Intercomparison of global ground-level ozone datasets for health-relevant metrics

Hantao Wang[1], Kazuyuki Miyazaki[2], Haitong Zhe Sun[3], Zhen Qu[4], Xiang Liu[5], Antje Inness[6], Martin Schultz[7], Sabine Schröder[7], Marc Serre[1], J. Jason West[1]

[1]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, Chapel Hill, 27599, USA
[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, 91125, USA
[3]Centre for Sustainable Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, 119228, SG
[4]Department of Marine, Earth and Atmospheric Sciences, North Carolina State University, Raleigh, 27606, USA
[5]Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts, 02138, USA
[6]ECMWF, Shinfield Park, Reading, RG2 9AX, UK
[7]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, 52428, Germany

Correspondence to: J. Jason West (jasonwest@unc.edu)

Abstract

Ground-level ozone is a significant air pollutant that detrimentally affects human health and agriculture. Global ground-level ozone concentrations have been estimated using chemical reanalyses, geostatistical methods, and machine learning, but these datasets have not been compared systematically. We compare six global ground-level ozone datasets (three chemical reanalyses, two machine learning, one geostatistics) ~~against one another and~~ relative to observations and against one another, for the ozone season daily maximum 8-hour average mixing ratio, for 2006 to 2016. Comparing with global ground-level observations, most datasets overestimate ozone, particularly at lower observed concentrations. In 2016, across all stations, grid-to-grid R^2 ranges from 0.50 to 0.75 and RMSE 4.25 to 12.22 ppb. Agreement with observed distributions is reduced at ozone concentrations above 50 ppb. Results show significant differences among datasets in global average ozone, as large as 5-10 ppb, multi-year trends, and regional distributions. For example, in Europe, the two chemical reanalyses show an increasing trend while ~~the~~ other datasets show no increase. Among the six datasets, the share of population exposed to over 50 ppb varies from 61% [28%, 94%] to 99% [62%, 100%] in East Asia, 17% [4%, 72%] to 88% [53%, 99%] in North America, and 9% [0%, 58%] to ~~77%~~76% [22%, 96%] in Europe (2006–2016 average). ~~These differences are large enough to impact assessments of health impacts and other applications. Comparing with Tropospheric Ozone Assessment Report (TOAR) II ground-level observations, most datasets overestimate ozone, particularly at lower observed concentrations. In 2016, across all stations, R^2 ranges among the six datasets from 0.35 to 0.63, and RMSE from 5.28 to 13.49 ppb. Agreement between modeled and observed ozone distributions is reduced at ozone concentrations above 50 ppb.~~ Although ~~some datasets share~~ sharing some of the same input data, we found important differences ~~among these datasets~~, likely from variations in approaches,

resolution, and other input data, highlighting the importance of continued research on global ozone distributions. These discrepancies are large enough to impact assessments of health impacts and other applications.

1. Introduction

Tropospheric ozone is a secondary pollutant that significantly impacts human health, plant life, and the climate system. Past studies have shown that ozone exposure can cause health effects ranging from mild subclinical symptoms to mortality (Balmes, 2022). The Global Burden of Disease 2021 (GBD) study estimated that ground-level ozone contributed to approximately 490,000 (95% UI: 107,000–837,000) global deaths in 2021, representing 0.72% (95% UI: 0.16% – 1.18%) of all deaths that year (Brauer et al., 2024). Ozone exposure is harmful not only to humans but also to plants. Ozone can enter plants through their stomata and cause oxidative damage, which reduces the global yields of major crops such as soybean, wheat, rice, and maize (Ainsworth, 2017; Mills et al., 2018a). Ozone is also an important greenhouse gas, ranking third behind carbon dioxide and methane in its contribution to anthropogenic climate change (Masson-Delmotte et al., 2021). Gaudel et al. find that since the mid-1990s, tropospheric ozone above the surface has increased across all 11 study regions in the Northern Hemisphere that they defined and analyzed (Western North America, Eastern North America, Southeast North America, Northern South America, Northeast China/Korea, ~~The~~ Persian Gulf, India, Southeast Asia, Malaysia/Indonesia, Europe, Gulf of Guinea) (Gaudel et al., 2020). In the United States, although extreme ground-level ozone concentrations have declined, winter ground-level ozone concentrations have increased in the Southwest and Midwest regions since 1990s (Chang et al., 2024). Using one global ozone dataset, from data fusion of ground observations and chemical model outputs, it is estimated that in 2017 21% of the global population was exposed to ozone concentrations above 65 ppb, and 96% lived in areas where concentrations exceeded the WHO guideline (30 ppb for annual metric) (Becker et al., 2023; Delang et al., 2021). Despite existing assessments, substantial uncertainties remain due to observational gaps, especially in remote and developing regions. The lack of knowledge of the ground-level ozone distribution in these regions limits our ability to accurately assess ozone impacts on human health and crops.

The Tropospheric Ozone Assessment Report (TOAR) aggregates ozone observations from thousands of monitoring stations worldwide, forming the most extensive ground-level ozone monitoring data compilation to date (Schultz et al., 2017). Using the TOAR dataset, researchers have analyzed the global distribution, trends, and impacts of surface level ozone (Gaudel et al., 2018). Currently, the second phase of the Tropospheric Ozone Assessment Report (TOAR-II) aims to include additional ground-based stations, especially new networks in China and India. However, despite significant progress, there remain large regions with limited ground-based monitoring, and a gap in understanding ground-level ozone variations over time and space. To bridge gaps in regions lacking ozone monitors, various methods, including chemical reanalysis based long-term data assimilation, machine learning, and geostatistical methods have been employed. Chemical reanalysis is an approach that integrates observations from various sources including satellites using data assimilation and chemical transport models (CTMs)

to reconstruct historical atmospheric chemical composition and understand long-term changes and trends in air quality and climate forcing (Miyazaki et al., 2020b). Tropospheric ozone records have been provided in recent chemical reanalyses including the Tropospheric Chemistry Reanalysis Version 2 (TCR-2, (Miyazaki et al., 2020b)), the Copernicus Atmosphere Monitoring Service (CAMS, (Inness et al., 2019)), and data assimilation using the GEOS-Chem adjoint model (GEOS-Chem, (Qu et al., 2020b)). In addition, two machine learning estimates of global ground-level ozone have been produced to date: one using a space-time Bayesian neural network trained on TOAR observations and CMIP6 simulations (Sun et al., 2022), and another with a cluster-enhanced ensemble learning method that utilizes various data sources (Liu et al., 2022). Finally, geostatistical methods were applied by DeLang et al. who used Bayesian Maximum Entropy (BME) to estimate ozone through a data fusion of TOAR observations and output from multiple CTMs (Delang et al., 2021). This approach was further enhanced by incorporating the Regionalized Air Quality Model Performance (RAMP) framework to correct model biases (Becker et al., 2023). These estimates of global ozone distributions and trends have supported analyses of health impacts. For example, ozone estimates of DeLang et al. (2021) were used in both the GBD 2021 study (Murray et al., 2020), and in a study of ozone health effects in urban areas globally (Malashock et al., 2022). However, there remains a lack of knowledge regarding the consistency of ground-level ozone estimates, distributions, and long-term trends across these global ozone mapping products.

Inconsistencies in these datasets could significantly impact public health research, especially in assessing the risks of ozone-related health impacts, and may impede the development of effective environmental policies and ozone management strategies (Post et al., 2012). Although each dataset incorporates a considerable amount of observational information and model simulations through various methodologies, each inherently incorporates biases from these input data sources during the fusion processes. While satellite measurements of precursor species can be used to constrain surface and lower tropospheric ozone in chemical reanalysis (Miyazaki et al., 2012), the performance of chemical reanalysis surface ozone is limited in part by the low sensitivities of satellite ozone measurements near the surface, as well as model simulation errors. Data fusion methods integrate outputs from multiple models with inherent biases, potentially propagating these biases to the final estimates (Delang et al., 2021). Furthermore, machine learning methods trained on observation data may yield inaccuracies in rural and remote areas due to the uneven distribution of ground-level ozone monitoring stations (Liu et al., 2022; Betancourt et al., 2022). Therefore, conducting comparisons and evaluations of various types of ground-level ozone mapping products is essential to understand the inconsistencies and biases in these datasets, ultimately benefiting global health studies.

This study aims to compare ground-level ozone concentrations estimated by six datasets, and to evaluate their accuracy over the 2006-2016 period, with a particular emphasis on their capacity to represent long-term ozone trends across different regions. The comparison and evaluation include three chemical reanalysis datasets, two machine-learning datasets, and one geostatistical dataset. The period 2006-2016 is chosen as the period over which the six datasets all produce ozone estimates. The ozone seasonal daily maximum 8-hour average mixing ratio (OSDMA8) was selected as the health-relevant metric for annual ozone evaluation (Turner et al., 2016). Our study specifically utilizes the OSDMA8 metric because we focus on

evaluating long-term ozone exposure, an aspect not comprehensively compared previously among global ozone mapping products. We employed a comprehensive set of indicators to assess the congruence between these datasets, globally and regionally, including for long-term population weighted ozone outdoor exposure. Relative to the latest TOAR-II observational dataset, this study also examines the six datasets' ability to estimate ground-level ozone concentrations across various regions for the years 2006-2016. This research endeavors to characterize differences among ground-level ozone datasets, including discrepancies in ozone estimates, distributions, and trends, that could hinder evaluation of ozone's effects on health and agriculture, as well as impede the formulation of effective environmental policies. Although the primary focus of this study is on health impacts, the results are also largely applicable to agricultural and ecosystem impacts.

2. Data

As shown in Table 1, this study compares and evaluates ground-level ozone estimates from six global ozone mapping products in three categories. We utilized ozone seasonal daily maximum 8-hour average mixing ratio (OSDMA8) as the yearly ozone metric across all datasets. OSDMA8 is defined here as the maximum of the six-month running monthly mean daily maximum 8-hr ozone (DMA8) from January of the current year wrapping to March of the following year (Delang et al., 2021). OSDMA8 is GBD's ozone metric for quantifying health effect from long-term ozone exposure (Brauer et al., 2024), and it is the metric used in the World Health Organization's air quality guidelines, with values of 30 ppb for the guideline and 50 ppb for the interim target (World-Health-Organization, 2021). All observations and model estimates are converted to OSDMA8 using the same algorithm. Details on the input data used to construct each dataset are available in the Supporting Information (SI).

2.1 Geostatistical ozone dataset

The BME dataset uses geostatistical methods to provide high-resolution global ground-level ozone estimates. First, M^3Fusion (Measurement and Multi-Model Fusion) is a statistical method developed to improve estimates of global surface ozone distributions by integrating observational data from TOAR and outputs from multiple chemistry models. Specifically, the method assigns weights to multiple global atmospheric chemistry models based on their regional accuracy compared to observed ozone values (Chang et al., 2019), creating a composite of multiple global atmospheric chemistry models by weights. The details of input data can be found in Table S1. Then BME data fusion integrates this multi-model composite with observations in space and time, and finally BME estimates are refined from 0.5° × 0.5° to 0.1° × 0.1° (Delang et al., 2021). The observations are from TOAR-I for 1990 to 2017, complemented by data from the Chinese National Environmental Monitoring Center (CNEMC) for 2013 to 2017. The latest version of this dataset employs RAMP for bias correction of M^3Fusion inputs (Becker et al., 2023). The BME ozone estimates are more accurate than the average outputs from multiple models, achieving an R^2 of 0.63 at 0.1° × 0.1° resolution, as evaluated against observations through cross-validation (Delang et al., 2021). Furthermore, incorporating RAMP into the BME process significantly improves R^2 by 0.15, especially in areas far from monitoring stations, as demonstrated through checkerboard cross-validation (Becker et al., 2023).

2.2 Machine learning ozone datasets

We utilized two machine learning global ground-level ozone datasets from the University of Cambridge, and Nanjing University. The University of Cambridge's machine learning (UKML) dataset was developed using a space-time Bayesian neural network, fusing various data sources including historical observations, CMIP6 multi-model simulations (AerChemMIP historical simulations and ScenarioMIP projections), population distributions, land cover properties, and emission inventories (Sun et al., 2022) (input data summarized in Table ~~S3~~S2). The UKML model categorized TOAR-I monthly ozone observations from 1990 to 2014 into urban and rural areas, and used these as labels for supervised learning. This model generates monthly global gridded ozone estimates from 1990 to 2019, downscaled to a 0.125° × 0.125° spatial resolution. It exhibited great performance in predicting urban and rural surface ozone concentrations, with R^2 values ranging from 0.89 to 0.97 and RMSE values between 1.97 and 3.42 ppb (Sun et al., 2022).

Nanjing University's machine learning (NJML) dataset was created using a cluster-enhanced ensemble machine learning method. This dataset integrates various data sources, including satellite observations, atmospheric reanalysis, land cover properties, emission inventories and meteorological features (Liu et al., 2022). The main input data for NJML include meteorological parameters from ERA5, atmospheric chemistry from the CAMS chemical reanalysis, aerosol concentrations from MERRA-2, satellite observations from OMI/Aura, and emissions data from CEDS, spanning 2003-2019 with varying spatial resolutions (input data summarized in Table ~~S2~~S3). It utilizes the monthly mean of daily maximum 8 h average (DMA8) data from TOAR-I and CNEMC observations from 2003–2019 as training data. The NJML dataset produces monthly global gridded ozone estimates from 2003 to 2019 with a 0.5° × 0.5° spatial resolution. The model demonstrates robust performance in both spatial and temporal predictions of ground-level ozone, with R^2 values of 0.909 and 0.925, respectively (Liu et al., 2022).

2.3 Chemical reanalysis products

We utilized surface ozone analysis fields from three chemical reanalysis products: the Tropospheric Chemistry Reanalysis Version 2 (TCR-2, (Miyazaki et al., 2020b)), the Copernicus Atmosphere Monitoring Service reanalysis (CAMS, (Inness et al., 2019)), and the GEOS-Chem reanalysis (GEOS, (Qu et al., 2020b)). Different from the machine learning and geostatistical ozone datasets, the chemical reanalysis products utilized satellite observations of atmospheric composition to produce three-dimensional profiles of atmospheric composition. In situ surface observations were not included in the global chemical reanalysis data assimilation. Because of the lack of direct observational constraints, challenges remain in estimating surface ozone in the current reanalysis products (Huijnen et al., 2020). Detailed comparisons of these reanalyses for ozone over the entire troposphere at finer timescales have been conducted by the TOAR-II chemical reanalysis working group (Sekiya et al., 2024; Jones et al., 2024; Miyazaki et al., 2024), but without a focus on the ground level and long-term metric as analyzed here.

TCR-2 was generated by assimilating multiple satellite observations into the MIROC-Chem model, that was developed as a part of the multi-model multi-constituent data assimilation (Miyazaki et al., 2020a). The meteorological fields were nudged to the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim Reanalysis meteorology. The data assimilation employed is an ensemble Kalman filter technique, which was used to effectively correct the emissions and concentrations of various chemical species (Miyazaki et al., 2020b). The assimilated data ~~includes~~include ozone, CO, NO_2, HNO_3 and SO_2 from satellite instruments such as OMI, MLS, GOME-2, SCIAMACHY and MOPITT over the period from 2005 to 2021 (input satellite data summarized in Table ~~S6~~S4). TCR-2 provides 2-hourly global ozone profiles at a 1.1° × 1.1° spatial resolution, with the regional mean ozone bias against global ozonesonde measurements ranging from -0.4 to 4.2 ppb in the lower troposphere (850-500 hPa) (Miyazaki et al., 2020b).

CAMS, operated by the European Centre for Medium-Range Weather Forecasts (ECMWF) on behalf of the European Commission, provides the global reanalysis dataset on atmospheric composition developed by ECMWF. The main inputs for the CAMS ECMWF Atmospheric Composition Reanalysis 4 (EAC4) chemical reanalysis are retrievals of CO, ozone, NO_2 and aerosol optical depth (AOD) from multiple satellite instruments including MLS, OMI, GOME-2, SCIAMACHY, MIPAS, SBUV/2 and MOPITT, covering various periods ranging from 2003 (input satellite data summarized in Table ~~S4~~S5). CAMS employed the four-dimensional variational data assimilation (4D-Var) method to integrate the satellite measurements under ECMWF's Integrated Forecasting System (IFS) CB05 model (Inness et al., 2019). It provides 3-hourly global profiles of ozone and other species at a 0.75° × 0.75° spatial resolution. While CAMS generally improves over previous analyses, challenges and biases remain, particularly at high latitudes and in accurately capturing seasonal variations (Inness et al., 2019).

The GEOS-Chem dataset is developed through 4D-Var data assimilation of NO_2 column densities using the GEOS-Chem adjoint model that includes updates in stratospheric and halogen chemistry (Henze et al., 2007). The GEOS-Chem model is driven by the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) meteorological fields from the NASA Global Modeling and Assimilation Office (GMAO). Prior anthropogenic emissions of NO_x, SO_2, NH_3, CO, NMVOCs (non-methane volatile organic compounds), and primary aerosols were obtained from the HTAP 2010 inventory version 2 (Janssens-Maenhout et al., 2015) (input data summarized in Table ~~S5~~S6). Operating at a 2° × 2.5° resolution, the assimilation estimates global ozone more accurately than the forward model from 2006 to 2016 by deriving emissions of NO_2 through inverse modelling. The GEOS-Chem dataset exhibits a small bias across all ozone metrics, and among metrics it has the best spatial consistency for DMA8 (R^2 = 0.88) (Qu et al., 2020b). However, the model has limitations in accurately capturing regional variations and seasonal trends in ozone concentrations.

