Carnegie Mellon University

Machine Learning Methods to Predict Air Pollution Concentration for Policymakers

for presentation at 19th Annual CMAS Conference 26-30 Oct 2020

Samuel Jones, Ph.D., Chief of Staff of the Air Force Strategic Ph.D. Fellow

Prof. Kun Zhang, Ph.D., *Philosophy & Machine Learning*

Prof. Nicholas Muller, Ph.D., *EPP & Tepper School of Business*

Prof. Peter Adams, Ph.D., *Civil and Environmental Engineering & EPP*

Disclaimer

The research and views presented are those of the author and do not necessarily represent the views of the United States Government, the Department of Defense or any of its components.

Introduction

- Policymakers want to make the best decisions possible
- Evidence-based policymaking is desired, but there are at least two limitations:
  - Data
  - Models – including heuristics (mental models)

- ML algorithms can make inroads by:
  - identifying and quantifying trends not deducible by other means
  - increasing the speed and ease at which analysis is performed

Aims and objectives

- Discover the causal relationship of Elemental Carbon emissions to concentrations as a step toward a better model to predict annual concentrations which have shown to have impact on human health
- Develop less computationally-expensive air pollution model than CTM

Why are these hard problems?

- Cannot perform large-scale true experiments with air pollution; unethical/harmful to intentionally pollute to measure human health response
- CTMs are very complicated and computationally expensive
- While carbon is well-understood, Volatile Organic Compound (VOC) relationships are not fully understood and therefore difficult to predict, difficult to trace Particulate Matter (PM), need a method to better understand VOCs and then apply it to Air Monitoring Data

CTMs are gold standard but expensive and there's still room for better understanding of chemical relationships

- Air pollution and $PM_{2.5}$ are linked to adverse health and mortality (Karydis et al. 2007; Peng et al. 2017; Stieb et al. 2002)

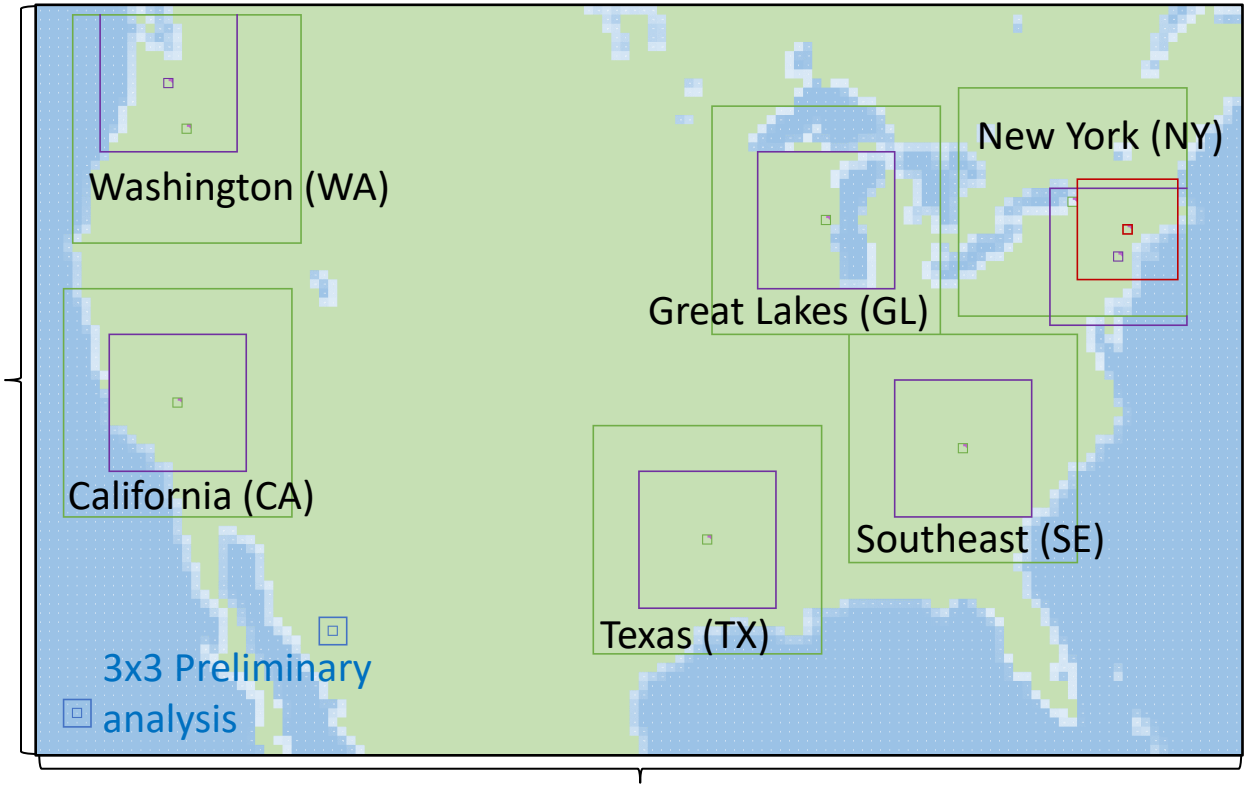
- Reduced Complexity Models are sufficiently accurate for use in policy exploration, but CTM required/recommended before implementation (Gilmore et al. 2019; Heo et al. 2016; Muller et al. 2011)

- ML techniques are being used in air pollution research (Feng et al 2015, Kleine Deters et al. 2017, Kelp et al. 2019, 2020, Xue et al. 2019, Bellinger et al. 2017)

- “Doing one of these requires both a lot of expertise and a lot of time,” says Adams. “These are the gold standard models, but the days or weeks of computer time is not even the big cost. The big cost is it takes months to prepare the inputs, and then months to analyze the data. It's definitely not user-friendly.”
  (“Researcher wants to put the power to model air pollution into your hands” 2017)

- Target use case: air quality engineers, policymakers, citizen scientists
  - Make a more accessible model in terms of time, hardware, knowledge

Lambert Conformal Projection with 36 x 36 km grid cells showing water land use and region samples



11x11 Preliminary analysis and Hyperparameter tuning

15x15 trained on PC using NVIDIA GeForce MX150 w/ 2 GB memory

25x25 trained on server using NVIDIA Tesla T4 w/ 15 GB memory

PMCAMx daily data files = 1.28 TB/year

Table 1. Technical information about PMCAMx daily data files.

