Modeling PM$_{2.5}$ Pollution from Wildfires in British Columbia using Machine Learning

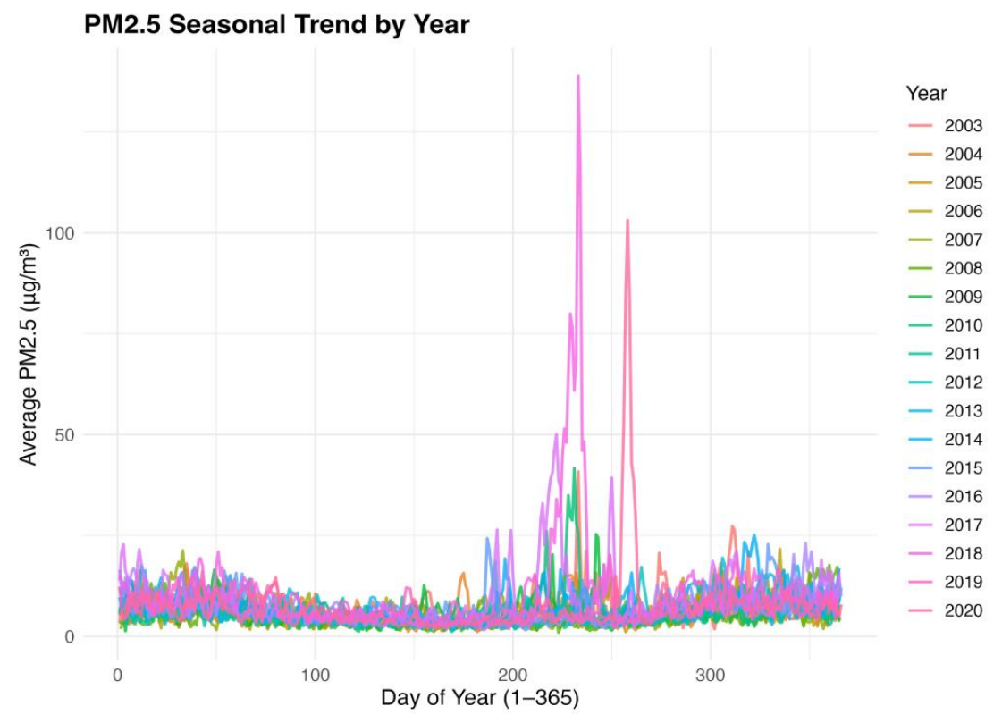
Statistical Sciences
UNIVERSITY OF TORONTO

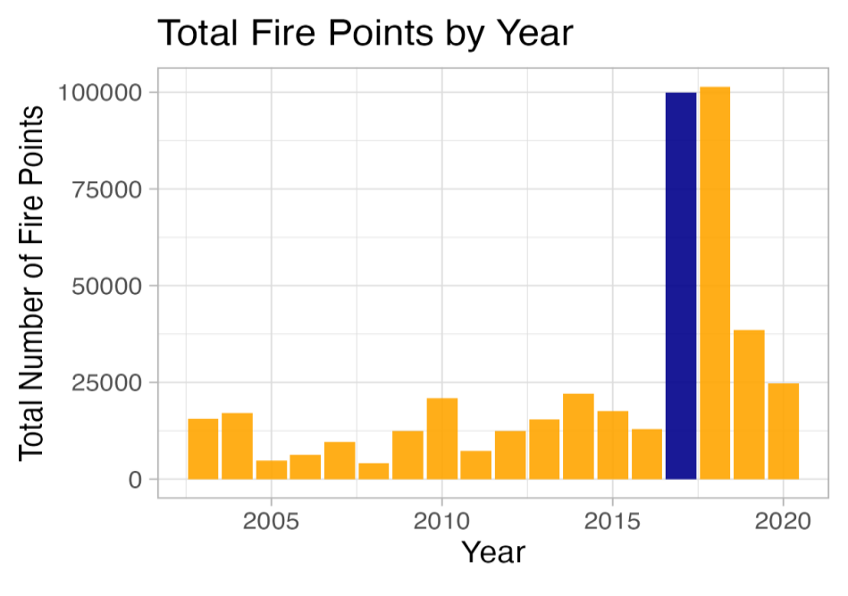
Hazel Buechner, Paula Mali, Madelyn Zhang, Morgan Huang
Supervisors: Meredith Franklin, Bernard Miskic

