A Machine Learning Approach for Ozone Forecasting and its Application for Kennewick, WA

Kai Fan[1], Brian Lamb[1], Ranil Dhammapala[2], Ryan Lamastro[3], and Yunha Lee[1]

[1]Laboratory for Atmospheric Research, Civil and Environmental Engineering, Washington State University

[2]Washington State Department of Ecology

[3]State University of New York at New Paltz

Motivation

- Kennewick, WA lies 32 km (20 mi) north of Washington's southern border, where high O_3 events occur during summer and fall.

- AIRPACT is a state-of-the-science CMAQ-based air quality forecasting system for Pacific Northwest. However, AIRPACT struggles to predict high O_3 concentrations in this area.

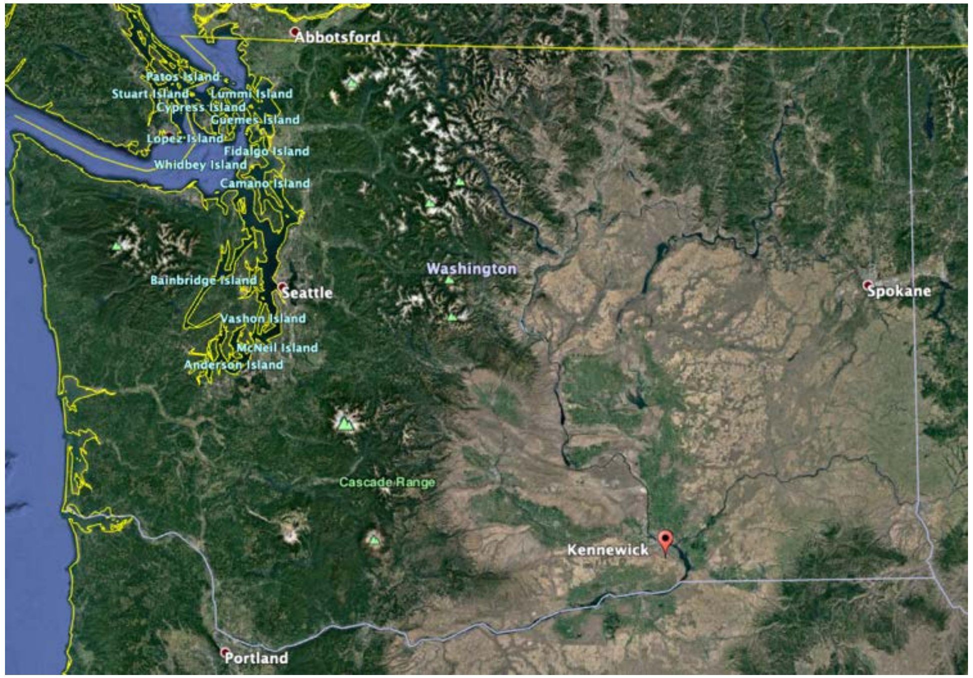
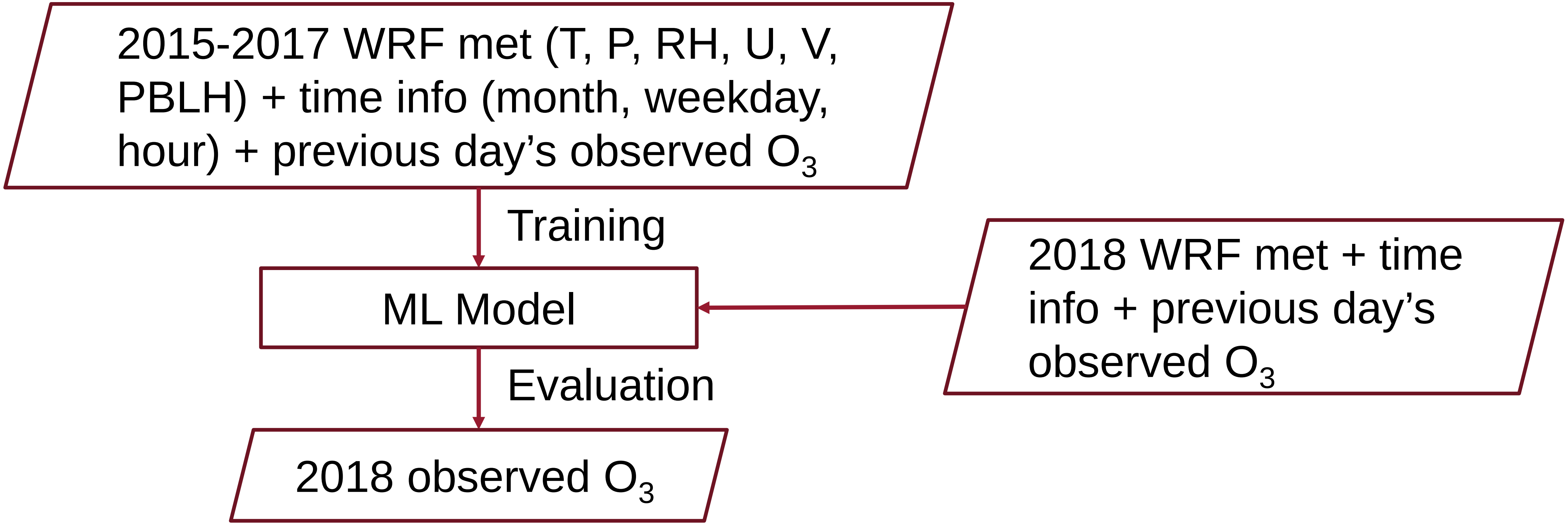


