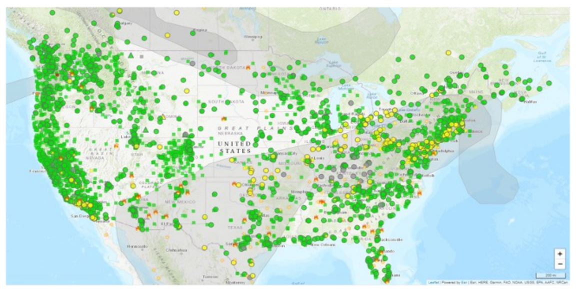
Bayesian Nonparametric Ensemble (BNE) algorithm for predictions of high spatiotemporal $PM_{2.5}$ concentrations

Vijay Kumar[1], Jaime Benavides[1], Carlos Carrillo-Gallegos[2], Arlene Fiore[2], John Paisley[3,4] and Marianthi-Anna Kioumourtzoglou[1,4]
[1]Environmental Health Sciences, Columbia University, New York, NY, [2]Earth and Planetary Sciences, MIT, Cambridge, MA,
[3]Electrical Engineering, Columbia University, New York, NY, [4]Data Science Institute, Columbia University, New York, NY

Background and Methods

Motivating Problem:

- Air pollution associated with several adverse health outcomes [1]
- Scarcity of ground level monitors means we must model exposure
- Satellite data essential in filling gaps, but requires translation from column to surface measurements
- Current health studies use predictions from a single dataset
- We present a novel ensembling method to incorporate several different exposure datasets (with various methodologies) and provide uncertainty