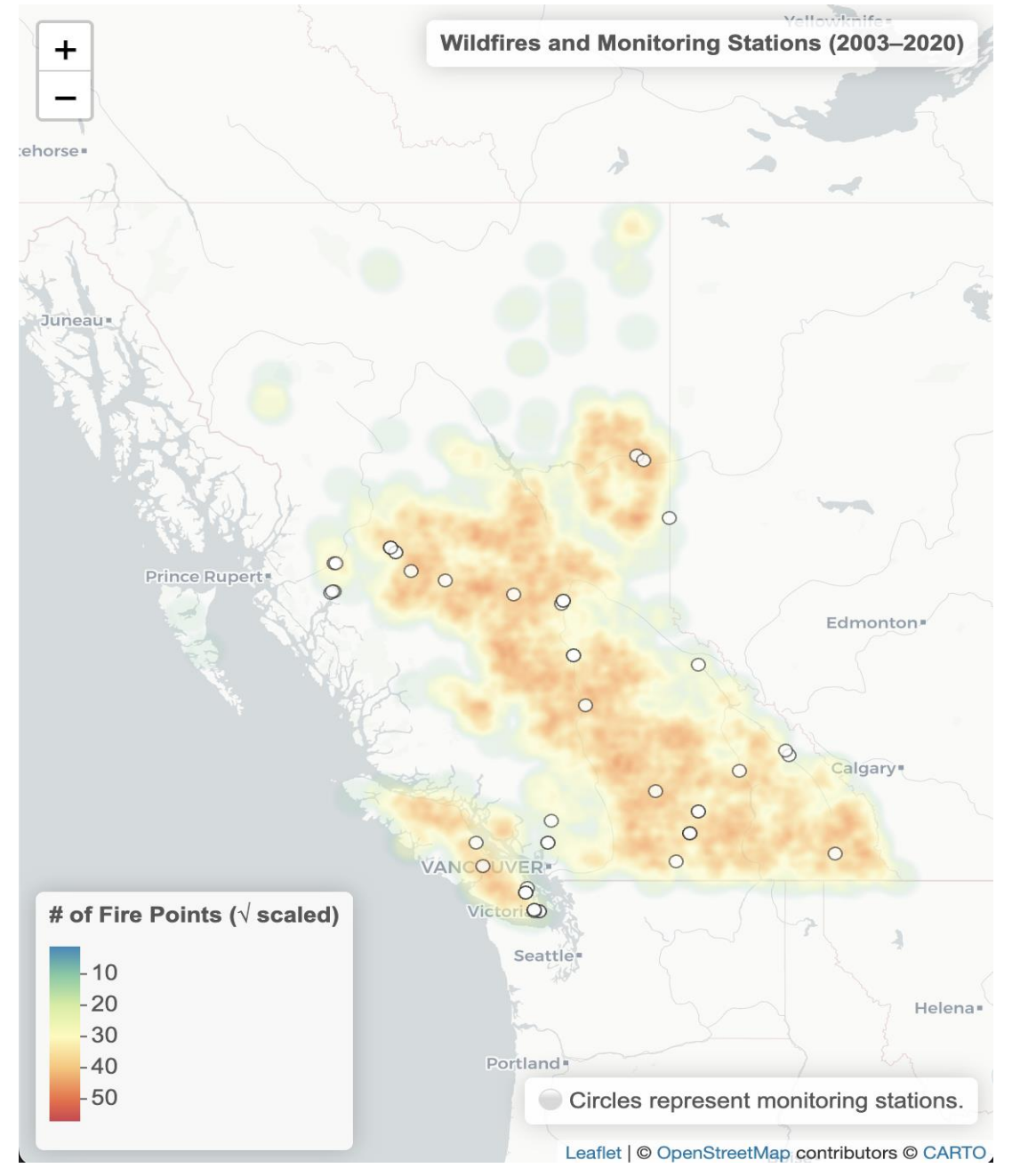
Objectives

Exposure to high concentrations of particle pollution, that is, fine particulate matter with a diameter less than 2.5µg/m^3 (PM$_{2.5}$), is associated with adverse health issues. In particular, the inhalation of PM$_{2.5}$ particles has been linked to respiratory effects. Our objective is to model the concentration of PM$_{2.5}$ as it relates to wildfires and wildfire smoke using statistical (GAM) and machine learning (XGBoost, Random Forest, GAMBoost, Neural Network) models. Our dataset includes meteorological and smoke plume factors spanning 2003-2020, collected from 40 monitoring stations in British Columbia.

We use 12 factors in our models, including meteorological variables such as daily measurements of temperature, wind speed, and wind direction, as well as spatial factors such as the distance to the nearest fire.






Wildfires and Monitoring Stations (2003–2020)
Juneau
Prince Rupert
Edmonton
Calgary
VANCOUVER
Victoria
Seattle
Portland
Helena
# of Fire Points (√ scaled)
10
20
30
40
50
Circles represent monitoring stations.
Leaflet | © OpenStreetMap contributors © CARTO

Map of British Columbia depicting the location and intensity of wildfires in 2017. Created using Leaflet widget (left). Graph demonstrating the seasonal PM2.5 trends by year (right.) Bar graph showing the total number of fire points per year, highlighting 2017 (above).

Methods

Data and Processing

Meteorological data were reported by the stations, while satellite imagery contributed smoke plume and fire observations. Extreme values of variables were removed or modified with linear interpolation. The PM$_{2.5}$ concentration was normalized with a log transformation. Smoke presence was classified numerically (0, 1, 2, or 3) based on the maximum smoke plume severity observed by each station on a given day. Temporal units were derived.

Methods and Modeling

Five models were used to predict daily PM$_{2.5}$ concentrations: Generalized Additive Model (GAM), GAMBoost, XGBoost, Random Forest, and GCN+LSTM. All classical models were trained and evaluated using a fixed 60/20/20 train - validation - test split,. The GCN+LSTM model, due to its sequential structure, was trained using a separate split with 15 complete stations. Model performance was evaluated on the test set using two metrics: R^2 and RMSE.

