organic carbon, and SO2 from 1980 to 2010 for hindcast model experiments, *Atmos. Chem. Phys.*, *12*, 24,895–24,954, doi:10.5194/acpd-12-24895-2012.

Donner, L. J., et al. (2011), The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component of the GFDL global coupled model CM3, *J. Clim.*, *24*, 3484–3519, doi:10.1175/2011JCLI3955.1.

Emmons, L. K., et al. (2010), Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4), *Geosci. Model Dev.*, *3*, 43–67, doi:10.5194/gmd-3-43-2010.

Eyring, V., et al. (2013), Long-term ozone changes and associated climate impacts in CMIP5 simulations, *J. Geophys. Res. Atmos.*, *118*, 5029–5060, doi:10.1002/jgrd.50316.

Fusco, A. C., and J. A. Logan (2003), Analysis of 1970–1995 trends in tropospheric ozone at Northern Hemisphere midlatitudes with the GEOSCHEM model, *J. Geophys. Res.*, *108*(D15), 4449, doi:10.1029/2002JD002742.

Granier, C., et al. (2011), Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980–2010 period, *Clim. Change*, *109*(1–2), 163–190, doi:10.1007/s10584-011-0154-1.

Griffies, S. M., et al. (2011), The GFDL CM3 coupled climate model: Characteristics of the ocean and sea ice simulations, *J. Clim.*, *24*, 3520–3544, doi:10.1175/2011JCLI3964.1.

Horowitz, L. W. (2006), Past, present, and future concentrations of tropospheric ozone and aerosols: Methodology, ozone evaluation, and sensitivity to aerosol wet removal, *J. Geophys. Res.*, *111*, D22211, doi:10.1029/2005JD006937.

Horowitz, L. W., et al. (2003), A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2, *J. Geophys. Res.*, *108*(D24), 4784, doi:10.1029/2002JD002853.

Horowitz, L. W., A. M. Fiore, G. P. Milly, R. C. Cohen, A. Perring, P. J. Wooldridge, P. G. Hess, L. K. Emmons, and J.-F. Lamarque (2007), Observational constraints on the chemistry of isoprene nitrates over the eastern Unites States, *J. Geophys. Res.*, *112*, D12S08, doi:10.1029/2006JD007747.

John, J., A. M. Fiore, V. Naik, L. W. Horowitz, and J. P. Dunne (2012), Climate versus emission drivers of methane lifetime against loss by tropospheric OH from 1860–2100, *Atmos. Chem. Phys.*, *12*, 12,021–12,036, doi:10.5194/acp-12-12021-2012.

Lamarque, J.-F., D. E. Kinnison, P. G. Hess, and F. Vitt (2008), Simulated lower stratospheric trends between 1970 and 2005: Identifying the role of climate and composition changes, *J. Geophys. Res.*, *113*, D12301, doi:10.1029/2007JD009277.

Lamarque, J.-F., et al. (2010), Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: Methodology and application, *Atmos. Chem. Phys.*, *10*, 7017–7039, doi:10.5194/acp-10-7017-2010.

Lamarque, J.-F., et al. (2012), CAM-chem: Description and evaluation of interactive atmospheric chemistry in the Community Earth System Model, *Geosci. Model Dev.*, *5*, 369–411.

Lamarque, J.-F., et al. (2013), The Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP): Overview and description of models, simulations and climate diagnostics, *Geosci. Model Dev.*, *6*, 179–206, doi:10.5194/gmd-6-179-2013.

Levy, H., II (1971), Normal atmosphere: Large radical and formaldehyde concentrations predicted, *Science*, *173*, 141–143.

Logan, J. A., et al. (2012), Changes in ozone over Europe: Analysis of ozone measurements from sondes, regular aircraft (MOZAIC) and alpine surface sites, *J. Geophys. Res.*, *117*, D09301, doi:10.1029/2011JD016952.

Mickley, L. J., D. J. Jacob, and D. Rind (2001), Uncertainty in preindustrial abundance of tropospheric ozone: Implications for radiative forcing calculations, *J. Geophys. Res.*, *106*, 3389–3399, doi:10.1029/2000JD900594.

