Bayesian Nonparametric Ensemble Method for $PM_{2.5}$ Prediction with Uncertainty

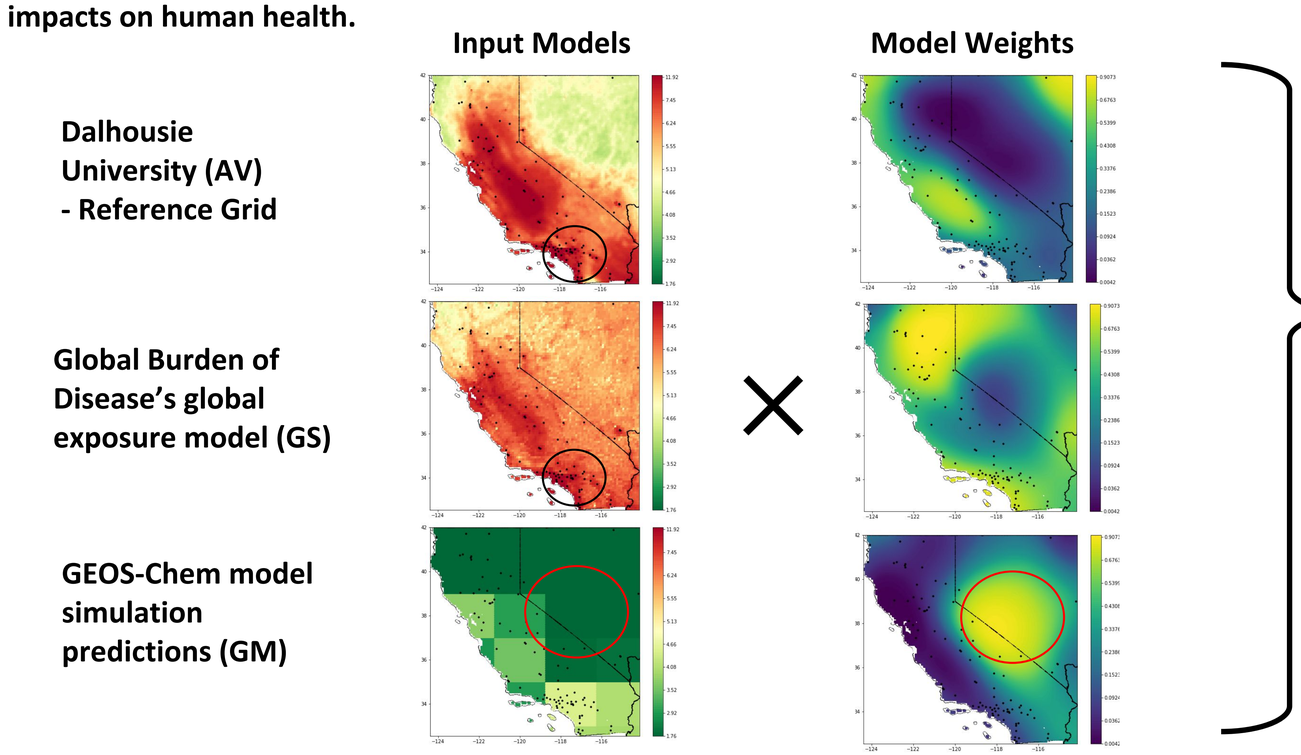
Data Science Institute
COLUMBIA UNIVERSITY

Hammaad Adam, Lea Collin, Ling (Kelly) He, Danielle Su
Marianthi-Anna Kioumourtzoglou, John Paisley

Data Science Capstone Project with Professors Kioumourtzoglou (Mailman) and Paisley (SEAS)

