Mapping Global Surface Ozone Concentrations through the Statistical Fusion of Observations and Models using Bayesian Maximum Entropy

Jacob Becker[1], Marissa DeLang[1], Kai-Lan Chang[2], Owen Cooper[2,3], Stephanie Cleland[1], Elyssa Collins[1], Marc Serre[1], J. Jason West[1]
[1]Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, NC
[2]NOAA Earth System Research Laboratory, Boulder, CO
[3]Cooperative Institute for Research in Environmental Studies, University of Colorado, Boulder, CO

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Background

The purpose of this project is to create accurate global surface ozone estimations utilizing both ozone measurement data and global models. The motivation for this project includes the following:

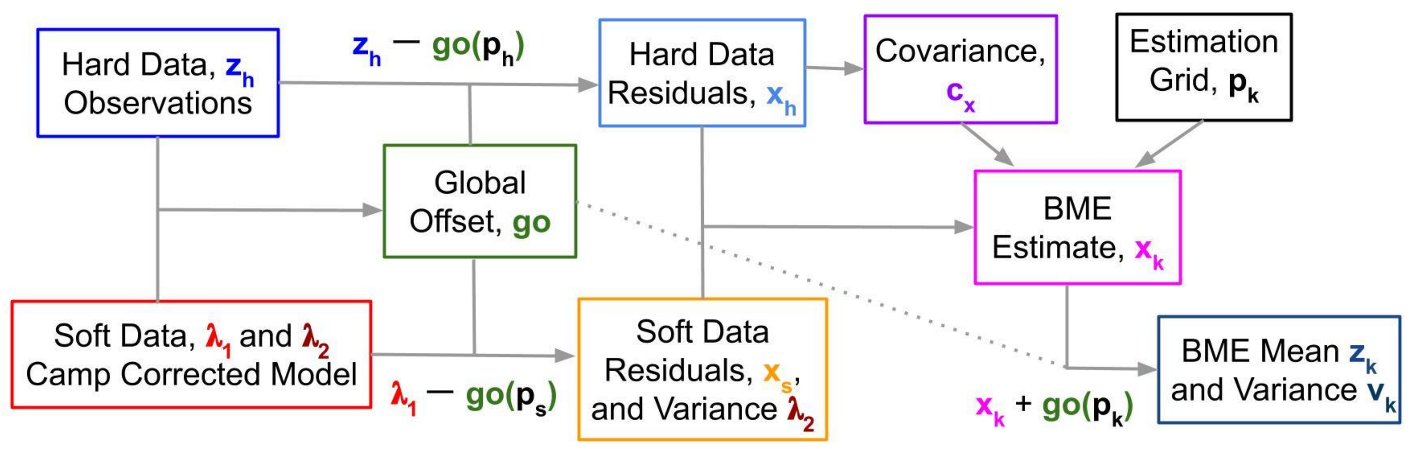
- Tropospheric ozone is an EPA criteria pollutant known to be detrimental to human, animal, and plant health.
- The Global Burden of Disease Assessment requests fine resolution global surface ozone estimations for the years 1990-present.
- Ozone measurements are well-documented in North America, Europe, and Japan, but other regions have very few measurements.
- Models provide information on ozone levels in data sparse regions, but model predictions are inaccurate.

Methods

Bayesian Maximum Entropy (BME) Framework:

BME is a geostatistical modeling tool that can be used to combine various knowledge bases for an air pollutant and combine them to create a single product. In this case, we use BME to combine site specific observations, modeled concentrations (treated as probabilistic data to account for discrepancies between models and observations), correlations between measurement locations, and trends in ozone pollution across time and space. Beyond combining these knowledge bases to provide an estimate of ozone pollution, BME also estimates a variance, which can be used to understand estimation confidence at different locations. In short, the steps are:

