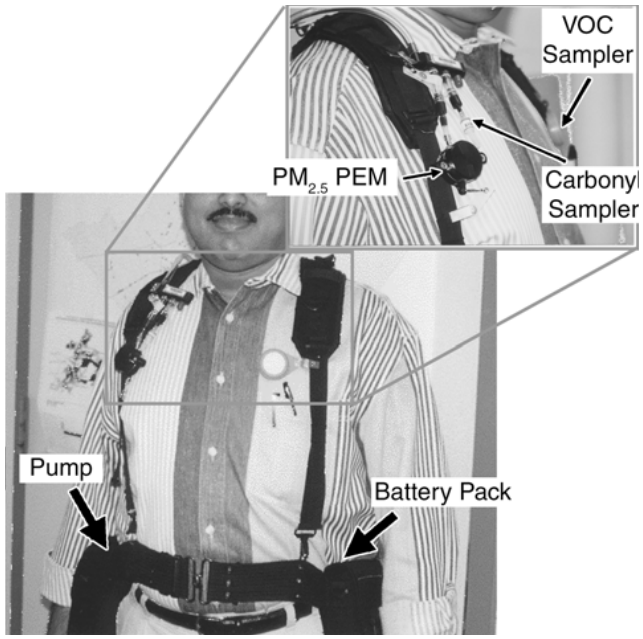
Bayesian Nonparametric Ensemble (BNE)

Marianthi-Anna Kioumourtzoglou
mk3961@cumc.columbia.edu
Environmental Health Sciences, Columbia University

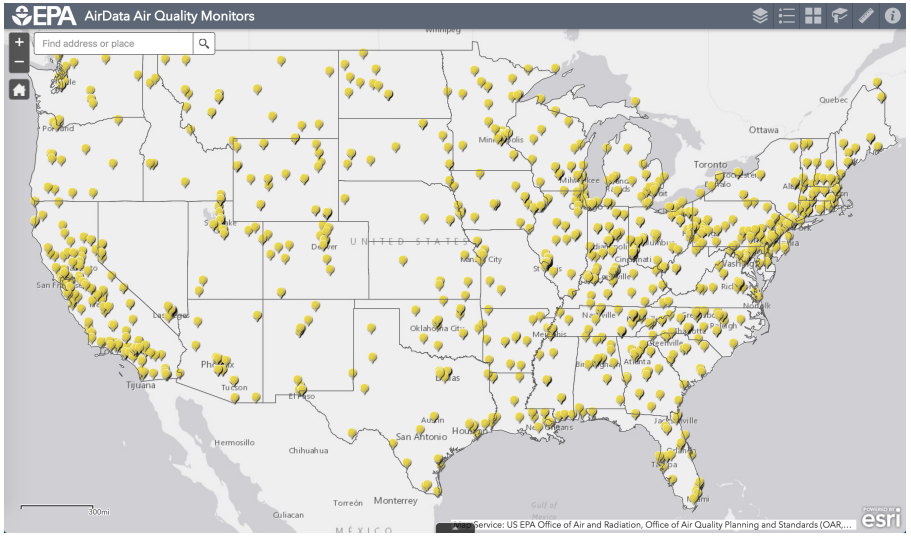
Aug 15, 2023

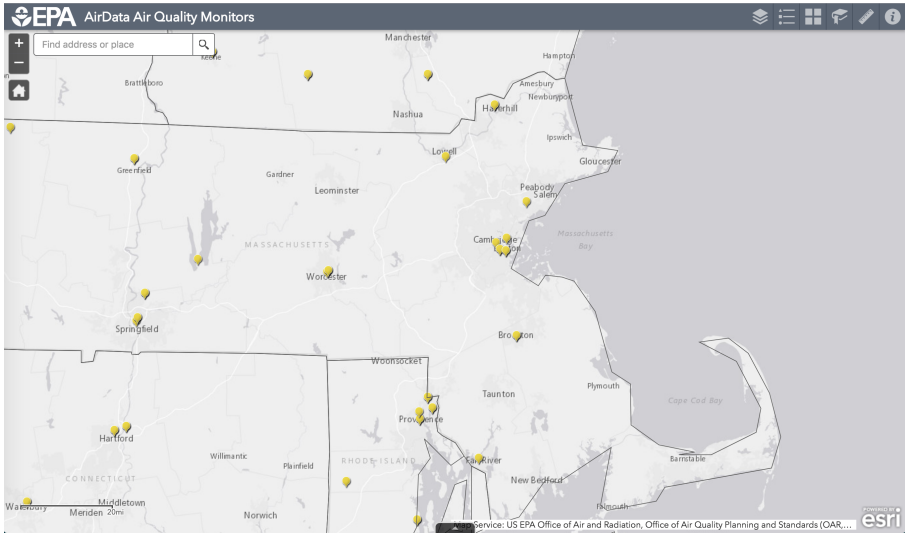
Air Pollution Exposure Assessment: Ideally···

