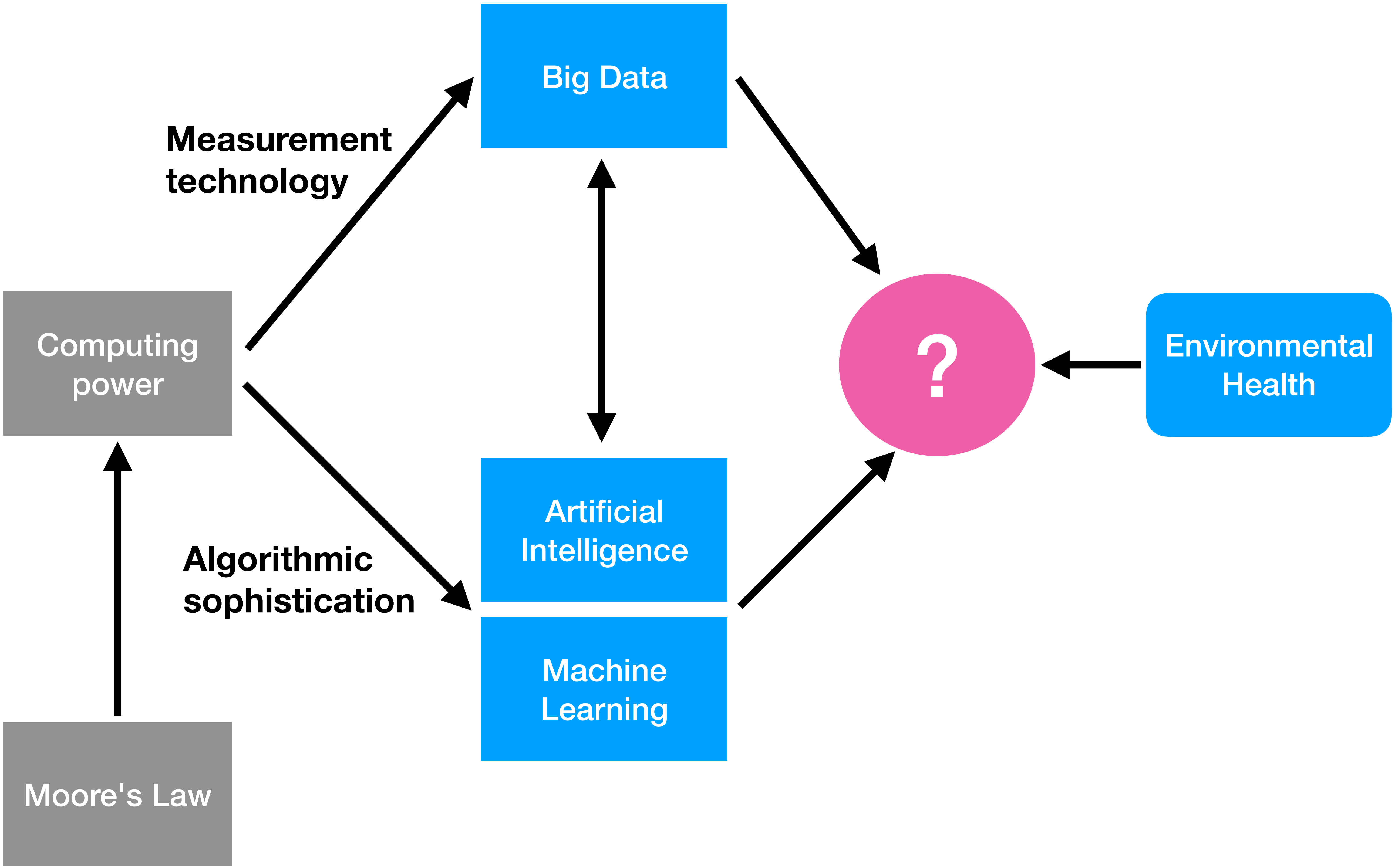
# Opportunities for Artificial Intelligence and Machine Learning in Environmental Health