*Image from Google Earth

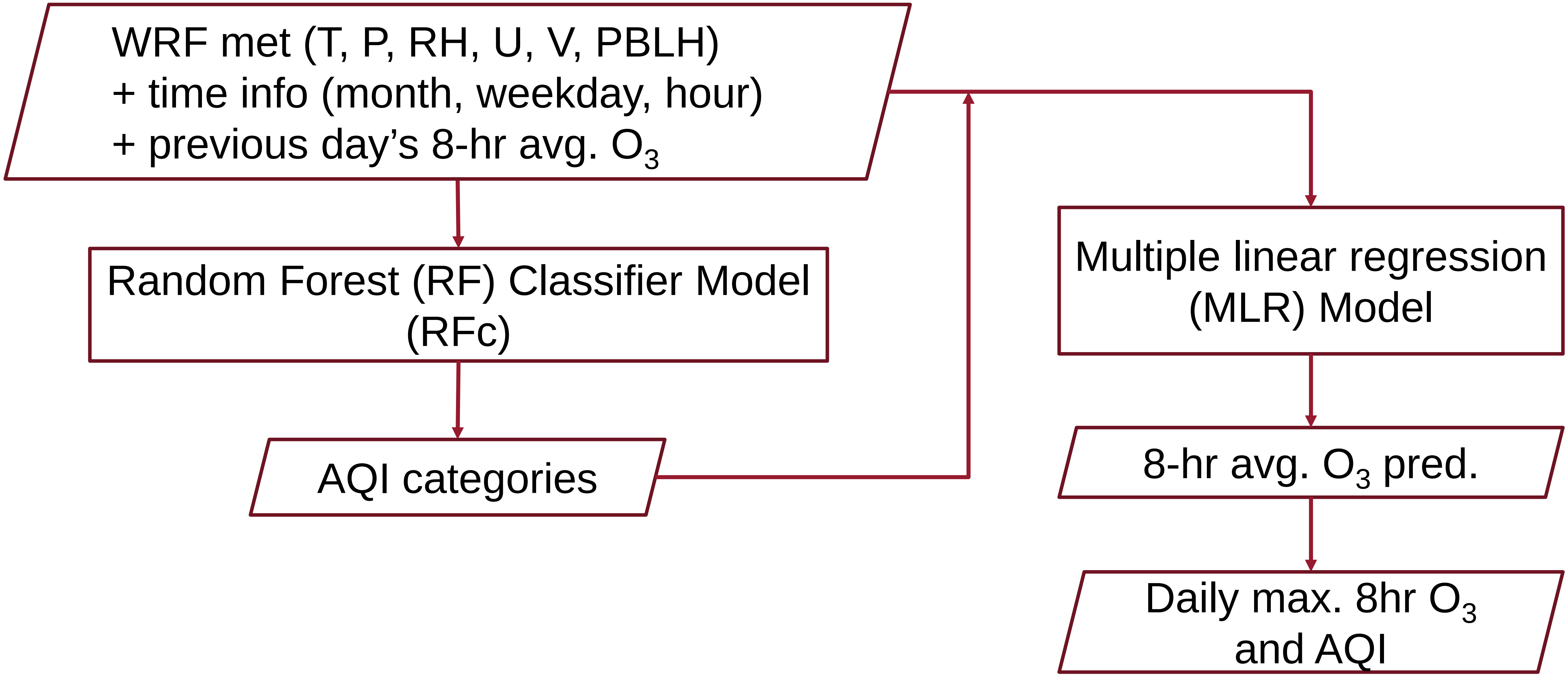
- The goal of our study is to provide a reliable forecast for high O_3 events using the machine learning (ML) models, which can learn from the historical data to make future forecasts.

