Using Regionalized Air Quality Model Performance and Bayesian Maximum Entropy data fusion to map global surface ozone concentration and associated uncertainty

Jacob S. Becker[1], Marissa N. DeLang[1], Kai-Lan Chang[2,3], Marc L. Serre[1], Owen R. Cooper[2,3], Martin G. Schultz[4], Sabine Schröder[4], Xiao Lu[5], Lin Zhang[5], Makoto Deushi[6], Beatrice Josse[7], Christoph A. Keller[8,9], Jean-François Lamarque[10], Meiyun Lin[11,12], Junhua Liu[8,9], Virginie Marécal[7], Sarah A. Strode[8,9], Kengo Sudo[13,14], Simone Tilmes[10], Li Zhang[11,12,15], Michael Brauer[16,17], J. Jason West[*,1]

[1]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA

[3]NOAA Chemical Sciences Laboratory, Boulder, CO, USA

[4]Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, Jülich, Germany

[5]Laboratory for Climate and Ocean-Atmosphere Studies, Department of Atmospheric and Oceanic Sciences, School of Physics, Peking University, Beijing, China

[6]Meteorological Research Institute (MRI), Tsukuba, Japan

[7]Centre National de Recherches Météorologiques, Université de Toulouse, Météo-France, CNRS, Toulouse, France

[8]NASA Goddard Space Flight Center, Greenbelt, MD, USA

[9]Universities Space Research Association, Columbia, MD, USA

[10]National Center for Atmospheric Research, Boulder, CO, USA

[11]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

[12]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA

[13]Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan

[14]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan

[15]Department of Meteorology and Atmospheric Science, Pennsylvania State University, University Park, PA. USA

[16]Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington, USA

[17]School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada

* jasonwest@unc.edu

Abstract

Estimates of ground level ozone concentrations have been improved through data fusion of observations and atmospheric models, but global estimates have been limited by model bias corrections that are linear and homogeneous within continents, including those we created for the Global Burden of Disease study. Here we use the Regionalized Air Quality Model Performance (RAMP) framework to correct model bias, accounting for the spatial inhomogeneity of bias and nonlinearity as a function of modeled ozone. RAMP bias correction is applied to a composite of nine global chemistry-climate models. Then estimates are fused with observations using the Bayesian Maximum Entropy (BME) framework, which matches observations at measurement stations and the influence of those observations declines with distance in space and time. We create global ozone maps for each year from 1990 to 2017 at fine spatial resolution. RAMP is shown to create unrealistic discontinuities due to the spatial clustering of ozone monitors, which we overcome by applying a weighting for RAMP based on the number of monitors nearby. While in BME the spatial influence of monitors is limited to a few degrees based on the spatial covariance, RAMP corrects biases over a much larger area. Incorporating RAMP before BME has little effect on model performance near stations, but strongly increases R^2 by 0.15 at locations without stations, shown through a checkerboard cross-validation. Corrections to estimates differ based on location in space and time, confirming heterogeneity. We use BME estimates of error variance to quantify the likelihood of exceeding selected ozone levels, finding that the western US, southern Europe, central Africa, the Middle East, India, and northern China are most likely to exceed 45 and 55 ppb in 2017. Our annual fine-resolution ozone estimates may be useful for several applications including epidemiology and assessments of impacts on health, agriculture and ecosystems.

1. Introduction

Ground level ozone is a pervasive air pollutant that detrimentally affects human health and plants. Ozone can cause a wide range of health problems in humans, and has been shown to cause premature mortality from daily exposures (Bell et al., 2004; Di et al., 2017; US EPA, 2020) and is likely to cause mortality from long-term chronic exposure (Jerrett et al., 2009; Turner et al., 2016; US EPA, 2020). Ozone concentrations above roughly 35 parts per billion (ppb) are associated with higher respiratory and cardiovascular mortality, with every 10 ppb increase increasing all-cause mortality by 2% (Turner et al., 2016). Ambient ozone is estimated to have caused about 365,000 deaths globally in 2019, or 0.65% of all global deaths (Murray et al., 2020). Unlike other air pollutants, ozone is purely secondary, created through photochemical reactions involving nitrogen oxides (NO_x), volatile organic compounds (VOCs), carbon monoxide, and methane in the atmosphere, in the presence of sunlight. Ozone concentrations are typically higher in the daytime and during summer months (NARSTO, 1999).

Understanding of ozone impacts on human health and plants has traditionally been limited in part by our understanding of ground level ozone distributions in space and time. Estimates of surface ozone distributions rely on monitoring station observations and chemical transport models, but both have limitations. While the US, Europe, and Japan have dense station networks that began prior to 1990 and China recently created a large network, station observations of ozone elsewhere are extremely limited (Schultz et al., 2017; Fleming et al., 2018; Lu et al., 2018). Models can help fill in these gaps in space and time, but have biases and overall are less accurate than observations (Cooper et al., 2014; Young et al., 2018).

In our previous work, we conducted a data fusion of ozone observations and multiple global atmospheric models, in two phases, to estimate global ground level ozone concentrations at fine spatial resolution (Chang et al., 2019; DeLang et al., 2021). These estimates of ozone were used to estimate global premature deaths from exposure to ambient ozone in the Global Burden of Diseases, Injuries, and Risk Factors (GBD) 2017 and 2019 Studies (Stanaway et al., 2018; Murray et al., 2020). GBD conducts a comparative risk assessment to estimate the global health burden caused by various risk factors from 1990 to the present.

Prior to GBD 2017, ozone in previous GBD studies was estimated solely by a single model with no observational bias correction (Brauer et al., 2016). The first global study to combine information from ozone observations and models developed and applied the M^{3}Fusion method to correct model bias, improving global ozone estimates from purely observation or model based approaches (Chang et al., 2019), in support of GBD 2017. M^{3}Fusion bias corrects and combines multiple chemical transport models by finding an optimal linear combination of models for each world region, using weighting based on performance when compared to available observations. The multimodel composite created was then corrected within two degrees of a monitoring station using a spatial interpolation of observations, creating fine resolution output.

We then improved on this using a novel combination of Bayesian Maximum Entropy (BME) along with M^{3}Fusion (DeLang et al., 2021), to support GBD 2019. BME is a framework for nonlinear geostatistics that performs the fusion of data from multiple sources (Serre and Christakos, 1999; Christakos et al., 2001; Christakos et al., 2004). BME used observations to correct the M^{3}Fusion multi-model composite smoothly in both space and time, so that ozone estimates match the observations at station locations. The influence each station exerts diminishes over time and space based on a calculated spatiotemporal covariance function. Before our study, BME had been used on smaller scales to fuse ozone observations and models (Christakos et al., 2004; de Nazelle et al., 2010; Xu et al., 2016), but not previously globally. The ability of observations in BME to influence estimates across time was also shown to be useful in informing earlier years before new stations were added (DeLang et al., 2021).

DeLang et al. (2021) showed that ozone estimates improved markedly over purely model- or observation-based estimates, and produced a global fine resolution (0.1°) dataset for each year from 1990 to 2017. However, like Chang et al. (2019), DeLang et al. (2021) rely on linear model bias corrections that are homogeneous across continents by M^{3}Fusion, where it does not correct based on a nearby observation. Previous research has shown that air pollution model performance and biases are non-homogenous (vary by location) and non-linear (vary as a non-linear function of the model estimate) (Reyes et al., 2017). For example, chemical transport models generally overpredict $PM_{2.5}$ when predicting high (>25 μg/m^3) values and underpredict elsewhere (Reyes et al., 2017).

Model errors for ozone stem from uncertainties in inputs, especially emissions of ozone precursors (NO_x and VOCs) from anthropogenic and natural sources, and in model processes including chemistry, model resolution, transport, and deposition (Young et al., 2018). Previous model evaluations have found that models have errors that vary by season and latitude (von Kuhlman et al., 2003), reflecting uncertainties in emissions inputs and in physical and chemical processes within the models. In short, we have imperfect knowledge of sources and sinks of ozone precursors. These errors could lead to overestimates in some locations and underestimates in others, indicating that model performance may be heterogenous (Liang and Jacobson, 2003). Ozone itself is also known to change non-linearly with emissions (NARSTO, 1999;

Cohan et al., 2005), and ozone model performance has also been shown to be non-linear with respect to ozone observations. The Community Multiscale Air Quality (CMAQ) model, for example, has been shown to overestimate maximum 8 hour average ozone levels where observations are below 35 ppb and underestimate where observations are above 85 ppb (Appel et al., 2007).

The goal of this work is to improve on the work of DeLang et al. (2021) to map global surface ozone concentrations each year from 1990 to 2017 at fine spatial resolution by adding a non-linear and heterogeneous bias correction using the Constant Air Quality Model Performance (CAMP) and Regionalized Air Quality Model Performance (RAMP) methods (Reyes et al, 2017), and evaluate the improvement in performance. CAMP and RAMP corrections are applied to the M^{3}Fusion multi-model composite prior to BME data fusion. While BME is not restricted spatially, ozone's steep covariance curve means that observations have little influence beyond one degree from an observation station (DeLang et al., 2021), while CAMP and RAMP corrections are not restricted by covariance. Specifically, we aim to use regional trends in model under/over estimation to correct the M^{3}Fusion results regionally and increase the fidelity of our estimation in areas with sparse or no ozone observation stations.

Both CAMP and RAMP bias correct models by comparing observed and modeled values at collocated points in space and time, and applying a non-homogenous, non-linear correction as a function of the modeled ozone concentration. CAMP assumes that model performance is constant across the study region, while RAMP improves on this by giving each model grid cell its own model bias correction based on nearby observations. Here, the RAMP method (Reyes et al., 2017) is applied globally for the first time, with each model grid cell being corrected based on a unique area that includes the nearest points in space/time. These areas are much smaller than the continental regions used in M^{3}Fusion, allowing us to better correct biases in the M^{3}Fusion multi-model composite at points far away from observations, while BME then applies corrections near them. In applying RAMP at a global scale, we also make a novel modification of the RAMP method because station observations are sparse in some regions and clustered in others. This modification prevents sharp spatial changes in corrections when transitioning between two different regions with dense observation stations. The CAMP and RAMP corrected estimates are then each used as global background ozone levels (the global offset) for BME data fusion with observations, as was done by DeLang et al. (2021) with the uncorrected M^{3}Fusion multimodel composite.

2. Methods

2.1 Ozone observations and model estimates

We use the ozone season daily maximum 8-hour mixing ratio (OSDMA8) as the annual ozone metric, as it is used for calculating health outcomes from ozone pollution by GBD (Murray et al., 2020) using concentration-mortality relationships from Turner et al. (2016). OSDMA8 is the maximum six-month running mean of monthly averages of the daily 8-hour maximum mixing ratios. Each defined year includes up to March of the following year to capture the Southern Hemisphere summer, as ozone is usually highest in the summer. All reported ozone values here, including observations, modeled values, and estimates are OSDMA8 values.