File Type (extension)	Layers	Rows	Columns	Variables	Time Steps
Output Files					
Daily Hourly Output	*1*	*82*	*132*	*16*	*24*
Emission Source Input Files					
Area - On Road Pollution	*1*	*116*	*152*	*12*	*24*
Area - Non-road Pollution	*1*	*116*	*152*	*12*	*24*

Time information is encoded using 6 variables and provided in VAR and input data:

Table 2. Time and seasonality variables

Variable	Possible Values
Year	1990, 2001, 2010
Month	1-12
Day of Year	1-366
Day of month	1-31
Weekday	1-7
Hour	1-24

Granger Causality with Neural Nets to model non-stationarity and non-linearity of temporal & chemical effects

Neural Net

$$C_t = G\left(\sum_{i=1}^{k} A_i^{(t)} C_{t-i}, \beta^{(t)}, E^{(t)}\right)$$

Neural Net

C is concentration

G is neural net that takes all inputs

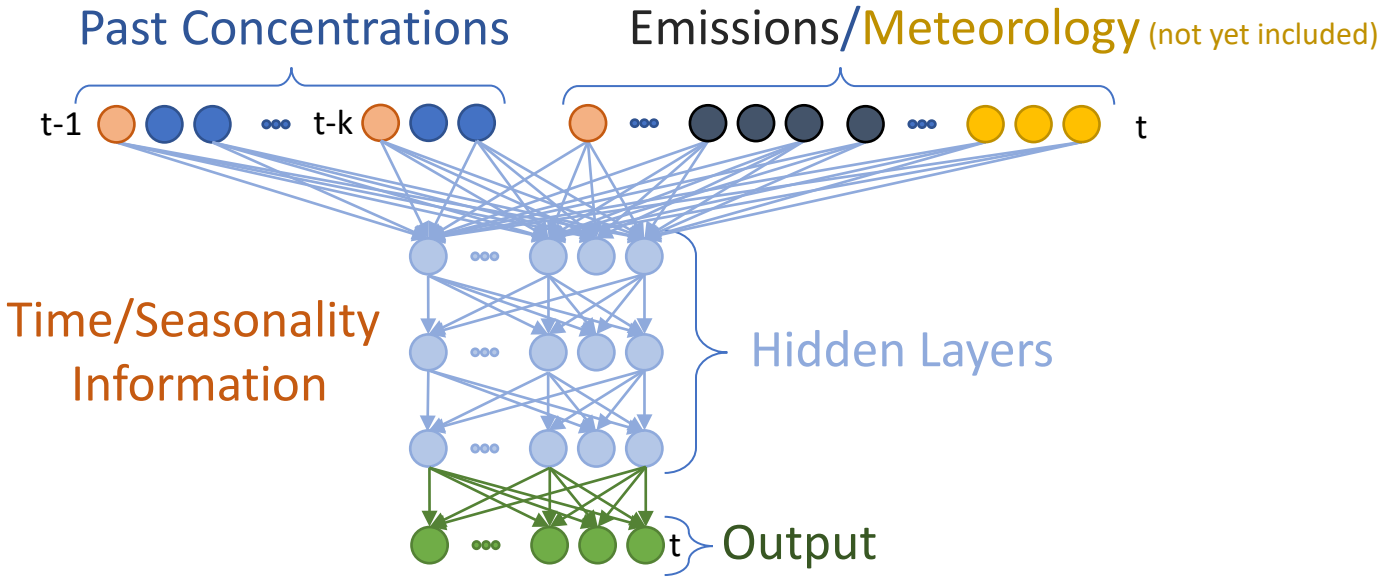
A is coefficient for concentration with different influence from different time (t) periods, causal relation between locations, can be represented by a neural network

β is temperature, wind, etc. (not used in initial model)

E is emissions

Builds upon Constraint-based causal Discovery from Nonstationary/heterogeneous Data (CD-NOD) (Zhang et al, 2019, "Causal discovery from nonstationary / heterogeneous data")

Hybrid Vector Autoregressive Neural Network with Time



For simplicity not all connections shown

Train/Test Cycle

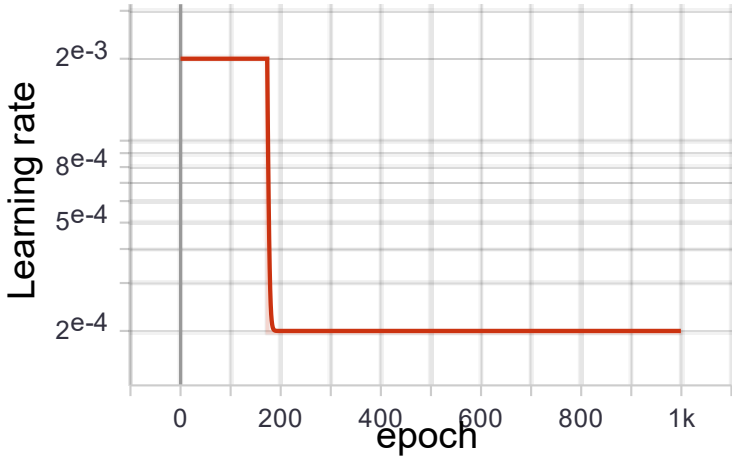
Gather and subset species and region from data files (input/output)