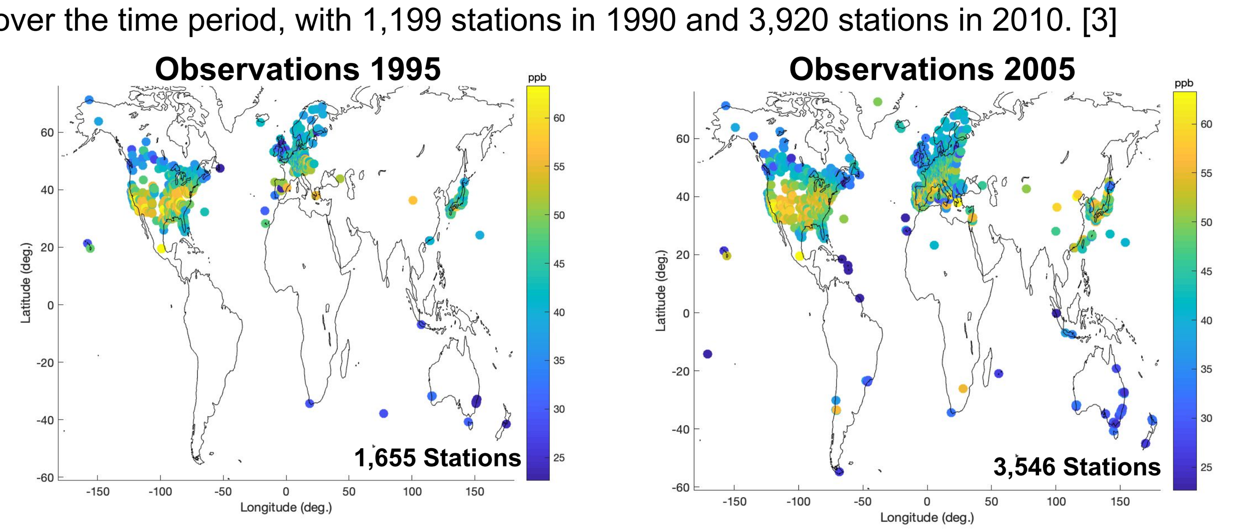
1. Let **Z(p)** be a field of ozone concentration estimations in space and time
2. Calculate the average trend over space and time of ozone, the "global offset": **go(p)**
3. Remove global offset from measurements (hard data), z_h, to obtain residuals $x_h = z_h - go(p_h)$.
4. Model the covariance (the correlation between locations in space and time) c_x based on the residuals x_h.
5. Subtract $go(p_s)$ from the mean of the probabilistic model-based data (λ_1) to obtain x_s.
6. Combine hard data residuals (x_h), soft data residuals (x_s), soft data variance (λ_2), covariance (c_x), and estimation parameters to get the BME estimation (x_k)
7. Obtain final values (z_k) by adding back the global offset $go(p_k)$ to BME estimation (x_k).



For the theory, derivation, and details of BME, see Christakos 1990 [1]. Implemented with BMElib [2]

Ozone Data

Ozone measurements were obtained from the Tropospheric Ozone Assessment Report (TOAR) for 1990-2015. All surface ozone measurements are the maximum six month average of the daily maximum eight hour average, or mda8 (in ppb), in each year. In total, there were 5,966 monitoring sites over the time period, with 1,199 stations in 1990 and 3,920 stations in 2010. [3]