Machine Learning (ML) Model Approach for the Kennewick Monitoring Site



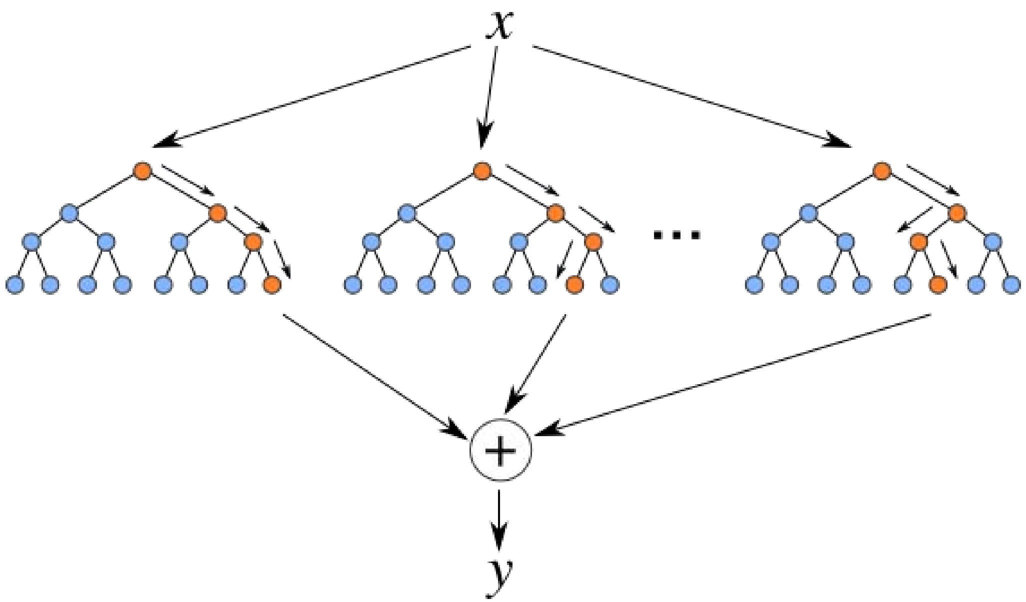
Machine Learning Model Framework 1: ML1

Combining Random Forest and Multiple Linear Regression methods



Random Forest (RF) classifier

- RF classifier is the consensus of many decision trees, which we use to predict the AQI categories.



* Image from https://blog.toadworld.com

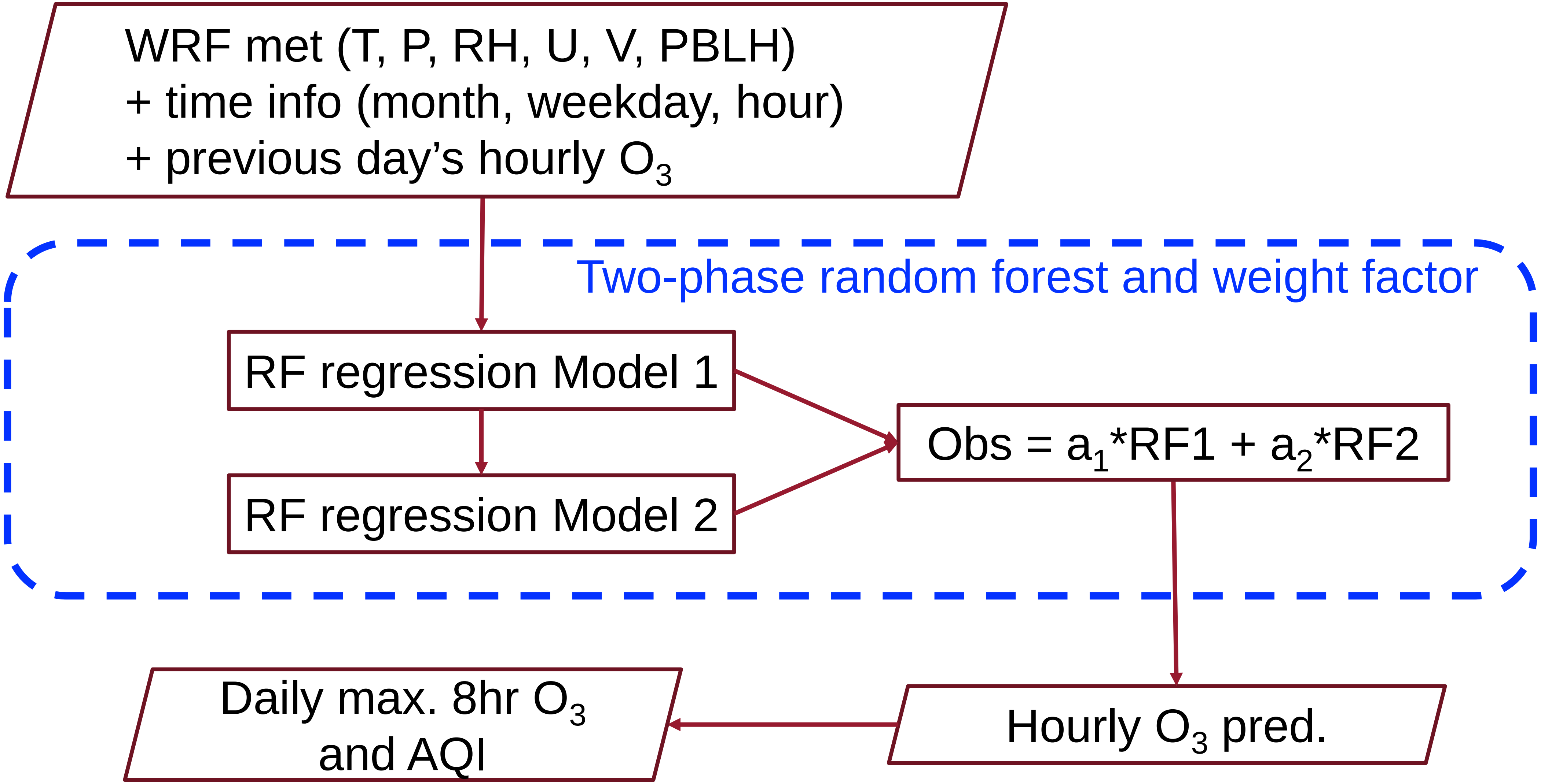
Multiple linear regression (MLR)