2.4 Ground-level ozone observations

For the evaluation in this project, we utilized both urban and non-urban ground-level ozone observations for the yearly OSDMA8 metric from the updated TOAR-II dataset, covering 2006 to 2016 (Schröder et al., 2021). This dataset represents

the most extensive collection of tropospheric ozone measurements available globally. Compared to TOAR-I (Schultz et al., 2017), TOAR-II incorporates an expanded dataset of ozone observations, notably including monitoring data from approximately 1,400 stations across China for the years 2015 to 2016 that are included in TOAR-II (https://toar-data.fz-juelich.de/gui/v2/dashboard/, last access: 15 November 2024). We require that at least 75% of the days in a month must have valid DMA8 values for that month to be included in the annual data calculations. The total number of observation sites used in our assessment varied from a minimum of 3715 in 2006 to a maximum of 7013 in 2016. Given that three ozone products in this study utilize the TOAR-I dataset for training or input, evaluations using the latest TOAR-II dataset for sites not included in TOAR-I can provide more objective results. Figure S1 illustrates the spatial distribution of TOAR-II monitoring stations in 2016. The version of the TOAR-II database employed in this analysis, as of ~~November 2024, may not represent its final version~~May 2025.

2.5 Population data

We analyzed ozone population exposures for each dataset using the globally gridded population data for the year 2019 from the Global Burden of Disease (GBD) 2019, which has a resolution of 0.1° × 0.1° (Lloyd et al., 2019). Since we use the same gridded population data for all years of the project, we focus on differences in exposure attributable to changes in ozone levels rather than changes in population. Therefore, population-weighted ozone over 2006 to 2016 can be biased even if the ozone data are unbiased.

3. Methodology

3.~~3~~1 Evaluation with ground-level observation

Previous research ~~has adopted a~~created 1º×1º grid-cell-averaged hourly ozone data from ~~TOAR~~ surface observations to evaluate global chemistry model performance over North America and Europe, which is suitable for analyzing extremes and validating seasonal and diel ozone cycles (Schnell and Prather, 2017; Schnell et al., 2015). We utilized OSDMA8 from TOAR-II observations covering 2006 to 2016 to evaluate the six datasets. During the evaluation process, we retained the original resolution of the six datasets (Table 1). ~~We~~

Considering that the six datasets have different resolutions and are designed for different applications, we adopted a dual evaluation strategy to provide a comprehensive assessment of their performance. The first method is a grid-to-grid evaluation. Similar to the approach of Schnell et al. (2015), we re-gridded TOAR-II observations to a 0.1° x 0.1° resolution by an inverse distance weighted method and then aggregated them to match the native resolution of each of the six datasets. In this approach, the sample size for each evaluation varies reflecting the varying resolution of the datasets; for 2016, BME had 173,718 grid cell pairs, NJML had 7,099, UKML had 162,419, CAMS had 4,614, GEOS-Chem had 782, and TCR-2 had 2,195. We also adopted the grid-to-grid evaluation method for regional evaluations, as it provides better spatial representativeness over large

areas. To quantify the uncertainty of the six datasets' estimates, we determined the lower and upper bounds (95% confidence interval), derived from the grid-to-grid regression analysis performed between the TOAR-II observations and each of the six datasets at their native resolutions.

The second method is a standard grid-to-point evaluation ~~approach, where the data from each TOAR-II~~. This approach ensures a consistent sample size across all datasets by comparing each dataset's estimate at the grid cell containing an observation ~~site was matched with a corresponding grid cell in each dataset.~~location. For grid cells ~~with~~containing a TOAR-II ~~observation~~site but no valid estimate ~~in a dataset~~ (NA value), we used the nearest valid estimate instead. ~~Table S12 displays the number of~~ This method captures a penalty for missing ~~values in each dataset in 2016 at TOAR-II locations, showing that~~ data and coarse resolution, only BME, NJML, and UKML ~~have~~had a small number of missing estimates~~. This method ensures the same sample sizes for evaluation across all datasets while accommodating their varied resolutions and avoiding the challenge of gridding TOAR-II observations. Our~~ at TOAR-II locations. The grid-to-point ~~evaluation approach~~method was used to evaluate model bias, as it ensures a consistent sample size ~~and captures penalties for missing data in~~across all datasets~~. We~~ when performing evaluations on different quantiles of the TOAR-II observations. For both methods, we assessed the performance of each dataset using the coefficient of determination (R^2) between ozone estimates and observations, and root mean square error (RMSE) as the primary metrics. We selected the 50 ppb as the threshold for high ozone concentration because it corresponds to the long-term air quality interim target of WHO. ~~These performance metrics should be interpreted considering the spatial representativeness uncertainty that is caused by the grid-to-point evaluation approach.~~

3.2 Pairwise spatial similarity comparison

Before comparing concentration estimates between datasets, we converted all ozone estimates from each dataset to OSDMA8, ensuring only one ozone estimate value per year for each grid cell (see the original temporal resolution in Table 1). The OSDMA8 metric is used for long-term ozone exposure given its utility and wide acceptance in health impact studies, despite the inherent loss of shorter temporal dynamics. We employed two quantitative metrics to classify how the datasets relate with one another: the Pearson correlation coefficient (R) and the root mean square difference (RMSD). The pairwise correlation R indicates the similarity in geographical distribution of ozone concentrations, and the RMSD quantifies the difference in ozone estimates between datasets. A higher R value suggests greater similarity in the spatial pattern between two datasets and a smaller RMSD indicates a less significant discrepancy in ozone concentration estimates between two datasets. We then group the six datasets, adopting a method that maximizes the difference between the correlation R within and outside the groups. The idea of this grouping is to distinguish the spatial similarity between the datasets, which is based on the pairwise correlation. For each grouping combination, 4 variables are computed: the sum of pairwise correlations within groups (C_i), the sum of pairwise correlations outside the groups (C_o), the number of dataset pairs within groups (N_i), and the number of dataset pairs outside the groups (N_o). The objective is to ascertain the grouping combination that maximizes the difference between C_i/N_i and C_o/N_o. More details of the calculation can be found in Text S1.

3.~~1~~3 Long-term exposure comparison

~~Before comparing concentration estimates between datasets, we converted all ozone estimates from each dataset to OSDMA8, ensuring only one ozone estimate value per year for each grid cell (see the original temporal resolution in Table 1). The OSDMA8 metric is used for long-term ozone exposure given its utility and wide acceptance in health impact studies, despite the inherent loss of shorter temporal dynamics.~~ Subsequently, we re-gridded all datasets and TOAR-II observations to 0.1° × 0.1° resolution to facilitate comparison at the same spatial scale. During re-gridding, we ensure that the average value of the finer grid cells matches that of the original coarse grid cell; for example, if a grid cell has a value of 30 ppb, then after re-gridding to finer grid cells, the average value of these grid cells will still be 30 ppb. Data over the ocean were excluded, retaining only land and populated islands for analysis. We calculated the yearly ozone trend using 50% quantile regression for each dataset using both population-weighted and area-weighted approaches, with details of the calculation methods provided in Text S2. In this study, the trend is interpreted from the slope of the quantile regression, and confidence in the trend is determined by its p-value: $p \leq 0.01$ is considered very high certainty; $0.01 < p \leq 0.05$, high certainty; $0.05 < p \leq 0.1$, medium certainty; $0.1 < p \leq 0.33$, low certainty; and $p > 0.33$, no evidence. We also regressed population-weighted mean ozone concentrations in different world regions of each dataset against the year to evaluate ozone long-term variations. For each grid cell we calculated the mean and standard deviation of the six OSDMA8 values obtained from each dataset to highlight regional differences and similarities. We also calculated the deviation from the ensemble mean for each dataset to assess geographic distribution variations.

Furthermore, we compared ozone exposure differences in various regions for each dataset to evaluate the potential for health impacts. Here we estimate exposure as the ambient concentration in 0.1° × 0.1° grid cells related to population at their residences, not including other factors that affect human exposure such as time-activity patterns. ~~Details of these calculations are available with parameters in the SI.~~To quantify the uncertainty in our exposure analysis, we established lower and upper bounds for all population exposure and share of population estimates. The OSDMA8 95% confidence interval (CI) for each dataset is determined through a grid-to-grid linear regression between each dataset and the re-gridded TOAR-II observations based on 0.1° × 0.1° grid cells. We use regional groupings defined by HTAP2 (Koffi et al., 2016), as detailed in the Table S7.

~~5~~4. Evaluation against TOAR-II observations

~~5~~4.1 Evaluation of ground-level ozone in 2016

We conducted regression and bias analyses for each dataset in comparison with TOAR-II observations for each year from 2006 to 2016. Fig. ~~7~~1(a) and Fig. 1(c) illustrates the scatterplot from the linear regression analysis of each dataset against the 7013 TOAR-II observations in 2016, accompanied by a density core that visualizes the data point distribution. The year 2016

is presented here because it has the highest number of TOAR-II observations from 2006 to 2016, and other years can be found in Figure ~~S11~~S2 and S3. For 2016, BME outperforms other datasets in both evaluation method, with the highest R^2 (0.75 for grid-to-grid, 0.63 for grid-to-point) and lowest RMSE (4.25 ppb for grid-to-grid, 5.28 ppb for grid-to-point), its density ~~core~~cores intersecting the y=x line. BME has an advantage in that its methods should nearly match the observed values for locations used as inputs to the data fusion. Consequently, we conduct another validation for TOAR-II sites not used as input for BME in 2016 (Figure ~~S13~~S4). After excluding all sites located at observation points previously used as BME input, using 3911 observations for validation, BME performs well compared to another datasets, though its R^2 decreases significantly to 0.65 for grid-to-grid and 0.53~~.~~ for grid-to-point. In Fig. ~~7~~1(a), all three chemical reanalysis datasets exhibit a moderate R^2 ranging from 0.51 to 0.60 for grid-to-grid and 0.35 to 0.41 for grid-to-point, comparable to the performance of the machine learning datasets, which have R^2 values of 0.50 and 0.56 for grid-to-grid, 0.37 and 0.38~~.~~ for grid-to-point. Among these five datasets, CAMS has the lowest RMSE (~~7~~6.00 ppb for grid-to-grid and 7.59 ppb for grid to point), which is better than other chemistry reanalysis products but relatively low R^2 (0.~~35~~).51 for grid-to-grid and 0.35 for grid-to-point). Its density ~~core~~cores slightly below the y=x line suggests CAMS estimates are marginally lower than TOAR-II observations. GEOS-Chem and TCR-2 demonstrate adequate performance, albeit with higher RMSE values of 8.47 ppb and 10.26 ppb for grid-to-grid, 10.27 ppb and 13.23 ppb for grid-to-point, respectively. Their density cores positioned above the y=x line indicate that these models tend to produce higher estimates compared to the TOAR-II observations. NJML~~, despite differing geographic distributions from other datasets (Fig. 5),~~ shows acceptable performance with higher R^2 (0.~~38~~56 for grid-to-grid and 0.38 for grid-to-point) than CAMS and lower RMSE (6.37 ppb for grid-to-grid and 8.63 ppb for grid-to-point) than TCR-2. UKML exhibits the highest RMSE of 12.22 ppb for grid-to-grid and 13.49 ppb for grid-to-point, and its density ~~core~~cores region ~~is~~are above the y=x dashed line, indicating an overestimation. This is because the UKML algorithm emphasizes higher ozone pollution levels in rural and remote areas compared to adjacent urban districts, which consequently leads to an overestimation especially in population-weighted metrics (Sun et al., 2024).

Fig. ~~7~~1(b) and Fig. 1(d) focuses only on TOAR-II grid cells or sites with OSDMA8 value above 50 ppb, showing that R^2 is reduced compared to the comparison of all ozone measurements (Fig. ~~7~~1(a) and Fig. 1(c)) for all six datasets, suggesting overall weaker agreement between modeled and observed ozone distributions at higher concentrations. All six datasets show decreasing performance from BME, NJML, and UKML to TCR-2, GEOS-Chem, and CAMS, with R^2 of 0.35, 0.33, 0.29, 0.25, 0.08, and 0.04 for grid-to-grid; 0.37, 0.30, 0.26, 0.25, 0.17, and 0.07 for grid-to-point, respectively. However, the change of biases varies among datasets at higher concentrations. Specifically, overestimation is reduced in the UKML, NJML, GEOS-Chem, and TCR-2 datasets when observations exceed 50 ppb~~.~~ in both evaluation methods. Conversely, we observe increased underestimation in the BME and CAMS datasets at ozone levels above 50 ppb. This proportional bias is consistent with the linear regression slope, which is less than 1 for all six datasets in Fig. 1. Fig. ~~Fig. 8~~2 shows the normalized mean bias for stratified concentration intervals in 2016, which provides insights into the average discrepancy between estimates and TOAR-II observations across ozone concentration ranges. All six datasets overestimate TOAR-II observations below the 40%

concentration interval. Only BME underestimates above the 40% concentration level, CAMS underestimates above the 80% concentration interval, and NJML underestimates above 90% concentration interval, aligning with the ~~observations~~density kernel presented in Fig. 41. BME demonstrates the smallest mean bias, particularly below the 50% concentration level and CAMS shows the smallest mean bias in the 50% to 90% concentration interval. In the 90% to 100% concentration interval, NJML and GEOS-Chem have the smallest mean bias. In summary, BME and CAMS perform better overall in terms of normalized mean bias, with other models tending to overestimate ozone at almost all concentrations. Detailed plots of normalized mean bias for stratified concentration intervals for each year from 2006 to 2015 are shown in Figure S5.

~~5~~4.2 Evaluation of ground-level ozone in different countries or regions

Fig. 3 presents the distribution of population exposure calculated from six datasets and the gridded TOAR-II observations in three world regions with a high density of observations, for 2016. Here we calculate the population-weighted kernel density for population exposure to OSDMA8 concentrations, based on the 0.1° x 0.1° resolution for each region, only for grid cells where the re-gridded TOAR-II data have a value. Corresponding plots for other years (2006 to 2015) are shown in Figure S6. Overall, the datasets are widely distributed, and the estimated exposure peaks vary. In East Asia (EAS), the population is exposed to high ozone concentrations. The concentration distribution is broad and has multiple peaks from TOAR-II observations, indicating a complex pollution environment, with a large population exposed to concentrations frequently exceeding 50 ppb, even 70 ppb. BME and NJML show a similar distribution as TOAR-II. Significant differences exist between UKML, CAMS and GEOS-Chem with the TOAR-II data for EAS. In Europe (EUR), exposure is concentrated between 40 and 50 ppb, indicating a more moderate and uniform exposure. The BME and CAMS have the best fit with the TOAR-II. NJML, UKML, GEOS-Chem, and TCR-2 show a peak at a higher ozone concentration range of 50–60 ppb. In North America (NAM), exposure peaks sharply in the 40 to 50 ppb range, which is slightly higher and more concentrated than in Europe. The NJML dataset agrees best with the shape of the TOAR-II distribution, and GEOS-Chem and BME capture the overall shape of the major exposure peaks well.

Table 42 presents the validation results for different countries or regions using re-gridded TOAR-II observations at each dataset's native resolution in 2016, focusing on the countries with the highest number of sites. Here we use ~~the~~ R^2 to assess the strength of the spatial correlation and RMSE to measure the bias across each country or region. The performance of each dataset varies by region, indicating that a dataset's overall performance does not guarantee its effectiveness in all regions. Reasonable R^2 and RMSE values are seen across all 6 datasets in the United States; BME leads with the highest R^2 (0.~~71~~75) and lowest RMSE (~~4.12~~3.48 ppb), and TCR-2 has the lowest R^2 (0.~~23~~43) with highest RMSE (~~10.58~~9.43 ppb). In Japan, BME leads with an RMSE of 4.~~59~~29 ppb, followed by CAMS at 4.~~95~~33 ppb, and UKML has the highest RMSE (~~18.25~~17.41 ppb). ~~Although there are over 1000 monitors in Japan, all datasets show poor R^2 values below 0.1. The six~~ The datasets also perform poorly in South Korea, where ~~TCR-2~~GEOS-Chem has the highest RMSE (~~18.53~~14.71 ppb~~), BME~~) and NJML has the lowest RMSE (~~7.33~~2.68 ppb). Although Japan and South Korea have a dense network of monitors, nearly all datasets show a weak correlation with observations, with R^2 below 0.2. Only the GEOS-Chem dataset has the highest R^2

value of 0.37 in Japan and 0.81 in South Korea, this result should be interpreted with caution, as the evaluation includes fewer than 30 grid-to-grid pairs. The performance of datasets within China exhibits significant variability, where BME and NJML demonstrate relatively good performance, and CAMS exhibits poor performance for R^2, while for RMSE, CAMS performs better than GEOS-Chem, TCR-2 and UKML. For other countries, which serve as a test of model performance in areas with sparse observations, nearly all datasets exhibit better R^2 ~~and RMSE~~ values than in South Korea and Japan, with TCR-2~~, NJML~~ and ~~BME~~NJML demonstrating particularly better performance than others. Overall, BME demonstrates strong performance in most countries, particularly in the United States, where it achieves the highest R^2 and the lowest RMSE, suggesting both strong spatial correlation with TOAR-II observations and high accuracy. NJML exhibits mixed performance, with relatively high R^2 values indicating good correlation in the United States and China, but it falls short in EU-27 ~~and Canada~~ with high RMSE ~~and low R^2~~. UKML presents consistently high RMSE values across countries suggesting high bias. CAMS displays variable performance with low R^2 values in China, indicating a lack of spatial correlation, yet its RMSE values are relatively small across all regions when compared to other chemical reanalysis datasets. Compared to CAMS, GEOS-Chem and TCR-2 exhibit ~~reasonable~~better spatial correlations in Europe, the United States, China, and Canada. ~~Notably, they outperform all other datasets in Canada, except for BME. TCR-2 demonstrates the best R^2 performance in other countries with less monitoring data.~~ However, TCR-2 also presents high RMSE values across all regions. ~~All six~~ Five datasets except GEOS-Chem exhibit lower spatial correlation compared to TOAR-II observations in countries with high monitoring density, such as Japan and South Korea, than in countries with lower monitoring densities. NJML, UKML, GEOS-Chem and TCR-2 show overestimates compared to the TOAR observations in every country in the Table ~~4~~2. Extending the analysis to the period from 2006 to 2016 (see tables in Table ~~S10~~S8), the percentage of underestimates from 6 datasets compared to TOAR observations in all countries is below 20%.