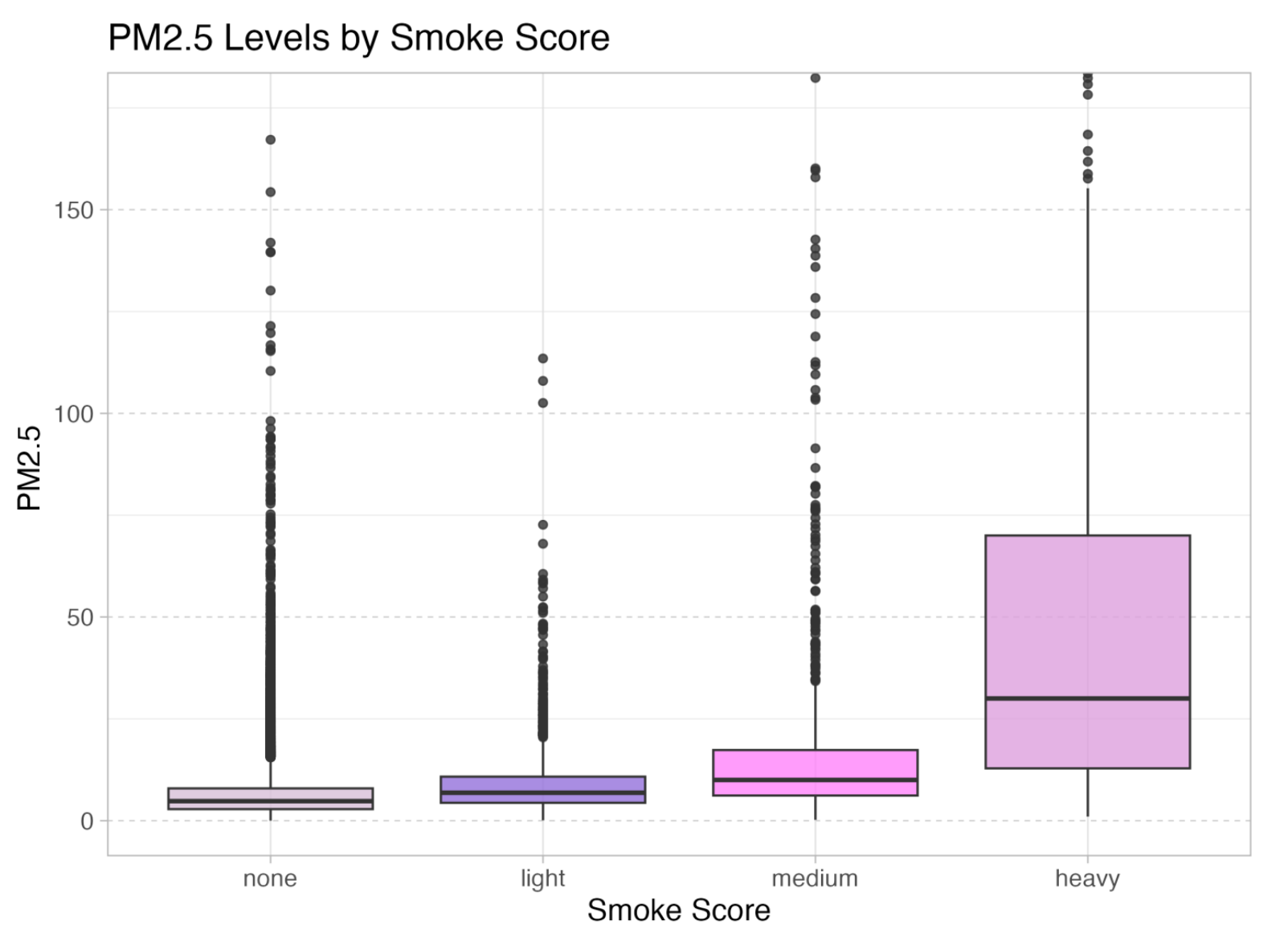
Generalized Additive Model (GAM): Continuous variables were smoothed to capture nonlinear temporal, spatial, and meteorological effects. Model specifications were tested by varying spline types and complexity to balance flexibility and prevent overfitting.

GAMBoost: We implemented a GAMBoost model using component-wise gradient boosting on both smooth and linear base-learners. The degrees of freedom and the number of boosting iterations were jointly tuned, and the final model was fitted using 2000 boosting iterations.

XGBoost: The model uses decision trees on a gradient-boosting algorithm. We tuned hyperparameters with 10-fold cross-validation. The learning rate (era), pre-split minimum loss reduction (gamma), complexity (max_depth and min_child_weight) and subsampling parameters were fitted with 1000 boosting iterations.

Random Forest: A random forest creates decision trees using different factors each time and combines their outputs to predict values, in this case PM$_{2.5}$ concentrations. We tuned the number of decision trees in the RF as well as the number of factors randomly chosen at each split.

GCN+LSTM Neural Network: We combined graph convolutional networks (GCN) and Long-Short-Term-Memory networks (LSTM) in our overall architecture. We used a rolling window technique to train on all possible timestamps, thus augmenting our data matrix. To account for missing data, we use a masking matrix of 1s and 0s for our loss function calculation.



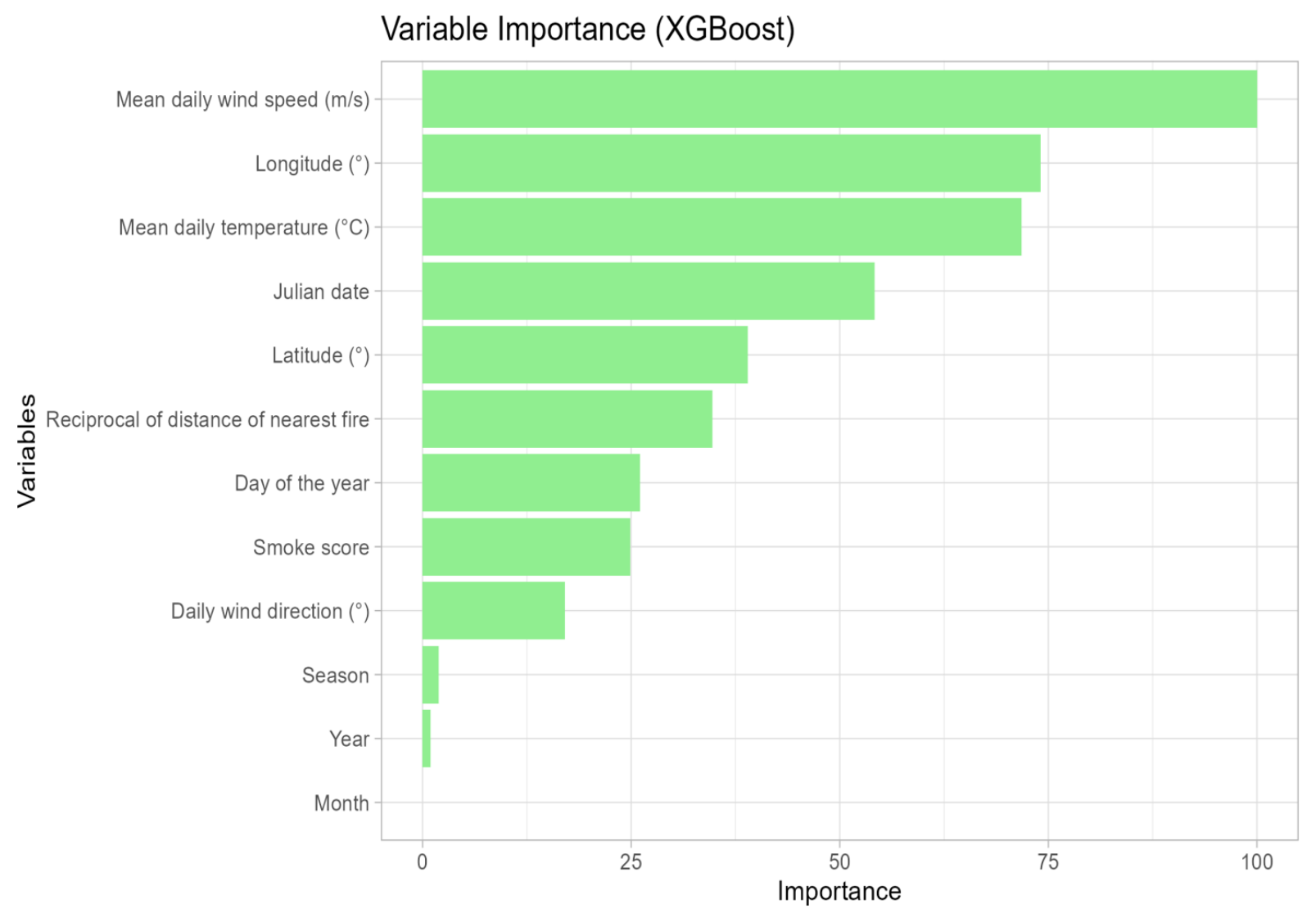
Boxplots showing the range of PM2.5 concentrations measured on days with smoke plumes recorded.