Ground level ozone measurements are taken from the Tropospheric Ozone Assessment Report (TOAR) and Chinese National Environmental Monitoring Center (CNEMC) (Schultz et al., 2017; Lu et al, 2018). The TOAR database is the largest collection of global hourly surface ozone concentrations and spans 1970-2015. To support this project, some national datasets were extended for 2015-2017 (DeLang et al., 2021). While observations are dense in North America, Europe, Japan, and South Korea, they are sparse to non-existent elsewhere (Figure 1). CNEMC provides 2013–2017 surface ozone observations in China (Lu et al., 2018). Both datasets were quality-controlled with the same algorithm developed for the TOAR database. The number of observation locations in the combined dataset is least in 1990 (with 1,190) while 2015 has the most (4,999).

We used surface concentration output from nine atmospheric chemistry model to create our M^{3}Fusion multi-model composite (Table S1). Models include four models from the Chemistry-Climate Model Initiative (CCMI; Morgenstern et al., 2017) that simulate 1990-2010, two additional CCMI models that extend the simulation beyond 2010, two CMIP6 models which cover years after 2010 (Collins et al., 2017), and MERRA2-GMI, which covers 1990-2017 (DeLang et al., 2021). The compilations of observations and models used here are the same as used by DeLang et al. (2021).

2.2. Data fusion methods

M^{3}Fusion was used to evaluate model performance and create a bias-corrected multi-model composite for each year 1990-2017 (Chang et al., 2019). This is the same composite used by DeLang et al. (2021). This method find a linear combination of the 9 models in each year and continental domain that minimizes the mean square error compared to interpolated observations, creating a single bias-corrected global composite for each year. However, M^{3}Fusion does not capture the non-linearity of model performance with respect to model value, nor how model performance varies within a continental domain (heterogeneity), both of which we address using RAMP.

The **Constant Air Quality Model Performance (CAMP)** method (de Nazelle et al., 2010) is a precursor to RAMP that bias corrects for non-linear model performance, but does not account for non-homogeneity as a single correction applies globally. It matches each observation point with the model estimate at that location. These matched pairs are then binned by the model estimate, and an average of model estimates and observations is set for each bin. The M^{3}Fusion composite is then corrected by interpolating between these values. Since CAMP is closely related to RAMP, we describe the method in depth in the next paragraphs. While CAMP works well for local applications in a single year, RAMP allows us to account for the heterogeneity in model performance we see at a global scale, by performing the model correction based on the nearest observations only.

Regionalized Air Quality Model Performance (RAMP) is a method to visualize and evaluate model performance, and can be used to bias correct models (Reyes et al., 2017). The correction accounts for non-linear and non-homogenous model performance (de Nazelle et al., 2010), in which the RAMP correction is not limited to a linear function with respect to model value, and it may correct differently in different geographic regions. Here we apply RAMP to the M^{3}Fusion composites so that we address residual non-linear and non-homogeneous biases. While previous studies have used RAMP to bias correct model estimates of air pollutants (de Nazelle et al., 2010; Xu et al., 2016; Reyes et al., 2017), none has done so at a global scale.

Let $\breve{y}(\boldsymbol{p})$ be the M^{3}Fusion multimodel composite prediction of ozone at space/time coordinate $\boldsymbol{p} = (\boldsymbol{s}, t)$, where $\boldsymbol{s}$ is the spatial location in longitude/latitude degrees, and t is time in years. Let $\hat{y}_i = \hat{y}(\boldsymbol{p}_i)$ be the ozone CENMC or TOAR observation at space/time monitoring points $\boldsymbol{p}_i$. M^{3}Fusion predictions are available throughout our entire global study domain, whereas observations are only available at certain locations. We pair each observation $\hat{y}_i$ with the underlying model prediction $\breve{y}_i = \breve{y}(\boldsymbol{p}_i)$, so that $(\hat{y}_i, \breve{y}_i)$ are the paired observation-model values. We let $\mathcal{R}(\boldsymbol{p})$ be the space/time region around $\boldsymbol{p}$ containing the *N*=250 spatially closest stations in years *t*, *t-1* and *t+1* (1990 does not use *t-1*, and 2017 does not use *t+1*). We chose 250 after trying other numbers as it was enough stations to maintain consistent patterns and prevent outliers from having significant effects, while giving a narrow enough spatial range to correlate with local trends. As we use three years, $\mathcal{R}(\boldsymbol{p})$ contains up to 750 collocated $(\hat{y}_i, \breve{y}_i)$ pairs. We sort these pairs by increasing model value and stratify them in 10 bins corresponding to increasing model decile values $\breve{y}_k,\ k = 1, \dots 10$. Then, we calculate the average observed value λ_1 for model decile value $\breve{y}_k$ in region $\mathcal{R}(\boldsymbol{p})$ as

$$\lambda_1(\breve{y}_k, \mathcal{R}(\boldsymbol{p})) = \frac{1}{n(\breve{y}_k, \mathcal{R}(\boldsymbol{p}))} \sum_{j=1}^{n(\breve{y}_k, \mathcal{R}(\boldsymbol{p}))} \hat{y}_j$$

where $n(\breve{y}_k, \mathcal{R}(\boldsymbol{p}))$ is the number of paired observed/modeled values $(\hat{y}_i, \breve{y}_i)$ for which $\breve{y}_i$ is in the k-th decile of modeled values, and $\hat{y}_j$ is the j-th observation in these pairs.

The above steps follow those outlined by Reyes et al. (2017); in this paper we further improve RAMP by ensuring that the slope between λ_1s does not become negative, or in other words, ensure the λ_1 RAMP curve for any $\mathcal{R}(\boldsymbol{p})$ is monotonically increasing. To do this, we define the mean value of all observed values $\hat{y}_i$ in $\mathcal{R}(\boldsymbol{p})$ as λ_{mean}. We compare λ_{mean} with $\lambda_1\left(\breve{y}_5, \mathcal{R}(\boldsymbol{p})\right)$**,** the $\lambda_1$in the 5th decile bin. If $\lambda_{mean} < \lambda_1\left(\breve{y}_5, \mathcal{R}(\boldsymbol{p})\right)$, we set $\lambda_1\left(\breve{y}_5, \mathcal{R}(\boldsymbol{p})\right) = \lambda_{mean}$. We then compare the 5th and 4th bin in the same way, and so on, ensuring that $\lambda_1\left(\breve{y}_k, \mathcal{R}(\boldsymbol{p})\right) \geq \lambda_1\left(\breve{y}_{k-1}, \mathcal{R}(\boldsymbol{p})\right)$**,** by setting them as equal when necessary. We do the same for bins k=6 through 10, first comparing bin 6 to λ_{mean} and setting the value of $\lambda_1\left(\breve{y}_6, \mathcal{R}(\boldsymbol{p})\right)$ equal to $\lambda_{mean}\ if\ \lambda_{mean} > \lambda_1\left(\breve{y}_6, \mathcal{R}(\boldsymbol{p})\right)$. This is a novel improvement to Reyes et al. (2017) as it maintains the ordinality of estimates from the original model with the same $\mathcal{R}(\boldsymbol{p})$.

By plotting $\lambda_1(\breve{y}_k, \mathcal{R}(\boldsymbol{p}))$ with respect to $\breve{y}_k$, we obtain the RAMP curve at location $\boldsymbol{p}$ showing how the average observation changes with respect to model value. Figure 2 visualizes the non-linear performance of the M^{3}Fusion composite, and by changing location $\boldsymbol{p}$, we can see how that performance varies across space and where it is non-linear. This visualization can, for example, be used to detect regions where the M^{3}Fusion prediction over-predicts high ozone values and under-predicts low ozone values. These plots also allow us to correct the model value by interpolating along $\lambda_1(\breve{y}_k, \mathcal{R}(\boldsymbol{p}))$ and selecting a new model values based on the value of λ_1 evaluated at the original $\breve{y}_k =$ M^{3}Fusion value. Therefore, the RAMP corrected model value is $\lambda_1(\breve{y}(\boldsymbol{p}_i), \mathcal{R}(\boldsymbol{p_i}))$.

A novel challenge posed by the implementation of the RAMP method at the global scale is that station locations are clustered in some countries or continents (e.g. the US, China, Japan, Europe), and are sparse in large areas in between. Previous applications of RAMP had more uniform distributions of observations (Reyes et al., 2017). As a result, globally the region $\mathcal{R}(\boldsymbol{p})$ containing the N=250 stations closest to $\boldsymbol{p}$ can change dramatically over a short distance, for example when shifting from a domain dominated by European observations to one dominated by China. This abrupt change in $\mathcal{R}(\boldsymbol{p})$ can result in a discontinuity in the RAMP corrected value $\lambda_1(\boldsymbol{p})$. To reduce this discontinuity, we introduce the RAMP-M^{3}Fusion weighted average $\lambda_1^*(\boldsymbol{p})$ calculated as

$$\lambda_1^*(\boldsymbol{p}) = w(\boldsymbol{p}) * \lambda_1(\boldsymbol{p}) + (1 - w(\boldsymbol{p})) * \breve{y}(\boldsymbol{p})$$

where $\lambda_1(\boldsymbol{p})$ and $\breve{y}(\boldsymbol{p})$ are the RAMP and M^{3}Fusion values, respectively, and $w(\boldsymbol{p})$ is the weight for RAMP at location $\boldsymbol{p}$. We want a weight that is high when a large fraction of the N stations used to construct the RAMP curve are close to $\boldsymbol{p}$ and low when this fraction is low. We therefore set the weight using the equation

$$w(\boldsymbol{p}) = \frac{N_q(\boldsymbol{p})}{N(\boldsymbol{p})}$$

where $N(\boldsymbol{p})$ is the number of stations used to calculate the RAMP curve at location $\boldsymbol{p}$ (250), $N_q(\boldsymbol{p})$ is the number of these stations that are within a radius q of $\boldsymbol{p}$, and $w(\boldsymbol{p})$ is the fraction of RAMP stations (between 0 and 1) that are within q degrees of $\boldsymbol{p}$ (Figure 3). We choose a radius q=25 degrees, so that the RAMP weight $w(\boldsymbol{p})$ allows RAMP to exert influence beyond the range of BME without extending into areas without representative observations. Areas that are more than 25 degrees away from these station clusters, like the area at the midpoint between China and Europe, will have a RAMP weight close to zero and a $\lambda_1^*(\boldsymbol{p}) \approx \breve{y}(\boldsymbol{p})$, thereby mitigating any RAMP discontinuity. We call the global output of λ_1^* values weighted RAMP, or wRAMP.