~~5~~4.3 Evaluation of ground-level ozone across different years

Fig. ~~9~~4 presents time series plots of R^2 and RMSE from ~~the evaluation~~grid-to-grid and grid-to-point evaluations of each ~~database~~dataset against TOAR-II observations from 2006 to 2016. It is important to note that the years 2015 and 2016 include observations from China. In Fig. 4(a) and Fig. 4(c) BME consistently shows the largest R^2, indicating its robust performance near the monitor locations due to the utilization of observational data as input and the effective exploitation of spatiotemporal autocorrelation among stations. Apart from BME, for both evaluation scenarios NJML outperforms other datasets in R^2 from 2010 to ~~2015,~~2014, and TCR-2 leads in 2007 and 2016~~, while UKML does so in 2008 and 2009. Five~~. In grid-to-point evaluation, five datasets, excluding NJML, demonstrate a drop in R^2 in 2010~~. All~~, and all datasets show an increase in R^2 from 2015 to 2016. In grid-to-grid evaluation, GEOS-Chem shows an overall better performance in R^2 than CAMS, TCR-2 and UKML. For both scenarios, BME maintains the lowest RMSE throughout the period, indicating the most accurate predictions. CAMS also performs well in terms of RMSE. From 2006 to 2013, GEOS-Chem consistently has lower RMSE than both TCR-2 and UKML. Meanwhile, NJML exhibits a decreasing RMSE ~~trend~~ from 2006 to 2016. The clear differences in time series of RMSE correspond with the yearly mean trends in Fig. ~~1~~5. Datasets with lower ozone values, BME and CAMS, also exhibit

lower RMSE, whereas those with higher estimates, specifically TCR-2 and UKML, have higher RMSE. From 2006 to 2016, the performance rankings derived from R^2 values varied significantly between the two evaluation scenarios, whereas the RMSE based rankings were nearly consistent.

45. Comparison between ozone mapping products

45.1 Temporal trends

Both the area-weighted and population-weighted mean trends of global OSDMA8 reveal substantial differences among global ozone mapping datasets (Fig. 15). Notably, BME and CAMS have lower ozone values than other datasets, for both metrics, while UKML and NJML have higher ozone estimates, with differences between these datasets exceeding 5 ppb. The higher values in GEOS-Chem and TCR-2 may be attributed to the remaining high bias in the forecast models, which is commonly found in CTMs (Travis and Jacob, 2019). The population-weighted mean is higher than the area-weighted mean, by 5-10 ppb across all datasets, and for UKML and BME, the disparity between population-weighted and area-weighted ozone concentrations appears to widen over time. The faster increase in the population-weighted mean compared to the area-weighted mean appears to be driven by rising ozone levels in highly populated regions. In Table 23, focusing on the period from 2006 to 2016, we find that NJML iswas the only dataset showingto exhibit a downward trend inwith very high certainty for both area-weighted and population-weighted mean ozone concentrations, with very high certainty.. In contrast, TCR-2 and UKML only show increasing trends in population-weighted mean ozone during this period with very high certainty. However, while the BME dataset shows a negative slope for the area-weighted mean, this downward trend has only low certainty; for the population-weighted mean, there is no evidence of a decreasing trend. Fig. Fig. 26 illustrates regional ozone changes per decade, weighted by population, across different regions in each dataset over 2006 to 2016. NJML, despite its overall decreasing trend in Table 23, does not uniformly show declines across all regions. The decrease in NJML is predominantly in North America, notably over 8 ppb per decade in the US and Canada, while Sub Saharan Africa and South America exhibit increases. BME and UKML, with the longest duration, both display decreasing trends in North America, and Europe, and increases in Southeast Asia and Middle East. Both datasets indicate greater decreases in North America than in Europe and more significant increases in the Middle East than in Southeast Asia. However, BME shows a downward trend in East Asia, while UKML exhibits the reverse. CAMS and TCR-2's trends in Fig. 26 are less distinct, except for the decrease in North America and the increase in East Asia, mirroring those of GEOS-Chem, which exhibits the smallest decadal ozone change, likely due to not directly assimilating ozone from satellite observations. From Table S11S9, we observe that some regions exhibit a clearer trend from 2006 to 2016, with very high certainty across six datasets. In East Asia, BME and NJML observe decreasing trends, whereas the other 4 datasets display increasing trends. In North America, all datasets display a downward trend, and in Europe, BME, NJML, UKML and TCR-2 show a decline, contrasting with increases in CAMS and GEOS-chem. Recent analyses using TOAR observations indicate that from 2006 to 2016, most sites in North America experienced decreasing ozone, while many sites in East Asia exhibited significant positive trends (Chang et al., 2024; Fleming et al., 2018;

Chang et al., 2017). These observed trends in North America, Europe and East Asia seem to agree best with the trends estimated by BME and UKML. Detailed plots of population weighted and area weighted trends for each dataset in each region are shown in Figures S7 and S8.

45.2 Difference maps

Fig. 37 shows the spatial maps of the 11-year (2006-2016) average of the annual multi-model means of OSDMA8 from the six datasets, and the associated standard deviations. India, China, and the Middle East are estimated to have the world's highest average ozone concentrations, exceeding 50 ppb in the multi-model average. High ozone levels are also found in parts of Europe and the eastern United States. Notably, regions in southern Africa near the Atlantic Ocean emerge as primary areas of ozone pollution, where some locations have average concentrations exceeding 60 ppb. Conversely, the Amazon Basin in South America, Central Africa, and Canada exhibit relatively lower ozone concentrations, with some areas below the WHO 30 ppb guideline. The six datasets show greater variation (high standard deviations above 10 ppb) in South America and Africa, particularly in rainforest regions, compared to North America and Europe, notably since these regions lack ozone monitors. The eastern coast of China also exhibits significant discrepancies with standard deviations above 15 ppb. Detailed plots of the annual multi-model means of OSDMA8 from the six datasets, and the associated standard deviations for each year (2006 to 2016) are shown in Figures S9 and S10. Fig. 48 compares the mean ozone concentration for each dataset with the multi-dataset average (Fig. 37(a)), showing wide variation in the magnitude and spatial distributions of ozone concentrations among the datasets. BME and CAMS display lower values than the average of six datasets in most regions, consistent with Fig. 1 and Fig. 5. BME records concentrations higher than average in central South America and central Africa near the Atlantic, while CAMS shows elevated levels in Southeast Asia and along the Middle East coast, contrasting TCR-2's lower coastal and higher inland concentrations. NJML and UKML report above-average values, except for NJML in southern China and UKML near the Sahara Desert and the Indian Ocean. Detailed plots of difference between annual ensemble mean and each dataset estimate for each year (2006 to 2016) are shown in Figure S11.

45.3 Pairwise spatial similarity

We calculated the correlation and RMSD between each pair of datasets for each year from 2006 to 2016. Fig. 59 displays the average correlation and RMSD values over these 11 years as heatmaps. Fig. 59(c) presents a scatter plot of the correlations and RMSD for each dataset pair. Using the correlation heatmap (Fig. 59(a)), we categorized the six datasets by the maximum difference method, identifying NJML as a distinct group (Group B) and the other five datasets as Group A. NJML's separation indicates its significant divergence in ozone geographic distribution compared to others. The scatter distribution in Fig. 59(c) reveals that most Group A data points cluster in regions of high correlation and low RMSD, suggesting broadly consistent ozone geographic distribution and concentration ~~estimation~~estimates within this group. Nevertheless, there is still substantial disagreement among the ~~current~~ reanalysis products, likely because of the differences in forecast model performance and data assimilation configuration. Conversely, Group B has lower correlations. Interestingly, RMSD does not consistently decrease

with increasing correlation, indicating that similar geographic distribution patterns can still yield significant differences in ozone concentration estimates. This is particularly evident with CAMS and GEOS-Chem, which exhibit the highest correlation with a large RMSD, suggesting substantial differences in ozone estimation.

45.4 Long-term ozone exposure

Fig. 610 illustrates the distribution of population in various regions exposed to average OSDMA8 from 2006 to 2016, as per each dataset. Detailed plotsWe also calculated the distribution of population exposure for each year (regarding the lower and upper bounds of OSDMA8 from 2006 to 2016) are for each dataset, as shown in Figure S10. S12. For the period 2006-2016, a majority of the population in most datasets is exposed to concentrations above 40-50 ppb. Populations in regions such as East Asia and South Asia appear to be exposed to higher ozone concentrations in all datasets compared to other regions., which supports our findings from exposure based on TOAR-II observations in Fig. 3. Conversely, populations in the Sub-Saharan Africa and Southeast Asia regions typically experienced concentrations below 50 ppb. The different regions show different distributions of population ozone exposure, and comparisons between datasets reveal considerable variations in the ozone distribution for each region. Some datasets (e.g., CAMS and TCR-2) show a wider distribution of population across ozone concentrations compared to others (e.g., NJML). In BME and CAMS, after South Asia, a significant fraction of the population in the East Asia region is exposed to levels above 50 ppb, while this proportion in North America, Europe, and the Middle East is less than in the other four datasets. When focusing on exposure above 70 ppb, South Asia dominates in BME, CAMS, and NJML, while East Asia leads in GEOS-Chem, UKML, and TCR-2. All six datasets clearly demonstrate a higher impact of ozone pollution in Asia compared to North America and Europe, aligning with previous findings based on TOAR observations (Chang et al., 2017).

Table 35 elucidates each region's population share above different ozone concentration levels.30 ppb, 50 ppb and 70 ppb thresholds from 2006 to 2016. Results are presented as the estimate with the lower and upper bound in parentheses (e.g., 42% [24%, 66%]). Detailed table of population share for each year (2006 to 2016) are shown in Table S10. For BME and CAMS, the global average of the population exposed to more than 50 ppb is 42.5%% [24%, 66%] and 48.1%,% [18%, 76%], respectively, indicating that more than half of the population us exposed to lower concentrations. Regional exposure estimates vary in East Asia, where the proportion of the population exposed to more than 50 ppb ranges from 61% [28%, 94%] in BME to over 90% 99% [62%, 100%] in UKML, 95% [58%, 100%] in GEOS-Chem, and 94% [63%, 100%] in TCR-2. The differences are stark in Europe, with BME and CAMS showing only 16% [0%, 56%] and 9% [0%, 58%] exposure, respectively, over 50 ppb, while NJML, UKML, and TCR-2 report overmuch higher exposures of 76% [22%, 96%], 77% [2%, 100%], 70%.% [5%, 100%]. Focusing on the highest threshold, TCR-2 and UKML project notably higher exposures in East Asia, withthat 41% [0%, 79%] and 31% [13%, 85%] of the population in East Asia exposed to levels above 70 ppb, respectively. In the Middle East, TCR-2's estimates are significantly higher than other datasets, indicating that 38% [0%, 86%] of the population is exposed to average concentrations above 70 ppb. TheDespite these regional differences, the six datasets

agree that a large majority of the global population is exposed to ozone above the WHO guideline for OSDMA8 (30 ppb) with percents ranging from 93% [74%, 99%] (CAMS) to 99% [96%, 100%] (NJML).

6. Discussion

When evaluating datasets against TOAR-II observations, differences in performance are seen among six datasets. BME performed well in the TOAR-II evaluation (Fig. 1), with minimal mean bias below the 50% concentration threshold (Fig. 2). Unlike the other databases, BME tends not to overestimate over the range of concentration, with a small underestimation bias. After removing TOAR sites that were used as inputs to BME (Fig. S13), BME's performance remains robust in both evaluation scenarios. NJML and UKML, both utilizing TOAR-I as a training set, showed overestimation in most areas (Table 2). NJML exhibits a higher R^2 from 2010 onward, especially at high ground-level ozone concentrations (above 50 ppb), where prediction accuracy generally declines across all datasets. However, NJML has missing data in some coastal regions, particularly in European coastal countries, which may contribute to its elevated RMSE in Europe compared to other datasets (Table 2), since missing data are substituted with the nearest model grid cell. UKML's performance after 2010 is not as good as NJML and is worse than the chemical reanalysis datasets. CAMS, GEOS-Chem and TCR-2 primarily rely on satellite data, suggesting that they might not compare favorably with other datasets that used observations as input or training data. Despite this, the three chemical reanalysis datasets unexpectedly outperform the machine learning datasets in R^2 (TCR-2, GEOS-Chem) and in RMSE (CAMS) over the full year 2016. In addition, for chemical reanalysis datasets, there is a clear trade-off between capturing the spatial pattern and the accuracy. As shown in Fig. 2, TCR-2, GEOS-Chem all have widespread overestimation, but they often capture spatial patterns more effectively (higher R^2). Conversely, CAMS exhibits low bias in RMSE but shows worse spatial correlation in China. All six datasets show a reduced performance at higher ozone concentrations (>50 ppb), which may complicate their accuracy for assessing long term high-pollution exposure. Furthermore, most datasets perform better in regions with lower monitoring density (e.g., the United States and China) than in those with higher density (e.g., Japan and South Korea), which suggests that resolving high-resolution local ozone distributions remains challenging even with a good amount of observational data. The performance of each dataset impacts the accuracy of trend analysis (Fig. 5 and Fig. 6) and population exposure assessment (Fig. 10), shown as uncertainty in these Figures, which may lead to different results when compared to the WHO guideline and interim target.

From the comparison, ~~we find there are~~the large ~~differences in ozone~~disagreements among the six datasets regarding ozone trends, population exposure, and concentration estimates ~~among datasets.~~are a direct consequence of the systematic biases and performance issues identified in the evaluation. Figure ~~1~~5(b) illustrates that BME and CAMS report lower ozone estimates compared to UKML and NJML, with differences exceeding 5 ppb. NJML demonstrates a very high certainty decreasing trend in global population-weighted and area-weighted yearly mean over the 2006-2016 period~~, while the five others~~. While TCR-2 and UKML exhibit ~~either~~very high certainty increasing trends ~~or no clear trend~~in global population-weighted mean which

relates to their overestimation. Divergence among datasets becomes even more evident in the analysis of regional ozone trends (Fig. ~~2). The ozone~~6). Ozone concentrations decreased in Europe from 2006 to 2016 according to BME, NJML, UKML, and TCR-2, yet increase in the other chemical reanalysis datasets. ~~Differences in regional distributions lead to variability in exposure estimates.~~These uncertainties critically undermine the reliability of population exposure assessment. Among the six datasets, the population exposed to more than 50 ppb of ozone in Europe from 2006 to 2016 spans a broad range, from as low as 9% for CAMS to over 70% for NJML, UKML, and TCR-2. ~~This highlights the importance of removing systematic biases from these data sets before applying them to exposure estimates.~~ In East Asia, exposure levels are consistently higher, with the percentage of the population affected ranging from 61% for BME to more than 90% for UKML, GEOS-Chem, and TCR-2 based on average OSDMA8 data over the same period. Global average exposures also vary, with the proportion of the population exposed to more than 50 ppb ranging from 42% to 70% across the six datasets. More importantly, the evaluation reveals that all datasets perform poorly at high ozone levels (> 50 ppb). This highlights the importance of removing systematic biases from these data sets before applying them to exposure estimates.