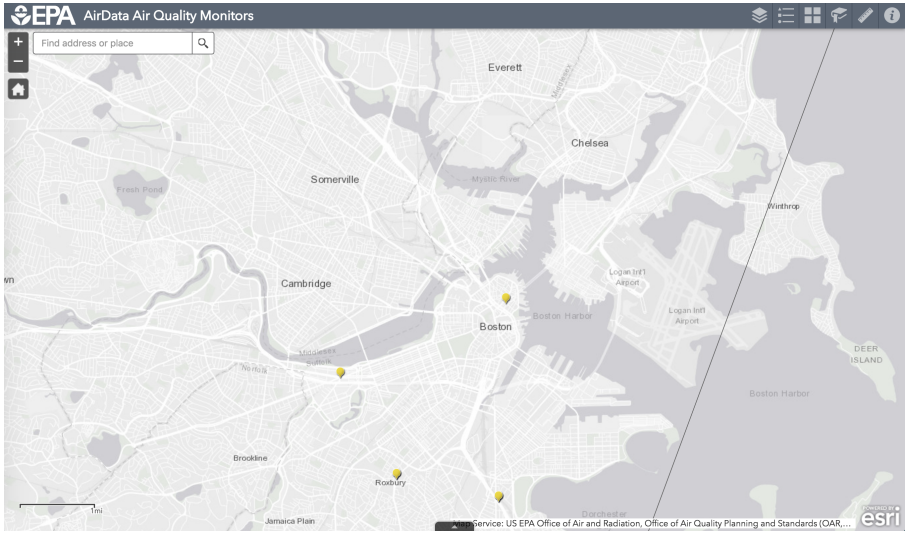
- Air pollution associated with numerous adverse health outcomes
- Ideally: Personal exposures
  - Expensive and intrusive
  - Not feasible for long time periods with many subjects



Air Pollution Exposure Assessment: Historically···



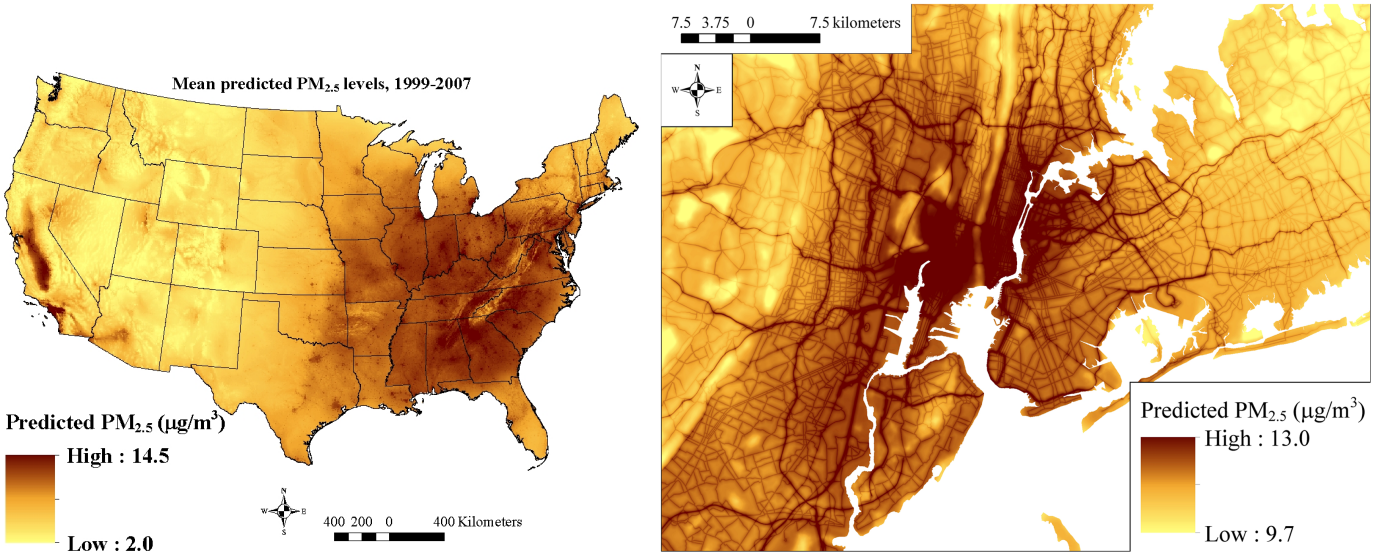




$PM_{2.5}$ Active Monitors on 01.26.2019

Air Pollution Exposure Assessment: More Recently···

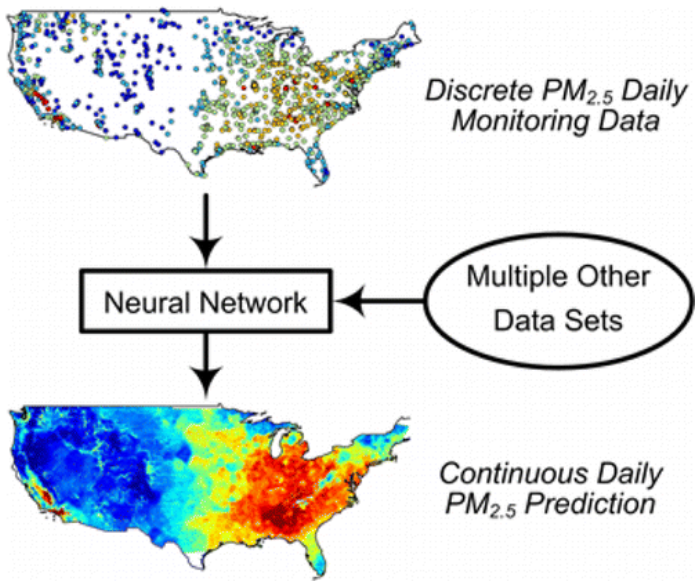
- Spatio-temporal prediction models
- In the beginning more "naïve" statistical models
  - (LUR – not time-varying)
  - Generalized Additive Mixed Models (GAMMs)
  - Inputs: met and GIS variables (e.g. land-use, population and road density, etc)
- Starting to get into "Big Data" territory and need for computationally scalable approaches