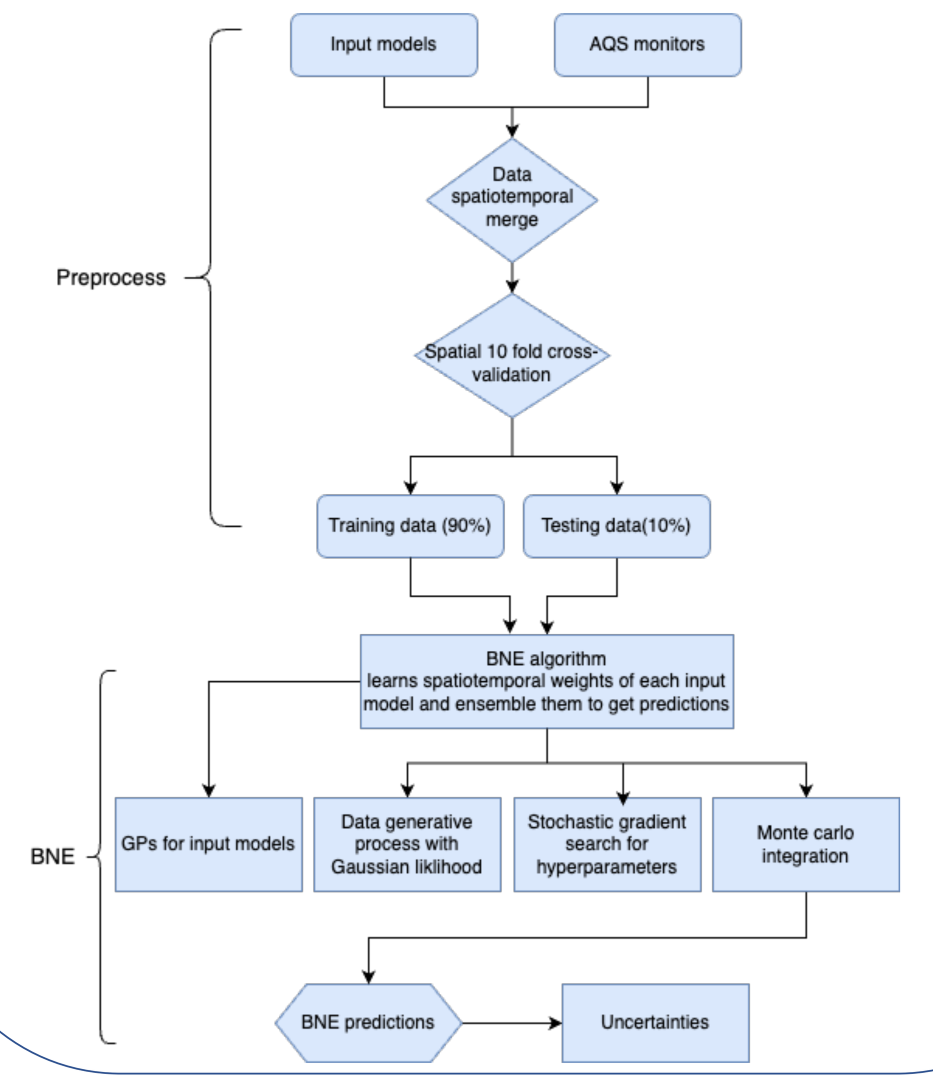


Environmental Protection Agency monitoring network

Bayesian Nonparametric Ensemble (BNE) algorithm

- Integrate information across existing spatiotemporal prediction models
- Weights each model by its spatiotemporal predictive accuracy
  - I.e., spatiotemporal weights
- Provide spatiotemporal uncertainty of predictions
  - Including due to existing model disagreement