Despite notable disparities in estimates, we still find some regional and temporal similarities across the six datasets. ~~In Table S13 an overall upward trend in ozone concentrations is evident across most datasets, particularly when examined as population-weighted means.~~ In Fig. ~~2~~6, all datasets exhibit a downward trend in North America over 2006 to 2016. And from the evaluation, we find that all datasets perform well in the United States, which makes the downward trend more reliable. In Fig. ~~3~~7(a) high ozone concentrations are predominantly found in regions with elevated anthropogenic and industrial emissions, while forests and sparsely populated areas have lower ozone concentrations, consistent with findings based on observations (Mills et al., 2018b; Fleming et al., 2018). In Fig. ~~3~~7(b) the standard deviation among six datasets is high in part of South America and Africa, especially in the rainforest areas, probably because of the lack of observational data in these areas and uncertainties in the emissions inventories (Pfister et al., 2019). However, for most regions it is low, such as North America and South Asia, indicating a good level of agreement on ozone estimates. The high pairwise correlation in Fig. ~~5~~9(a) supports that the geographical distributions of ground-level ozone are similar among most of datasets. The histograms of ground-level ozone exposure among the population (Fig. ~~6~~10) reveal the shared characteristic of widespread high ozone exposure in East Asia and Southeast Asia (Fleming et al., 2018).

~~When evaluating datasets against TOAR-II observations, differences in performance are seen among six datasets. BME performed well in the TOAR-II evaluation (Fig. 9), with minimal mean bias below the 50% concentration threshold (Fig. 8). Unlike the other databases, BME tends not to overestimate over the range of concentration, with a small underestimation bias. After removing TOAR sites that were used as inputs to BME (Fig. S13), BME's performance remains robust, with decreases in RMSE (from 5.28 to 5.15) and R^2 (from 0.63 to 0.53). NJML and UKML, both utilizing TOAR-I as a training set, showed overestimation in most areas (Table 4). Despite NJML's distinct spatial distribution in Fig. 5, its validation results are comparable to other datasets. NJML exhibits a higher R^2 from 2010 onward, especially at high ground-level ozone~~

~~concentrations (above 50 ppb), where prediction accuracy generally declines across all datasets. However, NJML has missing data in some coastal regions, particularly in European coastal countries, which may contribute to its elevated RMSE in Europe compared to other datasets in Table 4, since missing data are substituted with the nearest model grid cell. UKML's performance after 2010 is not as good as NJML and is worse than the chemical reanalysis datasets in 2011. CAMS, GEOS-Chem and TCR-2 primarily rely on satellite data, suggesting that they might not compare favorably with other datasets that used observations as input or training data. Despite this, CAMS unexpectedly outperforms the machine learning datasets in RMSE over the full year, especially for high ozone concentrations (50% to 90% range). In addition, as shown in Fig. 8, TCR-2, GEOS-Chem, NJML, and UKML all have widespread overestimation. The performance of each dataset can impact the accuracy of trend analysis (Fig. 1 and Fig. 2) and population exposure assessment (Fig. 6), which may lead to very different results when compared to the WHO guideline and interim target.~~

There are several possible explanations for the differences among the datasets, including several factors related to the characteristics, methodologies and input data for each dataset. BME has an unfair advantage in that it nearly matches observations at a monitoring location. But as mentioned earlier, BME still shows superior performance after removing its training data from the evaluation. BME's use of temporal autocorrelation to predict ozone in years where measurements are missing may help its good performance (Delang et al., 2021). The differing yearly ozone population-weighted mean trend in NJML compared to other datasets may be due to its unique input data, including land cover and satellite observations (Liu et al., 2022). The missing data near coastlines in NJML and relatively coarse resolution likely contribute to poorer performance in EU-27. For three chemical reanalysis datasets, previous studies have shown that significant challenges remain, particularly with respect to the representation of ozone in the lower troposphere, because of the limited sensitivity of satellite observations to ozone in the lower layers (Huijnen et al., 2020). Because of the lack of direct observational constraints at the surface in the chemical reanalyses, the better performance of CAMS may be attributable to the finer resolution that enables better representation of small-scale ozone distribution features than the other reanalysis datasets, and also to the better performance of the forecast model to predict surface ozone. Nevertheless, the assimilation of precursor measurements provides important constraints, particularly with respect to the spatial gradient and temporal variation of ground-level ozone. The low RMSE of GEOS-Chem compared to UKML and TCR-2 might be because it shares the same data assimilation method with CAMS (Qu et al., 2020a). Moreover, TCR-2, GEOS-Chem, and CAMS perform well in the United States, Canada and EU27, which may be because these regions have well-established emissions inventories for modeling (Schmedding et al., 2020) and because data assimilation is used to estimate key precursor emissions from satellite observations in TCR-2 and GEOS-Chem. Optimizing additional precursor emissions, such as VOCs, from satellite observations is considered to be important to better represent surface ozone (Miyazaki et al., 2019; Sekiya et al., 2024; Miyazaki et al., 2012). The poor performance in South Korea and Japan could be because the coarse resolution models may not accurately capture ozone gradients in a nation with a high density of monitors (Punger and West, 2013; Sekiya et al., 2021). This suggests a need for continued efforts to improve the mapping resolution to capture spatial variability in these regions. Since most of the current reanalysis products still suffer from large

systematic errors in their surface ozone analysis, it might be important to apply bias corrections while maintaining the detailed spatial and temporal variability of the original data using methods such as machine learning (Miyazaki et al., 2024) before performing exposure estimates. While these factors may help to explain differences between the datasets, we have not systematically tested them, and as discussed by Sekiya et al. (2024) and Jones et al. (2024), further inter-comparisons of reanalysis products and detailed discussions for improvement are required.

Although we conducted a comprehensive comparison and evaluation, this study still has some limitations. First, the comparison only focuses on land and inhabited islands, because of the focus on ground-level ozone impacts on health. Our estimates of population exposure are based on ambient concentration in each grid cell, ignoring other factors that impact ozone exposure, such as indoor ozone concentration. Also, using OSDMA8 as the metric to evaluate datasets might hide differences in model performance at hourly temporal resolution, which would need to be analyzed in a separate study. In instances of missing model estimates, we default to the nearest valid estimate to evaluate with TOAR-II observations~~.~~ or re-gridded grid cell. For datasets with coarse spatial resolution, this method may increase or reduce bias by double counting.

7. Conclusions

This study evaluates the consistency and accuracy of six ground-level ozone mapping products, developed using different methods. Substantial discrepancies among datasets are reflected in global and regional ozone trends, the spatial distribution of ozone, population exposure estimates, and model performance. Model performance evaluation based on TOAR-II observations varied.~~The global population-weighted average has a maximum span of 10 ppb among the six datasets. In terms of long-term trends over 2006 to 2016 period, UKML and TCR-2 show a consistent upward trend globally, while NJML shows a downward trend. Regionally, all datasets show a downward trend in North America, and only BME and NJML datasets demonstrate a downward trend in East Asia; In Europe, BME, UKML, NJML and TCR-2 report a downward trend, while the other two chemical reanalysis datasets reveal an upward trend that is not seen in observations. These differences among datasets are sufficiently large that assessments of health impacts of ozone would differ significantly when using different ozone datasets. Model performance evaluation based on TOAR-II observations varied; in 2016, R^2 values ranged from 0.35 to 0.63, and RMSE values ranged from 5.28 ppb to 13.49 ppb for all stations.~~ BME performs well near monitoring locations with good R^2 and small RMSE. All five datasets, except for BME, exhibit similar R^2 values in 2016. NJML performs well after 2010 and shows robust performance under high ozone concentrations. ~~Before 2010, UKML performs well, but after 2010, UKML shows decreased performance.~~ Machine learning datasets tend to overestimate. The chemical reanalysis datasets perform comparably with the geostatistical and machine learning datasets, which is somewhat surprising given that they were not designed to estimate ground-level ozone accurately and do not use ground-level observations as input. CAMS performs the best among the chemical reanalysis datasets in term of RMSE, although CAMS has difficulty capturing TOAR-II observations in China. In regions where TOAR-II observations are sparse, all datasets show RMSE values about 10 ppb, highlighting the difficulty

in mapping ground-level ozone ~~distributions~~magnitude in regions with little observational data. Conversely, in some regions with very dense TOAR-II observations, all datasets show R^2 values below 0.~~1~~2, highlighting the necessity for fine resolution mapping to accurately capture spatial variability. The global population-weighted average has a maximum span of 10 ppb among the six datasets. In terms of population-weighted mean trends over 2006 to 2016 period, UKML and TCR-2 show very high certainty upward trends globally, while NJML shows a very high certainty downward trend. Regionally, all datasets show a downward trend in North America, and the evaluation results make this trend more reliable. Only BME and NJML datasets demonstrate a downward trend in East Asia, and they also fit well with TOAR-II observations in population density distribution. In Europe, BME, UKML, NJML and TCR-2 report a downward trend, while the other two chemical reanalysis datasets reveal an upward trend that is not seen in observations. These differences among datasets are sufficiently large that assessments of health impacts of ozone would differ significantly when using different ozone datasets.

Given that some of the datasets used similar input data, it is somewhat surprising to find the large discrepancies shown here, suggesting that applications of these datasets to health burden assessments, epidemiology or similar applications for agricultural and ecosystem impacts may differ strongly based on the dataset selected. The coarse-resolution datasets, GEOS-Chem and TCR-2, perform well in grid-to-grid evaluations at their native resolutions, making them effective for studying long-term regional ozone effects. However, because of their coarser resolutions, these two datasets cannot capture site-level distributions and exhibit greater bias than the higher-resolution BME, CAMS, and NJML datasets. UKML, despite its relatively fine resolution (0.125°), shows larger biases and a lower R^2. The superior performance of BME and NJML should be noted with the fact that both datasets use observational data for input or training, which gives them an inherent advantage in these evaluations. More research will be needed before different methods converge on similar estimates. Such research can include more widespread ground observations, improved used of satellite observations, improved chemistry-climate modelling, and further development of different data fusion methods. Also, it is not clear whether differences among ~~different~~ datasets are due mainly to the methods used or to differences in input data. In addition, establishing a formal benchmark test based on the evaluation methods described in this study for the yearly OSDMA8 metric is essential. This would allow for new mapping products to be easily assessed. The general findings here of poor agreement among datasets may also be applicable to other air quality datasets or even datasets from other Earth system domains. According to this study, there is no clear consensus on the best ozone mapping methods. To further improve these ozone mapping products, researchers must update and adjust their methods and input data regularly and iteratively.

8. Code and data availability

Observational data are publicly available from the TOAR-II data portal (last accessed on 15 November 2024, toar-data.org) (Schröder et al., 2021). The BME dataset of global ground-level ozone estimates (Becker et al., 2023) is publicly available at zenodo.org/records/10498857. The NJML dataset is publicly available at doi.org/10.5281/zenodo.6378092. The CAMS

reanalyses data (Inness et al., 2019) are publicly available from https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4. The TCR-2 reanalyses data (Miyazaki, 2024) are publicly available from https://disc.gsfc.nasa.gov/datasets/TRPSCRO32H2D_1. Other datasets of global ozone concentrations can be obtained by contacting the creators of these datasets.

10. Author Contributions

This research was conceived by HW, JJW, and MLS. HW, KM, HZS, ZQ, XL, and AI provided ozone concentration datasets. MS and SS provided TOAR-II observational data. Data analyses and numerical results were generated by HW with input from MLS and JJW. HW, MLS and JJW wrote the paper and all authors provided edits and comments on drafts.

11. Competing Interests

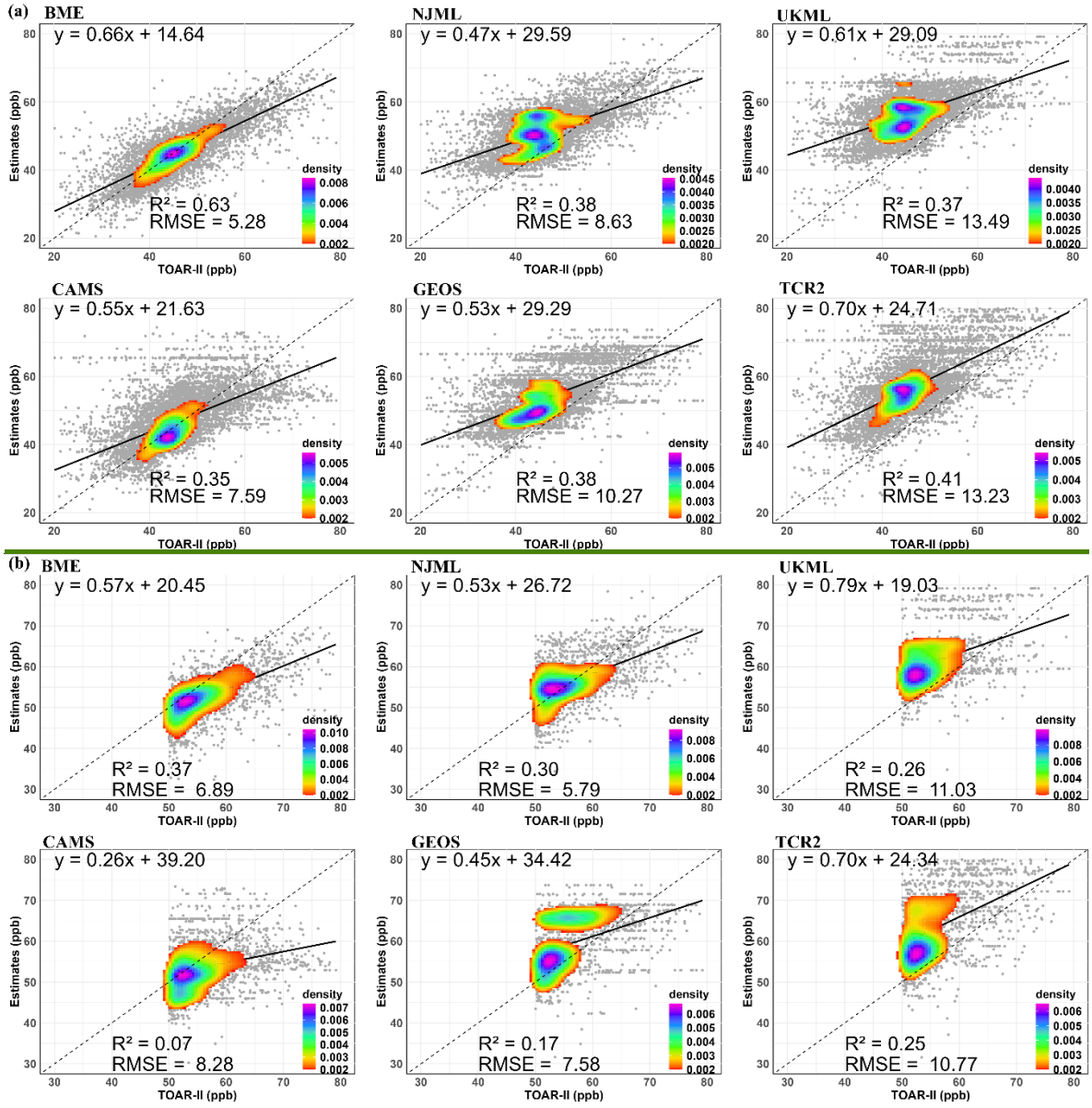
Some authors are members of the editorial board for *Atmospheric Chemistry & Physics*. The authors declare that they have no other conflicts of interest.

12. Financial Support

We gratefully acknowledge support for this work through National Aeronautics and Space Administration (NASA) grants #NNX16AQ30G and #80NSSC23K0930. Part of this work was conducted at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA. We also acknowledge the support of the NASA Atmospheric Composition: Aura Science Team Program (19-AURAST19-0044), Atmospheric Composition Modeling and Analysis Program (22-ACMAP22-0013), NASA Earth Science U.S. Participating Investigator program (22-EUSPI22-0005).

Table 1: Overview of six global ozone mapping products.