M^3 Fusion Model Combination

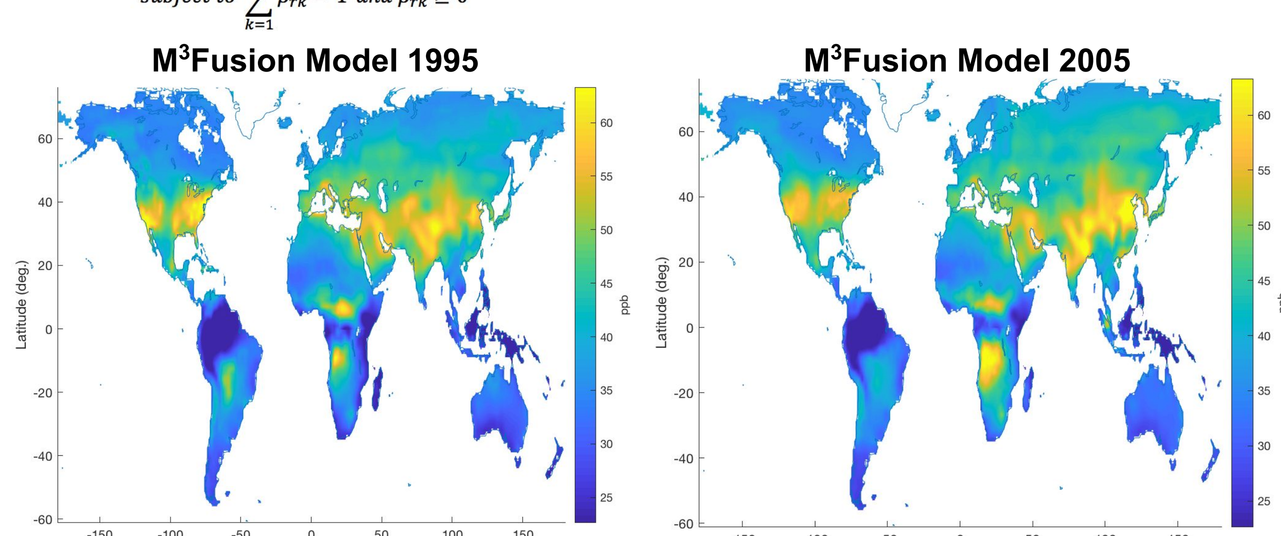
The M^3 fusion method was used to create a multimodel composite of seven specified-dynamics ozone models in each year from 1990 to 2010: CHASER, MOCAGE, MRI-ESM, NASA GEOSCCM, NASA MERRA2-GMI, NCAR CESM-Chem, and NCAR WACCM. Models were weighted in each region to minimize the difference between the bias-corrected multi-model average and observations as described by the following:

Let s_g be the grid cell at resolution 0.125° × 0.125°, $\hat{y}(s_g)$ be the interpolated observations, and $\{\eta_k(s_g); k = 1, \ldots, 7\}$ be the model output registered onto the same grid from the seven models. [4]

$$\underset{\{\alpha_r, \beta_{rk}; k=1,..,7\}}{\text{minimize}} \sum_{s_g \in Region\ r} \left(\hat{y}(s_g) - \alpha_r - \sum_{k=1}^{7} \beta_{rk} \eta_k(s_g) \right)^2,$$

$$\text{subject to} \sum_{k=1}^{7} \beta_{rk} = 1 \text{ and } \beta_{rk} \geq 0$$

	North America		Europe	
Year	*1995*	*2005*	*1995*	*2005*
CESM-Chem	0.38	0.29		
CHASER	0.04			
GFDL-AM3			0.18	0.62
MERRA		0.32	0.82	0.38
MOCAGE	0.36	0.01		
MRI-ESM	0.22	0.27		
WACCM		0.11		



Methods

Constant Air quality Model Performance (CAMP) [5]