Naik, V., L. W. Horowitz, A. M. Fiore, P. Ginoux, J. Mao, A. Aghedo, and H. IILevy (2013), Impact of preindustrial to present day changes in short-lived pollutant emissions on atmospheric composition and climate forcing, *J. Geophys. Res. Atmos.*, *118*, 8086–8110, doi:10.1002/jgrd.50608.

Oltmans, S., et al. (1998), Trends of ozone in the troposphere, *Geophys. Res. Lett.*, *25*, 139–142, doi:10.1029/97GL03505.

Oltmans, S., et al. (2006), Long-term changes in tropospheric ozone, *Atmos. Environ.*, *40*, 3156–3173.

Oltmans, S., et al. (2013), Recent tropospheric ozone changes—A pattern dominated by slow or no growth, *Atmos. Environ.*, *67*, 331–351.

Ordóñez, C., D. Brunner, J. Staehelin, P. Hadjinicolaou, J. A. Pyle, M. Jonas, H. Wernli, and A. S. H. Prévôt (2007), Strong influence of lowermost stratospheric ozone on lower tropospheric background ozone changes over Europe, *Geophys. Res. Lett.*, *34*, L07805, doi:10.1029/2006GL029113.

Parrella, J. P., et al. (2012), Tropospheric bromine chemistry: Implications for present and pre-industrial ozone and mercury, *Atmos. Chem. Phys.*, *12*, 6723–6740, doi:10.5194/acp-12-6723-2012.

Parrish, D. D., et al. (2012), Long-term changes in lower tropospheric baseline ozone concentrations at northern mid-latitudes, *Atmos. Chem. Phys.*, *12*, 11,485–11,504, doi:10.5194/acp-12-11485-2012.

Parrish, D. D., et al. (2013), Lower tropospheric ozone at northern midlatitudes: Changing seasonal cycle, *Geophys. Res. Lett.*, *40*, 1631–1636, doi:10.1002/grl.50303.

Royal Society (2008), *Ground-Level Ozone in the 21st Century: Future Trends, Impacts and Policy Implications*, RS Policy Document 15/08, The Royal Society, London.

Schnadt Poberaj, C., J. Staehelin, D. Brunner, V. Thouret, H. De Backer, and R. Stübi (2009), Long-term changes in UT/LS ozone between the late 1970s and the 1990s deduced from the GASP and MOZAIC aircraft programs and from ozonesondes, *Atmos. Chem. Phys.*, *9*, 5343–5369.

Schultz, M. G., et al. (2007), REanalysis of TROpospheric chemical composition over the past 40 years (RETRO)—A long-term global modeling study of tropospheric chemistry, Final Report, Published as report no. 48/2007 in the series "Reports on Earth System Science" of the Max Planck Institute for Meteorology, Hamburg, ISSN:1614-1199.

Shindell, D. T., et al. (2013), Interactive ozone and methane chemistry in GISS-E2 historical and future climate simulations, *Atmos. Chem. Phys.*, *13*, 2653–2689.

Staehelin, J., J. Thudium, R. Bühler, A. Volz, and W. Graber (1994), Surface ozone trends at Arosa (Switzerland), *Atmos. Environ.*, *28*, 75–87.

Stevenson, D. S., et al. (2006), Multimodel ensemble simulations of present-day and near-future tropospheric ozone, *J. Geophys. Res.*, *111*, D08301, doi:10.1029/2005JD006338.

Stevenson, D. S., et al. (2013), Tropospheric ozone changes, radiative forcing and attribution to emissions in the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), *Atmos. Chem. Phys.*, *13*, 3063–3085, doi:10.5194/acp-13-3063-2013.

van der Werf, G. R., J. T. Randerson, L. Giglio, G. J. Collatz, P. S. Kasibhatla, and A. F. Jr.Arellano (2006), Interannual variability in global biomass burning emissions from 1997 to 2004, *Atmos. Chem. Phys.*, *6*, 3423–3441, doi:10.5194/acp-6-3423-2006.

Wang, Y., and D. J. Jacob (1998), Anthropogenic forcing on tropospheric ozone and OH since preindustrial times, *J. Geophys. Res.*, *103*, 31,123–31,135, doi:10.1029/1998JD100004.

Young, P. J., et al. (2013), Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), *Atmos. Chem. Phys.*, *13*, 2063–2090, doi:10.5194/acp-13-2063-2013.