Yanosky et al, *Env Health*, 2014

Air Pollution Exposure Assessment: More Recently···

- "Fancier" spatio-temporal prediction models
- Fuse remote sensing data, predictions from chemical transport models, met and GIS variables
- Using more robust methods that yield higher predictive accuracy, e.g. random forests and neural networks
- Very high spatial and temporal resolution (i.e., daily predictions at 1×1 km^2 grids)
- Increasing importance of computational scalability



Di et al, *ES&T*, 2016

Improving Exposure Assessment: Ensemble Models

- To date, most health studies use predictions from a single model to assign exposures
- Very strong assumption: A single prediction model is *correct*

Improving Exposure Assessment: Ensemble Models

- To date, most health studies use predictions from a single model to assign exposures
- Very strong assumption: A single prediction model is *correct*

- Combining information from multiple prediction models leads to more accurate predictions

→ Ensemble Prediction models

Ensemble Modeling

Ensemble modeling $\rightarrow$ process during which multiple diverse models are created and combined to predict an outcome

- Different modeling algorithms
- Different training datasets

Goal: Improve generalizability/robustness of predictions over a single estimator

- i.e., limit high variance due to overfitting while improving prediction accuracy

Ensembling Algorithm Examples

- Bagging (bootstrap aggregation)
- Boosting
- Stacking
- Voting (classification)
- Bayesian Model Averaging

Improving Exposure Assessment: Ensemble Models

Limitations of existing ensemble models:

1. Strong assumption that the prediction errors are independent across the individual models
2. Model-specific contributions are constant across space and time
   - Spatio-temporal variability in prediction accuracy can result in erroneous identification of modifiers in the health model
     - OK, this is also true for single prediction models

Finally, One More Limitation of All Models

- Only very recently prediction models (single or ensemble) have started to provide spatio-temporal uncertainty
  - Di et al. 2019, Murray et al. 2019, Van Donkelaar 2021

$\Rightarrow$ In the health models we assume the *predicted* exposures to be "true" exposures, i.e. without bias or uncertainty

Finally, One More Limitation of All Models

- Only very recently prediction models (single or ensemble) have started to provide spatio-temporal uncertainty
  - Di et al. 2019, Murray et al. 2019, Van Donkelaar 2021

$\Rightarrow$ In the health models we assume the *predicted* exposures to be "true" exposures, i.e. without bias or uncertainty

- Bias in the predictions $\rightarrow$ biased effect estimates
- Failure to propagate uncertainty in predictions $\rightarrow$ invalid inferences of effect estimates

Developing a New Tool

- We have put together an amazing team to develop a novel Bayesian Non-parametric Ensemble (BNE, aka a super model)
- Atmospheric chemists and scientists, biostatisticians, data scientists, machine learning experts, environmental epidemiologists, environmental and electrical engineers, two DSIs, four Universities
- To address ALL the above-mentioned limitations

Developing a New Tool (cont'd)

We have developed a new tool that:

1. Integrates information across *existing* spatio-temporal prediction models
2. By weighing each candidate model by its spatio-temporal predictive accuracy
   - I.e., spatio-temporal weights
3. Provides *continuous* predictions in space and time
4. Provides spatio-temporal *uncertainty* of these predictions
5. Is (somewhat) computationally scalable

Bayesian Nonparametric Ensemble (BNE)

y(x): Pollution Spatial Process

$\{f_k(\boldsymbol{x})\}_{k=1}^{K}$: Predictions from K base models

We augment a classic ensemble model with nonparametric machinery:

$$\mu(\boldsymbol{x}) = \sum_{k=1}^{K} f_k(\boldsymbol{x}) * \boldsymbol{\omega}_k(\boldsymbol{x}) + \boldsymbol{\delta}(\boldsymbol{x})$$

$$y|\boldsymbol{x} \sim \boldsymbol{G_x}\Big[\Phi(y|\boldsymbol{x}, \mu)\Big],$$

- $\boldsymbol{\omega}$(x) spatially adaptive model combination
- $\boldsymbol{\delta}$(x) corrects spatially varying systematic bias in μ
- G_x models spatially varying predictive uncertainty

BNE technical details

2022 IEEE International Conference on Big Data (Big Data)

Bayesian Nonparametric Model Averaging Using Scalable Gaussian Process Representations

John Paisley
Dept. of Electrical Engineering
& Data Science Institute
Columbia University
New York, NY, USA

Sebastian Rowland
Environmental Health Sciences
Columbia University
New York, NY, USA

Jeremiah Zhe Liu
Google Research
San Francisco, CA, USA

Brent Coull
Department of Biostatistics
Harvard University
Boston, MA, USA

Marianthi-Anna Kioumourtzoglou
Environmental Health Sciences
Columbia University
New York, NY, USA