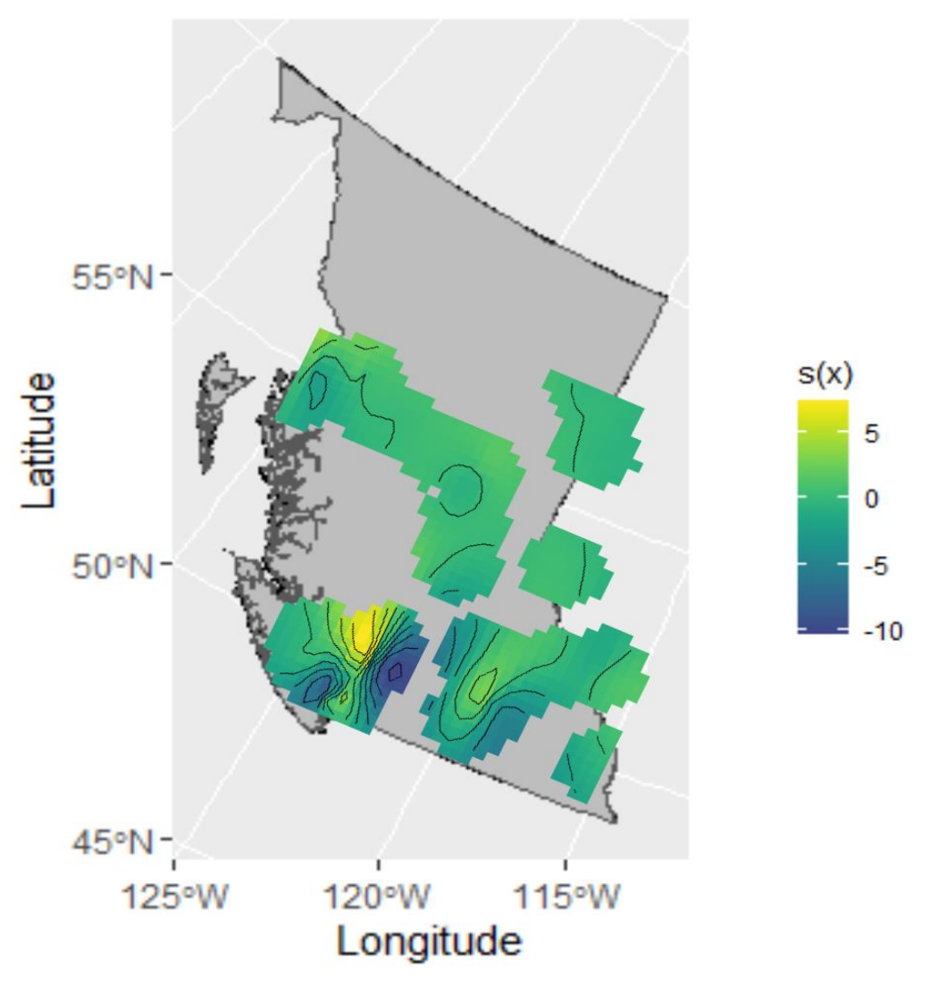
Results

The Generalized Additive Model (GAM) achieved an R^2 of 0.36 and an RMSE of 7.05 on the testing set. The GAMBoost model performed slightly worse, with an R^2 of 0.347 and an RMSE of 7.18. The Random Forest model achieved an R^2 of 0.60 and an RMSE of 5.86 on the testing set. XGBoost achieved an R^2 of 0.646, with an RMSE of 5.678. Our Neural Network achieved an R^2 of 0.54 with an RMSE of 7.758 on the testing dataset. Note that the testing dataset here is different from the other test dataset, so the results are not as comparable with the other models used in this paper.