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \ldots$$

- MLR approach is used to predict the 8-h average O_3, which shows good performance to predict high O_3 days.

Machine Learning Model Framework 2: ML2

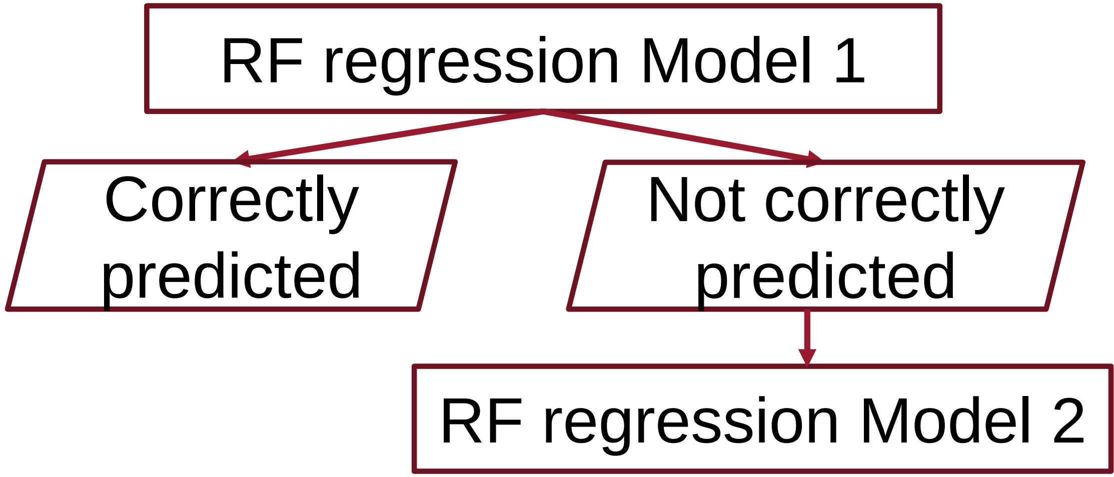
Two RF models weighted for optimal results



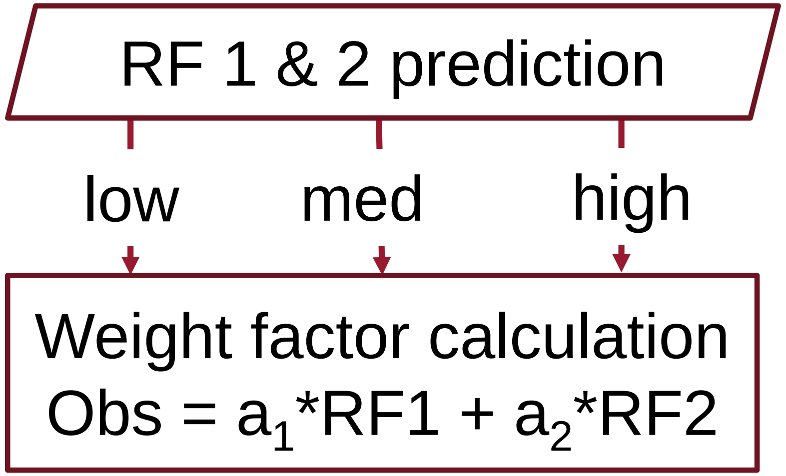
* Jiang, N., & Riley, M. L. (2015). Exploring the utility of the random forest method for forecasting ozone pollution in SYDNEY. *Journal of Environment Protection and Sustainable Development*, *1*(5), 245-254.