Bayesian Maximum Entropy (BME) data fusion is then applied after RAMP correction to fuse model prediction and observations, using the approach described by DeLang et al. (2021). Each BME estimate uses a different background assumption for global ozone levels at every grid cell, which we call the global offset, based on either the M^{3}Fusion composite, CAMP corrected M^{3}Fusion, or wRAMP corrected M^{3}Fusion. This global offset is corrected using BME so the final BME estimate matches observed values at each station location. Each station exerts an influence based on the difference between the station estimate and the global offset, which decreases as the space/time distance from observations increases, eventually matching the offset prediction away from observations. The rate at which this influence falls is based on a derived covariance function. BME has been used previously for the fusion of ozone observations and models (Christakos, 2000; Christakos et al., 2004; de Nazelle et al., 2010), though only once before at global scale (DeLang et al., 2021). While these papers provide the details of BME, we give here the main BME steps.

The fundamental step in BME data fusion is the definition of an offset function $o(\boldsymbol{p})$ at all points $\boldsymbol{p}$ across the study space/time domain. Here, we set $o(\boldsymbol{p})$ equal to either $\breve{y}(\boldsymbol{p})$ (M^{3}Fusion), $\lambda_1(\boldsymbol{p})$ (RAMP), or $\lambda_1^*(\boldsymbol{p})$ (weighted-RAMP). We calculate the offset-removed observations $\hat{x}_i$ as

$$\hat{x}_i = \hat{y}_i - o(\boldsymbol{p}_i), \qquad i = 1, \dots, n$$

where $\hat{y}_i = \hat{y}(\boldsymbol{p}_i)$ are the CENMC or TOAR observations at point $\boldsymbol{p}_i,\ i = 1, \dots, n$. We define $X(\boldsymbol{p})$ as a homogeneous/stationary Space/Time Random Field (S/TRF) with realizations $\hat{x}_i,\ i = 1, \dots, n$. $X(\boldsymbol{p})$ is a S/TRF representing the residual uncertainty and variability that is left in the offset-removed observations, and therefore its covariance function changes with the offset considered (either M^{3}Fusion, RAMP or weighted-RAMP). Finally, we define the S/TRF $Y(\boldsymbol{p})$ representing the ozone concentration as the sum of the residual field and the offset, i.e.

$$Y(\boldsymbol{p}) = X(\boldsymbol{p}) + o(\boldsymbol{p}).$$

We implement BME on the residual S/TRF $X(\boldsymbol{p})$ to obtain the BME estimate of $X(\boldsymbol{p}_k)$ at estimation points $\boldsymbol{p}_k$ across a global estimation grid. The general knowledge base characterizing $X(\boldsymbol{p})$ consists of a mean assumed zero within the estimation neighborhood, and a covariance function obtained from a variogram analysis (see supplementary information for details on the covariance model and its parameters). The site-specific knowledge consists in the offset-removed observations treated as hard data (data with no assumed uncertainty). We numerically implement BME using the BMElib library written in the MATLAB programming language (Serre and Christakos, 1999; Christakos et al., 2001), and as shown by DeLang et al. (2021), in this case the BME posterior pdf of $X(\boldsymbol{p}_k)$ is Gaussian with a mean $\tilde{x}_k$ equal to the simple kriging mean. Finally, the BME estimate $\tilde{y}_k$ of $Y(\boldsymbol{p}_k)$, representing ozone at the estimation point, is obtained as

$$\tilde{y}_k = \tilde{x}_k + o(\boldsymbol{p}_k),$$

where $o(\boldsymbol{p}_k)$ is the (M^{3}Fusion, RAMP or weighted-RAMP) offset at the estimation point. Estimation points are set on a 0.5 degree grid, giving a final BME estimation at 0.5 degree resolution.

2.3. Cross Validation

Leave One Out Cross Validation (LOOCV) was done by removing each observation one at a time and using various estimation methods to evaluate our ability to predict this observation based on the remaining data. LOOCV was performed by predicting ozone at each 0.5 degree grid cell containing an observation point, and comparing it with the observations ($\hat{y}(\boldsymbol{p}_i)$) in the grid cell. This was done for M^{3}Fusion, CAMP, and wRAMP both before and after data fusion with BME. For LOOCV of BME, BME was used to estimate each removed point in turn, and the aggregated errors were used to calculate R^2 and mean square error (MSE). For LOOCV on the offsets, the difference between the offset and observation point at each station location was used. We followed the same protocols for LOOCV as DeLang et al. (2021).

Whereas LOOCV tests the ability to predict based on nearby clustered observations, we use **Checkerboard Cross Validation (CBCV)** to better test each estimation method especially farther from nearby observations. This method is based on the radius-based validation methods of Xu et al. (2016) and

Cleland et al. (2020). In CBCV we create a "checkerboard" of boxes over the world with each box having a side length s latitude and side length $2*s$ longitude. For each box, we remove all observed values $\hat{y}(\boldsymbol{p}_i)$ within the box and use BME to re-estimate the ozone values at the location $y*(\boldsymbol{p}_i)$ of the removed observations within the box, using only observations outside of it. The validation error is defined as $e_h = \hat{y}(\boldsymbol{p}_i) - y^*(\boldsymbol{p}_i)$, which is then used to calculate R^2 values to quantify error for each observation in every box. We test CBCV with s ranging from 0.5 to 50 degrees. BME relies on observations to make corrections within the covariance range (1-2 degrees for ozone), so as s increases, observed values will have a smaller influence on correction. CBCV simulates the effect of sparse observations, while still having observations to validate the estimate. As most of the world does not have dense observation networks, the ability to correct away from observations is valuable to global estimations of air pollution.

3. Results

The M^3Fusion multimodel composite (Chang et al., 2019; DeLang et al. 2021) is used here as the basis for RAMP and CAMP corrections (Figure 4). We then bias correct the M^3Fusion composite with the RAMP method using TOAR and CNEMC observations (Figure 2). Using RAMP, we confirm that M^3Fusion bias varies at a finer scale than the continental regions used in M^3Fusion, supporting the value of RAMP's localized (non-homogenous) bias correction**.** While specific biases vary by region, some biases are more prevalent. M^3Fusion tends to overpredict ozone where it estimates high values and slightly underpredict low values, which is confirmed by CAMP (Figure S2). This has been demonstrated for individual models in previous studies of surface ozone (de Nazelle et al., 2010), but we are not aware that it has been demonstrated previously at a global level. We also find that model performance is not linear and tends to be better around the midrange of predicted values, again supporting the use of RAMP or CAMP for a non-linear approach to model correction.

We apply a RAMP bias correction to each model grid point, which results in a non-homogenous, non-linear correction. Corrections vary each year and at each location, but the largest changes generally occur where the M^3Fusion estimate is above about 55 ppb or below 35 ppb. While overall the M^3Fusion composite overestimates when it predicts high ozone, and underestimate where it predicts low, this is not true for all regions. Model bias was found to be non-homogenous and change based on space/time location. Figure 5a shows an area where the model consistently underpredicts ozone, and the RAMP correction has a steeper slope at high values. Figure 5b shows a nearby region in the same year where M^3Fusion overpredicts ozone at all but the lowest levels, and the ozone estimate at the specific point is lowered by the RAMP correction. Both Figures 5a and 5b are in the same correction region, showing that the M^{3}Fusion bias varies at finer spatial scales than continental. Figure 6 shows the heterogeneity in model performance and bias correction globally. While some areas like the Americas see primarily a RAMP correction in a single direction, others like northern Africa and eastern Europe have regions which are corrected upwards bordering regions corrected downward.

Figure 5b also shows a region where model performance is non-linear with respect to estimations, where the model overpredicts high values and underpredicts low values. Non-linearity is identifiable by an s-shaped λ_1 curve. In these areas, the M^{3}Fusion bias does not vary linearly with respect to the M^{3}Fusion estimate, and therefore our correction is not a linear function. This shows the value of the RAMP correction over a linear bias correction, as a linear correction could not replicate these non-linear curves. Figure 5c shows an example region where M^{3}Fusion consistently overestimates ozone. These RAMP curves show the trends in model performance in the region, as a function of modeled concentration, as well as correcting the individual points (the pink star) based on this evaluation.

At a global scale, RAMP creates “streaks” where the observations used to correct the model change from being dominated by one region (eastern Europe) to another far away region (Japan and South Korea) over a short spatial distance (Figure 7). This happens as there are no/few local observations for the RAMP correction in this area. Because of this, we weight RAMP (Figure 3) to allow a smooth transition between regions, using weights for RAMP and M^{3}Fusion that vary spatially and temporally (Figure 8). Weighted RAMP **(wRAMP)** heavily favors RAMP over M^{3}Fusion in areas with high density of observations stations, and RAMP maintains some influence up to 25 degrees from any station. This distance is long enough to give RAMP an influence in areas not reached by the BME correction, but short enough that it creates smooth transitions between regions and lessens the discontinuities seen in pure RAMP.

Using weighted RAMP as our global offset and station observations as hard data for BME, we obtain yearly estimates of global ozone and variance at 0.5-degree resolution (Figure 9). The ozone estimates match observations at any space/time location with an observation. The influence of observations decreases as a function space/time distance as the estimate moves further from an observation, based on the derived covariance (see supplemental equations). Temporally the influence of an observation over multiple years in BME is valuable in correcting areas with inconsistent observations. DeLang et al. (2021) explore the significance of the temporal factor in more detail. Variance is strongly influenced by proximity to observation stations, which are the only source of hard data in the BME estimate. Variance drops to 0 at stations and quickly rises as distance from stations increases. Therefore, variance is low in

Europe, North America, Japan, South Korea, and in some years parts of China, and high elsewhere. As variance approaches 60 ppb^2, the BME estimation approaches wRAMP.

Figure 10 shows the difference between our BME estimate using the weighted RAMP corrected model as our global offset and the BME estimate from DeLang et al. (2021) which uses the M^{3}Fusion composite as the global offset. The two methods differ most at distances more than 0.5 degrees from stations, but within the 25 degree bounds of wRAMP's influence. Whether RAMP increases or decreases estimates varies in time and space, and even nearby areas can have different signs of the correction. Changes in specific regions also vary year to year. General trends include decreases in the Korean peninsula, large changes in China once local data becomes available in 2014, overall increases in eastern China prior to 2014, increases in the northeastern United States in most years, slight increases in south eastern Europe, and overall better model performance in inland US and EU than on the coasts indicated by smaller corrections in those regions. The changes only extend 25 degrees from the nearest observation station, and are small in regions with few observation stations.

Evaluation and Cross Validation: We evaluate our results using leave one out cross validation (LOOCV) and checkerboard cross validation (CBCV). We test 7 scenarios using LOOCV:

- Simple Model Mean: an average of all models used in M^{3}Fusion where each is given equal weight
- M^{3}Fusion: multimodel composite of nine models using the M^{3}Fusion method
- CAMP: CAMP corrected M^{3}Fusion composite
- wRAMP: RAMP corrected M^{3}Fusion composite, weighted based on proximity to observations
- BME using M^{3}Fusion as Offset: BME data fusion using the M^{3}Fusion multimodel composite as the global offset, matching the results of DeLang et al. (2021)
- BME using CAMP as Offset: BME data fusion using CAMP as the global offset
- BME using wRAMP as Offset: BME data fusion using wRAMP as the global offset

Each scenario shows improved performance over the prior, with BME methods showing the same prediction capability in LOOCV (Table 1). CAMP and wRAMP provide clear improvements to M^{3}Fusion in estimating global ozone. The lack of a difference between BME methods is because most observations are clustered, and BME predicts accurately when observations are close together, similar to kriging on the observations.