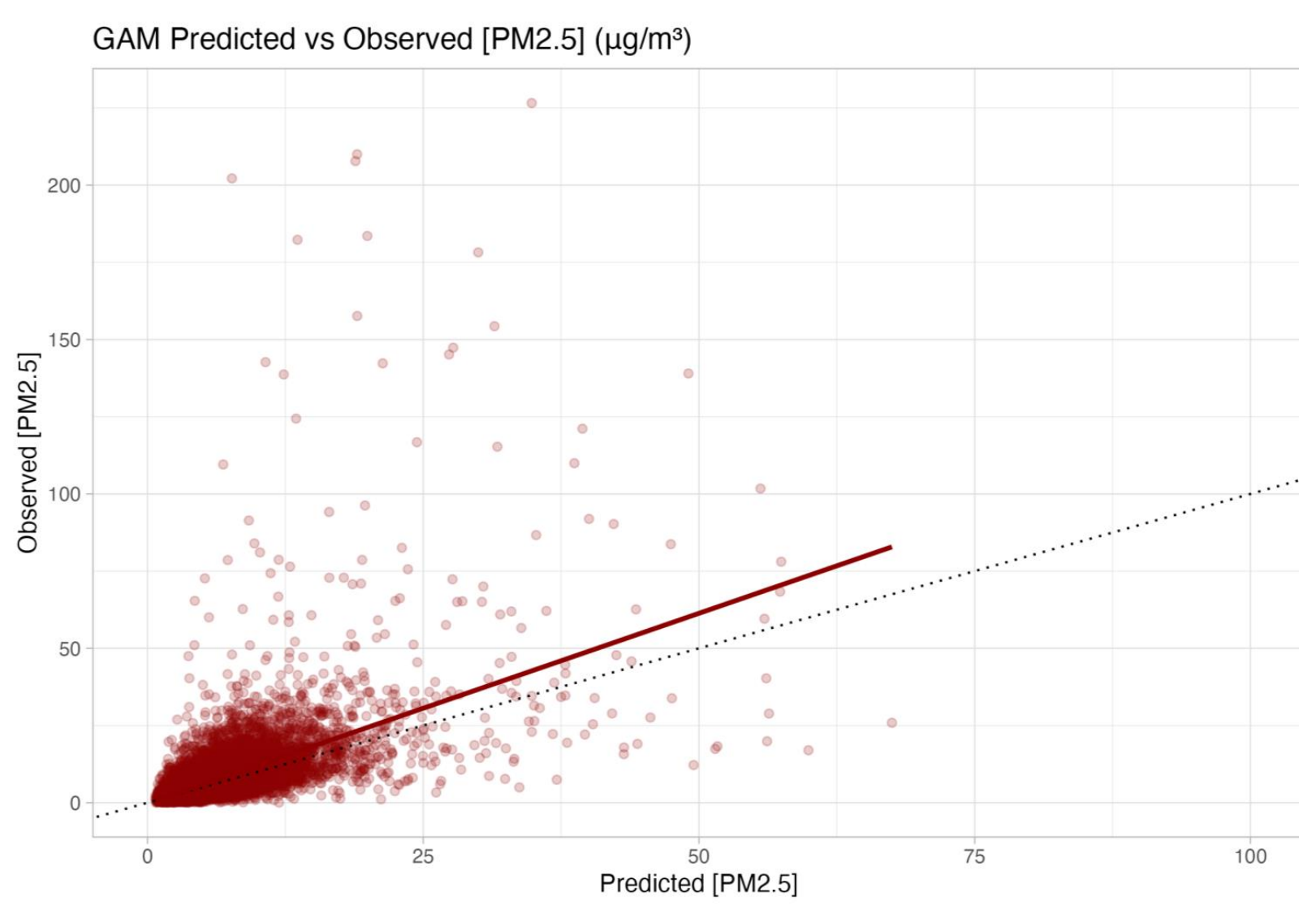
Model Diagnostics

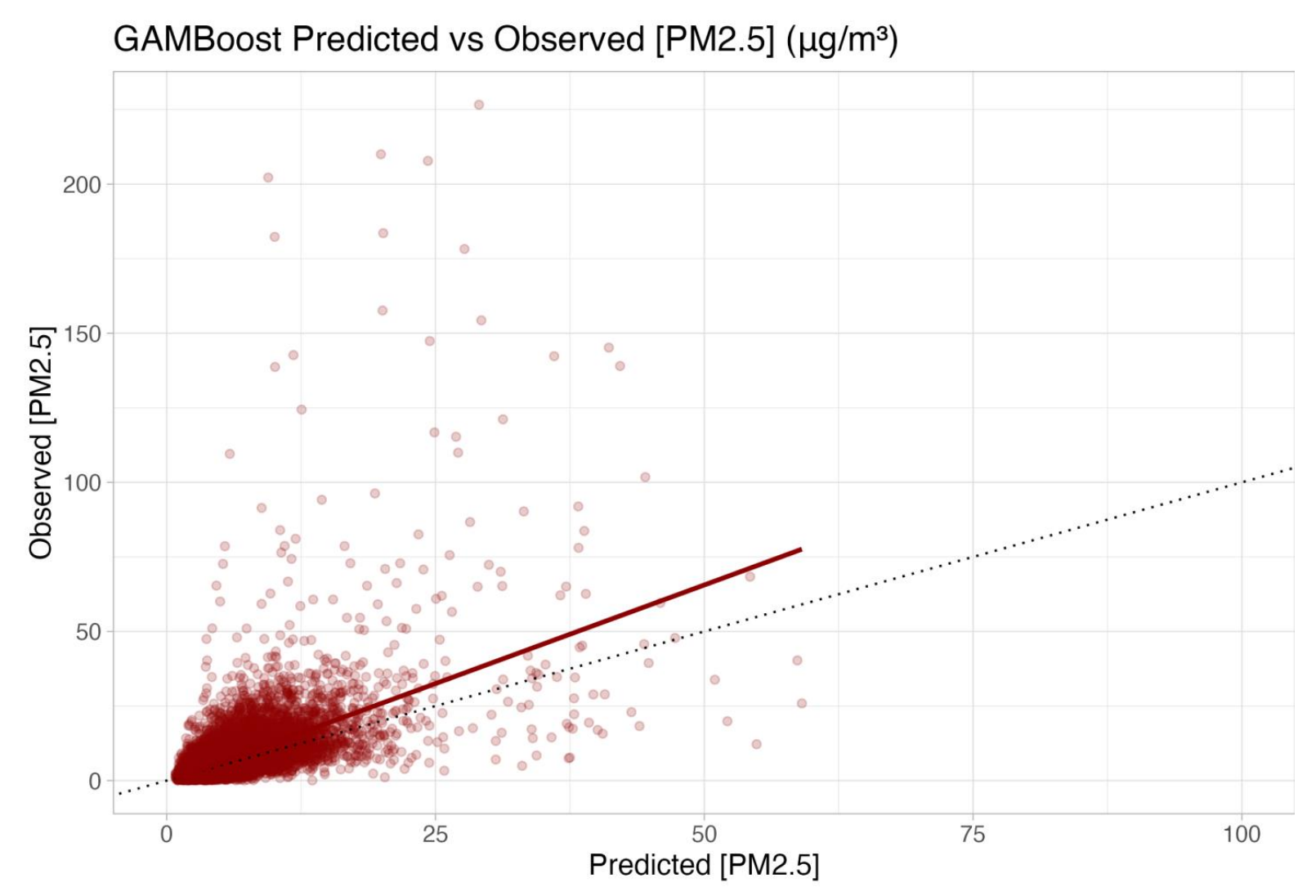
From the Predicted vs. Observed plots below, we can see that the models tend to underestimate predictions of PM$_{2.5}$ due to the slope of the regression line being bigger than 1 for each model. Out of all 5 models, however, XGBoost performs the best, with the highest R^2 and lowest RMSE. This is also evident qualitatively, as the XGBoost model's fitted line has a slope closest to 1, which indicates ideal agreement between predicted and observed values.