**Roger D. Peng, PhD**
*Department of Biostatistics*
*Johns Hopkins Bloomberg School of Public Health*

April 2020

# Big Data, AI/ML, Environmental Health



# AI, ML, EH

- NASEM hosted a 2-day workshop on implications of AI and machine learning in environmental health research and decisions (June 6-7, 2019)
- **Applications** - pollution source characterization, exposure assessment, predicting chemical toxicity
- **Challenges** - Data quality/uncertainty; transparency/reproducibility
- NASEM Summary - http://nap.edu/25520

# AI, ML, EH

- How might AI advance environmental health?
- Does AI change the standards used for conducting environmental health research?
- Does the use of AI allow us to change our established research principles?
- How does AI impact our training programs for the next generation of environmental health scientists?
- Are there barriers within the current academic incentive structures that are hindering the full potential of AI, and how might those barriers be overcome?
- Joint statement: https://tinyurl.com/v8fussz

# AI, ML, EH

- There is much we can "bring over" from the AI / ML world to advance environmental health research

- Will need to adapt AI / ML approaches to the specific needs of environmental health research

- Transparency and reproducibility

- Model evaluation methodologies

- Evidence for decision-making

# Decision-Making Levels

- Can AI / ML techniques used to automate "lower-level" modeling decisions?
- Reserve "higher-level" decisions for humans
- Low-level decisions can have large impacts on model results
- AI / ML techniques still require a substantial amount of manual tuning

# Measurement Technologies

- Wearables: accelerometers, sleep-tracking, heart rate
- Exposure monitors: personal monitors, low-cost stationary sensors, crowd-sourced monitoring
- Environment: GIS data, satellite monitoring
- All measured at higher frequencies, and higher spatial resolution

# AI / ML Approaches

- Data processing / transformation / filtering
- Feature selection / engineering
- Model building / evaluation / testing
- Out-of-sample prediction / minimize performance metric
- e.g. Neural network models, random forests, SVM, linear regression (!)

# AI / ML Approaches

- ML approaches generally thrive on large feature sets
- Computationally optimized for fitting complex models to large datasets
- Highly engineered platforms / libraries for executing more "routine" prediction problems (Tensorflow, PyTorch, Keras) at large scale
- Leverage very large datasets where nonlinearities and complex interactions can be observed

# Exposure Assessment: Augmenting Existing Approaches

## Extending the spatial scale of land use regression models for ambient ultrafine particles using satellite images and deep convolutional neural networks

Kris Y. Hong[a], Pedro O. Pinheiro[b], Laura Minet[c], Marianne Hatzopoulou[c], Scott Weichenthal[a,*]

[a] *McGill University, Department of Epidemiology, Biostatistics and Occupational Health, Montreal, QC, Canada*
[b] *Element AI, Montreal, Canada*
[c] *University of Toronto, Toronto, Canada*

# Exposure Assessment: Augmenting Existing Approaches

- Land use regression (LUR) models commonly used to predict levels of ambient PM
- Satellite images + Convolutional Neural Networks (CNN) can expand coverage of LUR models w/missing GIS data
- Satellite images obtained from Google Maps (via ggmap R package)
- CNN trained on LUR output in wider region
- Increased coverage vs. decreased precision