Impact of BNE use

1. Best possible predictions $\rightarrow$ minimize exposure measurement error

Impact of BNE use

1. Best possible predictions $\rightarrow$ minimize exposure measurement error
2. Uncertainty estimation $\rightarrow$ valid inferences of health effect estimates
   - Also developing computationally scalable statistical methods for uncertainty propagation into the health models

Impact of BNE use

1. Best possible predictions $\rightarrow$ minimize exposure measurement error
2. Uncertainty estimation $\rightarrow$ valid inferences of health effect estimates
   - Also developing computationally scalable statistical methods for uncertainty propagation into the health models
3. Reporting the estimated model-specific weights back to the teams that developed those $\rightarrow$ improvement of each single prediction model

Impact of BNE use

1. Best possible predictions $\rightarrow$ minimize exposure measurement error
2. Uncertainty estimation $\rightarrow$ valid inferences of health effect estimates
   - Also developing computationally scalable statistical methods for uncertainty propagation into the health models
3. Reporting the estimated model-specific weights back to the teams that developed those $\rightarrow$ improvement of each single prediction model
4. Identification of highest uncertainty locations $\rightarrow$ efficient placing of new monitoring stations

Impact of BNE use

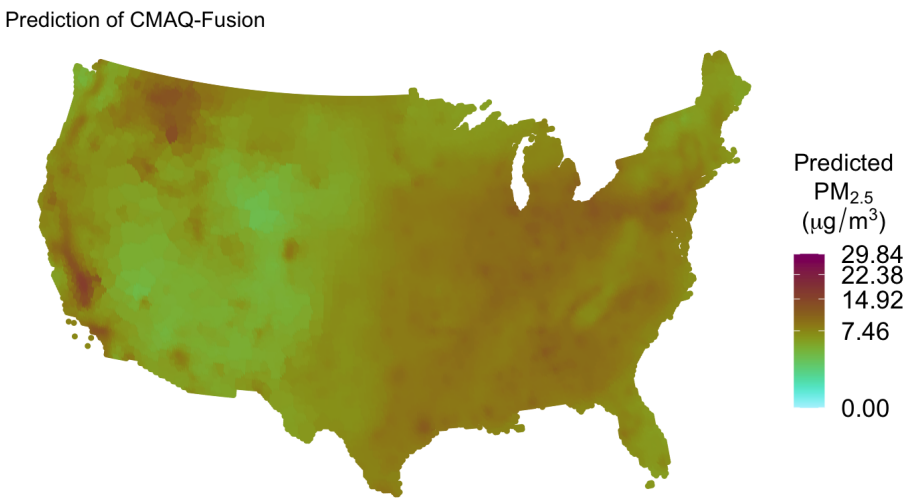
1. Best possible predictions $\rightarrow$ minimize exposure measurement error
2. Uncertainty estimation $\rightarrow$ valid inferences of health effect estimates
   - Also developing computationally scalable statistical methods for uncertainty propagation into the health models
3. Reporting the estimated model-specific weights back to the teams that developed those $\rightarrow$ improvement of each single prediction model
4. Identification of highest uncertainty locations $\rightarrow$ efficient placing of new monitoring stations
5. Our tool can also be used in other applications, beyond air pollution

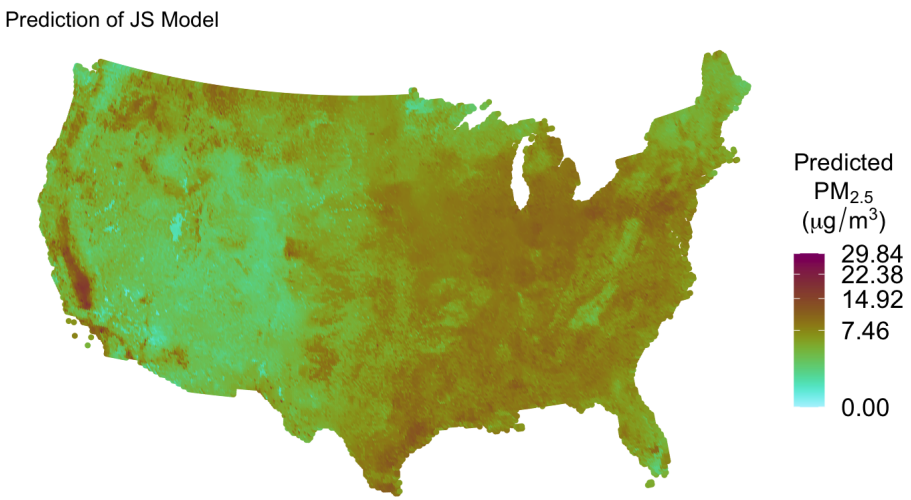
BNE Application: Annual $PM_{2.5}$ predictions