Global ozone dataset	Model type	Resolution	Period	Temporal Resolution
Bayesian Maximum Entropy Data Fusion (BME)(Delang et al., 2021)	Geostatistics	0.1° ×0.1°	1990-2017	OSDMA8
Cluster-Enhanced Ensemble Learning (NJML)(Liu et al., 2022)	Machine Learning	0.5° ×0.5°	2003-2019	Monthly DMA8
Space-Time Bayesian Neural Network Downscaler (UKML)(Sun et al., 2022)	Machine Learning	0.125° ×0.125°	1990-2019	Monthly DMA8
Copernicus Atmosphere Monitoring Service (CAMS)(Inness et al., 2019)	Chemical Reanalysis	0.75° ×0.75°	2003-2020	3-Hourly
GEOS-Chem (GEOS)(Qu et al., 2020b)	Chemical Reanalysis	2° ×2.5°	2005-2016	DMA8
Tropospheric Chemistry Reanalysis Version 2 (TCR-2)(Miyazaki et al., 2020b)	Chemical Reanalysis	1.125° ×1.125°	2005-2020	2-Hourly


(a) BME
y = 0.66x + 14.64
R² = 0.63
RMSE = 5.28
NJML
y = 0.47x + 29.59
R² = 0.38
RMSE = 8.63
UKML
y = 0.61x + 29.09
R² = 0.37
RMSE = 13.49
CAMS
y = 0.55x + 21.63
R² = 0.35
RMSE = 7.59
GEOS
y = 0.53x + 29.29
R² = 0.38
RMSE = 10.27
TCR2
y = 0.70x + 24.71
R² = 0.41
RMSE = 13.23
(b) BME
y = 0.57x + 20.45
R² = 0.37
RMSE = 6.89
NJML
y = 0.53x + 26.72
R² = 0.30
RMSE = 5.79
UKML
y = 0.79x + 19.03
R² = 0.26
RMSE = 11.03
CAMS
y = 0.26x + 39.20
R² = 0.07
RMSE = 8.28
GEOS
y = 0.45x + 34.42
R² = 0.17
RMSE = 7.58
TCR2
y = 0.70x + 24.34
R² = 0.25
RMSE = 10.77
Estimates (ppb)
TOAR-II (ppb)
density

675

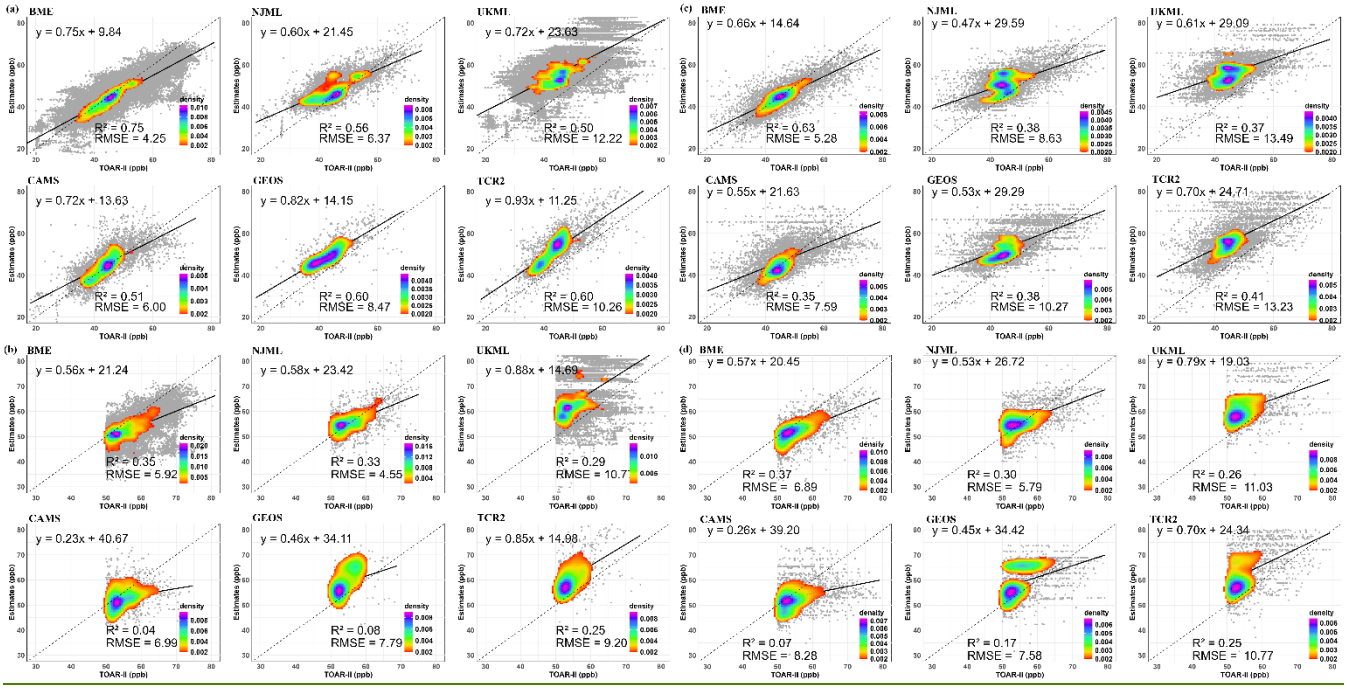


Figure ~~7~~1: Performance evaluations of six datasets with TOAR-II observations in 2016 for OSDMA8. The observation-prediction evaluations are presented ~~in scatter plots~~ with densities estimated by a Gaussian kernel function. ~~Determination (R^2)~~The coefficient of determination (R^2) and root mean squared error (RMSE) are ~~given.~~shown for four scenarios: (a) ~~The~~A grid-to-grid evaluation ~~includes all monitor stations in~~at the native resolution of each dataset using re-gridded TOAR-II ~~network in 2016,~~observations, (b) ~~The~~A grid-to-grid evaluation ~~includes~~, same as (a), but only ~~monitor stations~~for grid cells with observations above 50 ppb ~~in the~~ , (c) A grid-to-point evaluation using all TOAR-II ~~network in 2016~~sites, (d) A grid-to-point evaluation, same as (c), but only for sites with observations above 50 ppb. The dashed line marks where TOAR-II observations equal estimates (y=x line), and the solid black line represents the best-fit line. Performance evaluations for each year are shown in ~~Figure S11~~Figures S7 and ~~Figure S12~~S8.

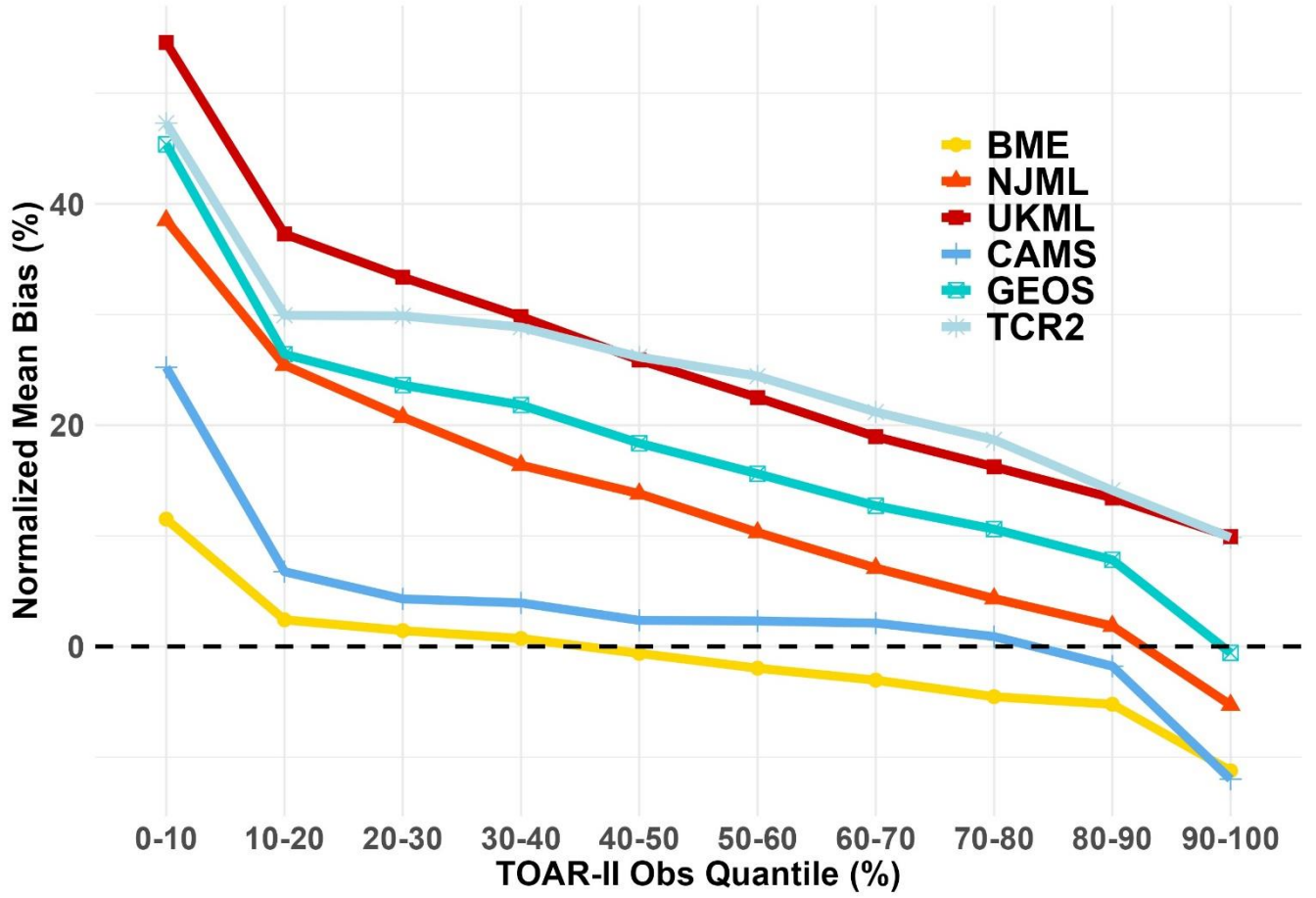


Figure ~~8~~2: Normalized mean bias of six databases against TOAR-II observations (OSDMA8) at different quantiles in 2016~~.~~, calculated based on the grid-to-point scenario. 0%: 13.46 ppb; 10%: 36.75 ppb; 20%: 39.80 ppb; 30%: 41.89 ppb; 40%: 43.57 ppb; 50%: 45.06 ppb; 60%: 46.82 ppb; 70%: 48.93 ppb; 80%: 52.18 ppb; 90%: 57.21 ppb; 100%: 86.25 ppb. Normalized mean bias for each year against TOAR-II observations are shown in Figure ~~S14~~S5. Different quantiles of TOAR-II observations for other years are shown in Table ~~S9~~S11.

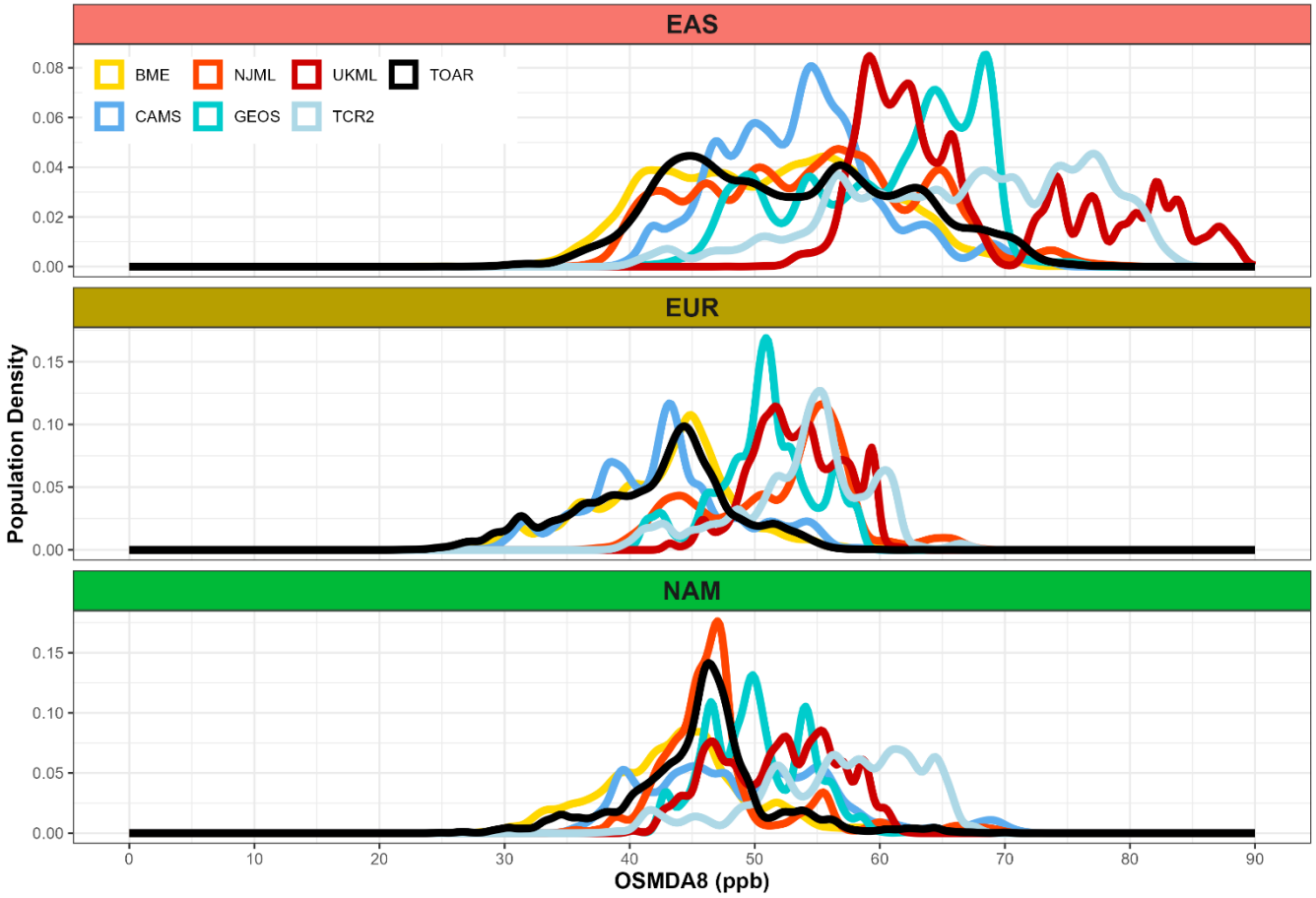


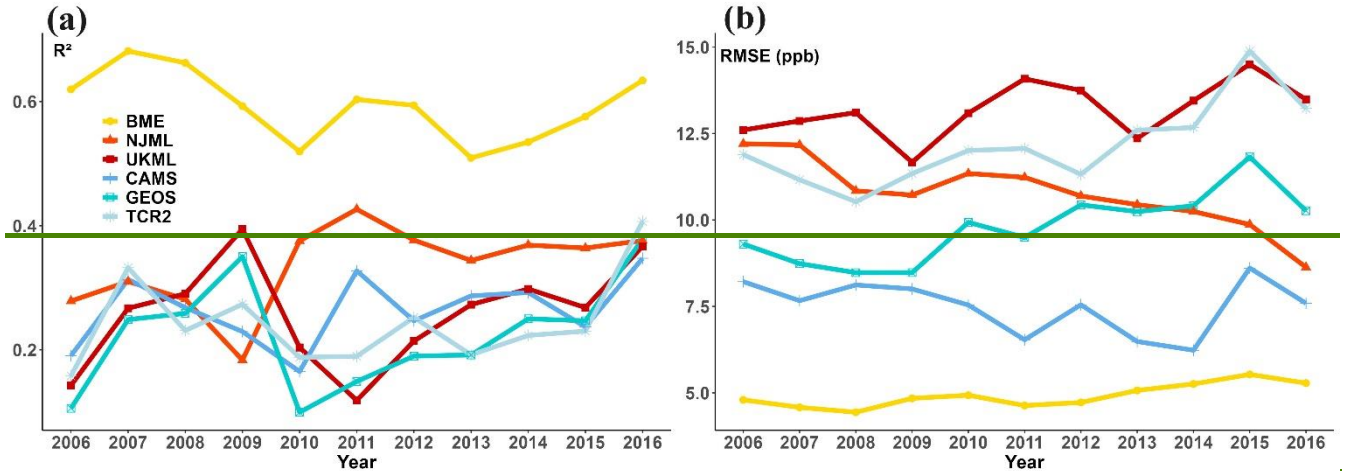
Figure 3: Population-weighted exposure distributions for OSDMA8 in 2016 in three regions: East Asia (EAS), Europe (EUR), and North America (NAM) (regions defined in Table 4S7). Each panel compares the distribution derived from the TOAR-II observations (black line) with estimates from six datasets (colored lines), calculating the population-weighted kernel density estimate, only for grid cells where TOAR-II measurements exist.

Table 2: Performance evaluation of six datasets for countries (~~union~~and the EU) with the most monitors in 2016 against TOAR-II observations of OSDMA8 based on the gird-to-grid scenario. Number is the number of the TOAR-II monitor stations in each country. Density (per km^2) is the density of the TOAR-II monitors in each country based on land area. Estimate is the average of the grid estimates for each dataset at the TOAR-II monitor locations in each country. Linear regression R^2 and root mean squared error (RMSE) against TOAR-II observations in each country are ~~presented.~~based on a grid-to-grid evaluation at each dataset's native resolution against re-gridded TOAR-II observations.. The Lower and Upper Bound represent the 95% confidence interval for the Estimate, calculated from the linear regression of each dataset against TOAR-II observations. Country names are United States of America (USA), China (CHN), Japan (JPN), South Korea (KOR), Canada (CAN). EU-27 includes Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden. Others is all other countries in TOAR-II apart from those listed. Performance evaluations for other years in these countries, are shown in Table ~~S10~~S8. (*) indicates the sample size of the comparison pair is less than 30.