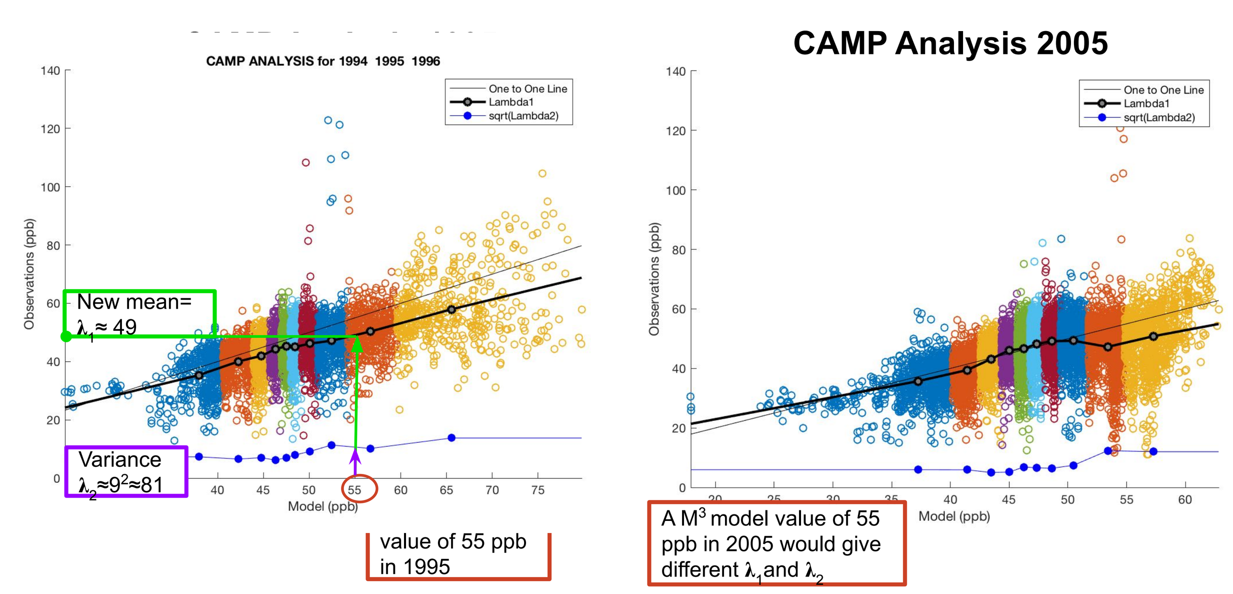
We use the M^3Fusion output as probabilistic data in the BME framework, in which each modeled value is given an estimated variance based on how closely the model values match the observed values (a lower variance is weighed higher in the final estimation). CAMP also removes model bias by basing new mean values on models and observations.

1) Match station measurement values (observations) with model estimations at each measurement location in space and time
2) Sort each paired value into 10 equally sized bins and calculate λ_1=mean and λ_2=variance of observations in each bin:

$$\lambda_1(\tilde{x}_i) = \frac{1}{n(\tilde{x}_i)} \sum_{j=1}^{n(\tilde{x}_i)} \hat{x}_j \qquad \lambda_2(\tilde{x}_i) = \frac{1}{n(\tilde{x}_i) - 1} \sum_{j=1}^{n(\tilde{x}_i)} \left(\hat{x}_j - \lambda_1(\tilde{x}_i)\right)^2$$

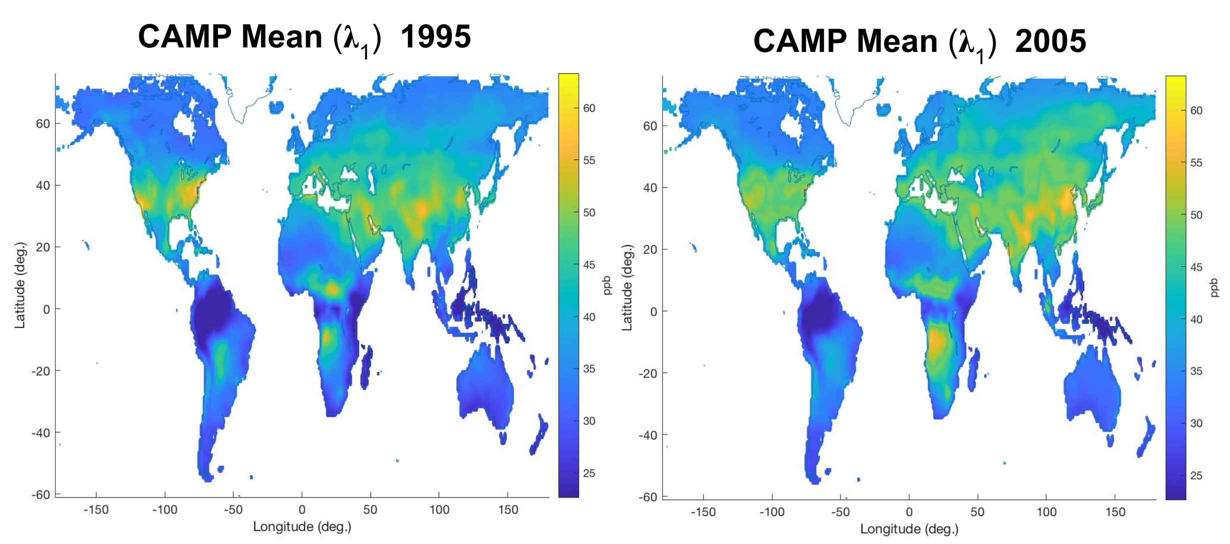
Where $n(\tilde{x}_i)$ is the number of collocated modeled ($\tilde{x}_i$) and observed ($\hat{x}_j$) values in each bin.