- Nationwide model, 2010 – 2015
- Trained on AQS data
- Seven base models

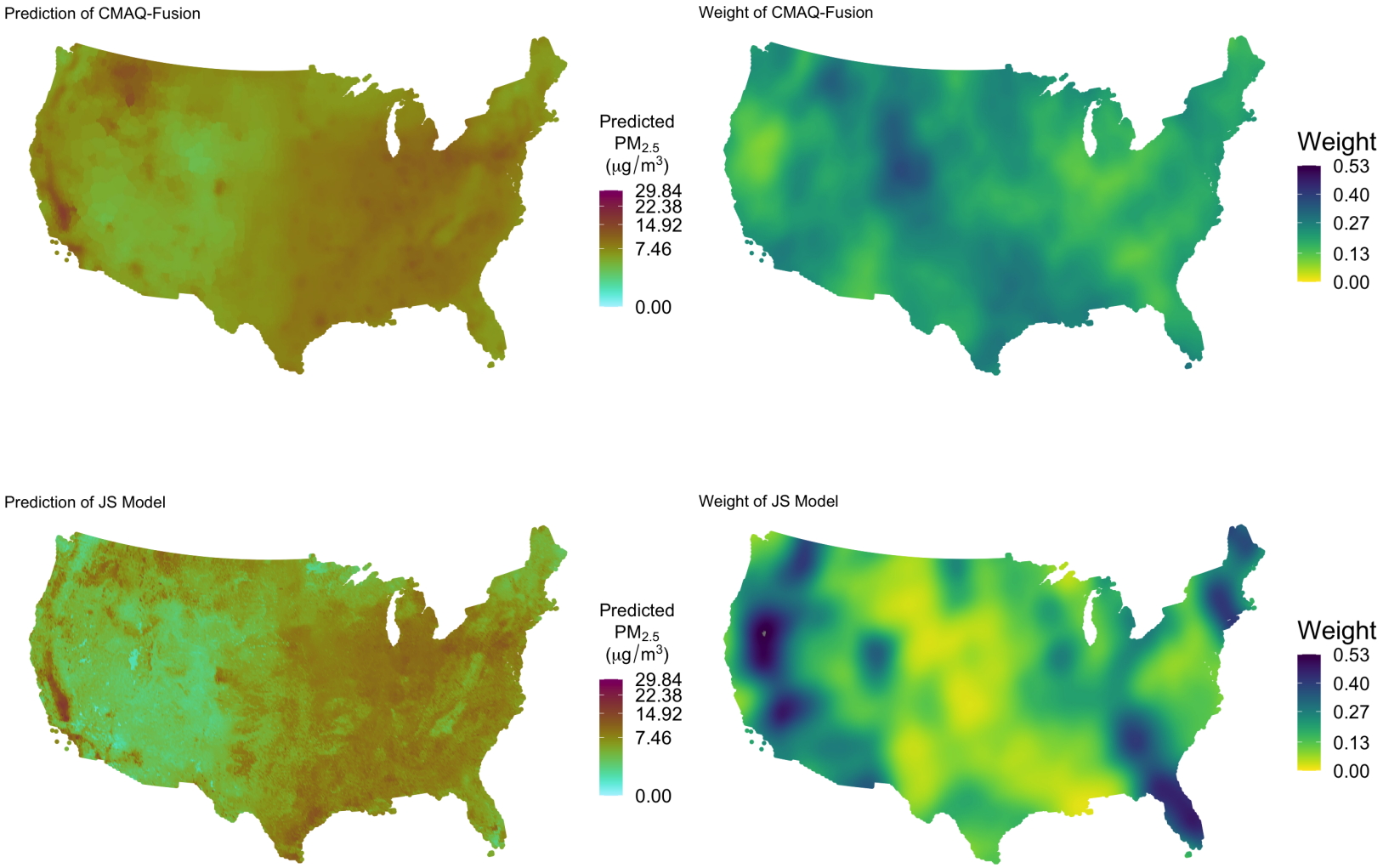
Model name	Method Summary	Inputs	Training Data
AV	Geographically weighted regression	AOD, GEOS-Chem	AQS + global measurements
CC	Land-use regression	AOD, Emissions Inventory, Land use	AQS
CM	Data-fused CTM	Emissions Inventory, Meteorology	AQS
GS	Data-fused CTM	AOD, GEOS-Chem, Land Use	AQS + global measurements
JS	Spatially flexible ensemble of three machine-learning models	AOD, CMAQ-AQS Fusion model, GEOS-Chem, lagged measurements, land use, meteorology	AQS
ME	CTM of $PM_{2.5}$ components	Emissions Inventory, Meteorology	n/a
RK	Data-fused CTM	AOD, Emissions Inventory, Meteorology	n/a

Predictions from Base models (examples)