We use CBCV to test each method's ability to estimate ozone where there is not a dense network of observations (Figure 11). At small boxes sizes, CBCV approximates LOOCV and all BME methods have similar R^2 (though a smaller R^2 than in LOOCV since CBCV removes observations in all years). As the box size increases, R^2 for wRAMP decreases less than the other cases, maintaining a minimum of 0.45. It also does not experience the dramatic performance drop-off that CAMP and M^3Fusion have at 4 degrees. CAMP also has consistently higher R^2 than M^3Fusion at box sizes greater than 4 degrees. BME with wRAMP shows great improvement in estimating where observations have been removed compared to the base M^3Fusion BME estimates. This indicates that it effectively captures local trends in model bias in regions where it has observations.

Using BME variance and estimations, we evaluate the likelihood that ozone values exceed specific thresholds. Specifically, using our best estimates using BME with wRAMP as the global offset, we analyze the likelihood to surpass 35, 45, 55, and 65 ppb (Figure 12). For reference, Turner et al. (2016) found that that for every 10 ppb increase over about 35 ppb, the risk of all-cause mortality increases by 2%, circulatory mortality by 3%, and respiratory mortality by 12%. Note that we do not estimate the likelihood of exceeding health-based standards, which are typically expressed for daily 8-hr. maximum ozone rather than OSDMA8. Areas with ozone estimates near the threshold and areas with high variance (few observations) are most likely to fall in the uncertain range. Certainty in exceeding or not exceeding a given value comes from extreme estimates and/or dense observations. For example, very few parts of the world are definitively below 45 ppb in 2017, but only areas with high estimations (central Africa, India, the Middle East and parts of China) and areas with dense sensor networks (EU and western US) are definitively above it. Similarly, comparing the likelihood of exceedance with our ozone estimates, we see some areas which have the same level of estimated ozone but have different likelihood of exceeding thresholds due to the difference in nearby observations (and therefore variance). For example, the hotspots in southern Africa are estimated to be above 65 ppb, but we are less than 90% certain that this area exceeds 55 ppb. Meanwhile the hotspot centered around Beijing, which has nearby observations, is above 55ppb with near certainty, and even above 65 ppb with 90% certainty in some areas.

Finally, following DeLang et al. (2021) we use global population data from GBD 2019 to analyze annual population weighted ozone in different regions (Figure S3). We use 2019 population data for all years, so all changes are due purely to ozone changes, not population changes. Trends in regions and years with sparse observations are less certain. Although there are small differences in individual years and regions, trends overall follow the same pattern as for DeLang et al. (2021). While results here differ from those of DeLang et al. (2021) far from monitors, they do not differ exactly at a monitoring location (Figure 10). Asia has a large upward trend, which along with a large increase in Africa drives an overall upward global trend in population weighted ozone. North America and Europe trend downward, though the European trend is much weaker. Russia begins to trend down in 2010, while South America and Oceania fluctuate but have no clear time trend. TOAR observational studies support the downward trend in North America 2000-2014 (Chang et al., 2017), and a study of CNEMC observations supports the increase in East Asia based on observational trends in China (Lu et al., 2018).

4. Analysis and Discussion

Here we improve upon the global ozone estimates of Chang et al. (2019) and DeLang et al. (2021) by providing an additional bias correction step to the M^{3}Fusion model composite before BME data fusion. This RAMP correction provides a local, non-linear, non-homogenous model bias correction, in which each point receives a different bias correction based on the M^{3}Fusion composite performance at the nearest stations, which leads to more accurate predictions of ozone when there are not nearby ozone stations. Using this corrected model as the basis for BME data fusion leads to improvements when simulating sparse observation station coverage, which are the areas BME provides the least certainty for. We found that performance of M^{3}Fusion varies by space/time location and is often nonlinear, making RAMP the ideal tool to further improve this composite. This method also takes full advantage of TOAR and CNEMC observations, as it allows them to both directly correct estimates locally through time with BME data fusion and inform model corrections at a larger regional scale through M^{3}Fusion and RAMP. Our final estimates provide yearly fine resolution global ozone estimates for 1990-2017, involving a data fusion of surface observations from global monitoring stations and nine chemistry-climate models.

The RAMP method demonstrates that model performance and biases have local variations, even after a uniform continental bias correction is applied in the M^{3}Fusion multi-model composite. RAMP therefore improves estimates over M^{3}Fusion or the global CAMP in accounting for heterogeneous model performance. RAMP also shows that model performance is non-linear with respect to observations in many areas, which often manifests as an overprediction at one extreme and an underprediction at the other. Overall, the multi-model composite is better at estimating ozone values near the average (often 40-55 ppb) and poorer at the extremes. RAMP's ability to account for non-linear model performance allows greater corrections where M^{3}Fusion predictions are worse.

As this is the first application of RAMP at a global scale, we find that RAMP alone creates “streaks” where the observations being used to inform the correction change over a short distance, showing the difficulty of using a single method over areas both rich and sparse in data. RAMP could potentially encounter this issue for any dataset where there are two or more heavily clustered regions of observations separated by areas with sparse observations. Therefore, we chose to weigh RAMP to create smooth transitions between regions, giving a much greater weight to the multi-model composite when corrections far from observations, while close to observations corrections are incorporated by BME data fusion. In areas with a more uniform distribution of observations, such as those from previous studies using RAMP (Xu et al., 2016; Reyes et al., 2017), weighing RAMP would not be necessary. Weighing RAMP by distance from observations preserves the correction and avoids such streaks. It also allows a smooth transition between RAMP, where observations support a regional model correction, and M^3Fusion, which bias corrects within each continent.

Overall, RAMP is more accurate than CAMP and M^3Fusion at estimating global ozone. When used in conjunction with BME, RAMP does not appreciably improve estimates in LOOCV and within one degree of another station. BME alone can correct the model within the range where observations co-vary with each other, especially if it can draw on observations at the same location in other years. The advantage of RAMP is seen in the CBCV, where there are few nearby observations. The improvements CAMP gives over M^3Fusion shows the value of a non-linear model correction alone, while RAMP's improvements over CAMP show the value of accounting for regional heterogeneity in model performance.

Because BME method provides both ozone estimates and the associated variance, we can evaluate the confidence that ozone is above or below selected values. We find that most of the world's population lives in areas very likely above 35 ppb in OSDMA8, and even above 45 ppb. Some regions estimated to have the highest ozone, including much of India, are very likely above 55 ppb. In the case of India, model estimates suggest high ozone that may be above 65 ppb, but the lack of ground observations decreases confidence in these regions. Regions with high modeled ozone but low confidence in results because of the distance from observations can be among those prioritized for increased monitoring. While RAMP improves estimation far from monitors, additional monitoring capacity in regions currently lacking monitors would be valuable for improving ozone estimates, and for better evaluation of chemistry-climate models, particularly the emission inventories input to global models. Currently much of the world's population lives far from ozone monitors in low- and middle-income nations, and the likely severity of ozone in these regions, the large populations exposed, and the fact that ozone is often growing fastest in these regions increases the urgency for expanding observations in these regions.

References

Appel, KW, Gilliland, AB, Sarwar, G, Gilliam, RC. 2007. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance: Part I—Ozone. *Atmospheric Environment* **41:** 9603–9615.

Bell, ML, McDermott, A, Zeger, SL, Samet, JM, Dominici, F. 2004. Ozone and short-term mortality in 95 US urban communities, 1987-2000. *Journal of the American Medical Association* **292**: 2372-2378.

Brauer, M. et al. 2016. Ambient air pollution exposure estimation for the Global Burden of Disease 2013. *Environmental Science & Technology* 50: 79–88.

Chang, K-L, Cooper, OR, West, JJ, Serre, ML, Schultz, MG, Lin, M, Marecal, V, Josse, B, Deushi, M, Sudo, K, Liu, J, Keller, CA. 2019. A new method (M^3Fusion v1) for combining observations and multiple model output for an improved estimate of the global surface ozone distribution. *Geoscientific Model Development*. **12**: 955-978, doi: 10.5194/gmd-12-955-2019.

Chang, K-L, Petropavlovskikh, I, Cooper, OR, Schultz, MG, Wang, T. 2017. Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia. *Elementa: Science of the Anthropocene* 5.

Christakos, G. 2000. *Modern Spatiotemporal Geostatistics*, Oxford University Press.

Christakos, G, Kolovos, A, Serre, ML, Vukovich, F. 2004. Total ozone mapping by integrating databases from remote sensing instruments and empirical models. *IEEE Transactions on Geoscience and Remote Sensing* **42**: 991–1008.

Christakos, G, Bogaert, P, Serre, M. 2001. *Temporal GIS: Advanced Functions for Field-Based Applications*. (Springer-Verlag). doi: 10.1007/978-3-642-56540-3.

Cleland, SE, West, JJ, Jia, Y, Reid, S, Raffuse, S, O'Neill, S, Serre, ML. 2020. Estimating wildfire smoke concentrations during the October 2017 California fires through BME space/time data fusion of observed, modeled, and satellite-derived $PM_{2.5}$. *Environmental Science & Technology* **54**:13439-13447, doi: 10.1021/acs.est.0c03761.

Cohan, DS, Hakami, A, Hu, Y, Russell, AG. 2005. Nonlinear response of ozone to emissions: source apportionment and sensitivity analysis. *Environmental Science & Technology* **39**: 6739–6748.

Collins, WJ, Lamarque, J-F, Schulz, M, Boucher, O, Eyring, V, Hegglin, MI, Maycock, A, Myhre, G, Prather, M, Shindell, D, Smith SJ. 2017. AerChemMIP: Quantifying the effects of chemistry and aerosols in CMIP6, *Geoscientific Model Development* **10**: 585–607, doi: 10.5194/gmd-10-585-2017.

Cooper, OR, et al. 2014. Global distribution and trends of tropospheric ozone: An observation-based review. *Elementa: Science of the Anthropocene* 2.

DeLang, MN, Becker, JS, Chang, K-L, Serre, ML, Cooper, OR, Schultz, MG, Schroder, S, Lu, X, Zhang, L, Deushi, M, Josse, B, Keller, CA, Lamarque, J-F, Lin, M, Liu, J, Marecal, V, Strode, SA, Sudo, K, Tilmes, S, Zhang, L, Cleland, S, Collins, E, Brauer, M, West JJ. 2021. Mapping yearly fine resolution global surface ozone through the Bayesian Maximum Entropy data fusion of observations and model output for 1990-2017. *Environmental Science & Technology* **55**: 4389-4398, doi: 10.1021/acs.est.0c07742.

de Nazelle, A, Arunachalam, S, Serre, ML. 2010. Bayesian Maximum Entropy integration of ozone observations and model predictions: an application for attainment demonstration in North Carolina. *Environmental Science & Technology* **55**: 5707–5713.