3) Find λ_1 and λ_2 for each model point, using three years of data to create the λs for a given year



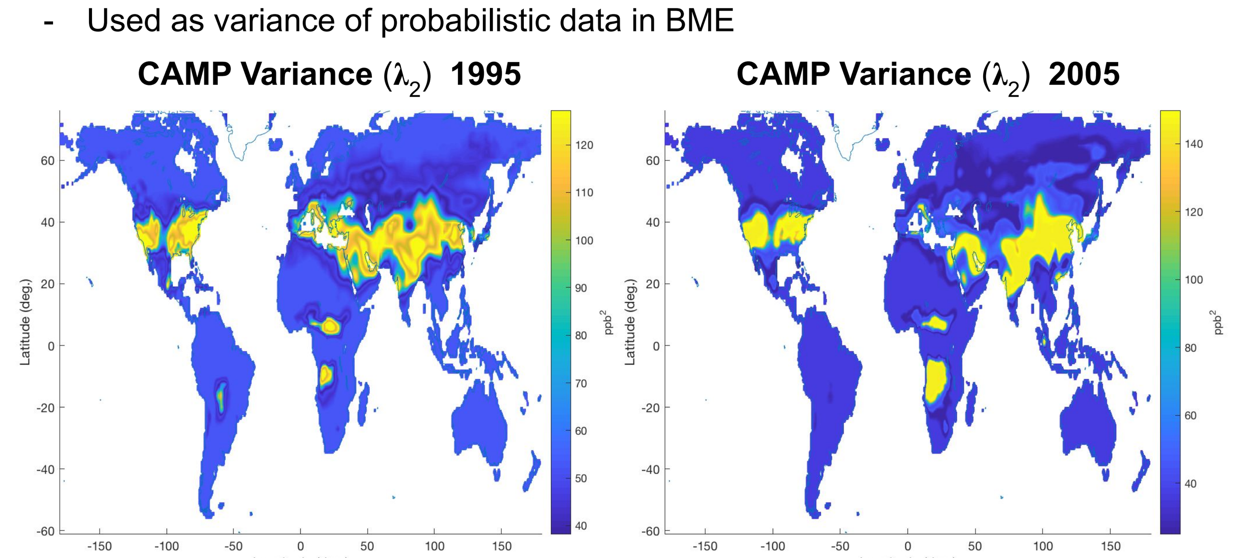
CAMP Mean (λ_1)

- CAMP corrected model values
- Used as the mean value when using model as probabilistic data in BME



CAMP Variance (λ_2)