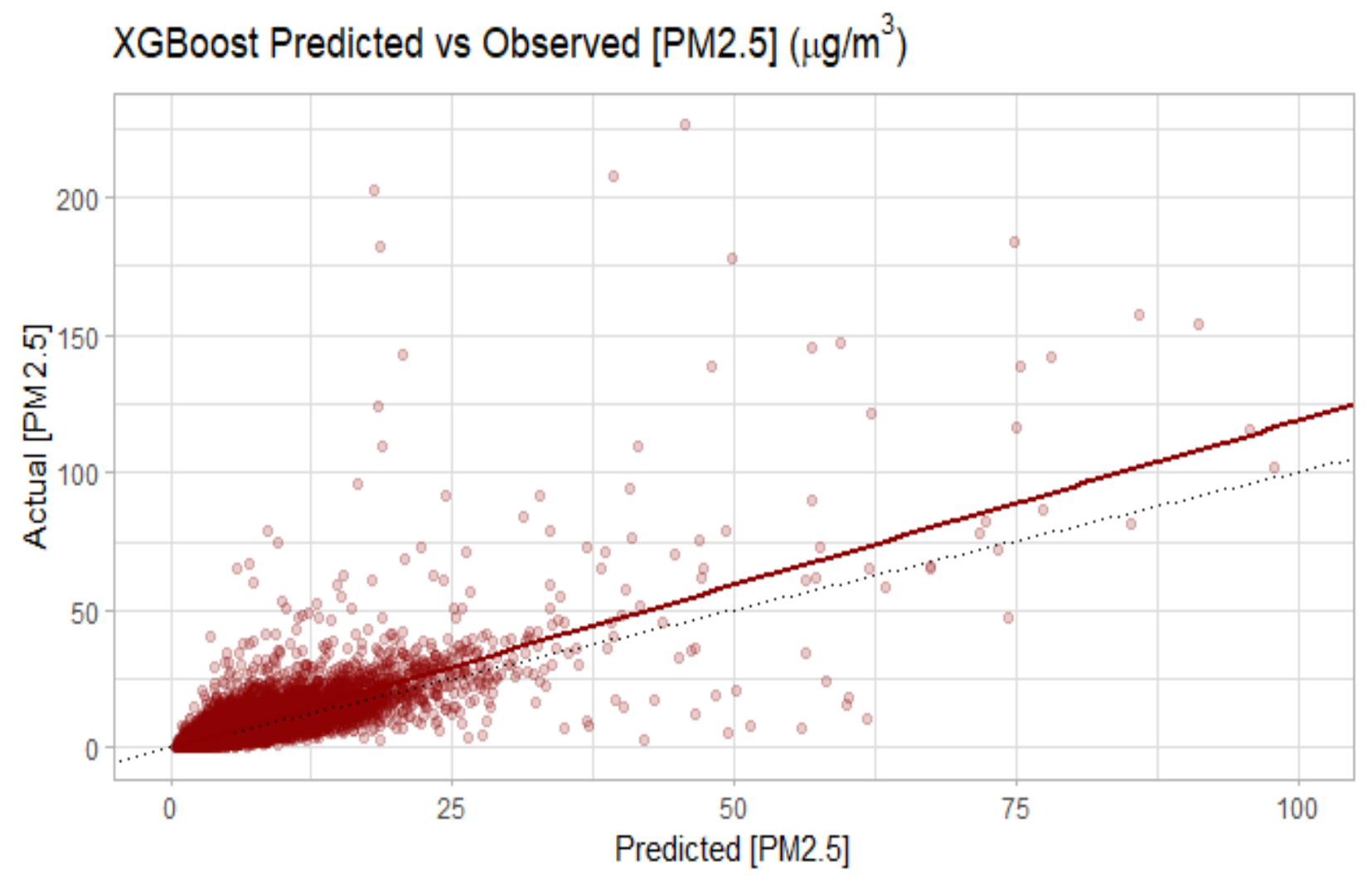


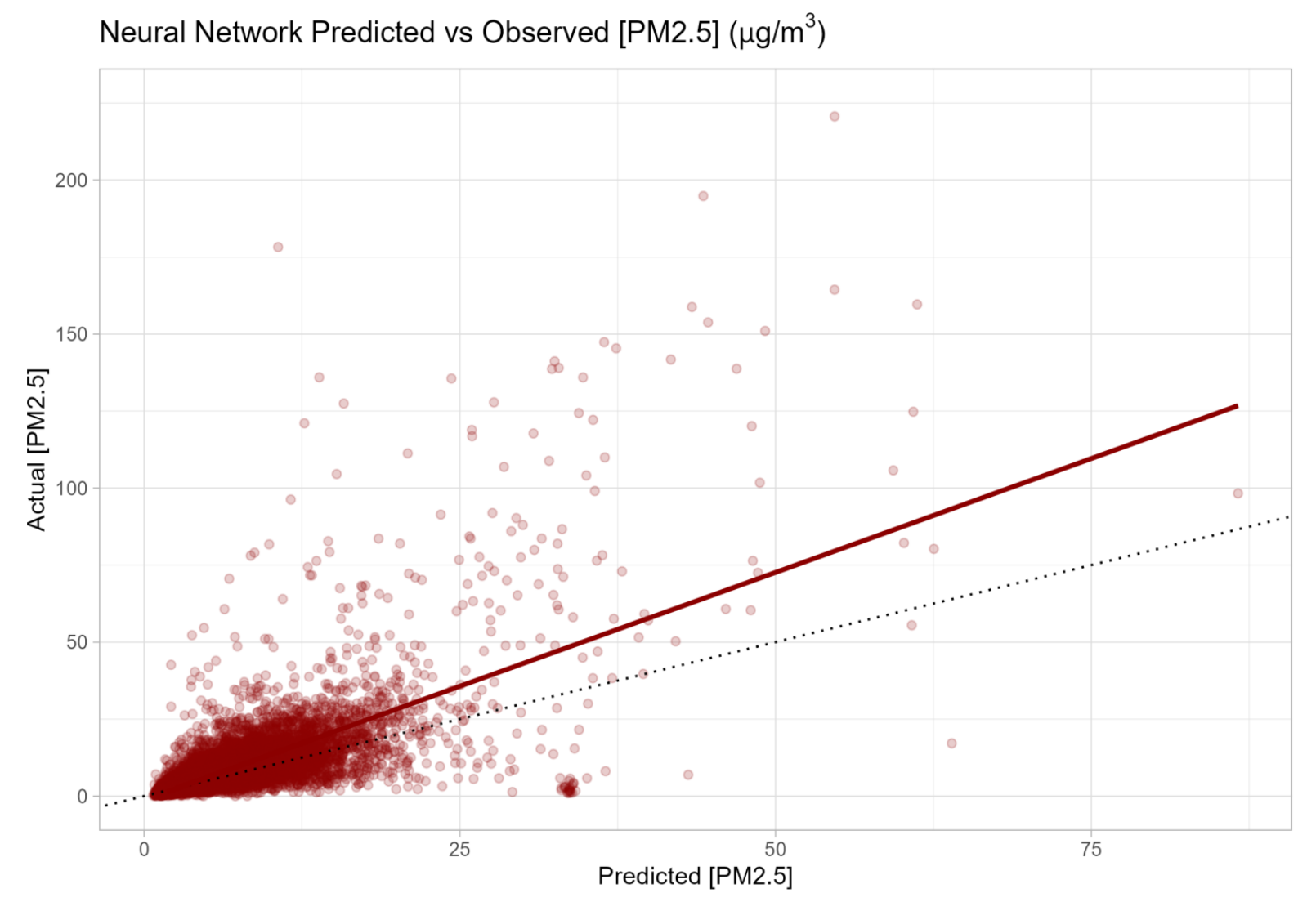


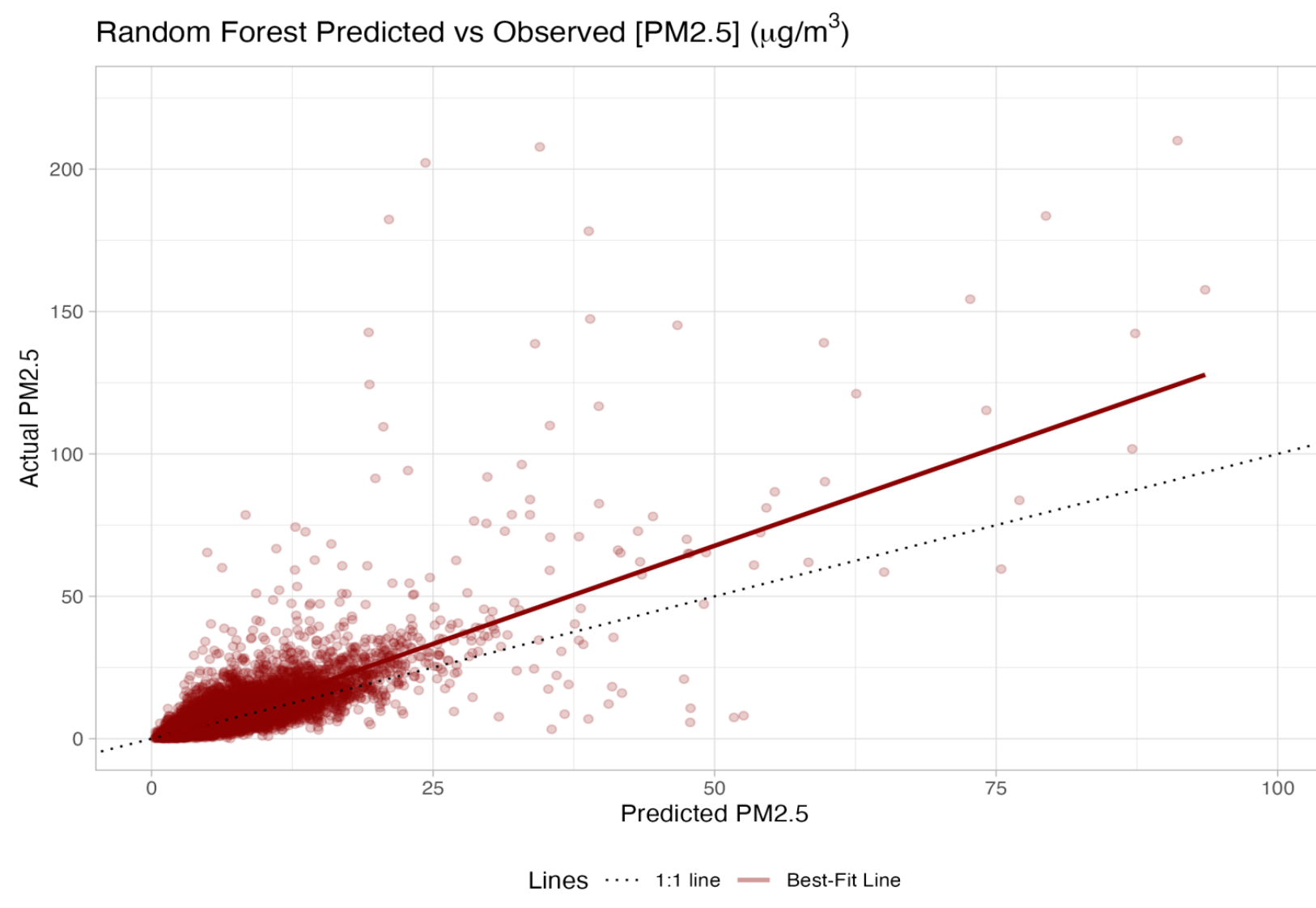
Graph showing the variable importance in the GAMBoost Model (above to the left.) PM2.5 heatmap of British Columbia (above right.) Scatterplots comparing predicted vs. observed values in the test set for each of our five statistical and machine learning models. Accompanying them is a chart displaying relevant test statistics (below).