Di, Q, Dai, L, Wang, Y, Zanobetti, A, Choirat, C, Schwartz, JD, Dominici, F. 2017. Association of short-term exposure to air pollution with mortality in older adults, *Journal of the American Medical Association*, **318**: 2446-2456. doi: 10.1001/jama.2017.17923.

Fleming, ZL, et al. 2018. Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health. *Elementa: Science of the Anthropocene* 6.

Jerrett, M, Burnett, RT, Pope, CA, Ito, K, Thurston, G, Krewski, D, Shi, Y, Calle, E, Thun, M. 2009. Long-term ozone exposure and mortality. *New England Journal of Medicine*, **360**: 1085–1095, doi: 10.1056/NEJMoa0803894.

Liang, J, Jacobson, MZ. 2000. Effects of subgrid segregation on ozone production efficiency in a chemical model. *Atmospheric Environment* **34**: 2975–2982.

Lu, X, Zhang, L, Wang, X, Gao, M, Li, K, Zhang, Y, Yue, X, Zhang, Y. 2020. Rapid increases in warm-season surface ozone and resulting health impact in China since 2013. *Environmental Science & Technology Letters* **7**: 240-247.

Morgenstern, O, Hegglin, M, Rozanov, E, O’Connor, F, Abraham, NL, Akiyoshi, H, Archibald, A, et al. 2017. Review of the global models used within Phase 1 of the Chemistry-Climate Model Initiative (CCMI). *Geoscientific Model Development* **10**: 639–671. https://doi.org/10.5194/gmd-10-639-2017.

Murray, CJL, et al. 2020. 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, doi: 10.1016/S0140-6736(20)30752-2.

North American Research Strategy for Tropospheric Ozone (1999) *An Assessment of Tropospheric Ozone Pollution: A North American Perspective*, cdiac.ess-dive.lbl.gov/programs/NARSTO/ozone_assessment.html.

Reyes, JM (2016) Geostatistical data fusion estimation methods of ambient $PM_{2.5}$ and polycyclic aromatic hydrocarbons. UNC Dissertation.

Reyes, JM, Xu, Y, Vizuete, W, Serre, ML. 2017. Regionalized $PM_{2.5}$ Community Multiscale Air Quality model performance evaluation across a continuous spatiotemporal domain. *Atmospheric Environment*, 148, 258–265.

Schultz, MG, Schröder, S, Lyapina, O, Cooper, O, Galbally, I, Petropavlovskikh, I, Von Schneidemesser, E, Tanimoto, H, Elshorbany, Y, et al. 2017. Tropospheric Ozone Assessment Report: database and metrics data of global surface ozone observations. *Elementa: Science of the Anthropocene* **5**: 58, doi: 10.1525/elementa.244.

Serre, ML, Christakos, G. 1999. Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge – the Equus Beds study. *Stochastic Environmental Research and Risk Assessment* **13**: 1–26.

Stanaway, JD, et al. 2018. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **392**: 1923–1994.

Turner, MC, *et al.* 2016. Long-term ozone exposure and mortality in a large prospective study. *American Journal of Respiratory and Critical Care Medicine* **193**: 1134–1142.

U.S. EPA. 2020. *Integrated Science Assessment for Ozone and Related Photochemical Oxidants*; U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-20/012.

von Kuhlmann, R, Lawrence, MG, Crutzen, PJ, Rasch, PJ. 2003. A model for studies of tropospheric ozone and nonmethane hydrocarbons: Model evaluation of ozone-related species. *Journal of Geophysical Research: Atmospheres* **108**.

Xu, Y, Serre, ML, Reyes, J, Vizuete, W. 2016. Bayesian Maximum Entropy integration of ozone observations and model predictions: a national application. *Environmental Science & Technology* **50**: 4393–4400.

Young, PJ, et al. 2018. Tropospheric Ozone Assessment Report: assessment of global-scale model performance for global and regional ozone distributions, variability, and trends. *Elementa: Science of the Anthropocene* 6.

Contributions

Contributed to conception and design: JSB, MND, KLC, MLS, ORC, MB, JJW

Contributed to acquisition of data: JSB, MND, KLC, ORC, MGS, SS, XL, LZ, MD, BJ, CAK, JFL, ML, JL, VM, SAS, KS, ST, LZ

Contributed to analysis and interpretation of data: JSB, MND, KLC, MLS, ORC, JJW

Drafted and/or revised the article: JSB, MND, KLC, MLS, ORC, JJW

Approved the submitted version for publication: All authors

Acknowledgments

None

Funding Information

We acknowledge funding from the NASA Health and Air Quality Applied Sciences Team (NNX16AQ30G) and the National Institute for Occupational Safety and Health (T42-OH008673). MD was supported by the Japan Society for the Promotion of Science (JP20K04070).

Competing interests

The authors have declared that no competing interests exist.

Supplemental material

The authors have declared that no competing interests exist.

Data accessibility statement

Input data from TOAR is publicly available through the TOAR web interface (https://join.fz-juelich.de/accounts/login/) and through PANGAEA (https://doi.pangaea.de/10.1594/PANGAEA.876108). Modeling output from CCMI used as input to this data fusion project can be obtained from CCMI (http://blogs.reading.ac.uk/ccmi/). Codes for the M^{3}Fusion method to create a multi-model composite were made available through the journal publication by Chang et al. (2019). BME codes are publicly available through the BME library (https://mserre.sph.unc.edu/BMElib_web/index.htm). Files containing full gridded results for our annual ozone maps produced by data fusion in this project are available from the authors upon request.

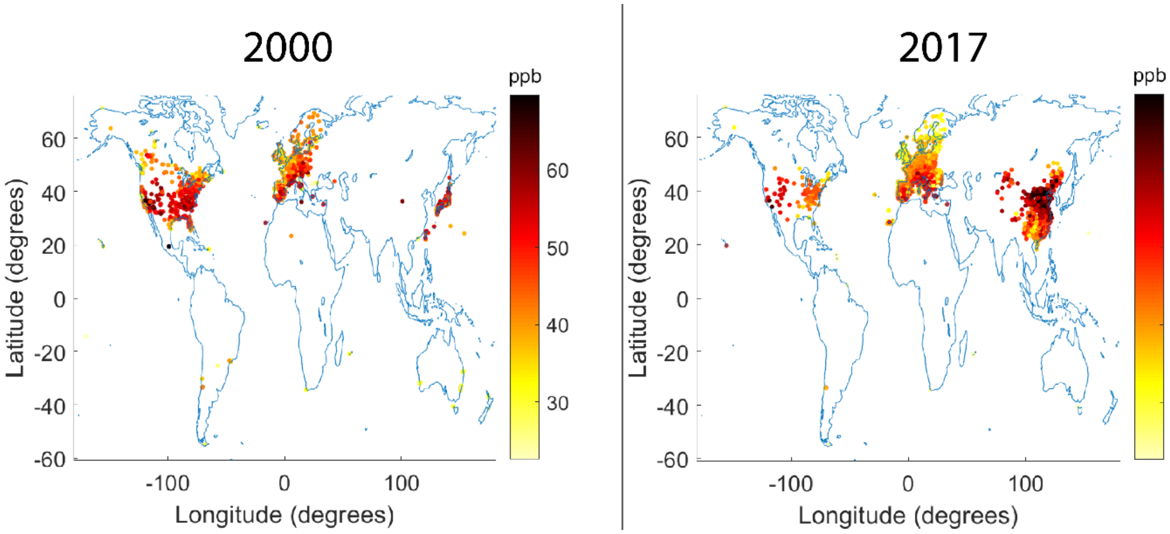

Figure 1. TOAR and CNEMC ozone observations. Ozone observations are shown for selected years 2000 and 2017.

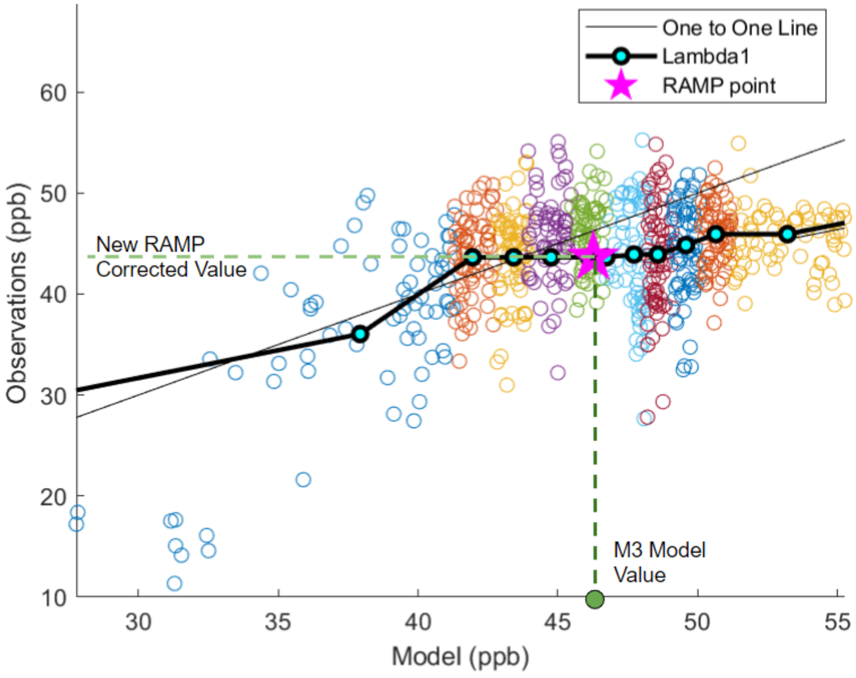

Figure 2. A visualization of the RAMP correction at a single point in North America for 2015. Three years of data (2014, 2015, and 2016) from the 250 nearest observation locations are used. Paired M^{3}Fusion/observation points are divided into deciles by the model value, and the M^{3}Fusion estimate at this gridpoint (x-axis) is corrected with RAMP to a new value (y-axis) using the $\boldsymbol{\lambda_1}$ line.

