Characterizing Highly Resolved Air Pollutant Concentrations and Their Spatio-temporal Uncertainty

Marianthi-Anna Kioumourtzoglou, ScD
mk3961@cumc.columbia.edu

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
ENVIRONMENTAL HEALTH SCIENCES

Introduction

- Air pollution exposure has been linked to numerous adverse health effects
- To improve exposure assessment sophisticated spatio-temporal prediction models are being built
- Incorporating satellite data
- Using machine learning methods that yield high 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)

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

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 (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 et al. 2021
- No standardized way to estimate uncertainties: provided estimates capture different uncertainty components

⇒ 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 et al. 2021
- No standardized way to estimate uncertainties: provided estimates capture different uncertainty components

$\Rightarrow$ 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

Bayesian Nonparametric Ensemble (BNE)

We developed a novel bayesian machine learning tool that:

1. Integrates diverse collection of *existing* model predictions made at different space/time resolutions
2. By adaptively combining predictions of each candidate model by its spatio-temporal predictive accuracy
   - I.e., spatio-temporal weights
3. Provides accurate estimates for the spatially varying predictive *uncertainty*, accounting for the model-to-model variability

Bayesian Nonparametric Ensemble (BNE)

y(x): Pollution Spatial Process

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

We augment a classic ensemble model with nonparametric machinery:

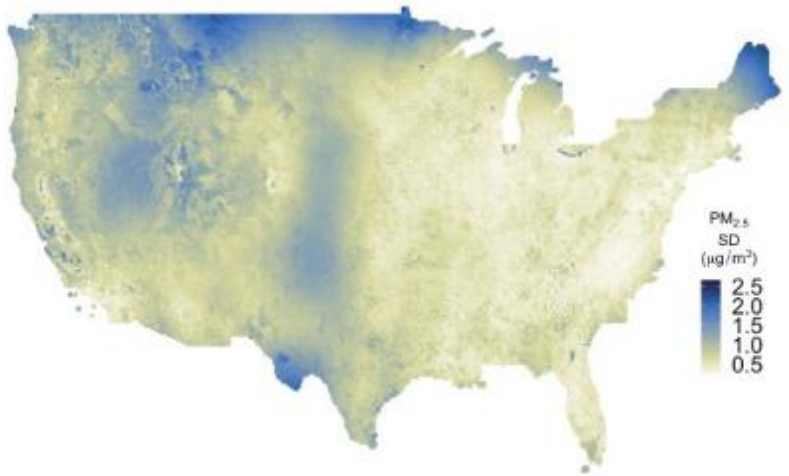
$$\mu(x) = \sum_{k=1}^{K} f_k(x) * \omega_k(x) + \delta(x)$$

$$y|x \sim G_x\left[\Phi(y|x,\mu)\right],$$

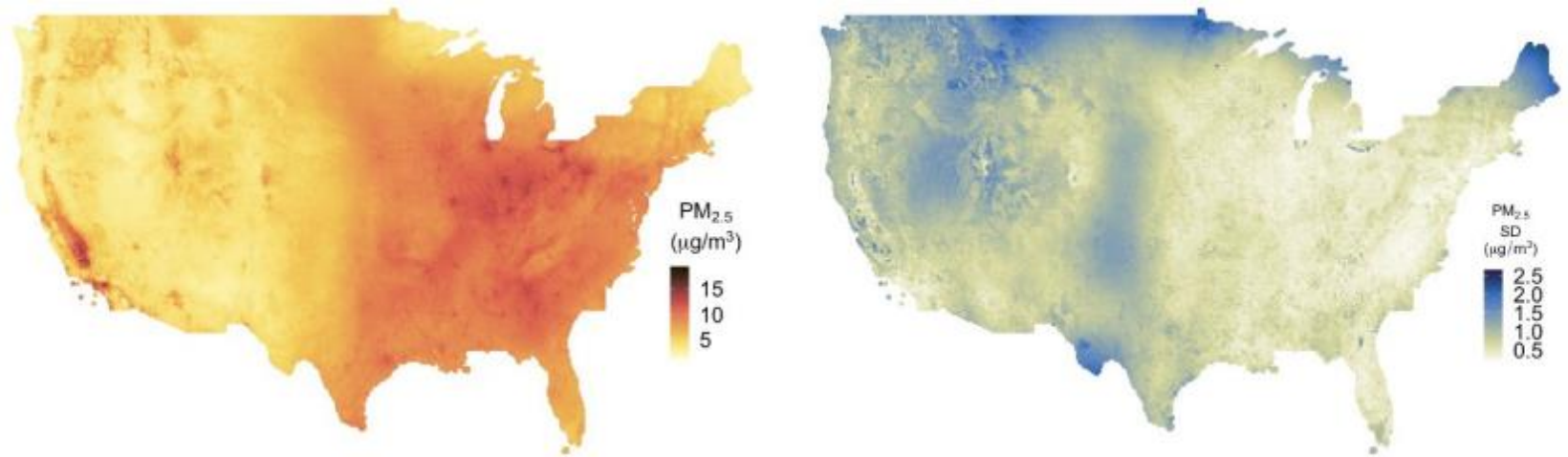
- $\omega(x)$ spatially adaptive model combination
- $\delta(x)$ corrects spatially varying systematic bias in μ
- G_x models spatially varying predictive uncertainty

Liu et al. NeurIPS 2019; Paisley et al. Big Data 2022

2010 BNE Estimates for $PM_{2.5}$



2010 BNE Estimates for $PM_{2.5}$



- For more information on BNE products:
  - Flash talk by Jaime Benavides (yesterday)
  - Poster by Kumar & Carrillo-Gallegos:

Taking BNE globally

- Plans to extend BNE globally
- Challenges for training the model
  - Limited/no monitoring sites at many countries
- Also a great opportunity
  - Identify high-uncertainty spots for optimal monitoring station placement
  - Report weights back to developers of input models → improve each input product
- Uncertainty estimates useful for health impacts assessments and GBD estimates etc

Acknowledgments

Funding:
NIEHS P30 ES009089; R01 ES028805; R01 ES030616; R21 ES028472
NASA 80NSSC21K0509 (HAQAST)