- CAMP determined variance
- Used as variance of probabilistic data in BME



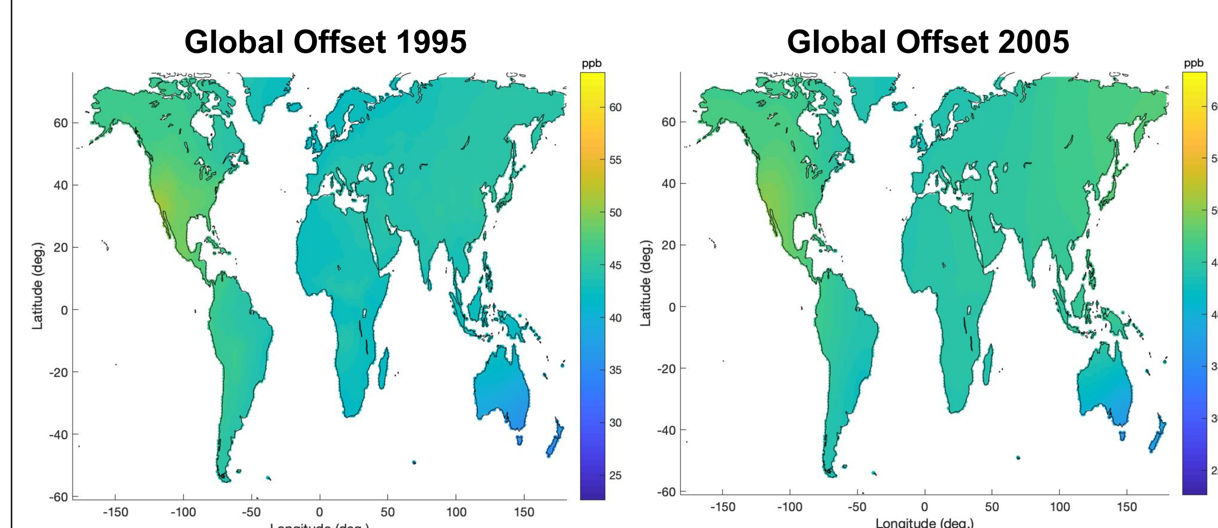
Global Offset

Mean trend of the data in space and time that is subtracted from data values in order to lower variability and improve estimation, calculated in two steps:

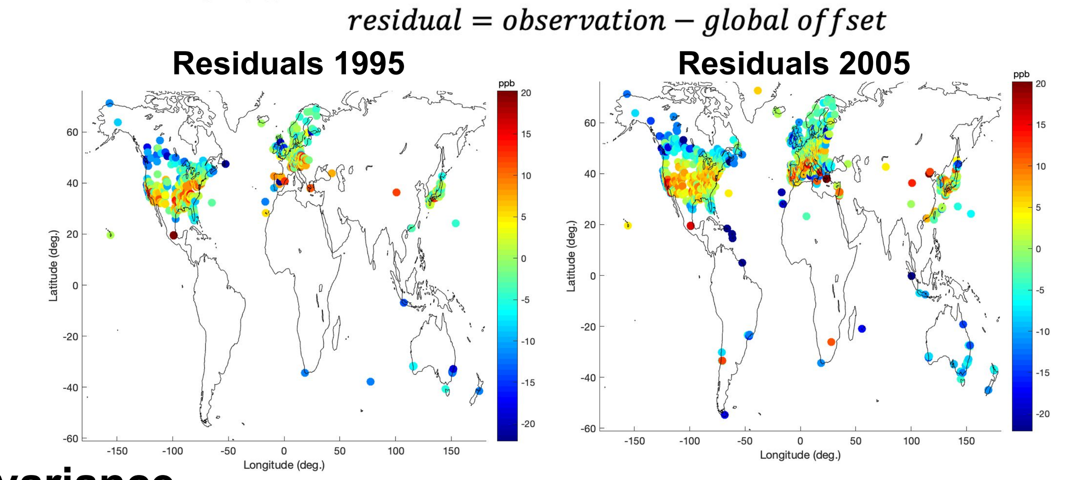
1. Exponential kernel smoothing of observations using domain wide parameters
2. Weighted sum of exponential kernel smoothing and CAMP corrected model (λ_1)

$$(global\ offset)_{weighted} = \alpha\,(global\ offset)_{kernelsmoothing} + (1 - \alpha) * \lambda_1$$

$$\alpha = \frac{\#\ of\ stations\ with\ a\ measurement\ within\ 5\ degrees}{total\ \#\ of\ stations}$$



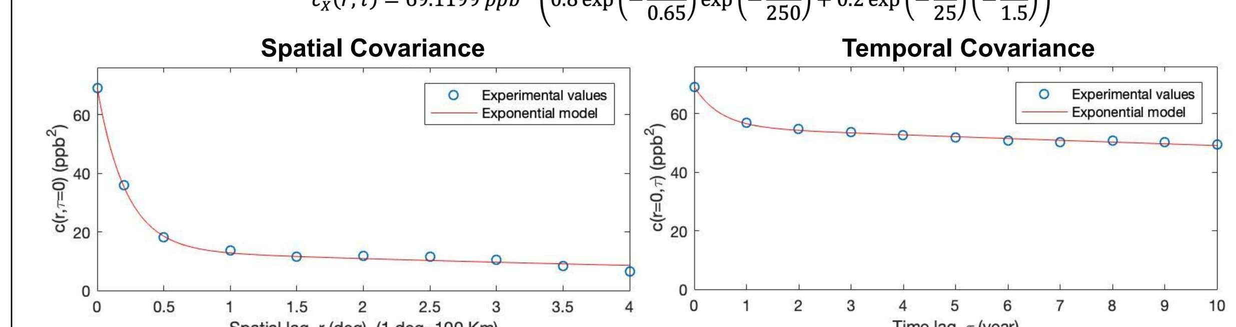
$$residual = observation - global\ offset$$