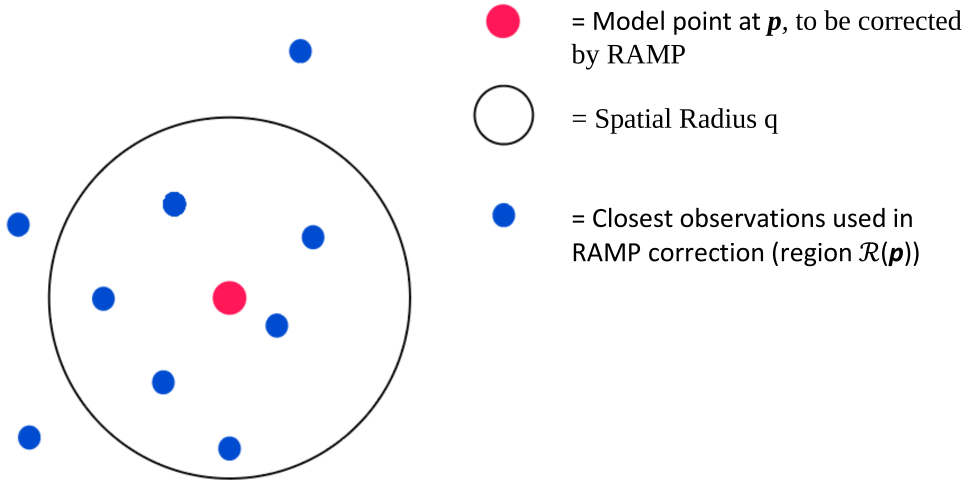
= Model point at $\boldsymbol{p}$, to be corrected by RAMP
= Spatial Radius q
= Closest observations used in RAMP correction (region $\mathcal{R}(\boldsymbol{p})$)

Figure 3. An example of RAMP weight at a model point. In this case $\boldsymbol{N_q(p) = 6}$**,** as 6 observations are within radius q. The weighted RAMP estimate at this location would be 6/250 times the RAMP-corrected composite value plus 244/250 times the M^{3}Fusion composite value without RAMP correction.

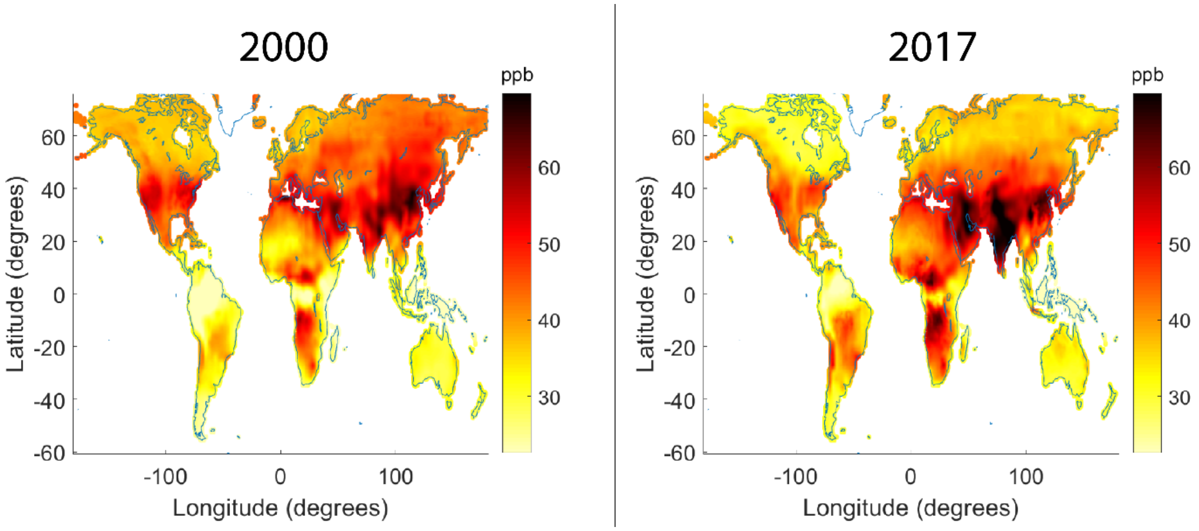

Figure 4. M^{3}Fusion multi-model composites expressed as OSDMA8. Composites are shown for 2000 (A) and 2017 (B). The 9 global models in Table S1 are linearly combined to minimize the difference with interpolated observations in each continental region. These composites are used as input to RAMP and BME here and are the same used by DeLang et al. (2021).

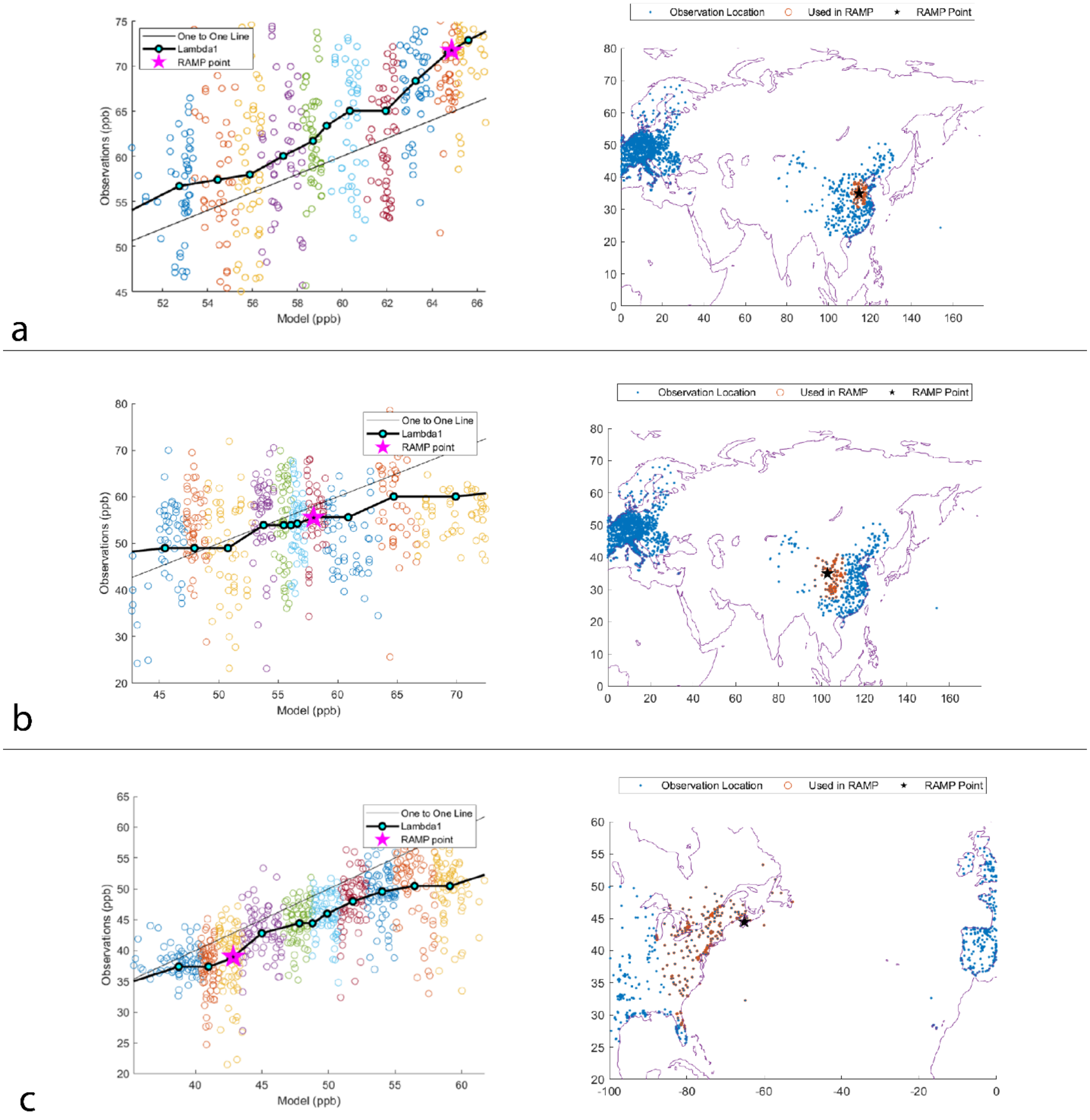

Figure 5. Examples of RAMP correction at three specific locations and years. The RAMP curve shows paired M^{3}Fusion composite values $\breve{\boldsymbol{y}}(\boldsymbol{p})$ and observations $\breve{\boldsymbol{y}}_{\boldsymbol{i}}$, and the locations of the selected points the nearest 250 observation locations is also shown. The RAMP corrected value $\boldsymbol{\lambda}_{\boldsymbol{1}}(\boldsymbol{p})$, the star, is estimated by replacing the M^{3}Fusion prediction $\breve{\boldsymbol{y}}(\boldsymbol{p})$ with its RAMP corrected value, i.e. $\boldsymbol{\lambda}_{\boldsymbol{1}}(\boldsymbol{p}) = \boldsymbol{\lambda}_{\boldsymbol{1}}(\breve{\boldsymbol{y}}(\boldsymbol{p})\,,\boldsymbol{\mathcal{R}}(\boldsymbol{p}))$ Each colored circle is a paired model/observation value $\boldsymbol{(\hat{\breve{y}}_{i},)}$ with the colors denoting which bin it falls into. If $\boldsymbol{\lambda}_{\boldsymbol{1}}(\boldsymbol{p})$ is below the 1:1 line, it indicates that M^{3}Fusion overpredicts ozone. Panel a shows an increase in estimation due to RAMP, b shows a non-linear correction and c shows a decrease in estimation.

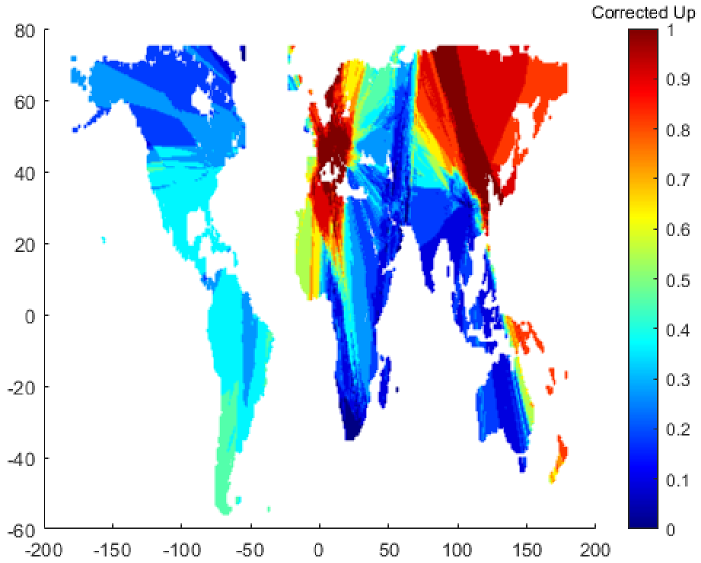

Figure 6. Fraction of λ_1 points above the one to one line in 2017. A higher number indicates a higher likelihood that a model point in this location would be increased with a RAMP correction, while a number less than 0.5 indicates a greater chance that RAMP lowers the M^{3}Fusion estimate. Results vary geographically, showing that the performance of the M^{3}Fusion composite is heterogeneous spatially, in some places varying strongly over a short distance.