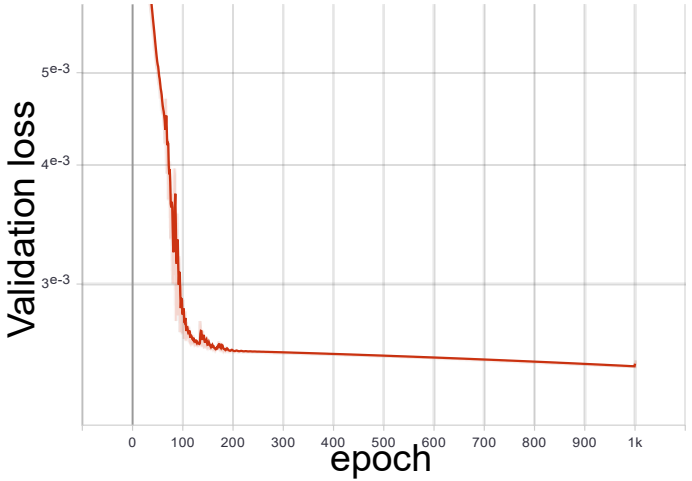
Add lags to create VAR data

Identically randomize data order in all 3 datasets

Split input, VAR and output data into matching 80%, 5%, 15% (train, validate, test sets)

Train model using: Adam optimizer with learning rate reducer, Huber (L-1 smooth) loss function, and an early stopping condition

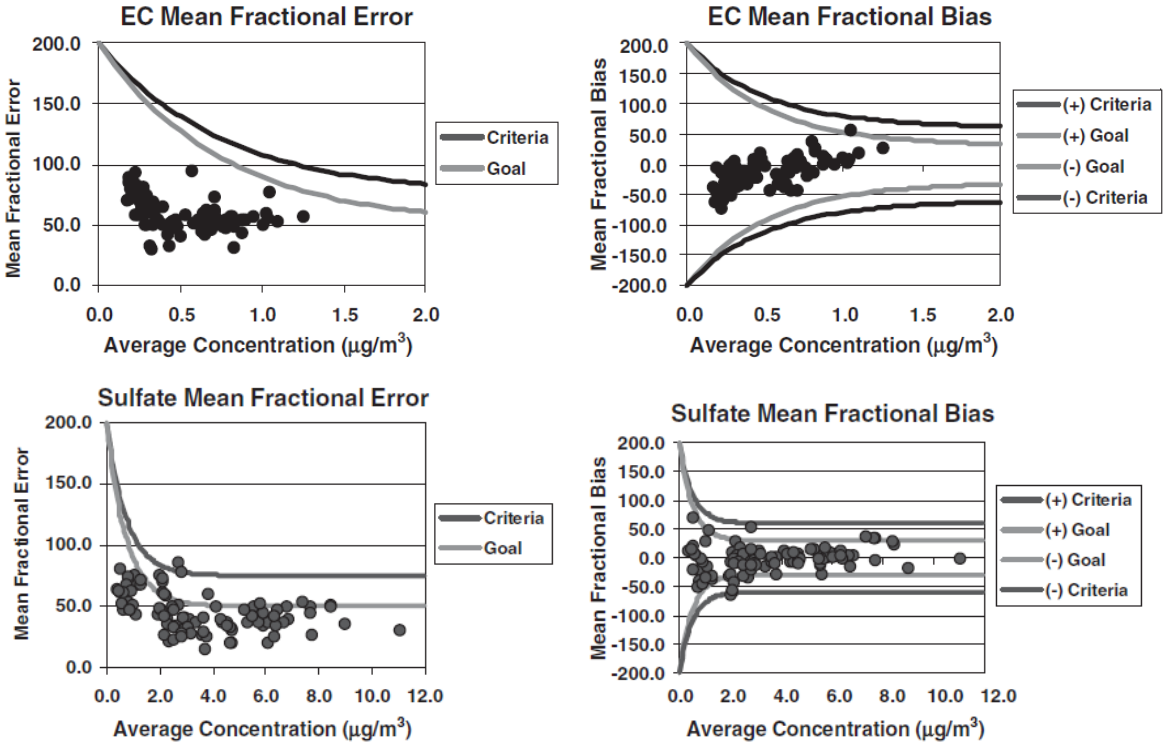




Hybrid VAR Neural Network with Time model training and test statistics for a 25 x 25 grid region for 2 sizes of EC on a server with a single NVIDIA Tesla T4 GPU with 15 GB of memory. Best value in **bold**.

Region	Hidden Layers	Neurons/ Hid Layer	Test Loss	Test MSE	All MSE	Time to train 1k epochs	Time to predict 3 yrs
CA	5	1262	0.0038	0.00774	0.0076	00:25:24	**00:00:02**
GL	5	1262	0.0053	0.011	0.0108	**00:23:24**	**00:00:02**
NY	5	1262	0.0084	0.0174	0.0174	00:26:09	**00:00:02**
SE	5	1262	0.0065	0.0132	0.0131	n/a	**00:00:02**
TX	5	1262	0.0050	**0.0102**	0.00997	00:26:50	**00:00:02**
WA	5	1262	**0.0014**	0.00276	**0.00276**	00:25:47	**00:00:02**
All 6	5	1262	0.0202	0.0435	0.0439	00:42:02	**00:00:13**

How good is good enough?