Two-phase random forest (RF)

- The first RF model can usually make right prediction for low O_3 events, and the second phase isolates the events incorrectly predicted to form a second training dataset.



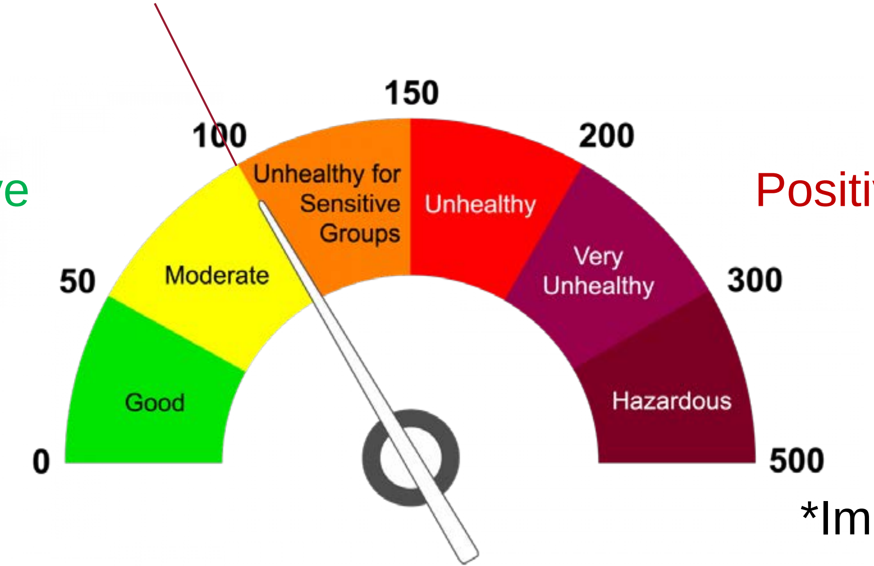
- We separate the initial predicted mixing ratios to three categories and give three sets of weight to two phases. The weight of two models are based on a simple linear regression equation.



Forecast evaluation metrics

Metric	Description
Hits	True positive/negative
False Alarms	False positive
Misses	False negative
FAR (False Alarm Ratio)	$\frac{\text{\# of false alarms}}{\text{total \# of events forecast}}$
POD (Probability of Detection)	$\frac{\text{\# of hits}}{\text{total \# of events forecast}}$

Negative



Positive

*Image from www.deq.ok.gov

Historical data summary

Year	Simulated days	# of days for each AQI			AQI > 2
		1	2	3	
2015	106	75	27	4	4%
2016	143	125	16	2	1%
2017	114	71	35	8	7%
2018	152	120	26	6	4%
Total	515	391	104	20	4%

More fires

ML1 Evaluation

Leave one out cross validation

Metric	2015	2016	2017	2018
Hits	94 (100)	127 (130)	99 (92)	138 (140)
False Alarms	8 (1)	4 (0)	4 (6)	5 (0)
Misses	1 (2)	1 (2)	2 (7)	2 (5)
FAR	8% (1%)	3% (0%)	4% (5%)	3% (0%)
POD	91% (97%)	96% (98%)	94% (88%)	95% (97%)

The numbers in parenthesis are the AIRPACT forecast performance.

- ML1 predicts more false alarms but fewer misses.
- For high O_3 year 2017, ML1 performs better than AIRPACT.