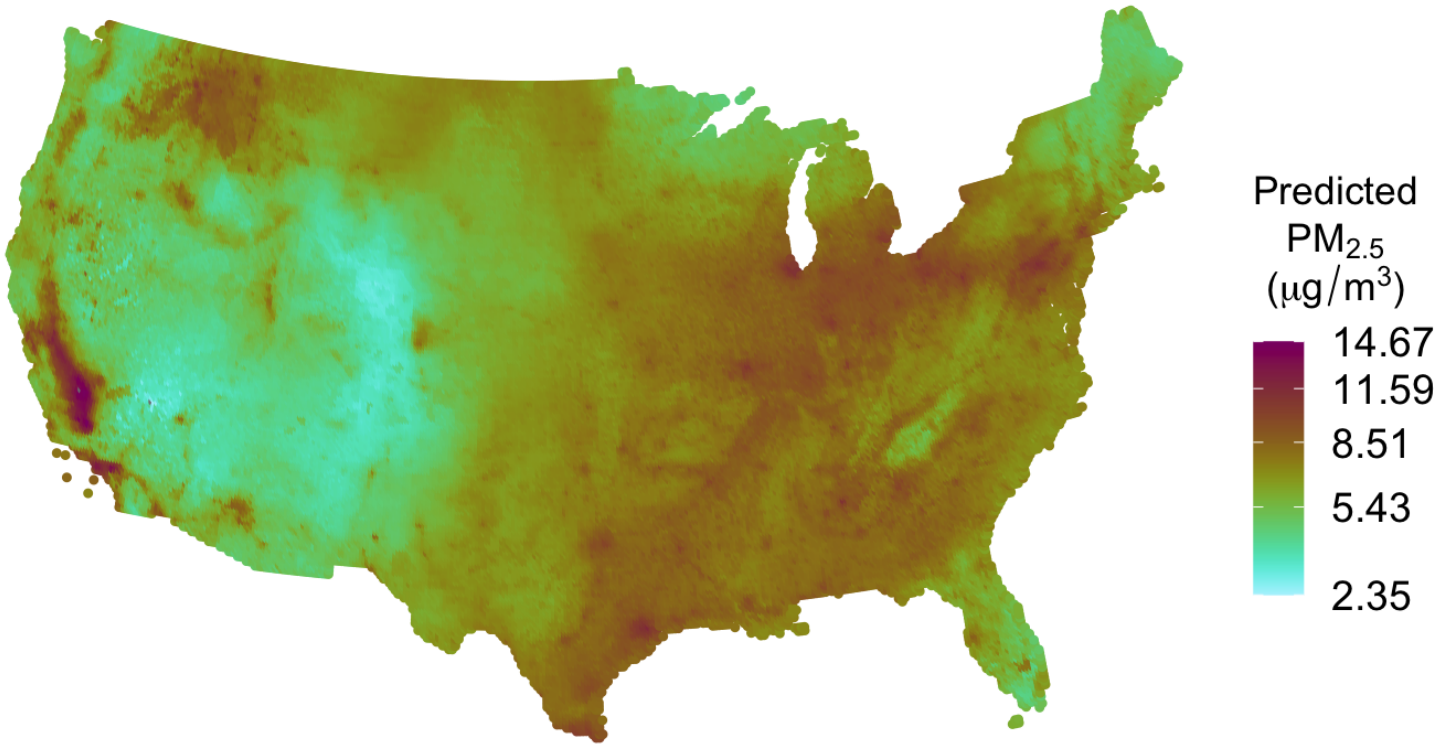




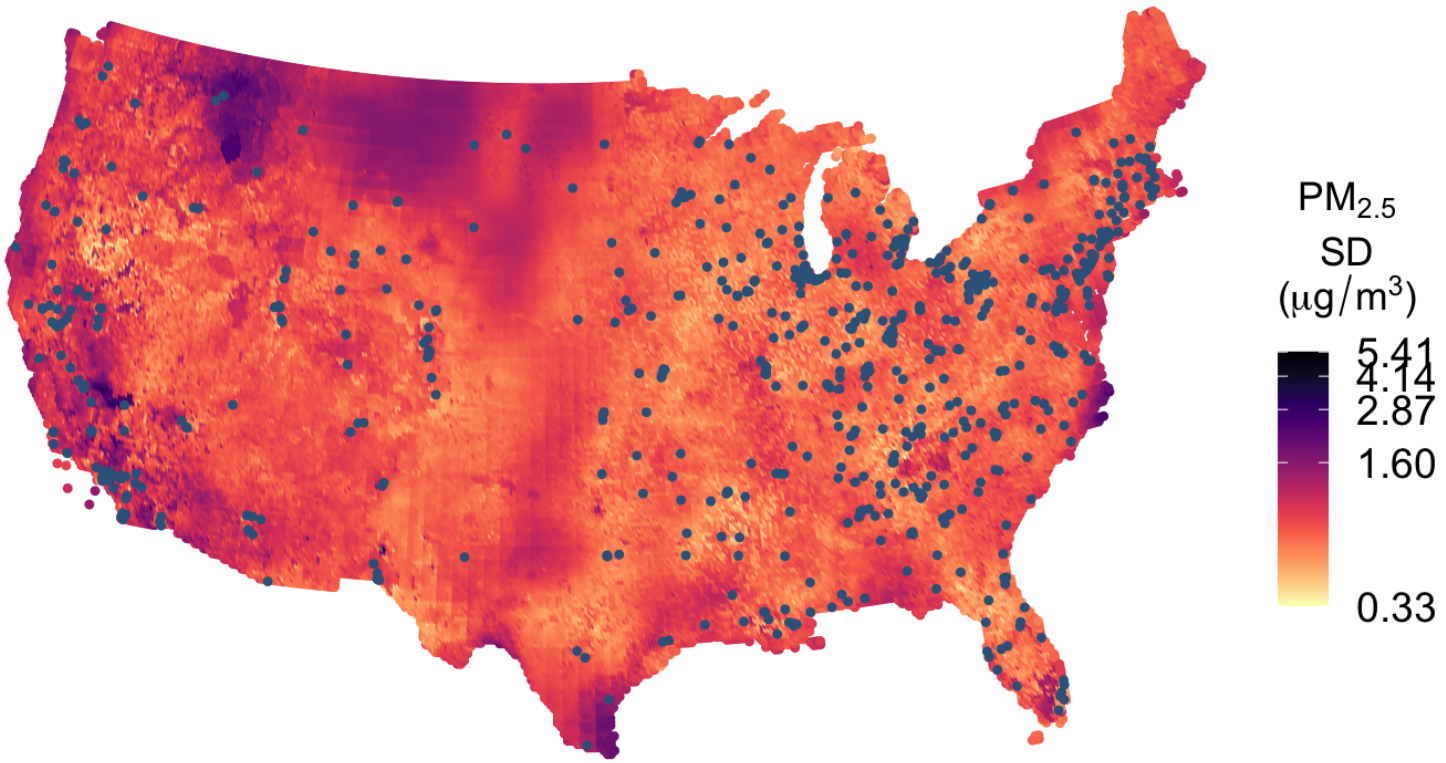
Weights for Base Models (examples)