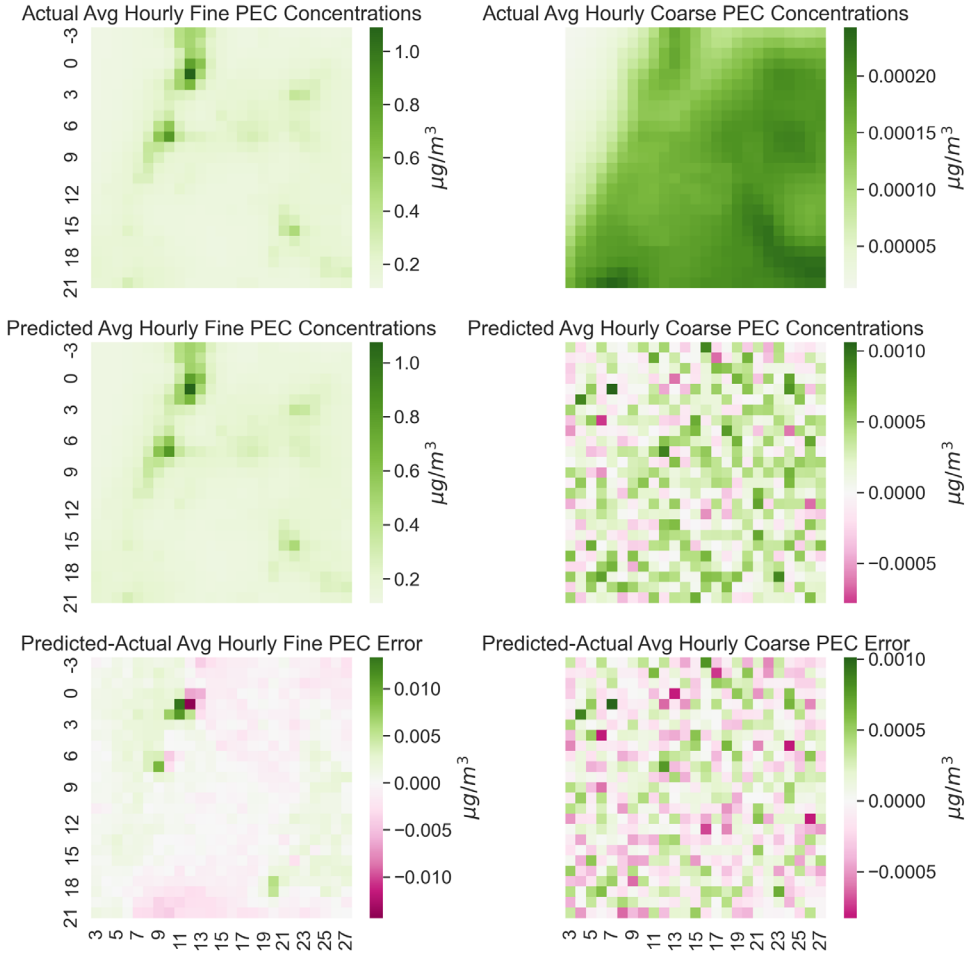
Elemental carbon ($\mu g/m^3$) and Sulfate($\mu g/m^3$) MFE (left) and MFB (right) for all benchmark runs compared to proposed performance goals and criteria from Boylan and Russell (2006) shown here as reference for criteria and goal performance measures.

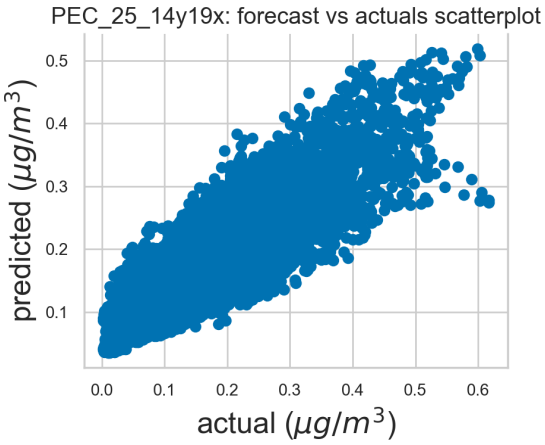
MFE and MFB average statistics meet goal standards for Fine EC

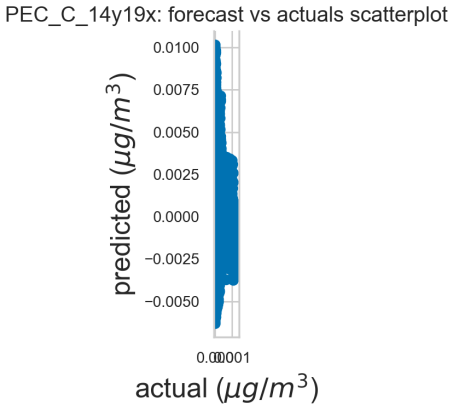
(Italicized MFE and MFB values meet accuracy goals set forth by Boylan and Russell (2006))

	$EC_{2.5}$				EC_C			
Region	**MFE ↓**	**MFB ↓**	**Pearson ↑**	**FAC2 ↑**	**MFE ↓**	**MFB ↓**	**Pearson ↑**	**FAC2 ↑**
CA	*0.268*	*0.043*	0.739	0.943	-0.399	2.059	0.013	0.031
GL	*0.249*	***0.030***	0.784	**0.955**	-0.309	1.964	-0.002	0.038
NY	*0.281*	*0.059*	**0.818**	0.928	0.175	1.392	0.008	0.060
SE	*0.268*	*0.065*	0.759	0.930	**-3.248**	4.840	0.002	0.052
TX	***0.241***	*0.053*	0.781	0.950	-1.468	3.063	0.010	0.050
WA	*0.248*	*0.047*	0.757	0.944	7.366	**-5.837**	**0.052**	**0.080**

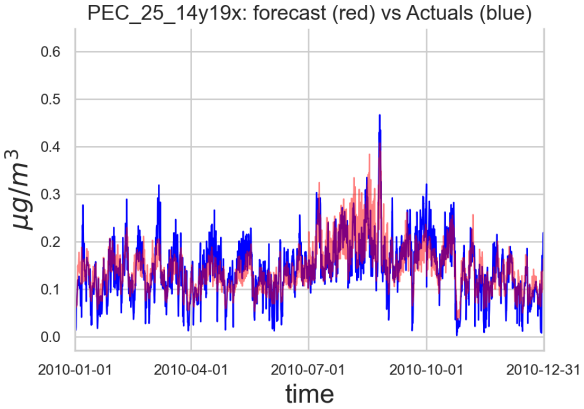
Better at predicting fine EC than coarse EC