ML2 Evaluation

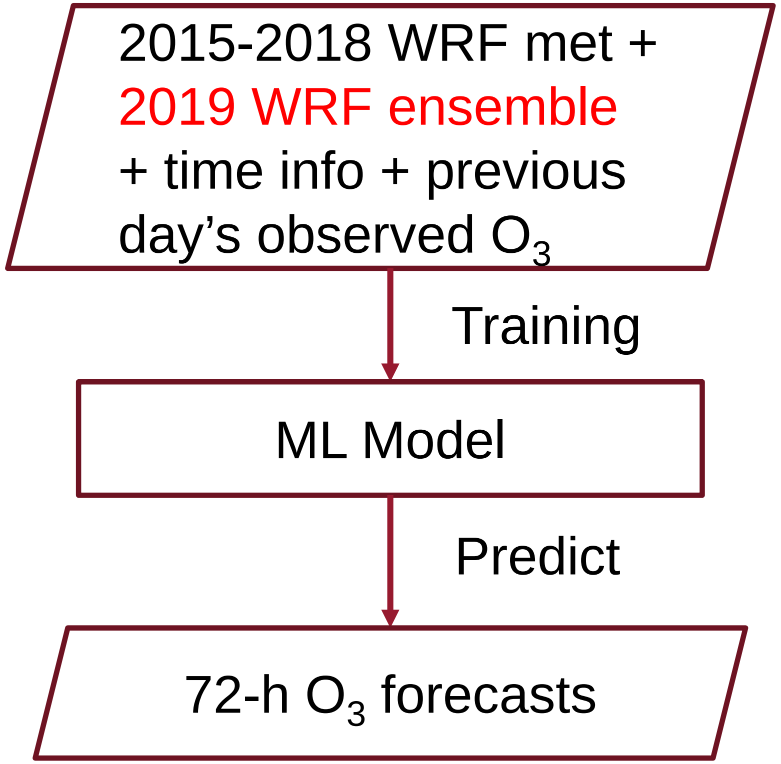
Leave one out cross validation

Metric	2015	2016	2017	2018
Hits	99 (100)	130 (130)	97 (91)	140 (141)
False Alarms	1 (1)	0 (0)	0 (6)	1 (0)
Misses	3 (2)	2 (2)	7 (7)	5 (5)
FAR	1% (1%)	0 (0)	0 (6%)	1% (0)
POD	96% (97%)	98% (98%)	93% (88%)	96% (97%)

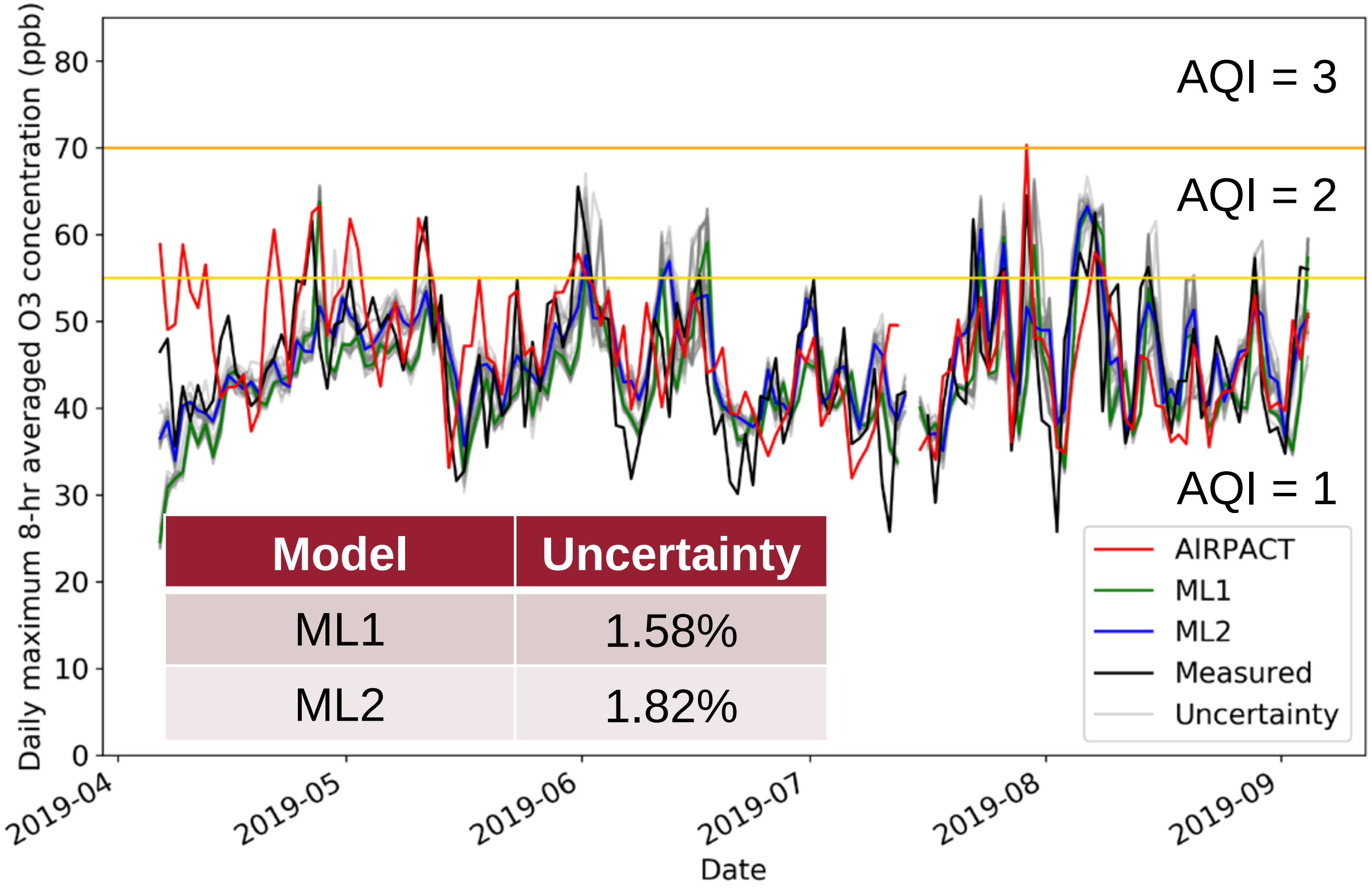
The numbers in parenthesis are the AIRPACT forecast performance.