Flow-chart

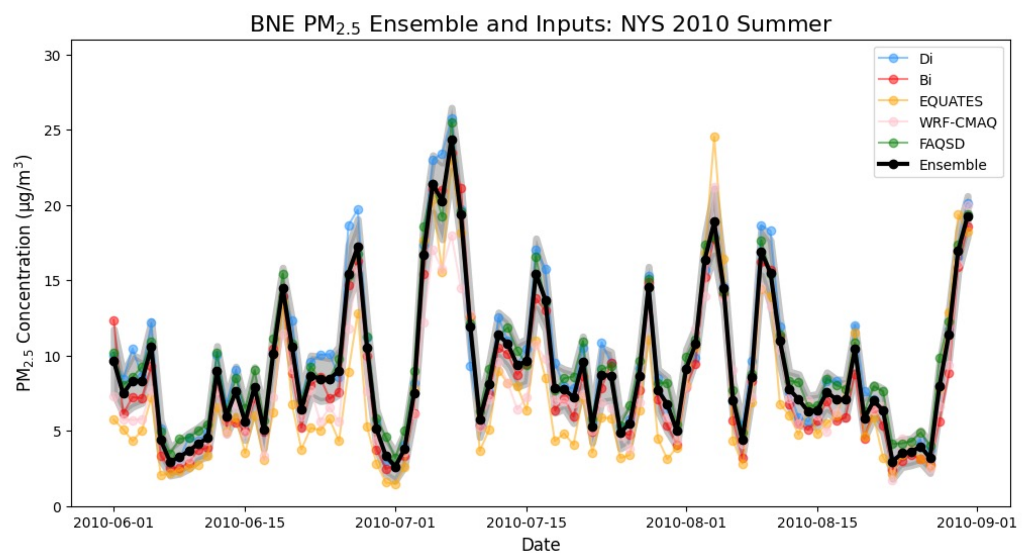


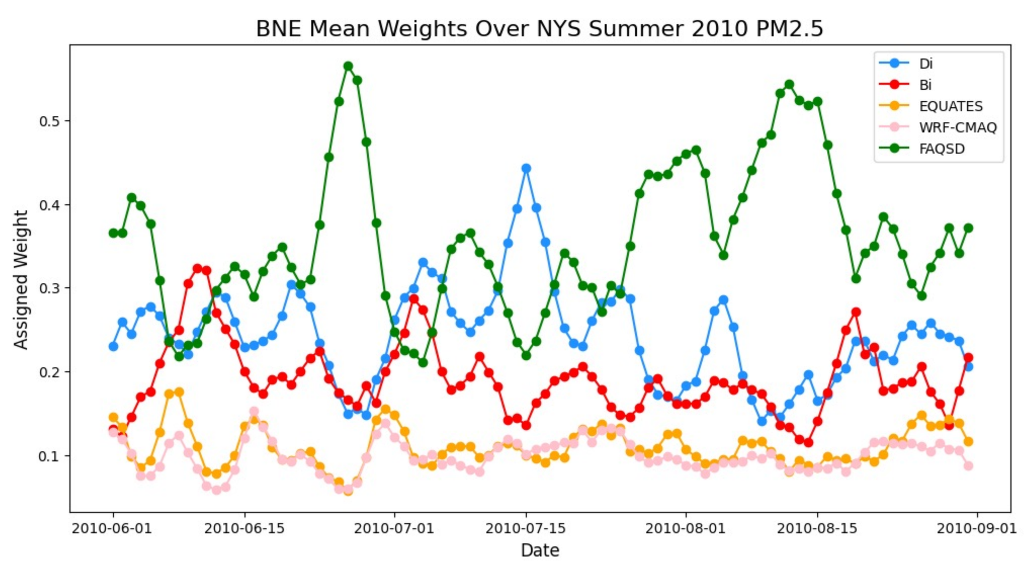
Data and Results

Input Exposure Models:

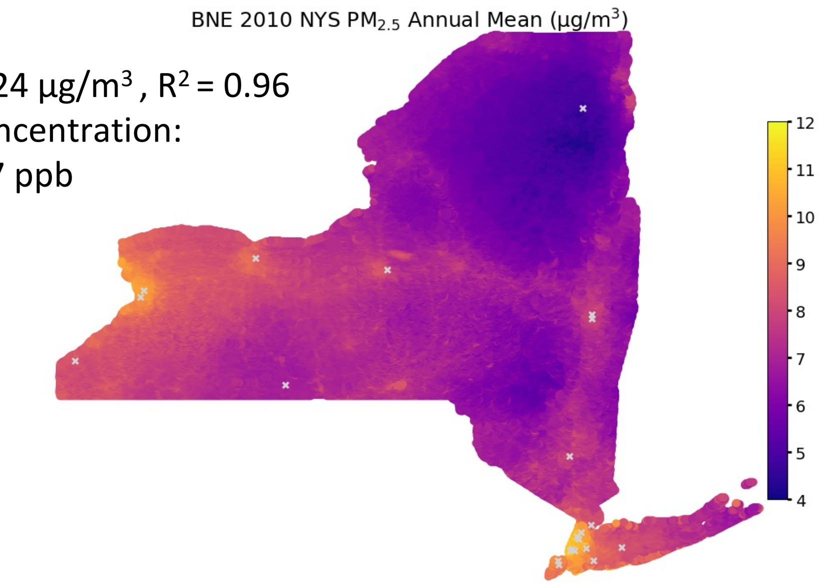
Dataset	Spatial Resolution	Methodology
Bi et. al 2019	1 x 1 km^2	AOD from MODIS, gap filled with random forest approach, incorporating AQS data, land use regression variables, and meteorology from NARR
Di et. al 2021	1 x 1 km^2	Generalized additive model of three ML methods incorporating AQS data, AOD from MODIS, GEOS-CHEM, CMAQ(v5.1), and land use variables
EQUATES (EPA Air QUAlity Time SEries)	12 x 12 km^2	U.S. EPA Air Quality Model: CMAQ (v5.3.2). driven by WRF (v4.1.1) meteorology and the National Emissions Inventory (NEI) for 2017
FAQSD (Fused Air Quality Surface Using Downscaling)	12 x 12 km^2	Fusion of CMAQ (v4.6) chemical transport model (incl. NEI and MM5(v3.6.3) meteorology) with ground level AQS data
WRF-CMAQ with AOD, CO (R. Kumar et. al 2021)	12 x 12 km^2	Reanalysis assimilating AOD retrievals from MODIS and CO retrievals from MOPITT into CMAQ, with WRF providing meteorological input