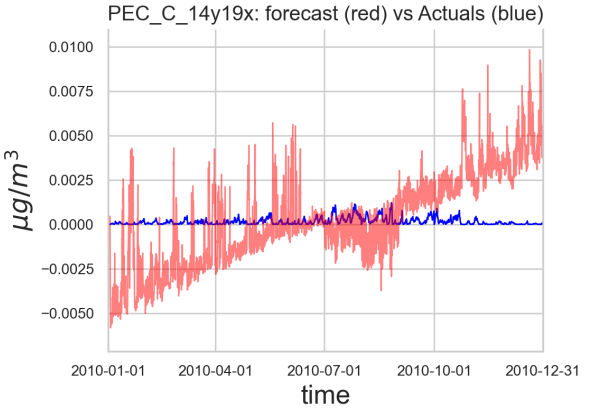






Better at predicting fine EC than coarse EC



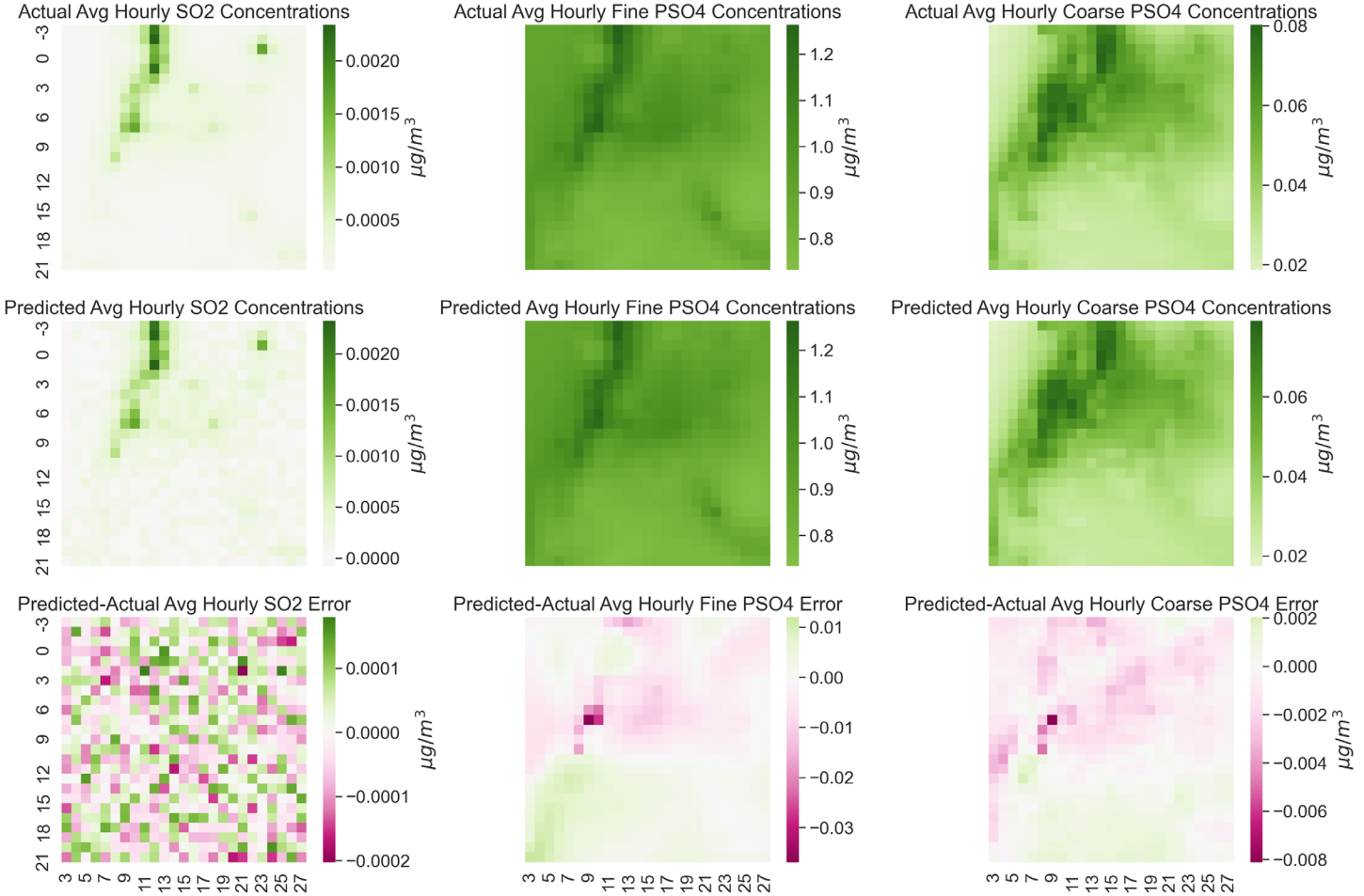


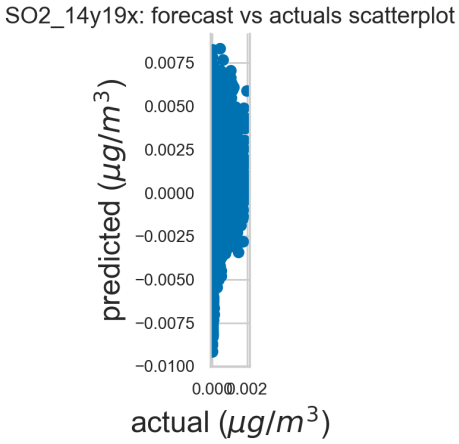
MFE and MFB average statistics meet goal standards for Fine PSO_4

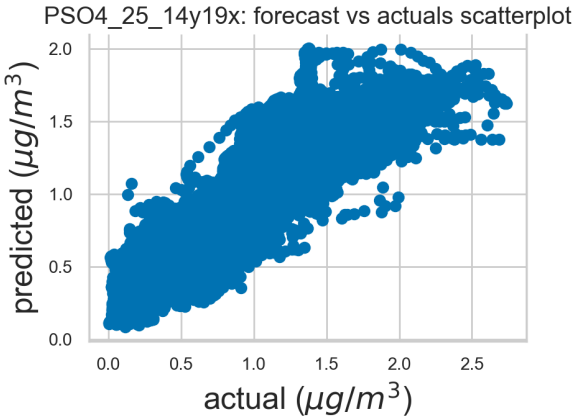
(Italicized MFE and MFB values meet accuracy goals set forth by Boylan and Russell (2006))