Covariance

The covariance of the residuals is the range of influence of a measurement to predict other concentrations in space and time.

$$c_X(r, \tau) = 69.1199\,ppb^2 \left(0.8 \exp\left(-\frac{3r}{0.65}\right) \exp\left(-\frac{3\tau}{250}\right) + 0.2 \exp\left(-\frac{3r}{25}\right)\left(-\frac{3\tau}{1.5}\right)\right)$$



Acknowledgements

We thank NASA HAQAST for funding, the Tropospheric Ozone Assessment Report, the Global Burden of Disease Assessment, and modeling teams who provided data: NASA GEOSCCM, NASA G5NR-Chem, NASA MERRA2-GMI, MOCAGE, NCAR CESM-Chem, NCAR WACCM, CHASER, and MRI-ESM.

Future Work

Cross Validation

We plan on doing a K-fold cross validation of our final BME product to compare the MSE and R^2 of the following:

- M^3Fusion model vs. CAMP corrected M^3Fusion model
- Various global offset smoothing parameters
- Exponential kernel smoothing vs. alpha weighted sum global offset
- Various BME estimation parameters

Add Fine Resolution

Most models used are approximately 2° by 2° resolution, but finer resolution is required for Global Burden of Disease Assessment. The NASA G5NR-Chem model contains surface ozone concentrations at 0.125° by 0.125° resolution for July 2013 to June 2014. We believe that the spatial distribution of this model can be used as the distribution pattern for other years, and we plan to scale our ozone estimations accordingly.

Recent Years

Repeat process for more recent years (2011-2017). The TOAR observations currently only go through 2015, but we are working with TOAR to have more observations available. Additionally, we have obtained ozone data from the China National Environmental Monitoring Center Network, which started collecting data in 2013. We were able to obtain four models for the recent time period: MRI ESM, NASA MERRA2-GMI, and NOAA GFDL-AM3 and GFDL-AM4.

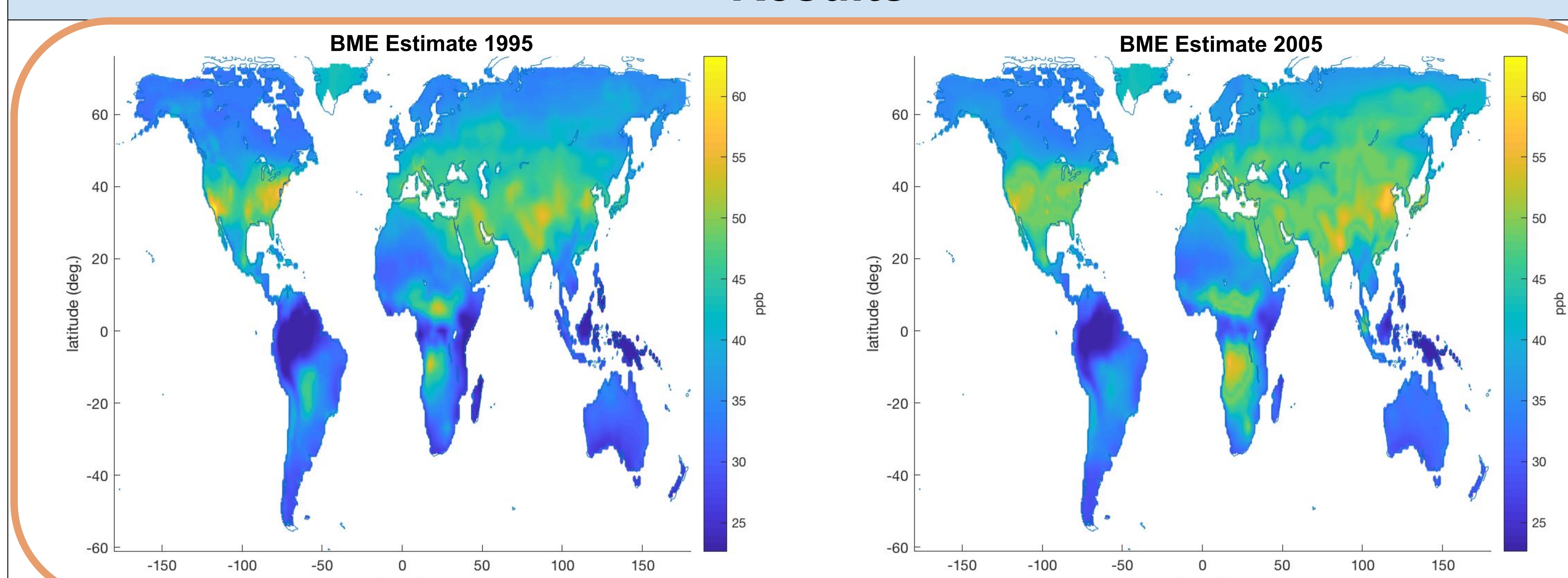
Regionalized Air Quality Model Performance