- ML2 predicts much fewer false alarms but similar miss number as AIRPACT.
- Both AIRPACT and ML2 fail to predict the high ozone days in 2017.

Tri-Cities Ozone "Ensemble" Forecast in 2019



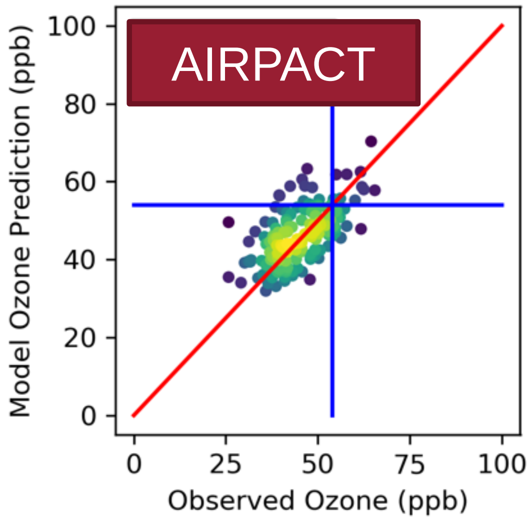
Time series of daily max. O_3

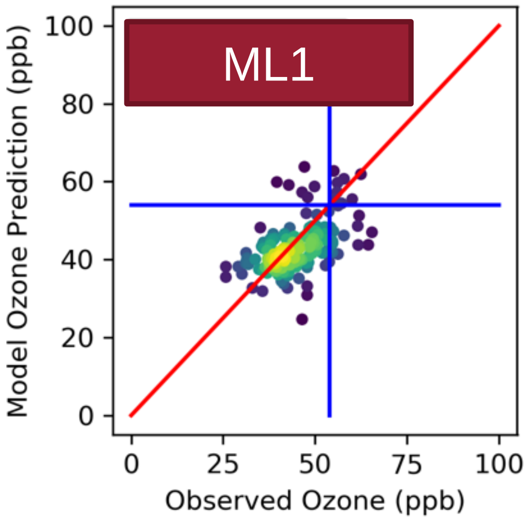


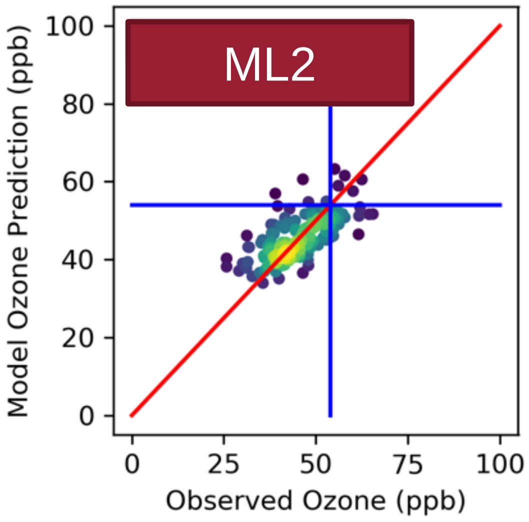
Model	Uncertainty
ML1	1.58%
ML2	1.82%

To get more data to train the model, we retrain our model everyday including previous day's measurements.