Dataset	Country	EU-27	USA	CHN	JPN	KOR	CAN	Others
	Number	2170	1425	1405	1108	315	260	330
	Density	5.43E-4	1.56E-4	1.50E-4	3.04E-3	3.23E-3	2.96E-5	1.07E-5
	TOAR	43.21	47.03	53.10	43.84	51.50	37.39	40.55
BME	Estimate	~~43.30~~42.42	~~45.12~~44.30	~~50.26~~48.84	~~44.69~~43.00	51.~~79~~67	35.~~26~~94	39.~~05~~62
	R^2	0.~~63~~76	0.~~71~~75	0.~~63~~78	0.~~03~~12	0.~~10~~08	0.~~46~~49	0.~~48~~44
	RMSE	~~3.91~~2.72	~~4.12~~3.48	~~64~~.97	4.~~59~~29	~~7.33~~3.72	~~4.39~~3.49	~~8.66~~7.45
	Lower	35.66	37.55	42.08	36.25	44.91	29.18	32.87
	Upper	49.17	51.05	55.59	49.76	58.42	42.69	46.37
NJML	Estimate	~~53.53~~50.97	~~48.44~~47.57	~~53.39~~51.13	~~49.40~~47.08	54.~~62~~84	43.~~79~~19	~~48.63~~47.09
	R^2	0.~~41~~63	0.~~58~~74	0.~~57~~72	0.~~00~~11	0.~~07~~14	0.~~30~~38	0.~~51~~55
	RMSE	9.43	2.96	4.72	5.93	2.68	6.19	9.30
	Lower	42.47	39.06	42.62	38.57	46.33	34.68	38.58
	~~RMSE~~Upper	~~11.49~~59.48	~~4.55~~56.08	~~6.86~~59.64	~~7.42~~55.59	~~8.01~~63.35	~~7.57~~51.70	~~11.59~~55.60
UKML	Estimate	53.~~27~~71	~~52.54~~53.27	~~66.78~~64.81	~~61.45~~60.63	65.~~02~~44	~~46.87~~48.40	49.~~01~~62
	R^2	0.~~21~~28	0.~~38~~49	0.~~37~~46	0.~~01~~18	0.~~01~~03	0.~~33~~19	0.~~32~~28
	RMSE	~~11.54~~12.17	7.~~52~~82	~~16.40~~15.05	~~18.25~~17.41	13.01~~15.54~~	~~10.32~~11.14	12.50~~13.01~~
	Lower	42.52	42.08	53.62	49.44	54.25	37.21	38.43
	Upper	64.90	64.47	76.00	71.82	76.63	59.59	60.82
CAMS	Estimate	~~42.17~~41.93	~~49.67~~47.12	~~53.85~~52.83	45.65~~44.91~~	~~58.93~~53.60	39.~~54~~03	39.~~84~~91
	R^2	0.~~32~~44	0.~~34~~47	0.~~07~~21	0.~~01~~16	0.~~01~~05	0.~~28~~25	0.~~39~~42
	RMSE	~~5.75~~4.92	~~6.65~~4.64	~~10.62~~7.99	4.~~95~~33	~~12.46~~4.35	~~4.63~~3.47	~~9.40~~8.09
	Lower	31.17	36.36	42.07	34.89	42.84	28.26	29.15
	Upper	52.69	57.88	63.59	56.41	64.37	49.79	50.66
GEOS	Estimate	~~49.76~~48.01	~~50.58~~48.94	~~60.48~~58.54	~~56.99~~54.58	65.~~94~~41	~~45.73~~44.83	~~44.54~~43.25
	R^2	0.~~30~~72	0.~~39~~55	0.~~37~~46	0.~~03~~37*	0.~~00~~81*	0.~~44~~34	0.~~31~~44
	RMSE	~~8.41~~7.62	~~6.08~~5.65	~~11.15~~10.19	~~13.94~~11.87	~~16.36~~14.71	~~9.08~~8.76	~~10.70~~9.11
	Lower	37.39	38.33	47.92	43.97	54.77	34.21	32.62

	Upper	58.62	59.55	69.17	65.20	76.06	55.45	53.87
TCR-2	Estimate	51.8349.60	55.5453.50	66.4361.24	58.3753.56	67.8762.46	45.9743.76	48.3245.47
	R^2	0.3355	0.2343	0.3646	0.0012	0.0201	0.4338	0.5452
	RMSE	10.159.33	10.589.43	15.9912.56	16.6912.63	18.5311.39	9.847.48	11.439.33
	Lower	37.77	41.67	49.40	41.72	50.62	31.92	33.63
	Upper	61.44	65.34	73.08	65.39	74.30	55.60	57.31

715



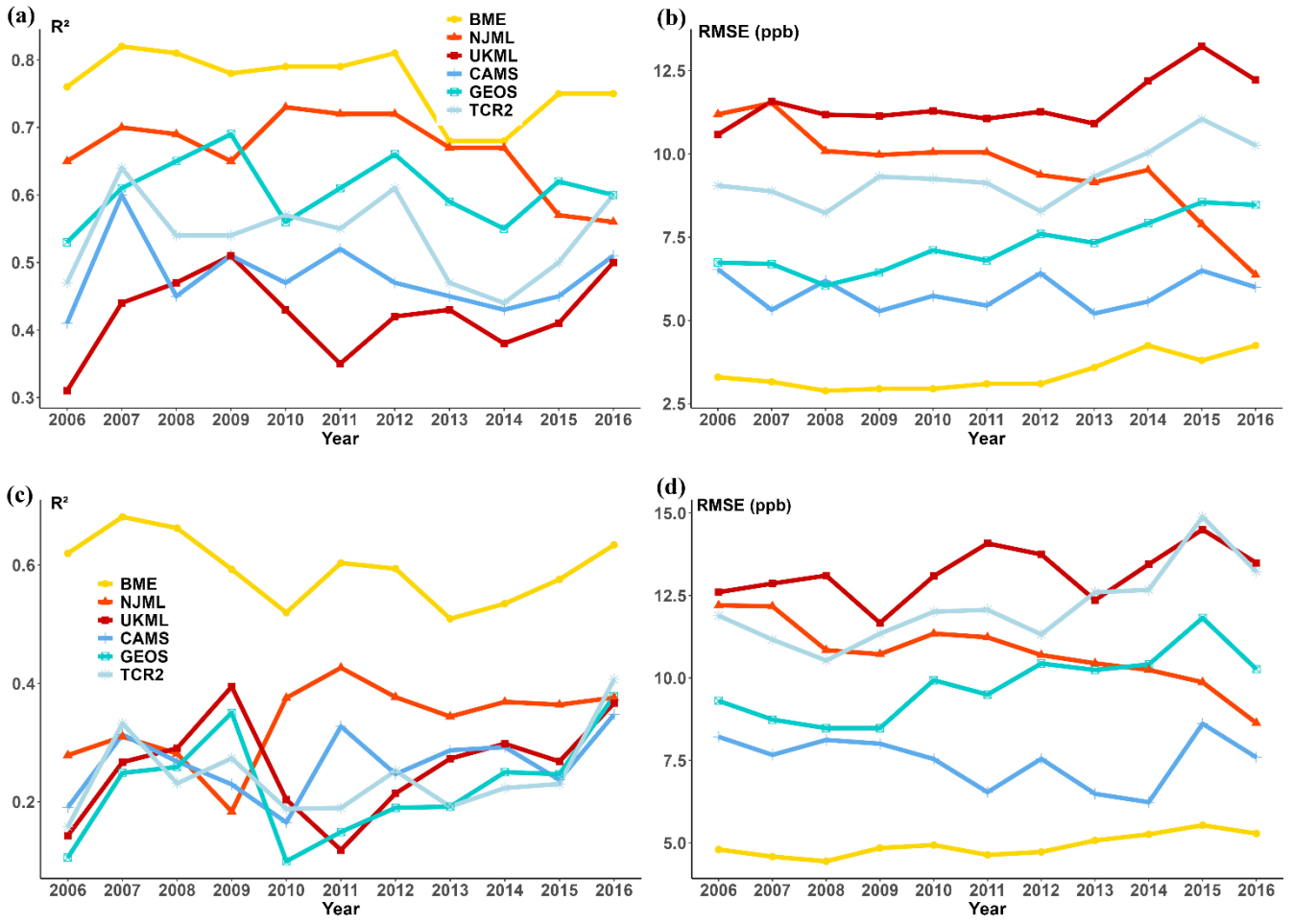


Figure ~~9~~4: (a) Time series of determination (R2) between each dataset and TOAR-II observations of OSDMA8 from 2006 to 2016~~.~~ based on grid-to-grid evaluation at the native resolution of each dataset using re-gridded TOAR-II observations. (b) Time series of root mean squared error (RMSE) between each dataset and TOAR-II from 2006 to 2016 based on grid-to-grid evaluation. (c) Time series of determination (R2) between each dataset and TOAR-II observations of OSDMA8 from 2006 to 2016 based on grid-to-point evaluation using all TOAR-II sites. (d) Time series of root mean squared error (RMSE) between each dataset and TOAR-II from 2006 to 2016 based on grid-to-point evaluation.

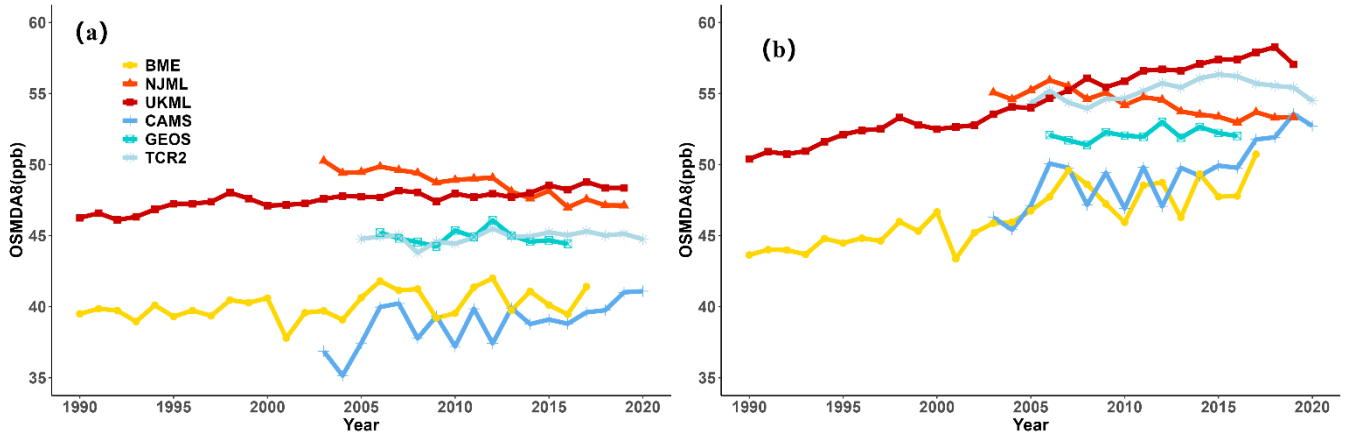
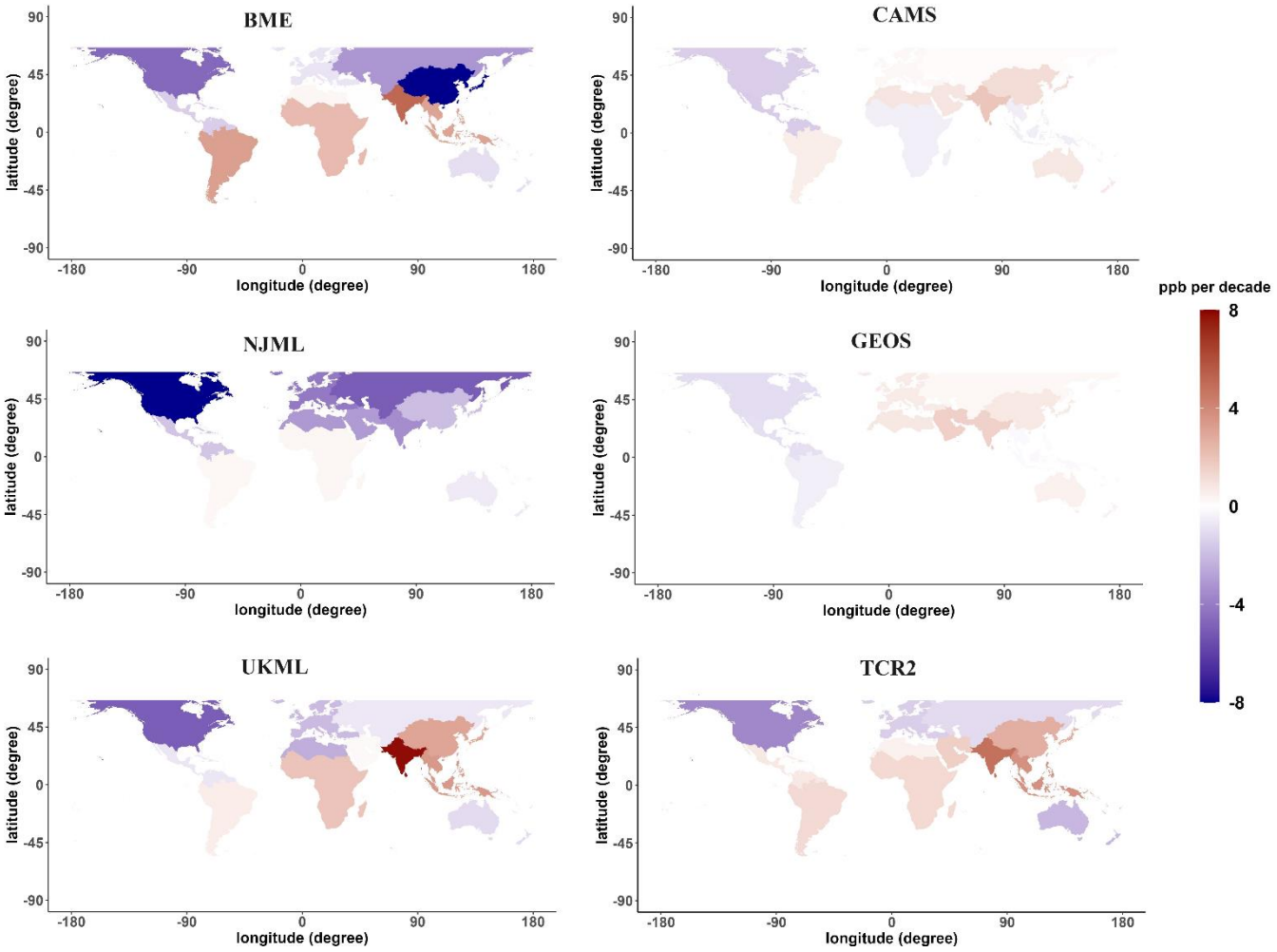


Figure ~~1~~5: Yearly trends of ground-level ozone for six datasets, shown for (a) the area weighted global mean ozone over land, and (b) population weighted global mean ozone, where ozone is expressed as OSDMA8. Yearly trends for individual world regions are shown in Figures S2 and S3. Mann-Kendall trend test for population weighted global mean over the full time series for each dataset: BME 0.688 ppb yr^{-1} trend with p-value < 0.0001, NJML -0.691 ppb yr^{-1} with p-value 0.0001, UKML 0.913 ppb yr^{-1} with p-value < 0.0001, CAMS 0.569 ppb yr^{-1} with p-value 0.0011, GEOS-Chem 0.164 ppb yr^{-1} with p-value 0.5334, TCR-2 0.4 ppb yr^{-1} with p-value 0.0343.

Table ~~2~~3: Yearly trends of area-weighted, and population-weighted global mean of ground-level ozone for six datasets with 95% confidence intervals (LowerCI and UpperCI) and p-values from 2006 to 2016.

Dataset	Slope (ppb/yr)	Lower CI (ppb/yr)	Upper CI (ppb/yr)	p-value	Weighted
BME	-0.12	-0.33	0.10	0.25	area
NJML	-0.24	-0.32	-0.16	0.00	area
UKML	0.04	-0.02	0.11	0.16	area
CAMS	-0.05	-0.29	0.18	0.62	area
GEOS	-0.02	-0.14	0.10	0.71	area
TCR-2	0.06	-0.03	0.15	0.18	area
BME	-0.04	-0.30	0.23	0.76	population
NJML	-0.26	-0.33	-0.19	0.00	population
UKML	0.26	0.20	0.32	0.00	population
CAMS	0.06	-0.23	0.34	0.67	population
GEOS	0.05	-0.04	0.14	0.23	population
TCR-2	0.20	0.10	0.30	0.00	population



740

Figure ~~2~~6: Population weighted ozone (OSDMA8) trends per decade for six datasets, calculated over the 2006-2016 period analyzed for each dataset. The different regions are defined in Table S7. Population weighted yearly trend of six datasets over priority regions (NAM, EUR, SAS, EAS, SEA, SAF, MDE) from 2006 to 2016 with 95% confidence intervals and p-values is shown in Table ~~S11. Population weighted ozone (OSDMA8) trends per decade for six datasets over the full period is shown in Figure S4.~~S9.