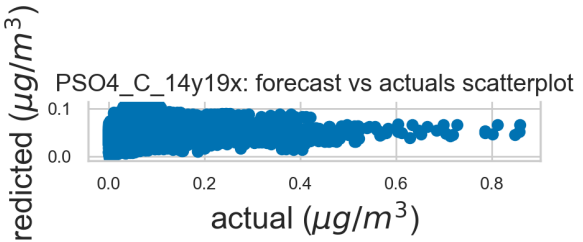
	SO_2				Fine PSO_4				Coarse PSO_4			
Region	**MFE ↓**	**MFB ↓**	**Pearson ↑**	**FAC2 ↑**	**MFE ↓**	**MFB ↓**	**Pearson ↑**	**FAC2 ↑**	**MFE ↓**	**MFB ↓**	**Pearson ↑**	**FAC2 ↑**
CA	-0.315	1.885	0.017	0.062	**0.121**	*0.016*	0.794	**0.988**	0.898	0.550	0.358	0.359
GL	**-2.893**	4.115	0.088	0.167	*0.147*	**0.014**	**0.952**	0.979	**0.628**	0.772	0.588	0.288
NY	-0.670	1.816	**0.107**	0.198	*0.169*	*0.016*	0.952	0.970	1.042	0.874	0.556	0.270
SE	0.862	**0.207**	0.087	**0.218**	*0.194*	*0.032*	0.928	0.954	0.803	0.535	**0.639**	0.377
TX	0.618	0.610	0.053	0.139	*0.173*	*0.023*	0.946	0.968	0.883	0.504	0.571	0.386
WA	-0.635	2.187	0.034	0.067	*0.161*	*0.031*	0.726	0.963	0.884	**0.473**	0.289	**0.391**

Better at predicting fine PSO_4





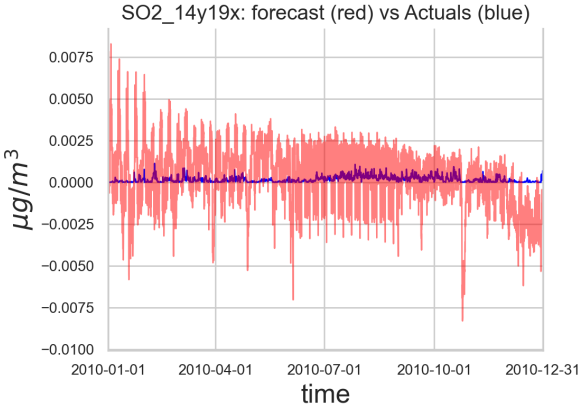


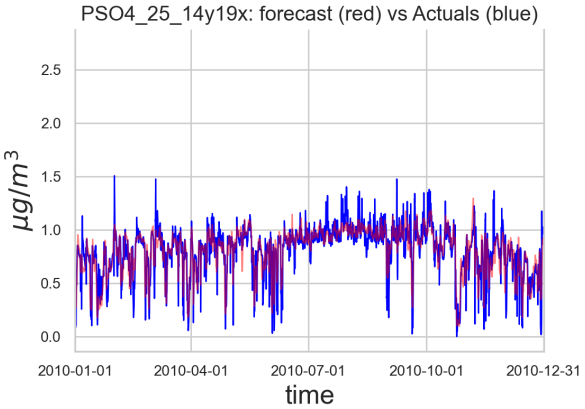


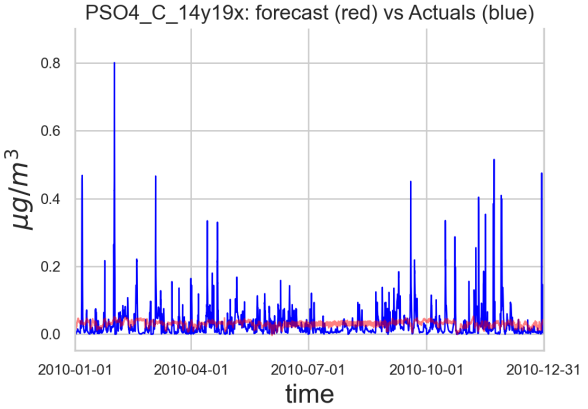
SO_2

Fine PSO_4

Coarse PSO_4







What's next?

- Continue air pollution research using remaining pollution and meteorological data (next slide) and then "real-world" data
  - Run model on each variable independently to determine absolute contribution of each variable
  - Integrate one at a time to determine differential benefit of each type

PMCAMx daily data files = 1.28 TB/year

Table 1. Technical information about PMCAMx daily data files. *Italics* indicate data used in model.

File Type (extension)	Layers	Rows	Columns	Variables	Time Steps
Output Files					
Daily Hourly Output	*1*	*82*	*132*	*509*	*24*
Meteorology Input Files					
Vertical Diffusivity	14	82	132	1	24
Land Use	1	82	132	11	1
Water Vapor	14	82	132	1	24
Temperature	14	82	132	2	24
Wind Speed	14	82	132	2	24
Height Pressure	14	82	132	2	24
Cloud/Rain	14	82	132	3	24
Snow	1	82	132	1	1
Emission Source Input Files					
Area - On Road Pollution	*1*	*116*	*152*	*114*	*24*
Area - Non-road Pollution	*1*	*116*	*152*	*114*	*24*
Point - Electricity Generating Units (EGU) pollution	1	8304	(x, y, z, coords included in data)	128	24
Point - Non Integrated Planning Module (IPM)	1	102951	(x, y, z, coords included in data)	128	24

Limitations and Applicability

- By using inputs and outputs of CTM, ML will be limited to:
  - being able to interpret or discover relationships included in the CTM
  - accuracy of CTM
- Must use real-world measurements to potentially discover as-yet-not-understood relationships or increase accuracy
- Dramatic shift in pollution regime (fuel type and quantity) may be significant enough to prevent algorithm from discovering patterns
- As weather variables are added, the time variables may become unnecessary
- Need to use predictions as pasts values to test stability of model

Summary and conclusions

- If similar performance was achieved with real-world data, the model would have met goal performance standards
- Potential of this model to further improve with additional input factors is high and should be pursued further
- The flexibility of the model to work at different geophysical scales to fit available hardware makes the model more accessible to the target audience

Supported By:

Engineering and Public Policy

Dissertation Committee:
Dr. D. Armanios, Co-Chair
Dr. K. Zhang, Co-Chair
Dr. P. Adams
Dr. N. Muller

Backups

Ethical Aspects of Study

Performing true experiments on people to discover the effects of infrastructure or $PM_{2.5}$ VOC pollution would be unethical and harmful ML methods often impute the ethics of their designers, therefore special care must be taken to ensure our own biases and ethics are not imputed to the methods used. Further, care must be taken when selecting data sets from which to train ML algorithms. Bias in the data can translate into biased results. Therefore a robust data set with appropriate variety should be used to train algorithms.