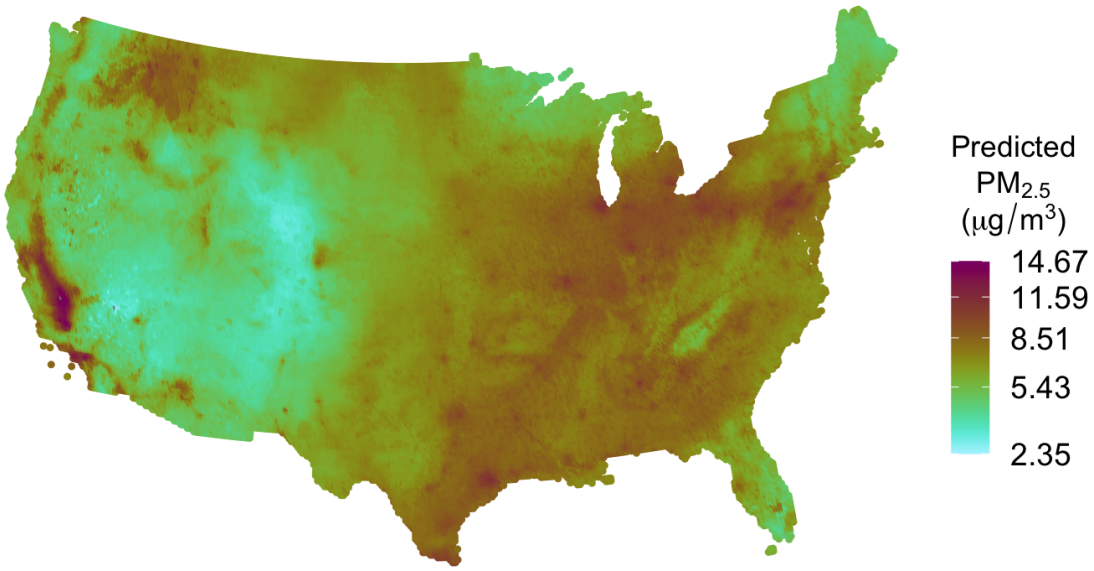
2015 BNE $PM_{2.5}$ Predictions

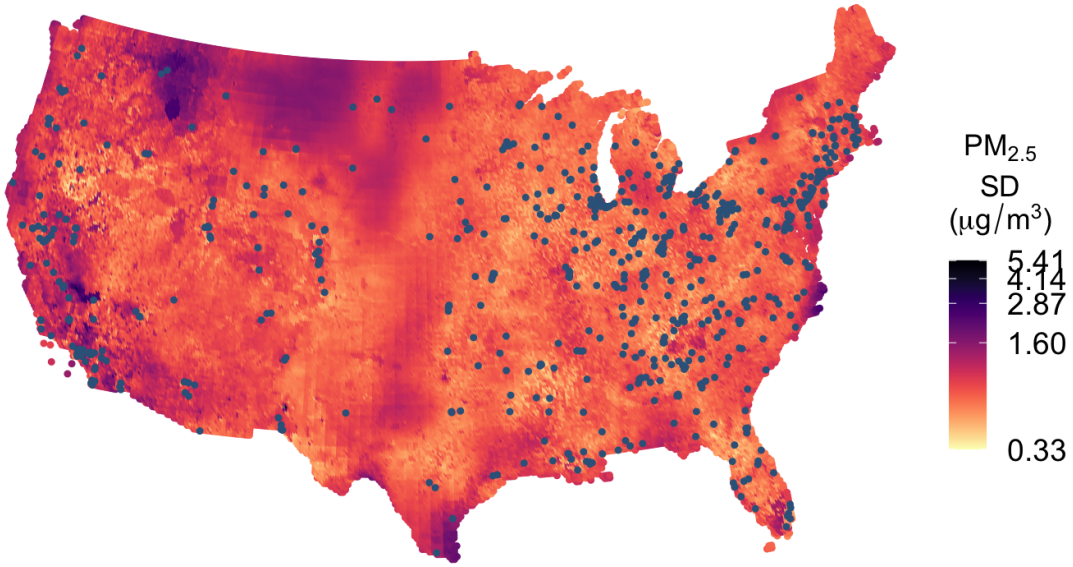


2015 BNE $PM_{2.5}$ Uncertainties



2015 BNE $PM_{2.5}$ Predictions & Uncertainties





How do I know if my prediction model is good?