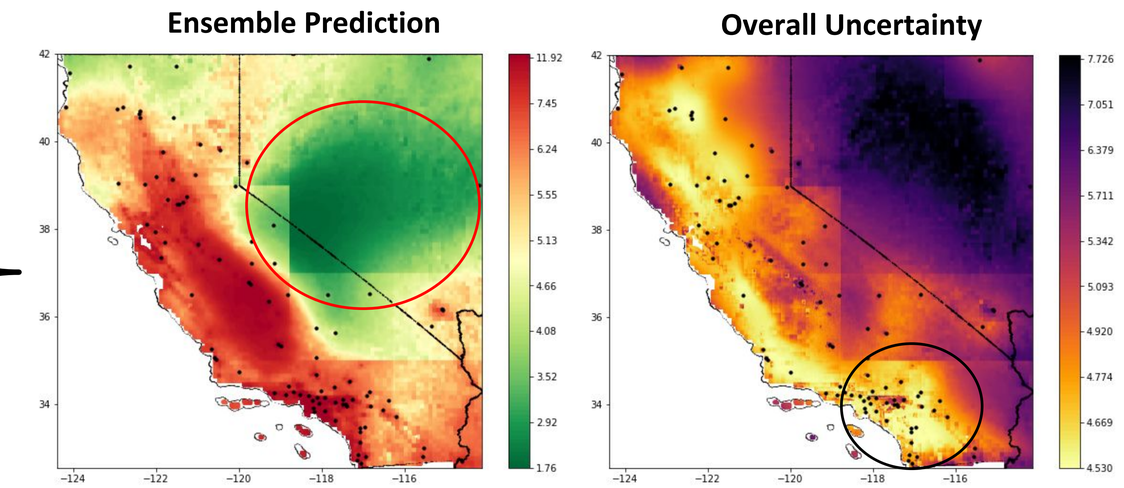
Introduction

The objective is to apply the Bayesian nonparametric ensemble (BNE) method to combine predictions from existing exposure models in order to generate better predictions and characterize uncertainty. The model predicts annual averages of fine particle ($PM_{2.5}$; particles with diameter ≤ 2.5 µm) concentrations. In 2015, exposure to $PM_{2.5}$ was the fifth-ranking mortality risk factor globally. Accurate predictions are key to estimating unbiased health effects from epidemiologic studies and understanding impacts on human health.



Results

The graphs below show how the BNE model takes different models as inputs, calculates the model weights based on training, and outputs both the ensembled predictions and uncertainty at each coordinate. The overall uncertainty is comprised of between-model uncertainty and prediction uncertainty. The example shown is the west coast, California area for 2011 with a model trained on data from 2005, 2010, and 2011.

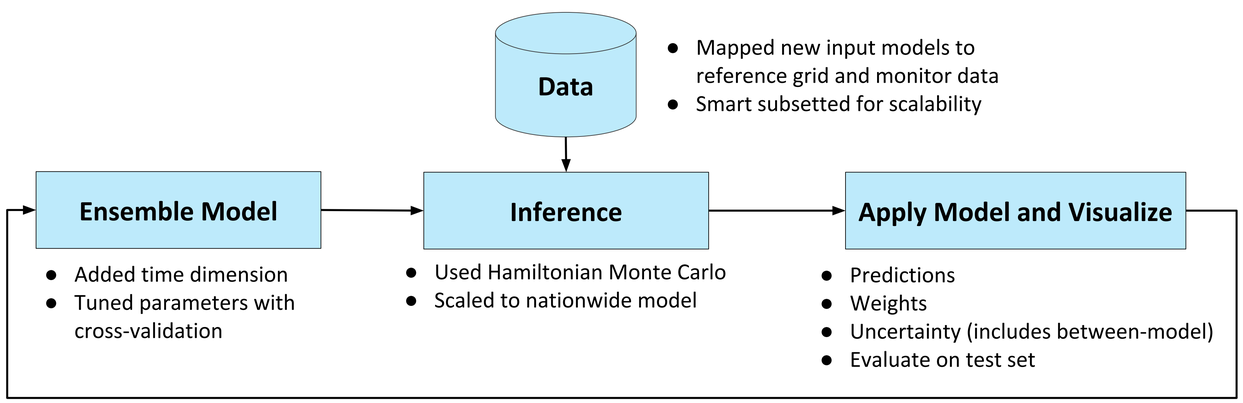


Ensemble predictions show larger "patches" where the GM model is weighted highly because it has the largest grid size.

Uncertainty is lowest around monitors, where we have ground truth, and/or when models agree. (e.g. LA county area)

Methodology

The methodology of using the BNE model to predict $PM_{2.5}$ was previously implemented in the Boston area. The diagram below details the process of taking that model and scaling it to the continental U.S., including different and more input models, and including a time dimension so that the model can utilize data from different years.



Conclusion

This project aims to build an ensemble that 1) produces accurate predictions of $PM_{2.5}$ concentrations and 2) for the first time ever, comprehensively characterizes prediction uncertainty. We have scaled this model to produce nationwide predictions over several years; our next steps include further tuning hyperparameters and evaluating the ensemble's performance on a test set. This model, once further developed and applied to other pollutants, will prove an effective tool both for research and the public.

Acknowledgments

We would like to thank our project mentors John Paisley and Marianthi-Anna Kioumourtzoglou, and Jeremiah Liu for their invaluable help and guidance throughout this semester.

References

Liu et al, *NIPS* (2018); Guo et al, *Atmos. Chem. Phys.* (2018); Shaddick et al, *Environ. Sci. Technol.* (2018); van Donkelaar et al, *Environ. Sci. Technol.* (2019); EPA (2019); IMPROVE (2019)