CTMs are gold standard but expensive and there's still room for better understanding of chemical relationships

- Air pollution and $PM_{2.5}$ are linked to adverse health and mortality (Karydis et al. 2007; Peng et al. 2017; Stieb et al. 2002)
- Due to temporal and spatial interactions, some relationships between chemical species is not yet well understood or accurately predicted (Fiore et al. 2003; Karydis et al. 2007; Stieb et al. 2002)
- Reduced Complexity Models are sufficiently accurate for use in policy exploration (Gilmore et al. 2019; Heo et al. 2016; Muller et al. 2011)
- ML techniques are being used in air pollution research (Feng et al 2015, Kleine Deters et al. 2017, Kelp et al. 2019, 2020, Xue et al. 2019, Bellinger et al. 2017)
- ML algorithms have been used successfully to model non-linear relationships not readily identified by other techniques (Hornik 1991; Hornik et al. 1989; Huang et al. 2016)

Table S9. LinVARNN hyperparameter tuning to determine optimal pairing of loss function and optimizer with 3x3 Mexico EC subset

Loss		Optimizer			Final	hh:mm:ss
Huber	MSE	RMSProp	Adam	SGD	Validation Loss	Time to train 1k
X		X			0.00002519	00:06:42
	X	X			0.00005232	00:06:09
X			X		0.00001754	00:06:41
	X		X		0.00003841	00:06:08
X				X	0.00836772	00:07:19
	X			X	0.00814624	00:06:41

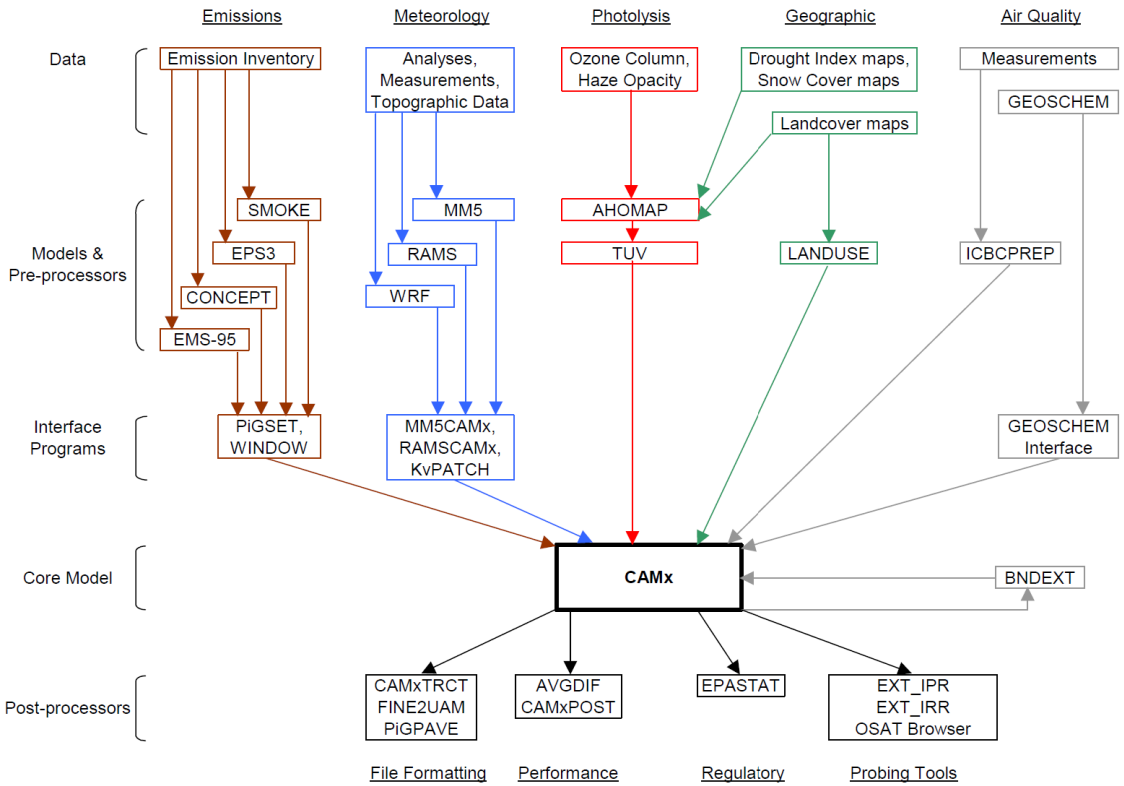


Fig. 2 Schematic Diagram of CAMx modeling system. Each of the five major data classes across the top of the figure require one or more input files. This diagram does not include any third-party models, pre- or post-processors.

Extensions)

5 categories with multiple files per category, multiple output files

Use Python package PseudoNetCDF to read files

1990, 2001, 2010 hourly average data for entire year (except 3 days in February and 1 day in December 1990)

References

Appel, K. W., Gilliam, R. C., Davis, N., Zubrow, A., and Howard, S. C. (2011). “Overview of the atmospheric model evaluation tool (AMET) v1.1 for evaluating meteorological and air quality models.” *Environmental Modelling & Software*, 26(4), 434–443.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.” *arXiv:1803.01271 [cs]*.

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., and Osornio-Vargas, A. (2017). “A systematic review of data mining and machine learning for air pollution epidemiology.” *BMC Public Health*, 17(1), 907.

Boylan, J. W., and Russell, A. G. (2006). “PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models.” *Atmospheric Environment*, 40(26), 4946–4959.