Table: Daily $PM_{2.5}$ base models

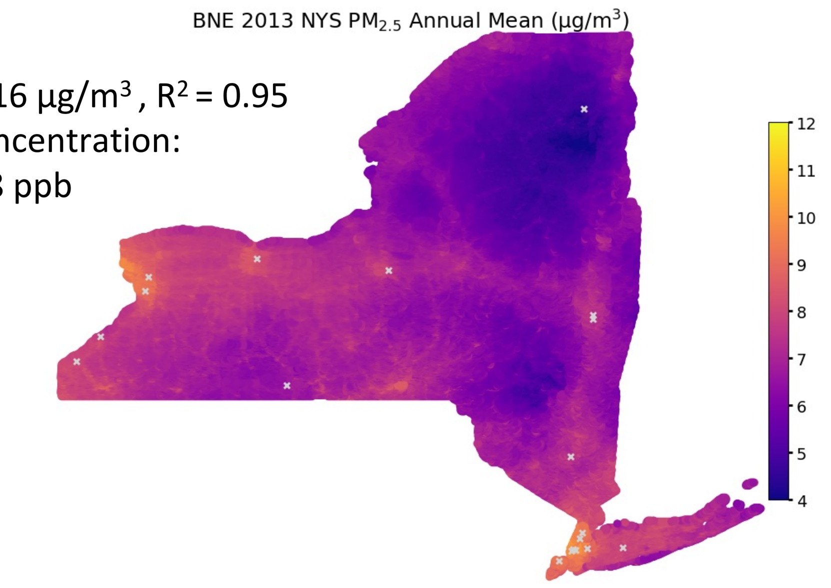




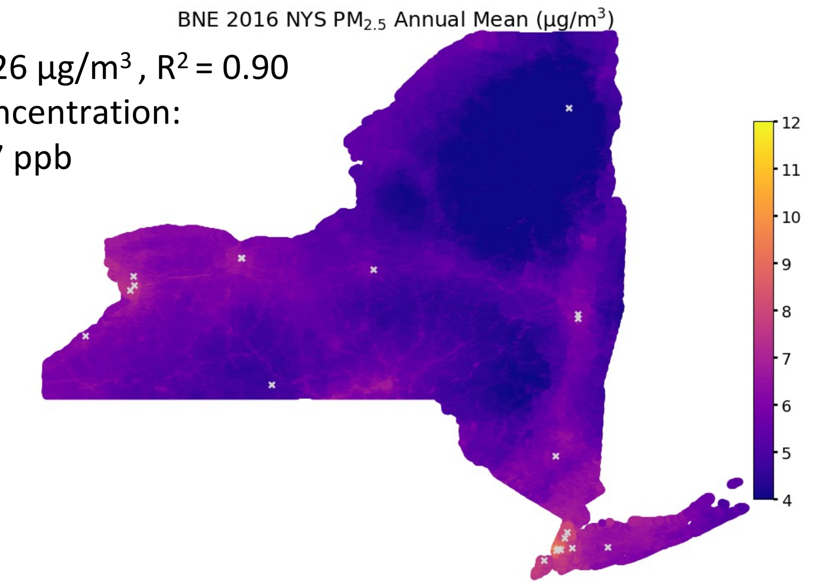
Results



RMSE=1.24 µg/m^3 , R^2 = 0.96
Mean Concentration:
6.84 ± 0.7 ppb



RMSE=1.16 µg/m^3 , R^2 = 0.95
Mean Concentration:
6.53 ± 0.8 ppb



RMSE=1.26 µg/m^3 , R^2 = 0.90
Mean Concentration:
5.03 ± 0.7 ppb

Completed 2010-2016; more years, pollutants coming!

References

- Fuller R. et. al, Pollution and health: a progress update. Lancet Planet Health. 2022 Jun;6(6):e535-e547
- Liu J, Paisley J, Kioumourtzoglou MA, Coull B. Accurate uncertainty estimation and decomposition in ensemble learning. Advances in neural information processing systems. 2019;32.
- Paisley J, Rowland S, Liu JZ, Coull B, Kioumourtzoglou MA. Bayesian nonparametric model averaging using scalable Gaussian process representations. In2022 IEEE International Conference on Big Data (Big Data) 2022 Dec 17 (pp. 55-64). IEEE.
- Icons by Google Inc., MaxxI, PICOL via Wikimedia Commons