**Hong *et al.* 2019, *Environ. Res.***

# CNN vs. LUR Performance

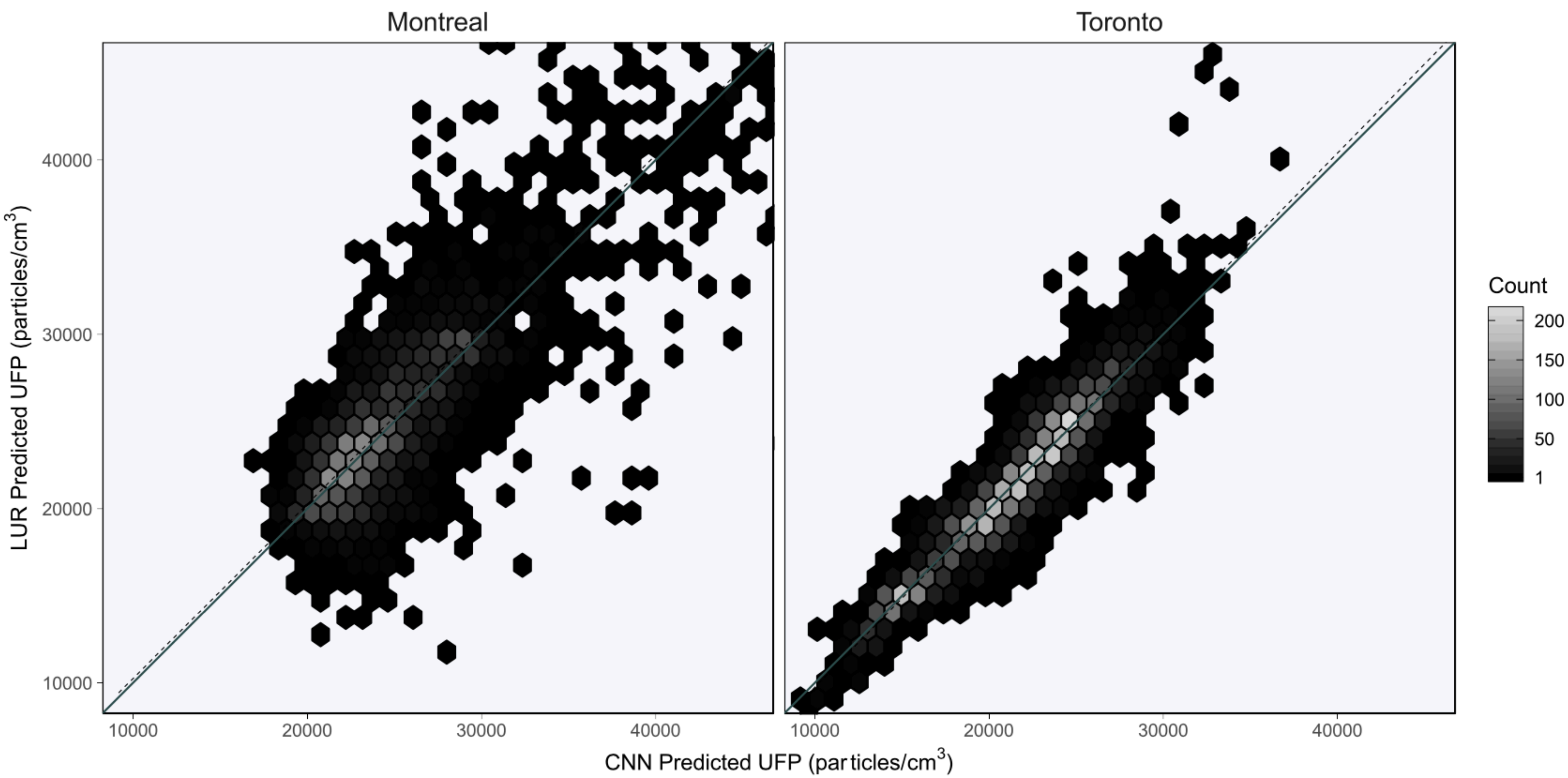



**Fig. 1.** Comparison of LUR-Predicted and CNN-Predicted UFP concentrations (particles/$cm^3$) in Montreal and Toronto, Canada.