Model	R^2	RMSE	Hyperparameters
GAM	0.36	7.05	smooths
GAMBoost	0.347	7.177	Mstop=2000, Degrees of freedom(df)
RF	0.60	5.86	mtry=3, numtrees= 1000

XGBoost	0.646	5.678	nrounds = 1000, eta = 0.015, max_depth = 7
			gamma = 0, colsample_bytree = 1
			min_child_weight = 7, subsample = 0.75

NN	0.54	7.758	# of Neurons = 64, # of LSTM layers = 2 # of GCN Layers = 1, Learning Rate = 0.001 Batch Size = 8, Epochs = 5000 Optimizer = Adam

Discussion

Extreme Outliers of PM$_{2.5}$

Visually, we can see that our machine learning models severely underestimate most of the outliers in the data. This is likely caused by the model being trained on logarithmic PM$_{2.5}$ data, which severely compresses extreme values. Thus, when training on the logarithmic data, small errors when predicting extreme values gets sharply inflated after re-exponentiating, rendering big differences between large outliers of PM$_{2.5}$ and its predicted counterparts. Extreme values of PM$_{2.5}$ are also difficult to predict, especially using only machine learning models, thus such models must settle for underestimation to not overfit the data. In the future we can try to implement the same machine learning models on the original PM$_{2.5}$ data and explore how these models handle log-normally distributed data.

Deep Learning Methods

Due to computational and time limitations, hyperparameter tuning was not performed on the neural network, and arbitrary values were assigned. Neural networks are known to be sensitive to hyperparameter tuning, so the performance of the GCN+LSTM in this study was greatly affected. Since we used a rolling window technique to our data which greatly augmented our data matrix, we were limited to only a batch size of 8 timestamps. Thus, the LSTM layers' potential to learn long-term seasonal trends was limited due to a small number of previous observations the model takes in as input in one forward pass.

Graph Convolutional Networks use graph networks between neighboring stations to model relationships accounting for nearby feature and PM$_{2.5}$ values. We used K-Nearest Neighbors with k=2 to define the edges between station. In the future, we can modify this graph structure by tuning the level of k and treating it as a hyperparameter. In addition, we can explore other graph structures; for example, our edges can be defined using a distance threshold, or we can use spatial autocorrelation (Moran's I) and define a weighted graph where the weights of the edges are determined by high autocorrelation of neighboring stations.

Prediction vs. Inference

We have so far focused on predicting PM$_{2.5}$ using station observations along with a couple of meteorological variables. However, prediction does not consider the uncertainty of future PM$_{2.5}$ values. With growing concern of wildfires occurring in recent years as well as the many uncertainties that go behind extreme PM$_{2.5}$ values, inference methods are becoming increasingly crucial. Therefore, to account for problems behind prediction methods, we can combine quantile regression methods to model confidence and prediction intervals with machine learning models. In addition, extreme value theory to predict extreme events using return periods or using EVT-based loss functions in machine learning algorithms can be useful in improving the prediction and inference of future PM$_{2.5}$.

References

Adetona, O., Reinhardt, T. E., Domitrovich, J., Broyles, G., Adetona, A. M., Kleinman, M. T., Ottmar, R. D., & Naeher, L. P. (2016). Review of the health effects of wildland fire smoke on wildland firefighters and the public. Inhalation Toxicology, 28(3), 95–139. https://doi.org/10.3109/08958378.2016.1145771

Childs, M. L., Li, J., Wen, J., Heft-Neal, S., Driscoll, A., Wang, S., Gould, C. F., Qiu, M., Burney, J., & Burke, M. (2022). Daily local-level estimates of ambient wildfire smoke PM2.5 for the contiguous US. Environmental Science & Technology, 56(19), 13607–13621. https://doi.org/10.1021/acs.est.2c02934

Liao, L., Li, H., Shang, W., & Ma, L. (2022). An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of Deep Neural Networks. ACM Transactions on Software Engineering and Methodology, 31(3), 1–40. https://doi.org/10.1145/3506695

Panumasvivat, J., Sapbamrer, R., Sittitoon, N., Khacha-ananda, S., Kiratipaisarl, W., Sirikul, W., Insian, W., & Assavanopakun, P. (2024). Exploring the adverse effect of Fine Particulate matter (PM2.5) on Wildland Firefighters' pulmonary function and DNA damage. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-58721-4

Yang, H.-C., Yang, M.-C., Wong, G.-W., & Chen, M. C. (2023). Extreme event discovery with self-attention for PM2.5 anomaly prediction. IEEE Intelligent Systems, 38(2), 36–45. https://doi.org/10.1109/mis.2023.3236561

Yang, S., & Wu, H. (2022). A novel PM2.5 concentrations probability density prediction model combines the least absolute shrinkage and selection operator with quantile regression. Environmental Science and Pollution Research, 29(52), 78265–78291. https://doi.org/10.1007/s11356-022-21318-3

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 prediction based on Random Forest, XGBoost, and deep learning using Multisource Remote Sensing Data. Atmosphere, 10(7), 373. https://doi.org/10.3390/atmos10070373