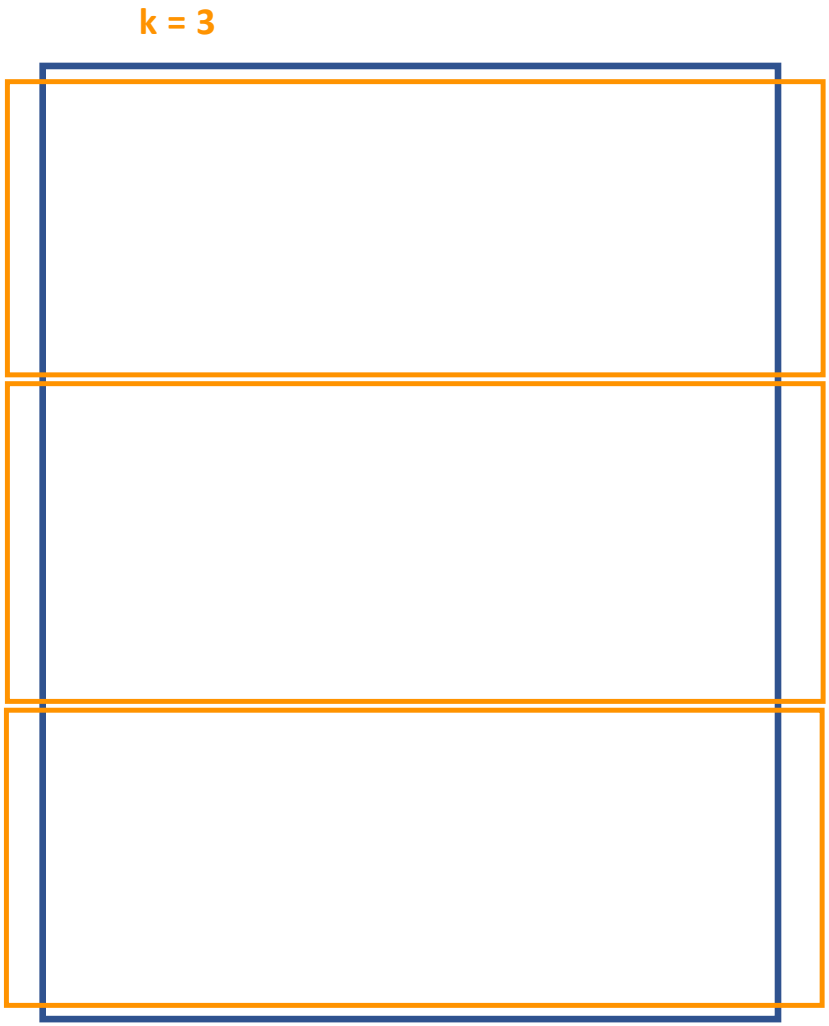
Validation!!

1. External validation dataset
   - Needs to have info on both predictors and dependent variable of interest
2. Cross-validation

k-fold Cross-validation

1. Randomly break the dataset into k unique sets
2. For each of the k sets:
    1. Take the kth set out as the test set
    2. Use the remaining $k-1$ sets as the training set
    3. Fit a model on the training set and evaluate it on the test set
    4. Repeat for all sets

k-fold Cross-validation (graphically)



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

(Doesn't have to be linear regression….)

$$\widehat{\beta_0}, \widehat{\beta_1}$$

Use X from held-out k^{th} sample → Predict Y in k^{th} sample → Compare with observed Y in k^{th} sample

Repeat for all k!

How good is the model? Evaluation

- Using information from external validation, training/testing or cross-validation
- R^2 and β between observed and predicted values
  - We want $R^2 \to 1$ and $\beta \to 1$
- Root mean square error (RMSE)
  $$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_i - \hat{y}_i)^2}$$
- Mean Absolute error (MAE)
  $$\frac{1}{n}\sum_{j=1}^{n}|y_i - \hat{y}_i|$$

Exposure Uncertainty Propagation Into Health Models

- Not to be confused with exposure measurement error correction
  - Although some correction can occur
- Valid inference of health effect estimates

Exposure Uncertainty Propagation Into Health Models

- Not to be confused with exposure measurement error correction
  - Although some correction can occur
- Valid inference of health effect estimates

- Several different ways to do this
- The most straightforward:
  - Run a Bayesian formulation of the health model
  - For the exposure, draw values from the exposure posterior predictive probability

Exposure Uncertainty Propagation Into Health Models

- Not to be confused with exposure measurement error correction
  - Although some correction can occur
- Valid inference of health effect estimates

- Several different ways to do this
- The most straightforward:
  - Run a Bayesian formulation of the health model
  - For the exposure, draw values from the exposure posterior predictive probability

- Application: Long-term $PM_{2.5}$ – mortality in Medicare enrollees (nationwide)

Acknowledgments

Jeremiah Z. Liu	Google
John Paisley	Columbia EE & DSI
Brent A. Coull	HSPH Biostats
Arlene Fiore	Columbia LDEO
Daven Henze	University of Colorado at Boulder
Sebastian Rowland	PSE
Jaime Benavides	Columbia MSPH
Robbie Parks	Columbia MSPH

Funding: NIEHS R01 ES030616

Getting ready for the lab

The lab for this session

The goal of this lab is to take some concepts from the Introduction to Bayesian Nonparametric Ensemble (BNE) lecture and introduce you to the way BNE works.

During this lab session, we will:

1. Explore several air pollution concentration prediction models;
2. Understand how air pollution concentration models can disagree;
3. Understand what BNE does;
4. Explore output of BNE; and
5. Understand why BNE is useful for estimating uncertainty of exposures.

Questions?