Tri-Cities Ozone (ensemble mean) Forecast in 2019







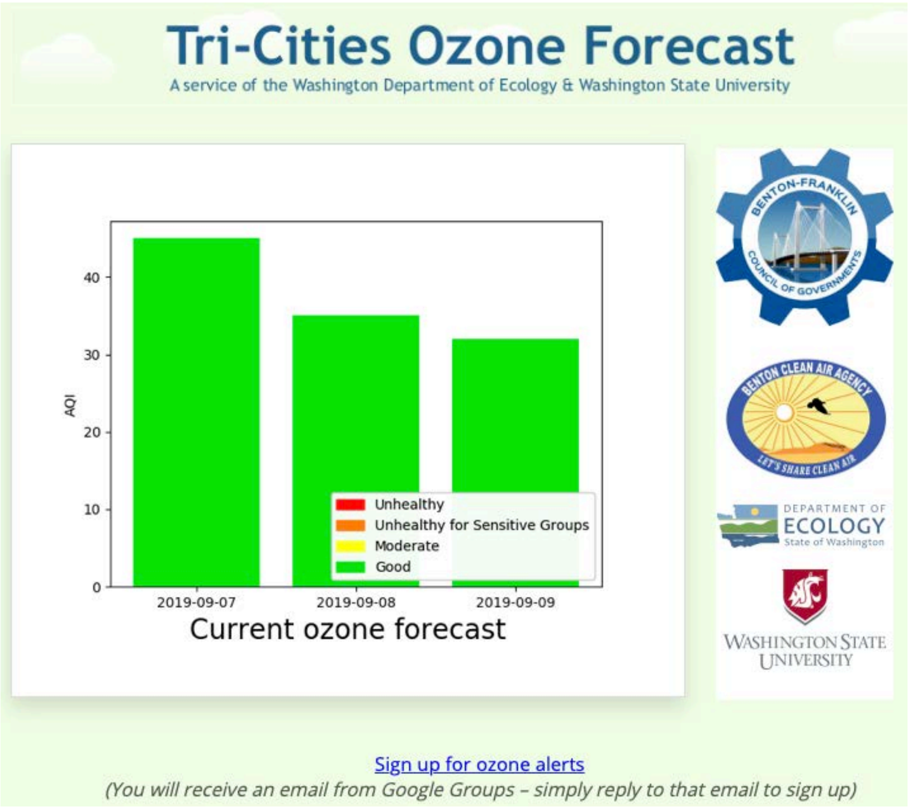
		Observation	
		AQI 1	AQI 2
AIRPACT	AQI 1	123	8
	AQI 2	10	9

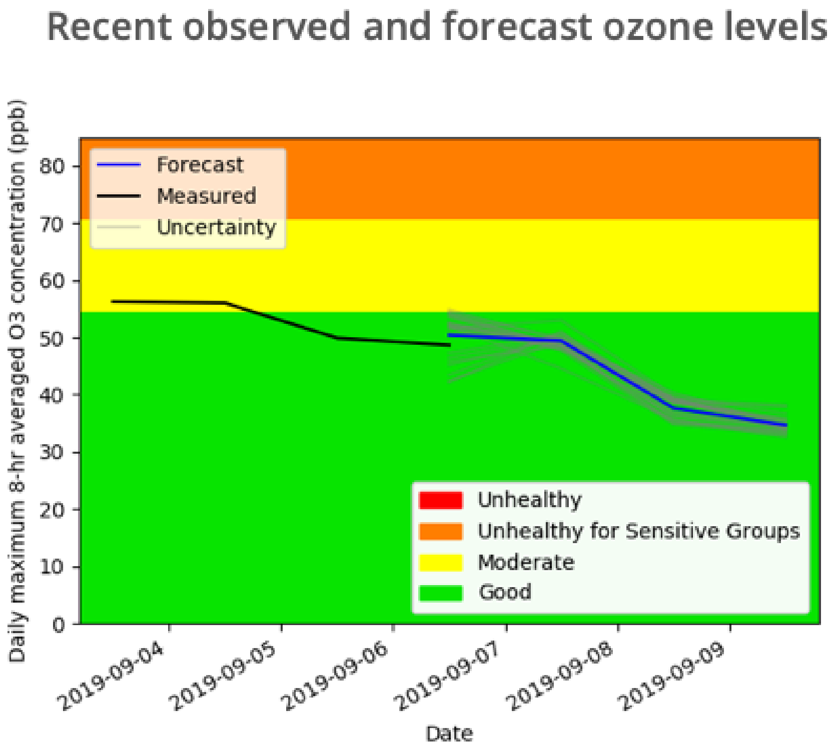
		Observation	
		AQI 1	AQI 2
ML1	AQI 1	127	10
	AQI 2	6	7

		Observation	
		AQI 1	AQI 2
ML2	AQI 1	129	12
	AQI 2	2	5

ML2 performs the best to reduce false AQI2 days (in red cells). Thus we chose ML2 to run our operational daily ozone forecasting for Kennewick.

Our Machine Learning O_3 forecasts go public everyday!





http://ozonematters.com/

Summary

- The ML1 model raised more false alarms than AIRPACT, but performed better in the high ozone year.
- Both ML2 and AIRPACT missed some high ozone events, but ML2 raised fewer false alarms than AIRPACT.
- Our training dataset contains only a few high O_3 days, which makes it difficult to predict a high O_3 day using a ML approach. To overcome that issue, we updated the training dataset each day.
- We plan to apply our ML models to other cities that has a well-distributed AQI (Air Quality Index) values.

Thank you!