745

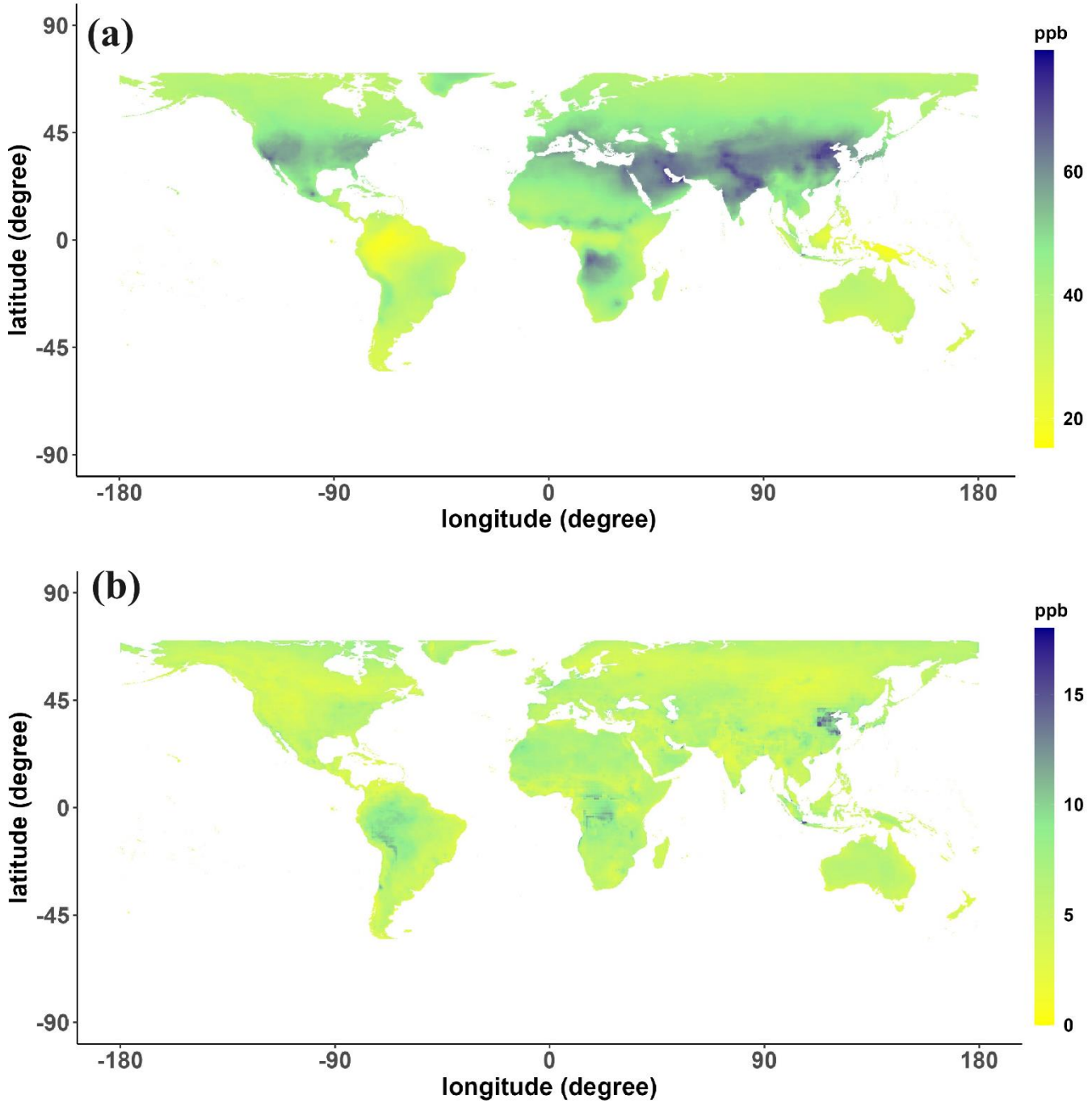
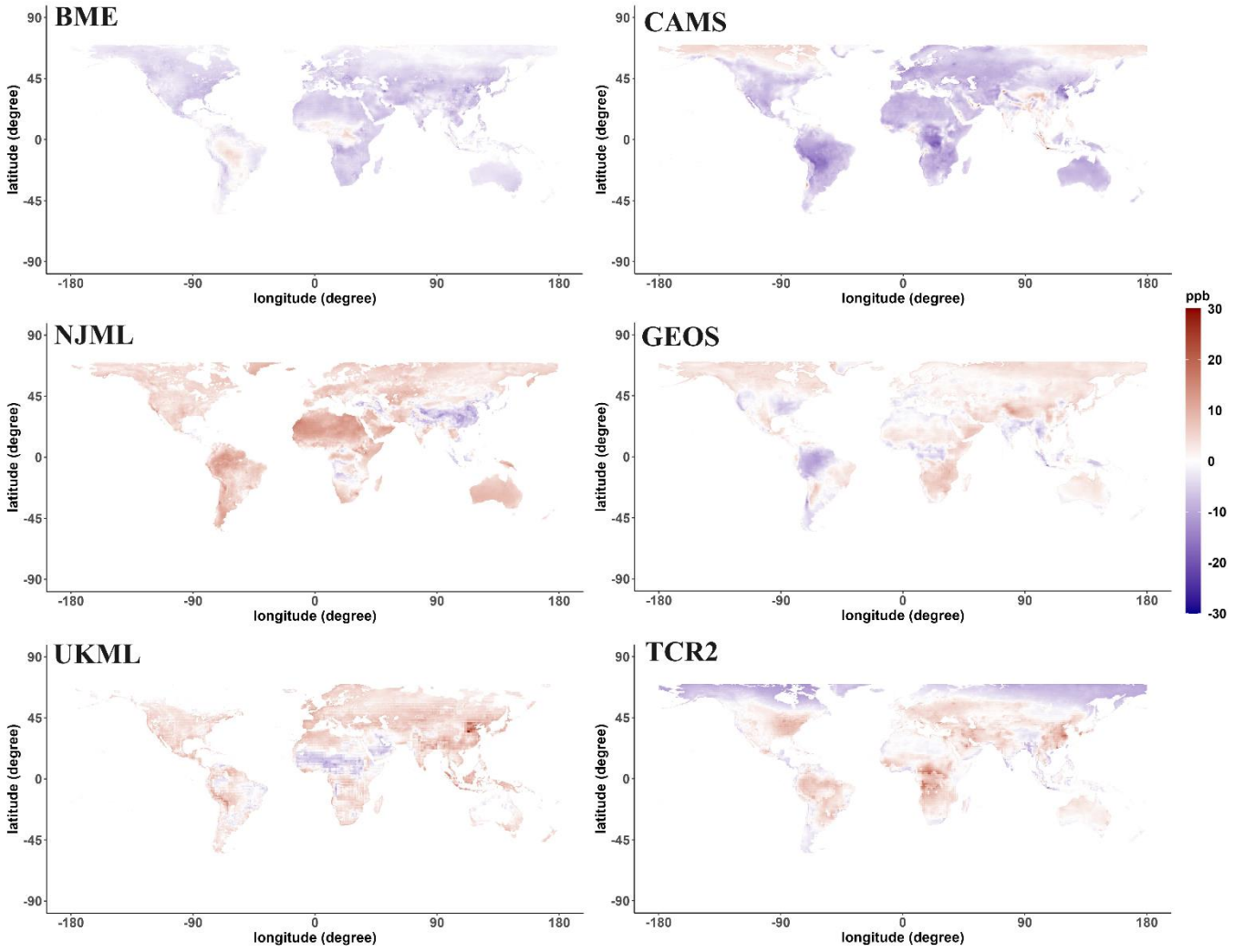


Figure ~~37~~: For six datasets from 2006 to 2016, (a) the 11-year ensemble mean, and (b) the average of annual standard deviations. Ozone data are reported as OSDMA8. The mean and standard deviation for each year are shown in Figures S5 and S6.



750

Figure 48: The difference of OSDMA8 in each grid cell between the 11-year (2006-2016) mean of each of six datasets and the ensemble mean (Figure 3). Positive values indicate that the average estimate of the dataset is higher than the ensemble mean. Negative values indicate that the average estimate of the dataset is lower than the ensemble mean of the six datasets. Difference maps for each year are shown in Fig. S7.

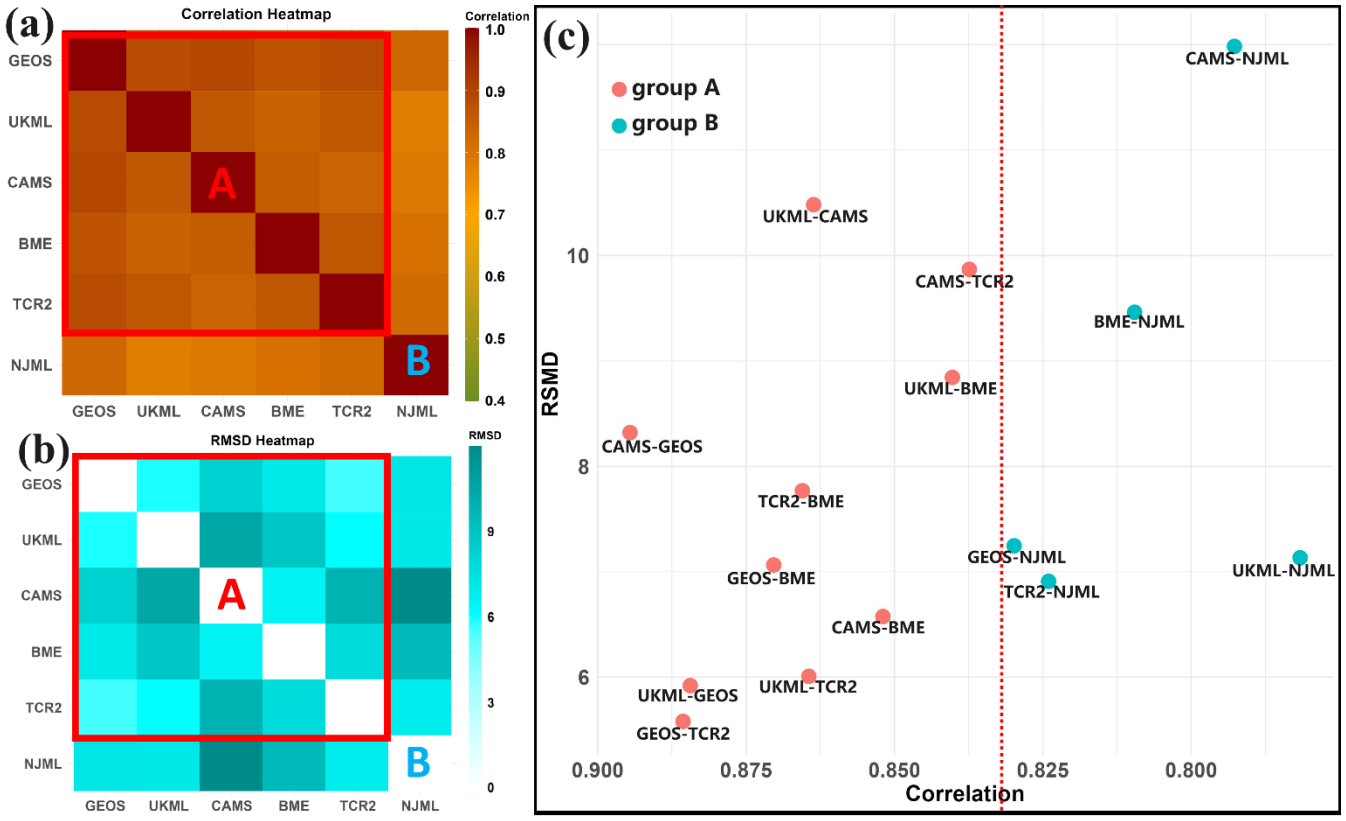


Figure ~~5~~9: Heatmaps of similarity among the six datasets, including (a) heatmaps of average of pairwise correlation (Pearson R) between each dataset from 2006 to 2016. (b) heatmaps of average of pairwise Root mean square difference (RMSD) between each dataset from 2006 to 2016. Group A designates five datasets with strong similarity, while Group B is composed of one dataset with lower similarity with the rest. (c) Scatterplot of correlation and RSMD between each pair of datasets. The datasets with greatest similarity are in the lower left of panel c, and comparisons with the Group B dataset have lower correlation. ~~Heatmaps for each year are shown in Figure S8 and Figure S9.~~

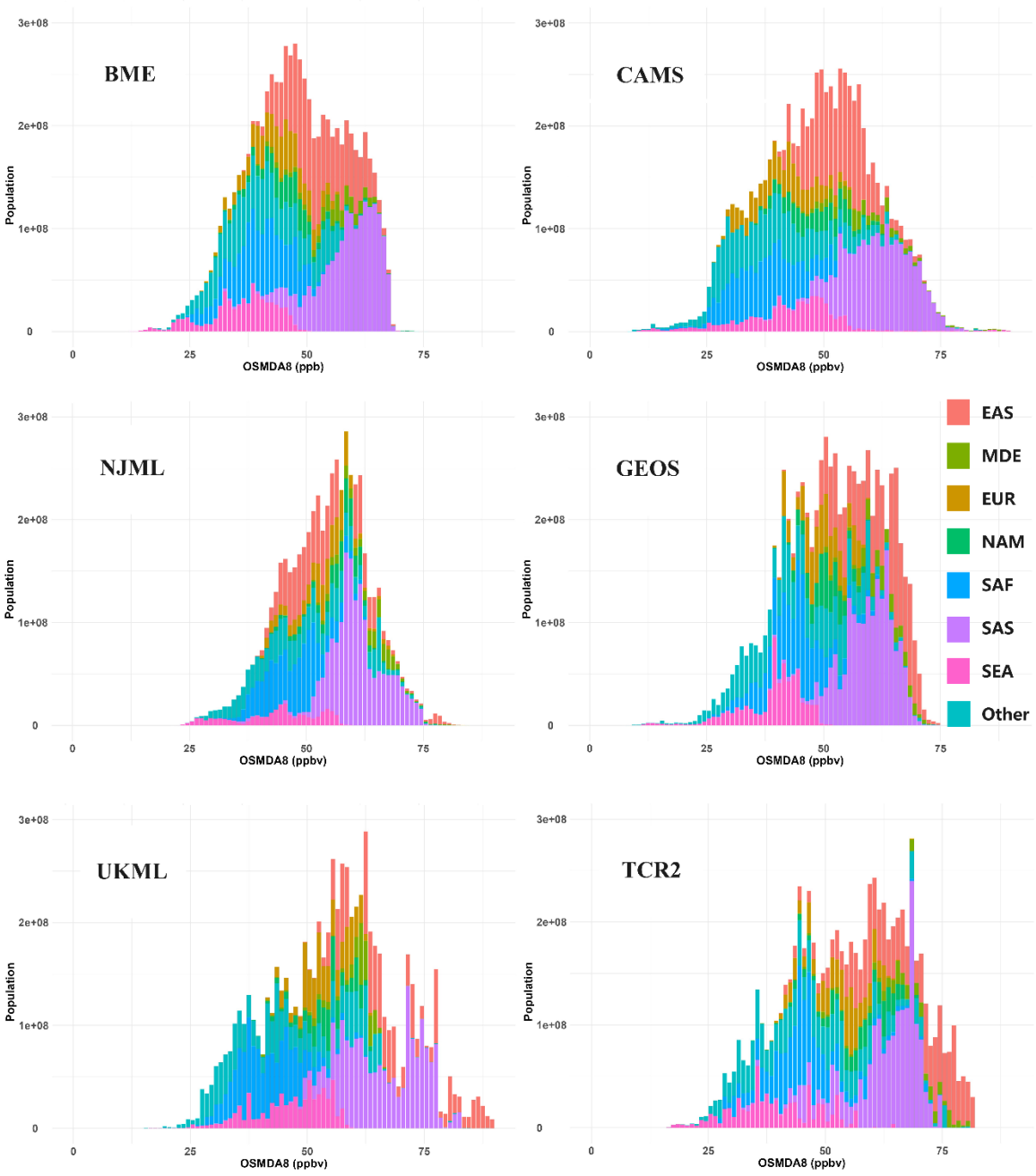


Figure ~~6~~10: Population exposed to 11-year average ozone (OSDMA8) from 2006 to 2016 in different regions. The horizontal axis represents ozone concentrations, and the vertical axis represents population size. ~~Concentrations and population for each year are presented in Figure S10.~~ The definitions of different regions are included in Table S7. The Lower and Upper Bound of population exposure, which represent the 95% prediction interval for the estimate, are presented in Figure S12.

Table ~~34~~: The share of population in percentage (%) exposed to ozone above three particular thresholds (ppb) in each world region, for the 2006 to 2016 average OSDMA8 for six ozone datasets. ~~0% means greater than 0 but less than 0.5%, 0 means no~~Each region shows the share of the population ~~share greater than this ozone concentration.~~exposed at each threshold, calculated using the estimate, the lower bound and the upper bound of the OSDMA8 from each dataset, respectively. The bounds represent the 95% prediction interval for the estimate, derived from the linear regression of each dataset against TOAR-II observations. Population shares for each year are shown in Table ~~S8~~S10. The definitions of different regions are included in Table S7.