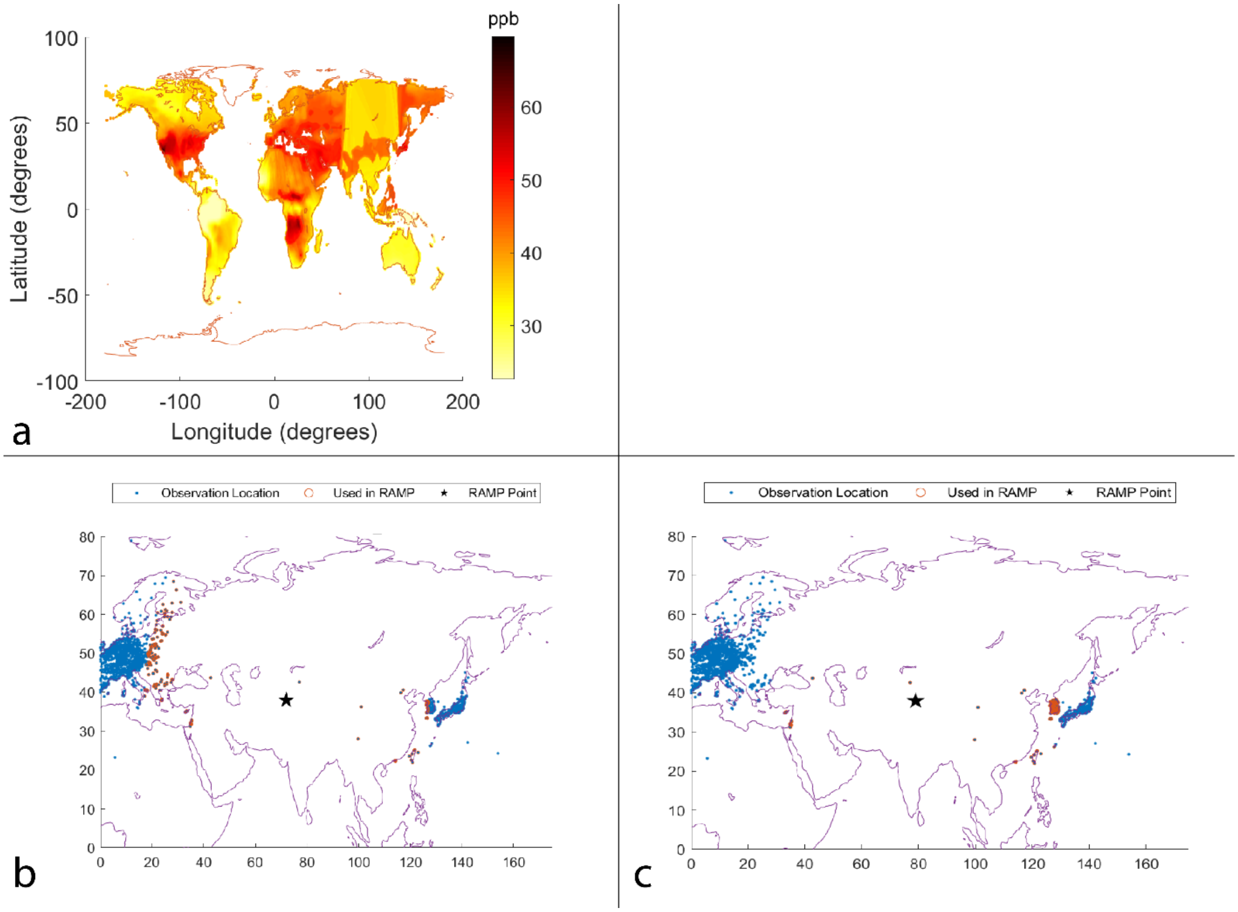

Figure 7. RAMP corrected estimate of OSDMA8 ozone for 2005 with a streak in central Asia. Applying a RAMP correction to the M^{3}Fusion multimodel composite produces discontinuities that appear as streaks in central Asia (a). Examination of which observations are used for RAMP correction ($\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ in $\boldsymbol{\mathcal{R}}(\boldsymbol{p})$) at two nearby model points ($\tilde{\boldsymbol{y}}(\boldsymbol{p})$) (b and c) shows a large shift in points comprising $\boldsymbol{\mathcal{R}}(\boldsymbol{p})$**,** which causes these large changes over short spatial ranges. Weighting RAMP prevents this from occurring, as areas far from the stations used for RAMP corrections will default to the M^{3}Fusion composite.

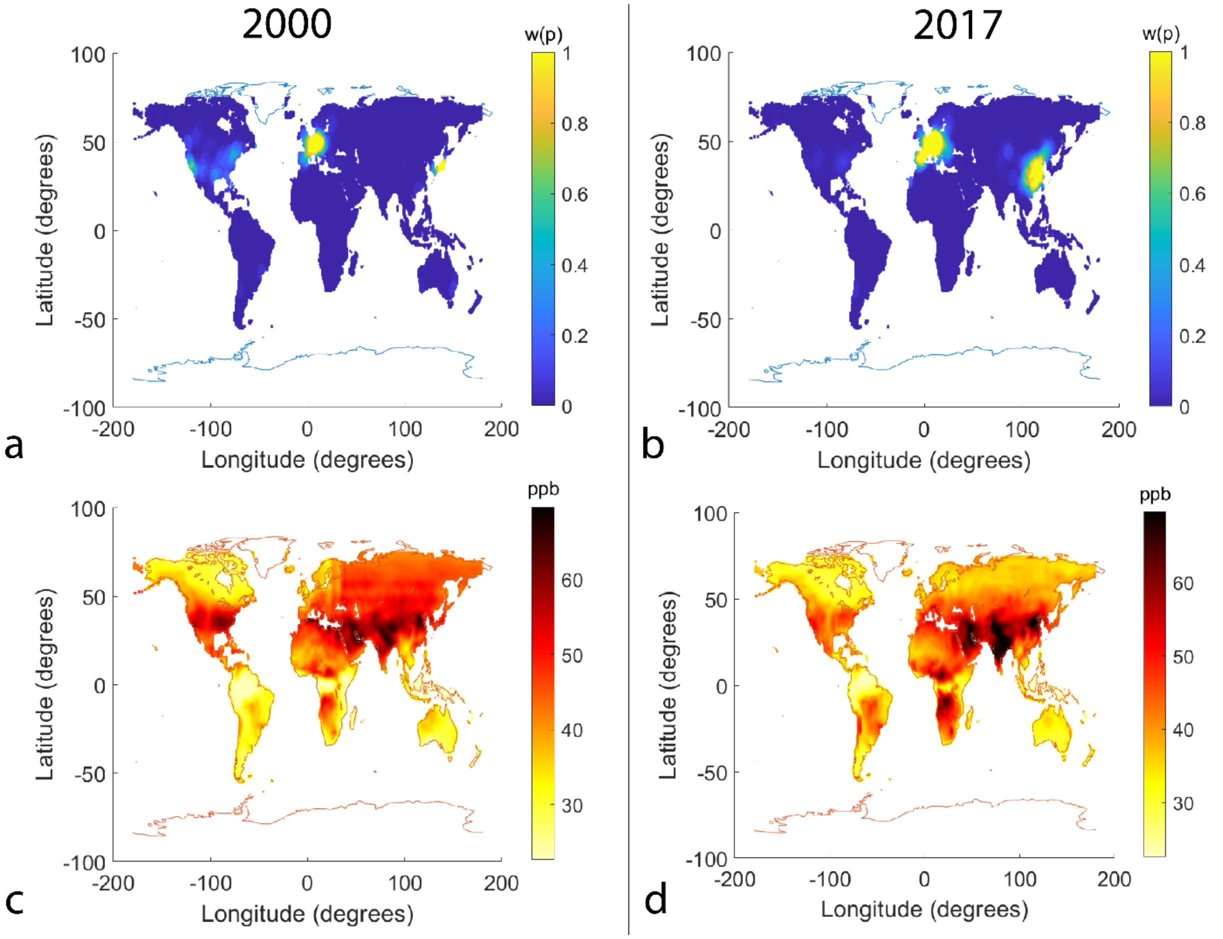

Figure 8. Weights applied in weighted RAMP and weighted RAMP ozone estimates. The weighs applied in weighted RAMP, $\boldsymbol{w(p)}$, for 2000 (A) and 2017 (B) corresponds to the percent of the estimate at that location that is based on RAMP, with $\mathbf{1}-\boldsymbol{w(p)}$ being the weight applied to the M^{3}Fusion composite. Weighted RAMP ozone estimates, for 2000 (C) and 2017 (D) show that the streaks such as those shown in Figure 7 are eliminated, as those areas have little or no RAMP weight.

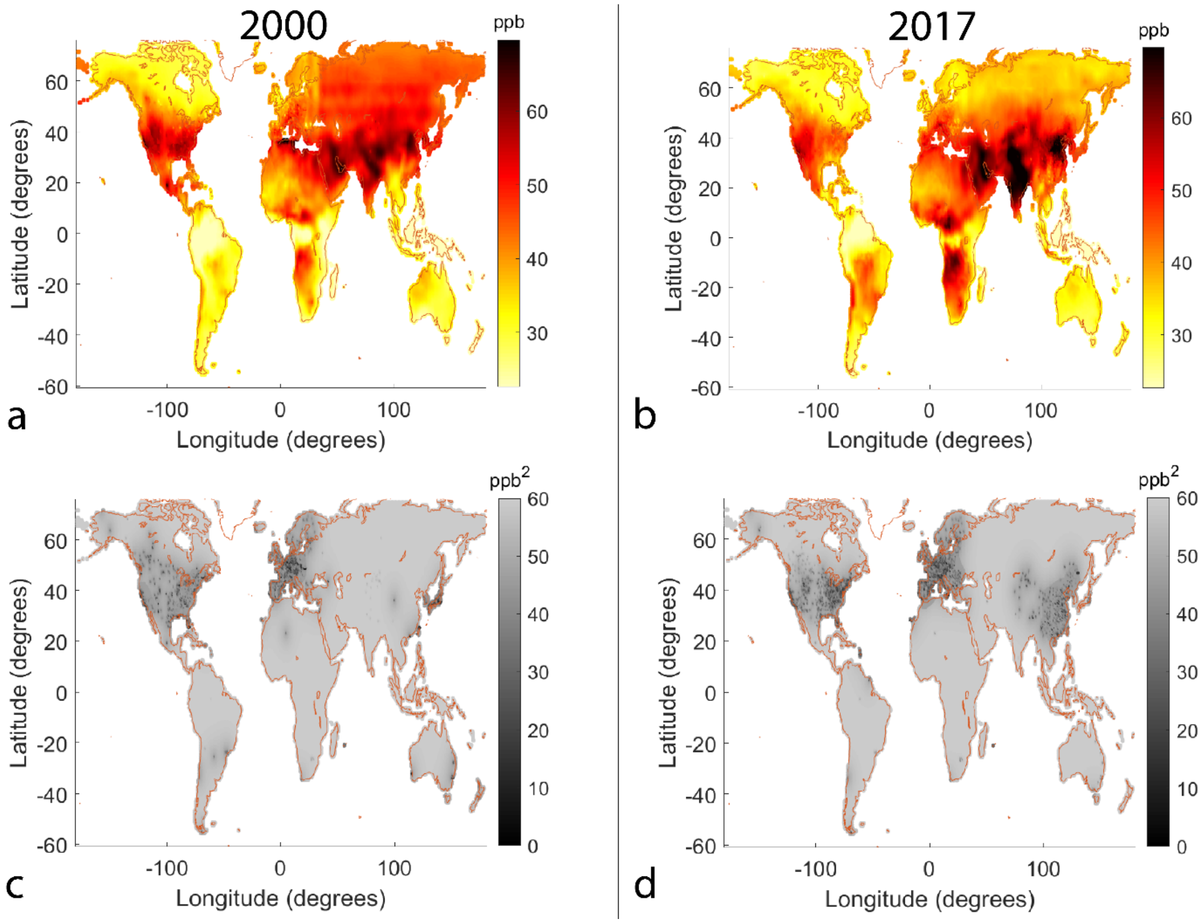

Figure 9. Final BME estimates of OSDMA8 ozone and variance in 2000 and 2017. BME estimates for 2000 (A) and 2017 (B) are obtained using the multimodel composite bias corrected with weighted RAMP (Figure 8) as the global offset in BME data fusion. Ozone values match observations (Figure 1) at each station location, with an observational influence that decreases as space/time distance from the observations increases. The variance of BME estimates for 2000 (C) and 2017 (D) is obtained as a function of spatial/temporal distance from observation locations. Variance is zero at any observation location.