Using the Regionalized Air Quality Model Performance (RAMP) method proposed by Reyes et al. would account for non-homogeneity in model performance across space, which could provide even better model correction in data sparse areas [6].

References

[1] Christakos, George. 1990. "A Bayesian/Maximum-Entropy View to the Spatial Estimation Problem." *Mathematical Geology* 22 (7): 763–77. https://doi.org/10.1007/BF00890661.
[2] Serre, M. L., and G. Christakos. 1999. "Modern Geostatistics: Computational BME Analysis in the Light of Uncertain Physical Knowledge - the Equus Beds Study." *Stochastic Environmental Research and Risk Assessment (SERRA)* 13 (1–2): 1–26. https://doi.org/10.1007/s004770050029.
[3] Reyes, Jeanette M., Yadong Xu, William Vizuete, and Marc L. Serre. 2017. "Regionalized PM2.5 Community Multiscale Air Quality Model Performance Evaluation across a Continuous Spatiotemporal Domain." *Atmospheric Environment* 148 (January): 258–65. https://doi.org/10.1016/J.ATMOSENV.2016.10.048.
[4] Chang, Kai-Lan, Owen R. Cooper, J. Jason West, Marc L. Serre, Martin G. Schultz, Meiyun Lin, Virginie Marécal, et al. 2019. "A New Method (M3 Fusion v1) for Combining Observations and Multiple Model Output for an Improved Estimate of the Global Surface Ozone Distribution." *Geoscientific Model Development* 12 (3): 955–78. https://doi.org/10.5194/gmd-12-955-2019.
[5] Nazelle, Audrey de, Saravanan Arunachalam, and Marc L. Serre. 2010. "Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions: An Application for Attainment Demonstration in North Carolina." *Environmental Science & Technology* 44 (15): 5707–13. https://doi.org/10.1021/es100228w.
[6] Schultz, Martin G., Sabine Schröder, Olga Lyapina, Owen Cooper, Ian Galbally, Irina Petropavlovskikh, Erika Von Schneidemesser, et al. 2017. "Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations." *Elem Sci Anth* 5 (0): 58. https://doi.org/10.1525/elementa.244.

Results



Features of Our BME Data Fusion Approach

- Combines observations with seven weighted and corrected models
- Observations take precedence in estimation near observation locations, but BME weight falls off based on covariance and space/time distance
- BME outputs variance at each estimation point as well as mean
- Accounts for non-linear, non-homoscedastic, and non-homogenous (in time) relationship between model and observations
- Model impact on BME map weighted by variance and difference from global offset
- Model variance was higher when predicting high values due to a tendency to overpredict when compared with observations