2006-2016		EAS	EUR	MDE	NAM	SAF	SAS	SEA	GLO
BME	>30	100 [100,100]	99 [87,100]	100 [99,100]	99 [93,100]	93 [71,100]	100 [99,100]	84 [52,93]	96 [85,99]
	>50	61 [28,94]	16 [0,56]	79 [43,92]	17 [4,72]	3 [0,31]	89 [68,98]	0 [0,16]	42 [24,66]
	>70	0 [0,5]	0 [0,0]	0 [0,6]	0 [0,1]	0 [0,0]	0 [0,28]	0 [0,0]	0 [0,8]
NJML	>30	100 [100,100]	100 [100,100]	100 [100,100]	100 [100,100]	99 [96,100]	100 [100,100]	89 [76,100]	99 [96,100]
	>50	72 [37,98]	76 [22,96]	99 [81,100]	88 [53,99]	36 [5,78]	99 [77,100]	27 [0,63]	70 [41,90]
	>70	3 [2,12]	0 [0,5]	5 [1,58]	3 [0,9]	0 [0,1]	8 [0,34]	0 [0,0]	4 [0,17]
UKML	>30	100 [100,100]	100 [100,100]	100 [99,100]	100 [99,100]	98 [46,100]	100 [100,100]	97 [74,100]	98 [83,100]
	>50	99 [62,100]	77 [2,100]	94 [42,100]	84 [3,100]	10 [0,67]	99 [64,100]	41 [0,80]	69 [32,88]
	>70	31 [13,85]	0 [0,27]	0 [0,79]	0 [0,24]	0 [0,2]	40 [1,82]	0 [0,1]	16 [3,48]
CAMS	>30	100 [100,100]	98 [53,100]	100 [99,100]	100 [88,100]	86 [37,99]	100 [100,100]	88 [66,97]	93 [74,99]
	>50	67 [9,100]	9 [0,58]	88 [33,100]	40 [2,91]	8 [0,41]	96 [58,100]	24 [8,71]	48 [18,76]
	>70	0 [0,12]	0 [0,0]	8 [4,37]	0 [0,3]	0 [0,0]	12 [0,62]	6 [5,8]	4 [1,20]
GEOS	>30	100 [100,100]	100 [100,100]	100 [100,100]	100 [100,100]	99 [72,100]	100 [100,100]	89 [49,98]	98 [85,100]
	>50	95 [58,100]	44 [0,100]	99 [60,100]	55 [0,100]	14 [1,78]	95 [47,100]	0 [0,56]	59 [26,87]
	>70	4 [0,62]	0 [0,0]	4 [0,69]	0 [0,0]	0 [0,2]	0 [0,54]	0 [0,0]	1 [0,29]
TCR-2	>30	100 [99,100]	100 [96,100]	100 [100,100]	99 [99,100]	98 [67,100]	99 [99,100]	85 [38,100]	97 [83,100]
	>50	94 [63,100]	70 [5,100]	94 [79,100]	86 [31,100]	18 [4,85]	90 [62,99]	13 [1,51]	64 [35,89]
	>70	41 [0,79]	0 [0,21]	38 [0,86]	0 [0,51]	1 [0,7]	10 [0,79]	0 [0,1]	13 [0,46]

References

Ainsworth, E. A.: Understanding and improving global crop response to ozone pollution, The Plant Journal, 90, 886-897, https://doi.org/10.1111/tpj.13298, 2017.

Balmes, J. R.: Long-Term Exposure to Ozone and Small Airways: A Large Impact?, American Journal of Respiratory and Critical Care Medicine, 205, 384-385, 10.1164/rccm.202112-2733ED, 2022.

Becker, J. S., DeLang, M. N., Chang, K.-L., Serre, M. L., Cooper, O. R., Wang, H., Schultz, M. G., Schröder, S., Lu, X., Zhang, L., Deushi, M., Josse, B., Keller, C. A., Lamarque, J.-F., Lin, M., Liu, J., Marécal, V., Strode, S. A., Sudo, K., Tilmes, S., Zhang, L., Brauer, M., and West, J. J.: Using Regionalized Air Quality Model Performance and Bayesian Maximum Entropy data fusion to map global surface ozone concentration, Elementa: Science of the Anthropocene, 11, 10.1525/elementa.2022.00025, 2023.

Betancourt, C., Stomberg, T. T., Edrich, A.-K., Patnala, A., Schultz, M. G., Roscher, R., Kowalski, J., and Stadtler, S.: Global, high-resolution mapping of tropospheric ozone–explainable machine learning and impact of uncertainties, Geoscientific Model Development, 15, 4331-4354, 2022.

Brauer, M., Roth, G. A., Aravkin, A. Y., Zheng, P., Abate, K. H., Abate, Y. H., Abbafati, C., Abbasgholizadeh, R., Abbasi, M. A., and Abbasian, M.: Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021, The Lancet, 403, 2162-2203, 2024.

Chang, K.-L., Petropavlovskikh, I., Cooper, O. R., Schultz, M. G., and Wang, T.: Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia, Elem Sci Anth, 5, 50, 2017.

Chang, K.-L., Cooper, O. R., West, J. J., Serre, M. L., Schultz, M. G., Lin, M., Marécal, V., Josse, B., Deushi, M., and Sudo, K.: A new method (M 3 Fusion v1) for combining observations and multiple model output for an improved estimate of the global surface ozone distribution, Geoscientific Model Development, 12, 955-978, 2019.

Chang, K. L., McDonald, B. C., and Cooper, O. R.: Surface ozone trend variability across the United States and the impact of heatwaves (1990-2023), EGUsphere, 2024, 1-43, 10.5194/egusphere-2024-3674, 2024.

DeLang, M. N., Becker, J. S., Chang, K. L., Serre, M. L., Cooper, O. R., Schultz, M. G., Schroder, S., Lu, X., Zhang, L., Deushi, M., Josse, B., Keller, C. A., Lamarque, J. F., Lin, M., Liu, J., Marecal, V., Strode, S. A., Sudo, K., Tilmes, S., Zhang, L., Cleland, S. E., Collins, E. L., Brauer, M., and West, J. J.: Mapping Yearly Fine Resolution Global Surface Ozone through the Bayesian Maximum Entropy Data Fusion of Observations and Model Output for 1990-2017, Environ Sci Technol, 55, 4389-4398, 10.1021/acs.est.0c07742, 2021.

Fleming, Z. L., Doherty, R. M., von Schneidemesser, E., Malley, C. S., Cooper, O. R., Pinto, J. P., Colette, A., Xu, X., Simpson, D., Schultz, M. G., Lefohn, A. S., Hamad, S., Moolla, R., Solberg, S., and Feng, Z.: Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health, Elementa: Science of the Anthropocene, 6, 10.1525/elementa.273, 2018.

Gaudel, A., Cooper, O. R., Ancellet, G., Barret, B., Boynard, A., Burrows, J. P., Clerbaux, C., Coheur, P.-F., Cuesta, J., and Cuevas, E.: Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation, Elem Sci Anth, 6, 39, 2018.

Gaudel, A., Cooper, O. R., Chang, K.-L., Bourgeois, I., Ziemke, J. R., Strode, S. A., Oman, L. D., Sellitto, P., Nédélec, P., Blot, R., Thouret, V., and Granier, C.: Aircraft observations since the 1990s reveal increases of tropospheric ozone at multiple locations across the Northern Hemisphere, Science Advances, 6, eaba8272, doi:10.1126/sciadv.aba8272, 2020.

Henze, D. K., Hakami, A., and Seinfeld, J. H.: Development of the adjoint of GEOS-Chem, Atmospheric Chemistry and Physics, 7, 2413-2433, 2007.

Huijnen, V., Miyazaki, K., Flemming, J., Inness, A., Sekiya, T., and Schultz, M. G.: An intercomparison of tropospheric ozone reanalysis products from CAMS, CAMS interim, TCR-1, and TCR-2, Geosci. Model Dev., 13, 1513-1544, 10.5194/gmd-13-1513-2020, 2020.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A. M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V. H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, Atmos. Chem. Phys., 19, 3515-3556, 10.5194/acp-19-3515-2019, 2019.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Dentener, F., Muntean, M., Pouliot, G., Keating, T., Zhang, Q., Kurokawa, J., and Wankmüller, R.: HTAP_v2. 2: a mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution, Atmospheric Chemistry and Physics, 15, 11411-11432, 2015.

Jones, D. B. A., Prates, L., Qu, Z., Cheng, W. Y. Y., Miyazaki, K., Inness, A., Kumar, R., Tang, X., Worden, H., Koren, G., and Huijnen, V.: Assessment of regional and interannual variations in tropospheric ozone in chemical reanalyses, 2024.

KOFFI, L. B., DENTENER, F., JANSSENS-MAENHOUT, G., GUIZZARDI, D., CRIPPA, M., DIEHL, T., GALMARINI, S., and SOLAZZO, E.: Hemispheric Transport of Air Pollution (HTAP): Specification of the HTAP2 experiments: Ensuring harmonized modelling, 2016.

Liu, X., Zhu, Y., Xue, L., Desai, A. R., and Wang, H.: Cluster-Enhanced Ensemble Learning for Mapping Global Monthly Surface Ozone From 2003 to 2019, Geophysical Research Letters, 49, e2022GL097947, https://doi.org/10.1029/2022GL097947, 2022.

Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., Gaughan, A. E., Nieves, J. J., Hornby, G., MacManus, K., Sinha, P., Bondarenko, M., Sorichetta, A., and Tatem, A. J.: Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets, Big Earth Data, 3, 108-139, 10.1080/20964471.2019.1625151, 2019.

Malashock, D., DeLang, M., Becker, J., Serre, M., West, J., Chang, K.-L., Cooper, O., and Anenberg, S.: Estimates of ozone concentrations and attributable mortality in urban, peri-urban and rural areas worldwide in 2019, Environmental Research Letters, 17, 10.1088/1748-9326/ac66f3, 2022.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., and Gomis, M.: Climate change 2021: the physical science basis, Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change, 2, 2391, 2021.

Mills, G., Sharps, K., Simpson, D., Pleijel, H., Frei, M., Burkey, K., Emberson, L., Uddling, J., Broberg, M., Feng, Z., Kobayashi, K., and Agrawal, M.: Closing the global ozone yield gap: Quantification and cobenefits for multistress tolerance, Global Change Biology, 24, 4869-4893, https://doi.org/10.1111/gcb.14381, 2018a.

Mills, G., Pleijel, H., Malley, C. S., Sinha, B., Cooper, O. R., Schultz, M. G., Neufeld, H. S., Simpson, D., Sharps, K., Feng, Z., Gerosa, G., Harmens, H., Kobayashi, K., Saxena, P., Paoletti, E., Sinha, V., and Xu, X.: Tropospheric Ozone Assessment Report: Present-day tropospheric ozone distribution and trends relevant to vegetation, Elementa: Science of the Anthropocene, 6, 10.1525/elementa.302, 2018b.

Miyazaki, K.: TROPESS Chemical Reanalysis Surface O3 2-Hourly 2-dimensional Product V1, Goddard Earth Sciences Data and Information Services Center (GES DISC) [dataset], 10.5067/NN87W53OVGUS, 2024.

Miyazaki, K., Bowman, K., Marchetti, Y., Montgomery, J., and Lu, S.: Drivers of regional surface ozone bias drivers in chemical reanalysis air quality revealed by explainable machine learning, 2024.

Miyazaki, K., Bowman, K. W., Yumimoto, K., Walker, T., and Sudo, K.: Evaluation of a multi-model, multi-constituent assimilation framework for tropospheric chemical reanalysis, Atmospheric Chemistry and Physics, 20, 931-967, 2020a.

Miyazaki, K., Eskes, H., Sudo, K., Takigawa, M., Van Weele, M., and Boersma, K.: Simultaneous assimilation of satellite NO 2, O 3, CO, and HNO 3 data for the analysis of tropospheric chemical composition and emissions, Atmospheric Chemistry and Physics, 12, 9545-9579, 2012.

Miyazaki, K., Sekiya, T., Fu, D., Bowman, K., Kulawik, S., Sudo, K., Walker, T., Kanaya, Y., Takigawa, M., and Ogochi, K.: Balance of emission and dynamical controls on ozone during the Korea-United States Air Quality campaign from multiconstituent satellite data assimilation, Journal of Geophysical Research: Atmospheres, 124, 387-413, 2019.

Miyazaki, K., Bowman, K., Sekiya, T., Eskes, H., Boersma, F., Worden, H., Livesey, N., Payne, V. H., Sudo, K., Kanaya, Y., Takigawa, M., and Ogochi, K.: Updated tropospheric chemistry reanalysis and emission estimates, TCR-2, for 2005–2018, Earth Syst. Sci. Data, 12, 2223-2259, 10.5194/essd-12-2223-2020, 2020b.

Murray, C. J., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., and Abdollahpour, I.: Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, The lancet, 396, 1223-1249, 2020.

Pfister, G., Wang, C.-T., Barth, M., Flocke, F., Vizuete, W., and Walters, S.: Chemical Characteristics and Ozone Production in the Northern Colorado Front Range, Journal of Geophysical Research: Atmospheres, 124, 13397-13419, https://doi.org/10.1029/2019JD030544, 2019.

Post, E. S., Grambsch, A., Weaver, C., Morefield, P., Huang, J., Leung, L. Y., Nolte, C. G., Adams, P., Liang, X. Z., Zhu, J. H., and Mahoney, H.: Variation in estimated ozone-related health impacts of climate change due to modeling choices and assumptions, Environ Health Perspect, 120, 1559-1564, 10.1289/ehp.1104271, 2012.

Punger, E. M. and West, J. J.: The effect of grid resolution on estimates of the burden of ozone and fine particulate matter on premature mortality in the USA, Air Quality, Atmosphere & Health, 6, 563-573, 2013.

Qu, Z., Daven, K. H., Owen, R. C., and Jessica, L. N.: Global (2°x2.5°) top-down NOx emissions from OMI NASA product (2005-2016) (V1), Harvard Dataverse [dataset], doi:10.7910/DVN/HVT1FO, 2020a.

Qu, Z., Henze, D. K., Cooper, O. R., and Neu, J. L.: Impacts of global NOx inversions on NO2 and ozone simulations, Atmos. Chem. Phys., 20, 13109-13130, 10.5194/acp-20-13109-2020, 2020b.

Schmedding, R., Rasool, Q. Z., Zhang, Y., Pye, H. O., Zhang, H., Chen, Y., Surratt, J. D., Lopez-Hilfiker, F. D., Thornton, J. A., and Goldstein, A. H.: Predicting secondary organic aerosol phase state and viscosity and its effect on multiphase chemistry in a regional-scale air quality model, Atmospheric chemistry and physics, 20, 8201-8225, 2020.

Schnell, J., Prather, M., Josse, B., Naik, V., Horowitz, L., Cameron-Smith, P., Bergmann, D., Zeng, G., Plummer, D., and Sudo, K.: Use of North American and European air quality networks to evaluate global chemistry–climate modeling of surface ozone, Atmospheric Chemistry and Physics, 15, 10581-10596, 2015.

Schnell, J. L. and Prather, M. J.: Co-occurrence of extremes in surface ozone, particulate matter, and temperature over eastern North America, Proceedings of the National Academy of Sciences, 114, 2854-2859, 2017.

Schröder, S., Schultz, M. G., Selke, N., Sun, J., Ahring, J., Mozaffari, A., Romberg, M., Epp, E., Lensing, M., Apweiler, S., Leufen, L. H., Betancourt, C., Hagemeier, B., and Rajveer, S.: TOAR Data Infrastructure, 2021.

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O. R., Galbally, I., Petropavlovskikh, I., Von Schneidemesser, E., Tanimoto, H., Elshorbany, Y., and Naja, M.: Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations, Elem Sci Anth, 5, 58, 2017.

Sekiya, T., Miyazaki, K., Ogochi, K., Sudo, K., Takigawa, M., Eskes, H., and Boersma, K. F.: Impacts of horizontal resolution on global data assimilation of satellite measurements for tropospheric chemistry analysis, Journal of Advances in Modeling Earth Systems, 13, e2020MS002180, 2021.

Sekiya, T., Emili, E., Miyazaki, K., Inness, A., Qu, Z., Pierce, R. B., Jones, D., Worden, H., Cheng, W. Y., and Huijnen, V.: Assessing the relative impacts of satellite ozone and its precursor observations to improve global tropospheric ozone analysis using multiple chemical reanalysis systems, EGUsphere, 2024, 1-35, 2024.

Sun, H., Shin, Y. M., Xia, M., Ke, S., Wan, M., Yuan, L., Guo, Y., and Archibald, A. T.: Spatial Resolved Surface Ozone with Urban and Rural Differentiation during 1990–2019: A Space–Time Bayesian Neural Network Downscaler, Environmental Science & Technology, 56, 7337-7349, 10.1021/acs.est.1c04797, 2022.

Sun, H. Z., van Daalen, K. R., Morawska, L., Guillas, S., Giorio, C., Di, Q., Kan, H., Loo, E. X.-L., Shek, L. P., Watts, N., Guo, Y., and Archibald, A. T.: An estimate of global cardiovascular mortality burden attributable to ambient ozone exposure reveals urban-rural environmental injustice, One Earth, 7, 1803-1819, 10.1016/j.oneear.2024.08.018, 2024.

Travis, K. R. and Jacob, D. J.: Systematic bias in evaluating chemical transport models with maximum daily 8 h average (MDA8) surface ozone for air quality applications: a case study with GEOS-Chem v9. 02, Geoscientific Model Development, 12, 3641-3648, 2019.

Turner, M. C., Jerrett, M., Pope, C. A., 3rd, Krewski, D., Gapstur, S. M., Diver, W. R., Beckerman, B. S., Marshall, J. D., Su, J., Crouse, D. L., and Burnett, R. T.: Long-Term Ozone Exposure and Mortality in a Large Prospective Study, Am J Respir Crit Care Med, 193, 1134-1142, 10.1164/rccm.201508-1633OC, 2016.

World-Health-Organization: WHO global air quality guidelines: particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, World Health Organization2021.