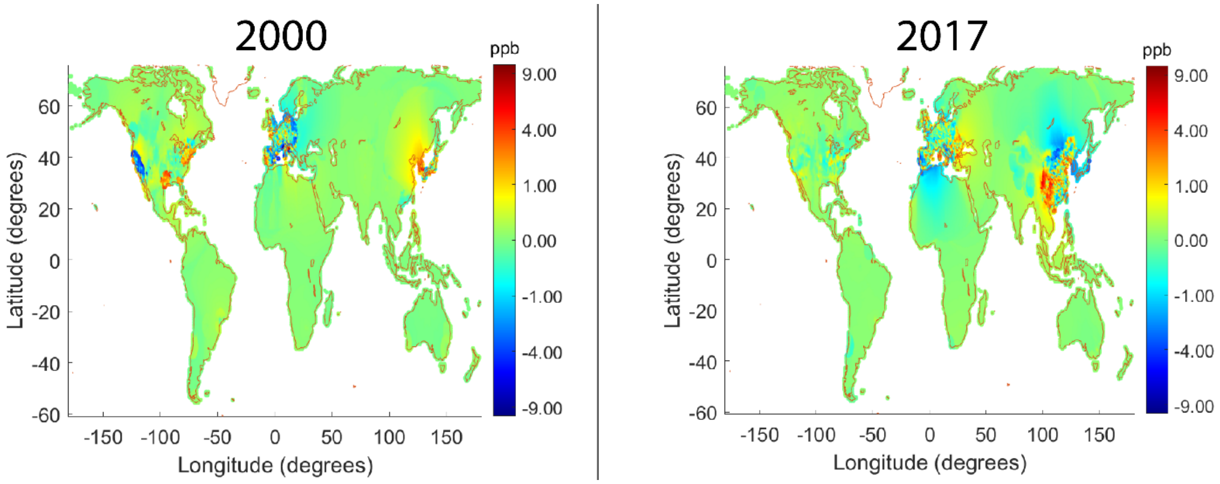

Figure 10. Differences in final BME estimates caused by applying weighted RAMP. Results show the difference as BME with weighted RAMP as the global offset minus BME with the M^3Fusion composite as the offset (the results of DeLang et al., 2021), for 2000 and 2017. Red indicates that RAMP corrected the M^3Fusion composite value up, while blue indicates that RAMP lowered the M^3Fusion value.

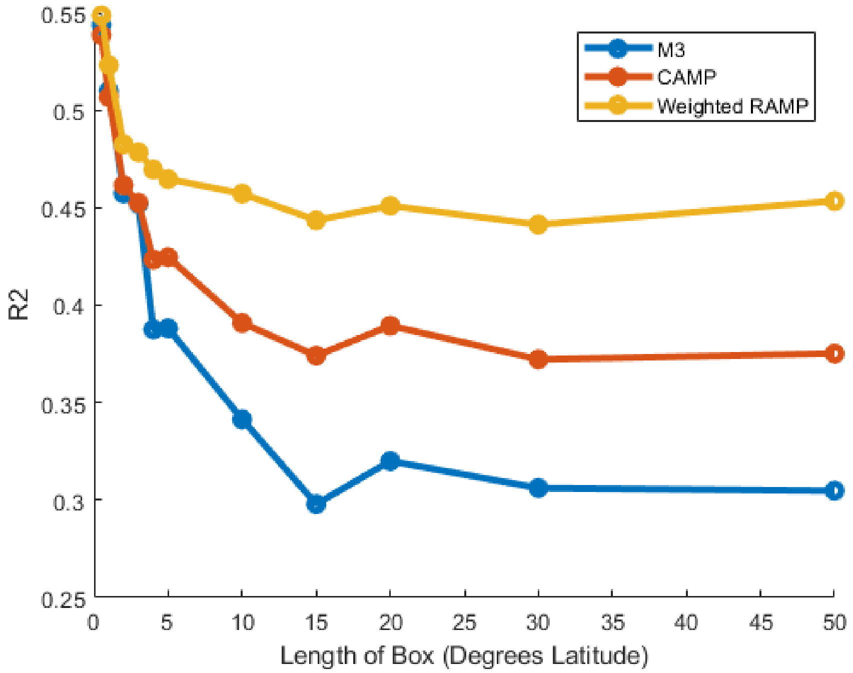

Figure 11. Checkerboard Cross Validation Results. Results are for BME predictions with box latitude length (s) indicated on the x-axis (longitude length is 2s), which indicates the size of the area devoid of observation stations while performing BME. Results are shown for BME data fusion using the M^{3}Fusion composite, and that composite corrected with CAMP and weighted RAMP (wRAMP) as the global offsets pooling results over all years. While the differences between M^{3}Fusion and wRAMP are small at small box sizes (similar to LOOCV), RAMP greatly outperforms M^{3}Fusion and CAMP at large box sizes, where BME has less influence as there are fewer nearby observations to aid estimation.

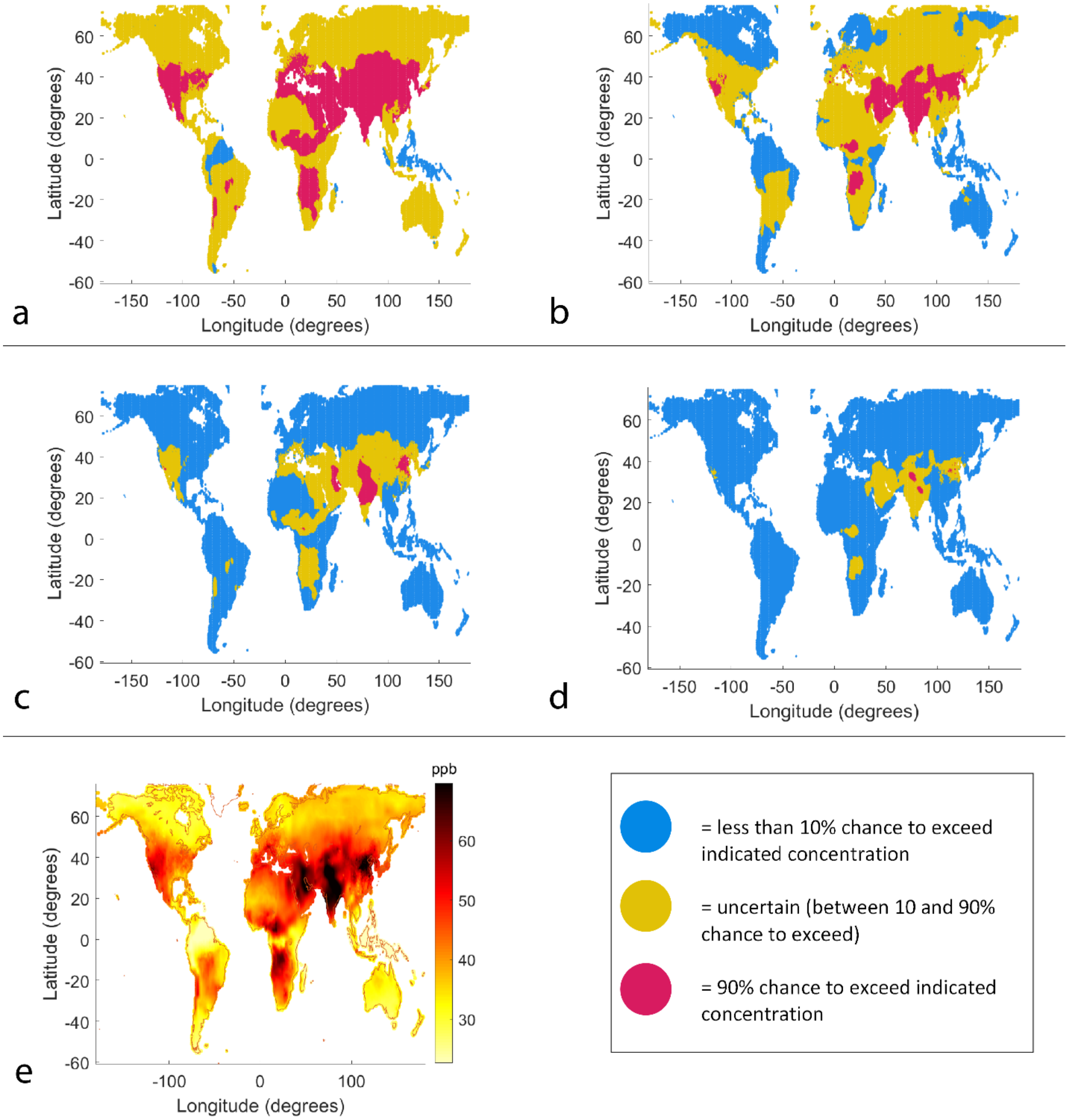

Figure 12. Likelihood of exceedance of four ozone levels in 2017. Results are shown for BME with wRAMP as the global offset, relative to four ozone levels: 35 (a), 45 (b), 55 (c), and 65 (d). Also show are ozone estimates for 2017 (e). Areas with low variance (near station observations) have more certainty, as do areas where BME estimates are very high or low compared to the levels selected.

Table 1. Leave one out cross validation results

Scenario	MSE(ppb^2)	R^2
Simple multi-model mean	189.23	0.28
M^{3}Fusion	61.14	0.30
CAMP	53.54	0.35
wRAMP	46.79	0.43
BME using M^{3}Fusion as offset (DeLang et al., 2021)	14.5	0.83
BME using CAMP as offset	14.5	0.83
BME using wRAMP as offset	14.5	0.83