Hong *et al.* 2019, *Environ. Res.*

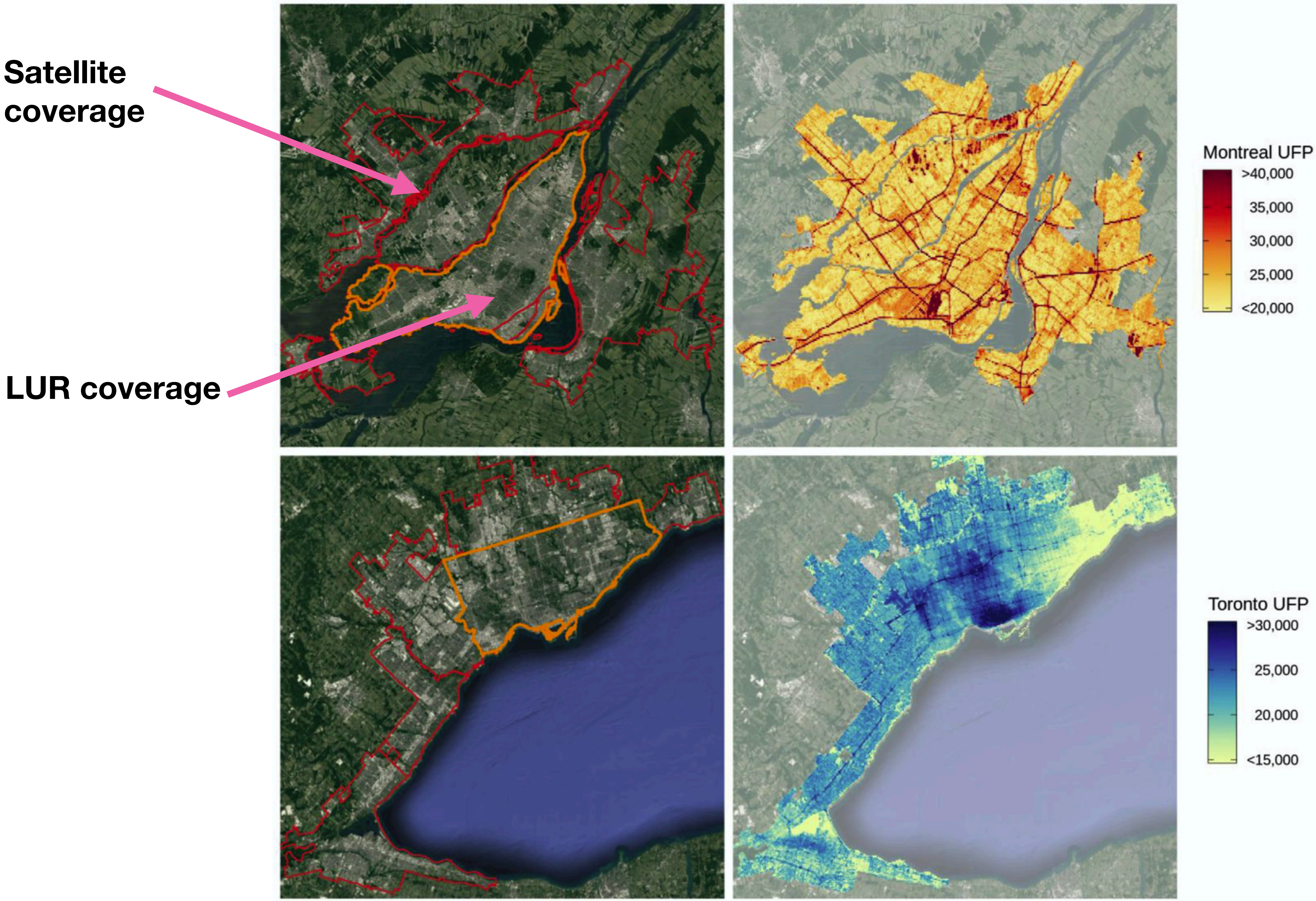


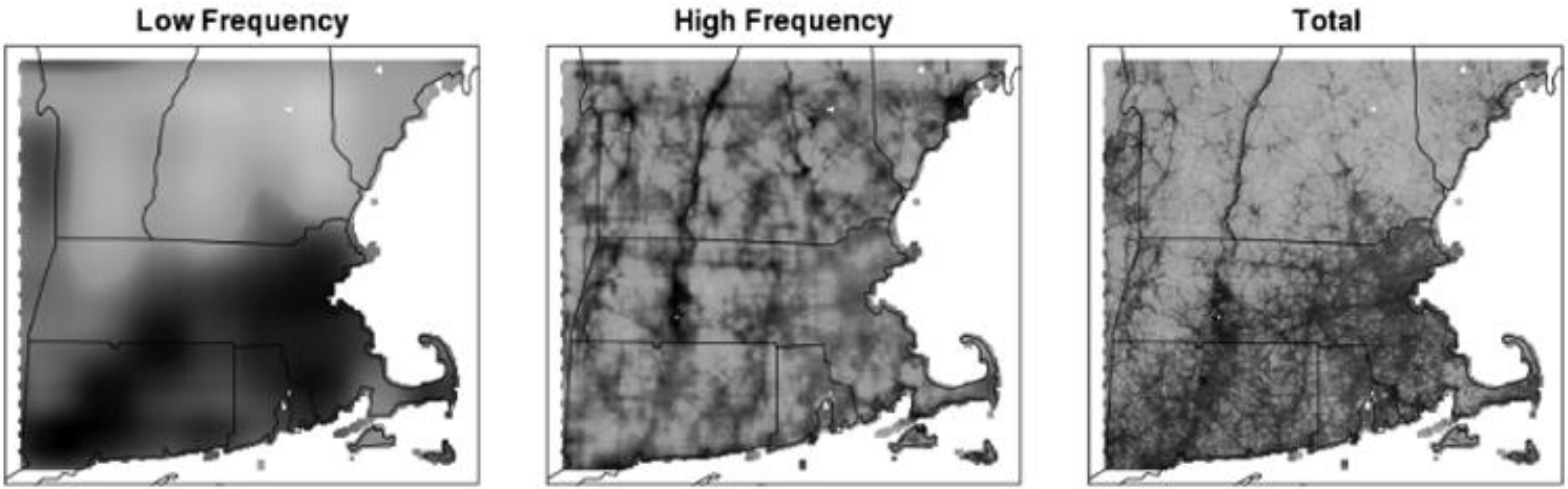

Figure 2

Hong *et al.* 2019, *Environ. Res.*

# Model Evaluation

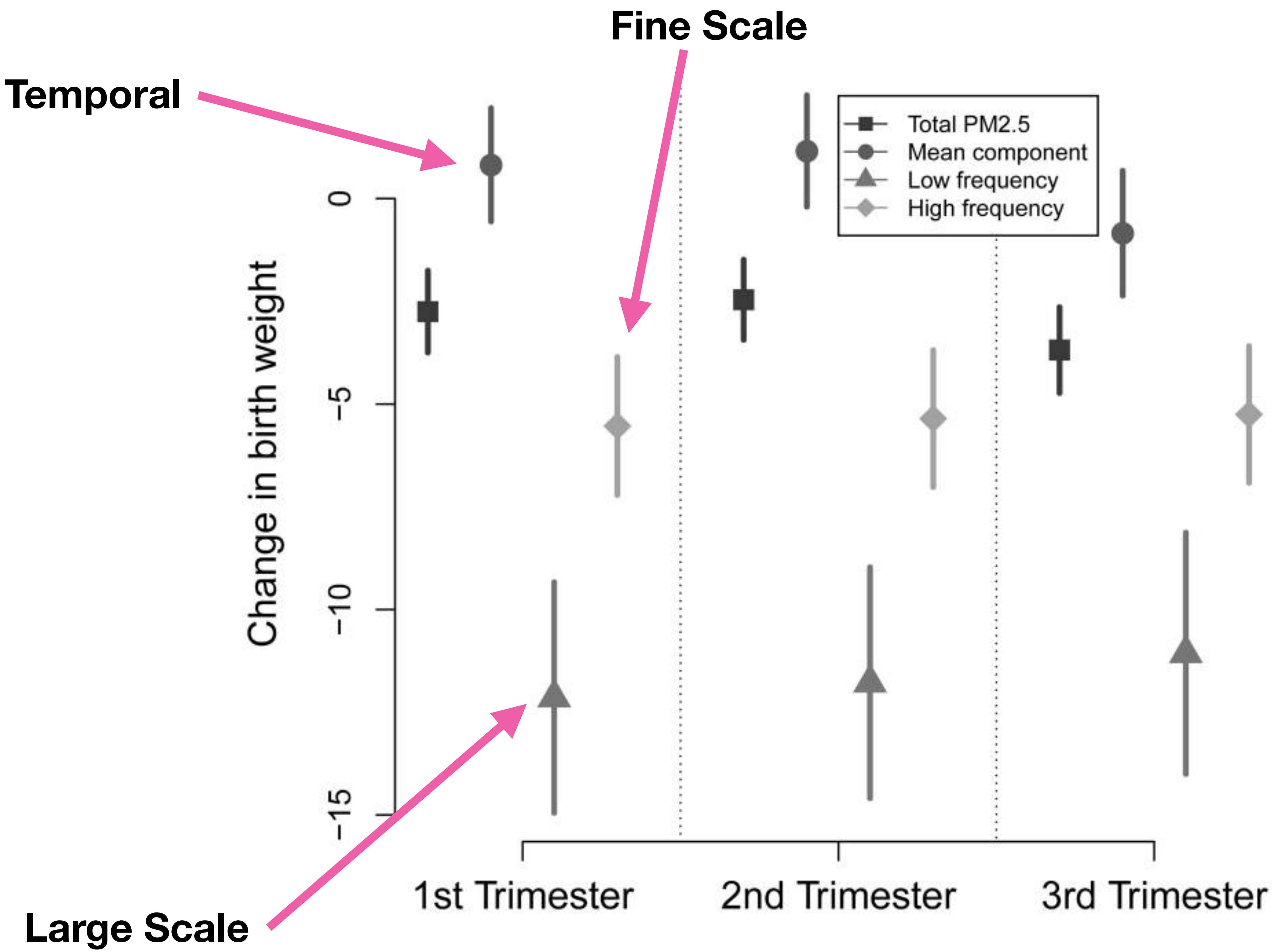
- Environmental health data often measured over space and time
- Estimates of health effects often focus on particular spatial or temporal scales of variation
- AI / ML model evaluation metrics / tools (e.g. $R^2$, RMSE) tend to be more global in nature
- Global metrics can hide errors that may exist at specific temporal / spatial scales critical for air pollution studies