Chang, J. C., and Hanna, S. R. (2004). “Air quality model performance evaluation.” *Meteorology and Atmospheric Physics*, 87(1), 167–196.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. (2017). “Language Modeling with Gated Convolutional Networks.” *arXiv:1612.08083 [cs]*.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., and Wang, J. (2015). “Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation.” *Atmospheric Environment*, 107, 118–128.

Fiore, A. M., Jacob, D. J., Mathur, R., and Martin, R. V. (2003). “Application of empirical orthogonal functions to evaluate ozone simulations with regional and global models.” *Journal of Geophysical Research*, 108(D14).

Gilmore, E. A., Heo, J., Muller, N. Z., Tessum, C. W., Hill, J., Marshall, J., and Adams, P. J. (2019). “An inter-comparison of air quality social cost estimates from reduced-complexity models\ast \ast This article was modified on 23 April 2019. Formatting errors in the online version were corrected.” *Environmental Research Letters*.

Granger, C. W. J. (1969). “Investigating Causal Relations by Econometric Models and Cross-spectral Methods.” *Econometrica*, [Wiley, Econometric Society], 37(3), 424–438.

Heo, J., Adams, P. J., and Gao, H. O. (2016). “Reduced-form modeling of public health impacts of inorganic PM2.5 and precursor emissions.” *Atmospheric Environment*, 137, 80–89.

Hornik, K. (1991). “Approximation capabilities of multilayer feedforward networks.” *Neural Networks*, 4(2), 251–257.

References (continued)

Hornik, K., Stinchcombe, M., and White, H. (1989). “Multilayer Feedforward Networks are Universal Approximators.” *Neural Networks*, 2, 359–366.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2016). “Densely Connected Convolutional Networks.” *arXiv:1608.06993 [cs]*.

Johansen, S. (1991). “Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models.” *Econometrica*, [Wiley, Econometric Society], 59(6), 1551–1580.

Jones, S. H., and Zhang, K. (n.d.). “Vector Autoregression Neural Networks.”

Kalchbrenner, N., Oord, A. van den, Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., and Kavukcuoglu, K. (2016). “Video Pixel Networks.” *arXiv:1610.00527 [cs]*.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V. (2017). “Machine Learning for the Geosciences: Challenges and Opportunities.” *arXiv:1711.04708 [physics]*.

Karydis, V. A., Tsimpidi, A. P., and Pandis, S. N. (2007). “Evaluation of a three-dimensional chemical transport model (PMCAMx) in the eastern United States for all four seasons.” *Journal of Geophysical Research*, 112(D14).

Kelp, M., Jacob, D. J., Kutz, J. N., Marshall, J. D., and Tessum, C. W. (2020). “Toward stable, general machine-learned models of the atmospheric chemical system.” EarthArXiv.

Kelp, M. M., Tessum, C. W., and Marshall, J. D. (2019). “Orders-of-magnitude speedup in atmospheric chemistry modeling through neural.” *draft*, 23.

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., and Rybarczyk, Y. (2017). “Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters.” *Journal of Electrical and Computer Engineering*.

Krewski, D., Jerrett, M., Burnett, R. T., Ma, R., Hughes, E., Shi, Y., Turner, M. C., Pope, C. A., Thurston, G., Calle, E. E., Thun, M. J., Beckerman, B., DeLuca, P., Finkelstein, N., Ito, K., Moore, D. K., Newbold, K. B., Ramsay, T., Ross, Z., Shin, H., and Tempalski, B. (2009). “Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality.” *Research Report (Health Effects Institute)*, (140), 5–114; discussion 115-136.

References (continued)

Lee, L. A., Carslaw, K. S., Pringle, K. J., Mann, G. W., and Spracklen, D. V. (2011). "Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters." *Atmospheric Chemistry and Physics*, 11(23), 12253–12273.

Lepeule, J., Laden, F., Dockery, D., and Schwartz, J. (2012). "Chronic Exposure to Fine Particles and Mortality: An Extended Follow-up of the Harvard Six Cities Study from 1974 to 2009." *Environmental Health Perspectives*, 120(7), 965–970.

Muller, N. Z., Mendelsohn, R., and Nordhaus, W. (2011). "Environmental Accounting for Pollution in the United States Economy." *American Economic Review*, 101(5), 1649–1675.

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016a). "WaveNet: A Generative Model for Raw Audio." *arXiv:1609.03499 [cs]*.

Oord, A. van den, Kalchbrenner, N., and Kavukcuoglu, K. (2016b). "Pixel Recurrent Neural Networks." *arXiv:1601.06759 [cs]*.

Stieb, D. M., Judek, S., and Burnett, R. T. (2002). "Meta-Analysis of Time-Series Studies of Air Pollution and Mortality: Effects of Gases and Particles and the Influence of Cause of Death, Age, and Season." *Journal of the Air & Waste Management Association*, 52(4), 470–484.

Tenney, I., Das, D., and Pavlick, E. (2019). "BERT Rediscovers the Classical NLP Pipeline." *arXiv:1905.05950 [cs]*.

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Triebe, O., Laptev, N., and Rajagopal, R. (2019). "AR-Net: A simple Auto-Regressive Neural Network for time-series." *arXiv:1911.12436 [cs, stat]*.

US EPA, O. (2014). "What is Particle Pollution?" *US EPA*, Collections and Lists, <https://www.epa.gov/pmcourse/what-particle-pollution> (Apr. 13, 2020).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention Is All You Need." *arXiv:1706.03762 [cs]*.

References (continued)

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). "Grandmaster level in StarCraft II using multi-agent reinforcement learning." *Nature*, Nature Publishing Group, 575(7782), 350–354.

Xing, J., Pleim, J., Mathur, R., Pouliot, G., Hogrefe, C., Gan, C.-M., and Wei, C. (2013). "Historical gaseous and primary aerosol emissions in the United States from 1990 to 2010." *Atmospheric Chemistry and Physics*, 13(15), 7531–7549.

Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., and Zhang, Q. (2019). "Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations." *Environment International*, 123, 345–357.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. (2017). "Causal Discovery from Nonstationary/Heterogeneous Data: Skeleton Estimation and Orientation Determination." *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 1347–1353.