# Spatial Scales of Variation



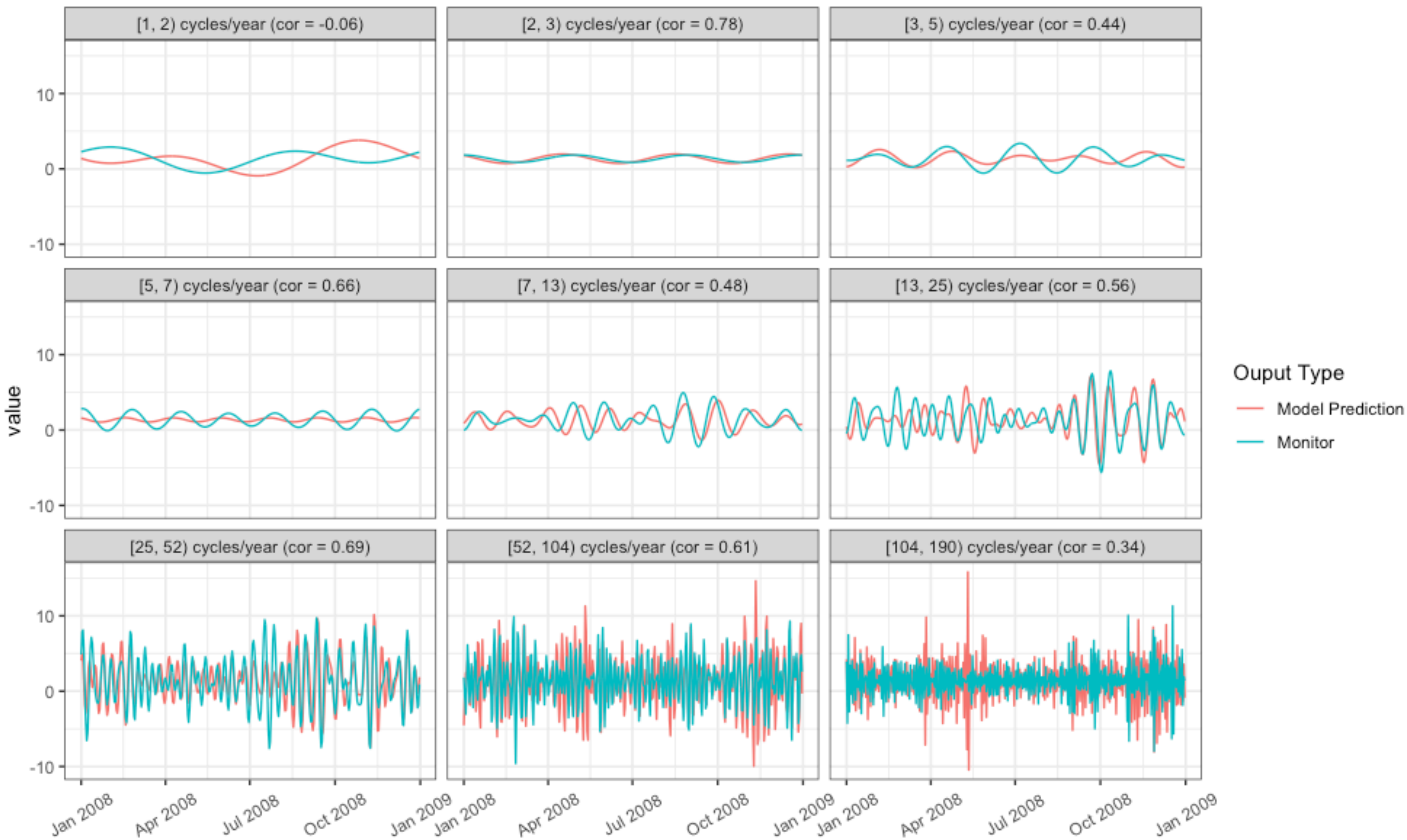

Antonelli *et al.* (2017) *Ann. Appl. Stat.*

# Spatial Scales of Variation




**Antonelli *et al.* (2017) *Ann. Appl. Stat.***

# Temporal Scales of Variation



# Transparency and Reproducibility

- AI / ML methods introduce dramatic increase in complexity for both the data and the methods
- Would you accept a paper that did a logistic regression, but did not publish the weights? (J. Muschelli - https://tinyurl.com/rm3kzv5)
- Many more details must be disclosed for reproducibility, increased complexity for disclosure too
- Minor variations on standard ML platforms can be difficult to reproduce
- Reproducibility = understanding what is going on, **not** a badge of quality
- "Trust me, it just works" $\neq$ Science

# Transparency and Reproducibility

- McKinney, S. M., *et al.* International evaluation of an AI system for breast cancer screening. *Nature* (2020).

- "*The code used for training the models has a large number of dependencies on internal tooling, infrastructure and hardware, and its release is therefore not feasible.*" (authors all employees at Google, Inc.)

- Haibe-Kains, *et al.* - "Even with sufficient description, reproducing complex computational pipelines based purely on text is a subjective and challenging task."

# Details, Details...

- Training pipeline / data transformation / feature engineering
- Hyperparameters defining model structure
- Stochastic data transformations / model elements
- Fitting algorithm details (stochastic gradient descent) / custom tuning
- Proprietary datasets

# Social/Ethical Considerations

- AI / ML research in EH can lead to highly consequential decisions being made
- AI / ML experts and EH stakeholders need aligned interests; trust in the process of evidence generation
- Accountability - reduce information asymmetries between various stakeholders
- Justification for the problem addressed; methods used; limitations of methods and training data
- Understanding of consequences of computation

# Social/Ethical Considerations

- Data scientists have a "fiduciary duty" to use data in a way that does not betray end users and/or harm them
- Data are not abstract, not a "natural resource" -- they are produced by and have an impact on humans
- Tools (e.g. checklists) can be developed to implement AI principles, but ethics is ultimately a socio-cultural concept
- Ethical considerations unify areas of product development and scientific research

**Madaio, *et al.* (2020), *CHI***
**Stark and Hoffmann (2019), *J. Cultural Analytics***

# Summary

- AI / ML approaches have potential to be used widely in environmental health research
- AI / ML methods need to adapt to specific issues in environmental health research
- Transparency and reproducibility is critical for building trust and for aligning stakeholders
- Decision-making in environmental health typical relies on numerous pieces of evidence; AI / ML findings can have a place in that framework