A Machine Learning Approach for Ozone Forecasting and its Application for Tri-Cities

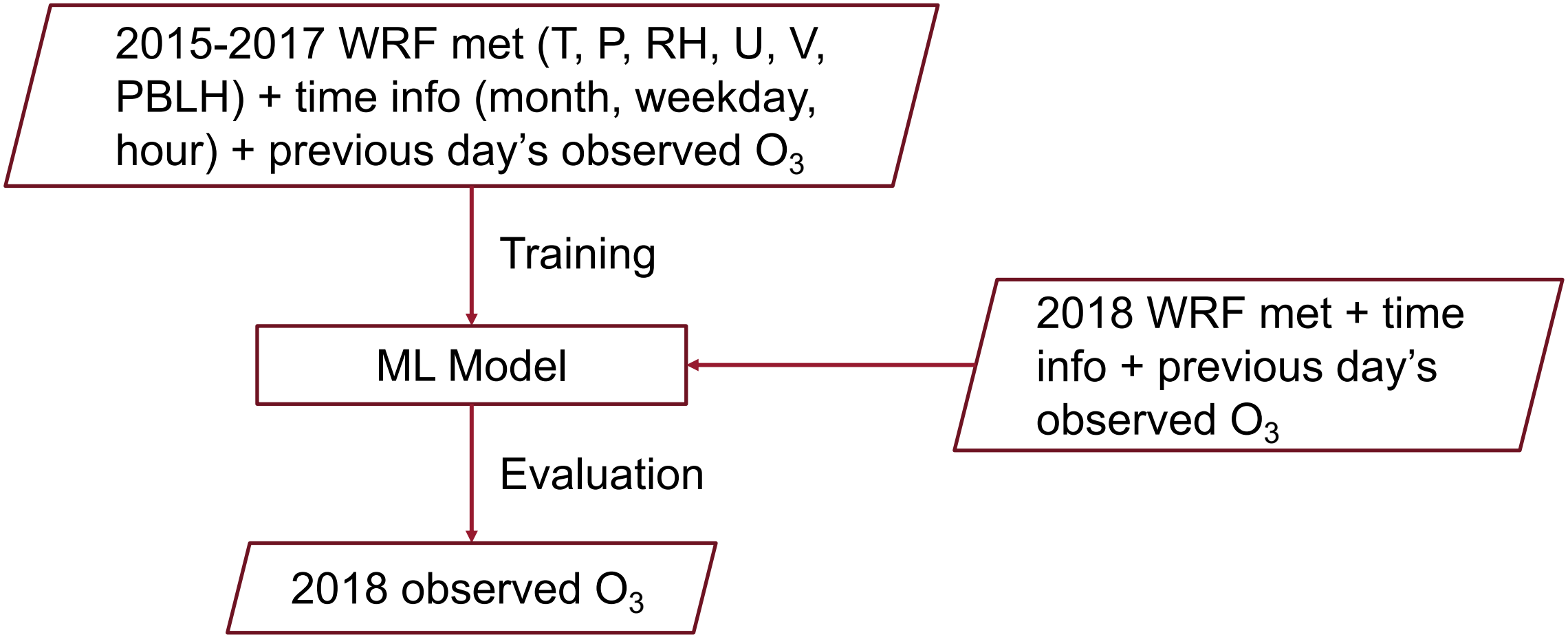
Kai Fan[1], Ranil Dhammapala[2], Brian Lamb[1],
Ryan Lamastro[3], Yunha Lee[1]

[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

Machine Learning Models

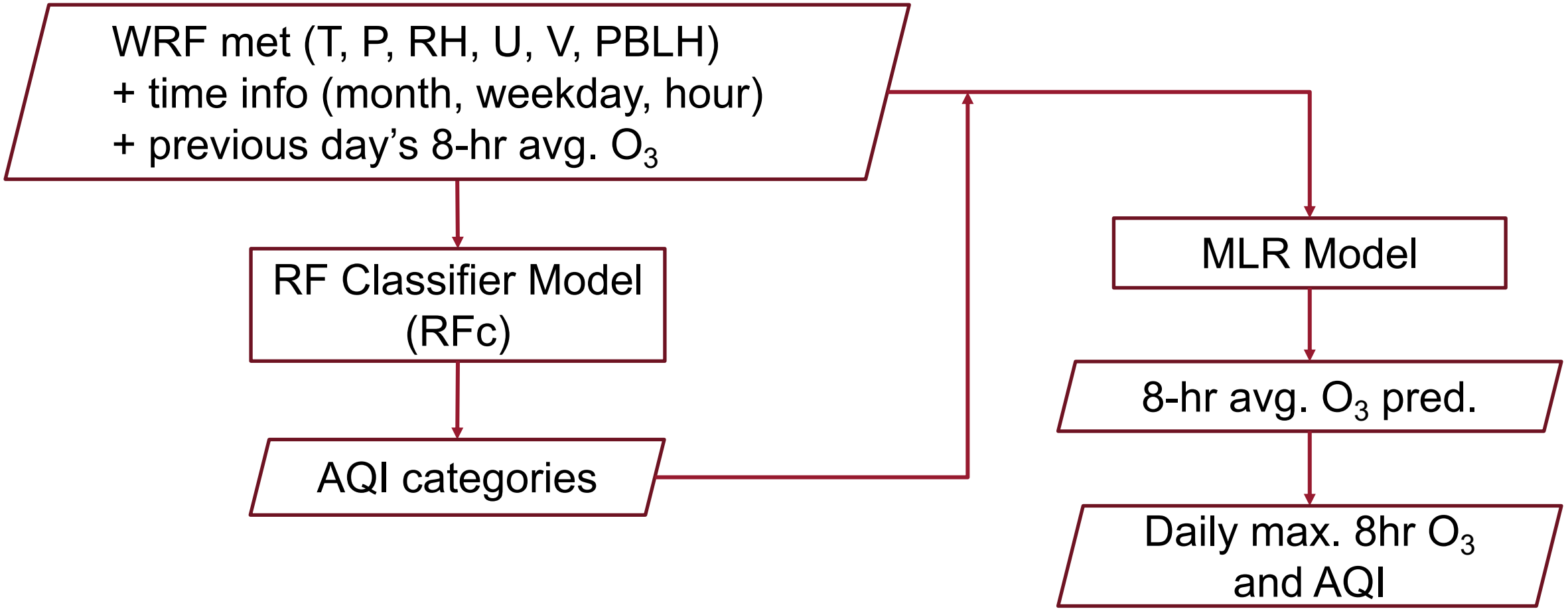
- Machine Learning is an application of artificial intelligence that lets the model learn from historical data and then make future forecasts
- Our approach uses multiple linear regression (MLR) model and random forest (RF) model
- Multiple linear regression fits a line between multiple independent variables and one dependent variable
- Random forest is the average result of many decision trees

Machine Learning Scheme for the Kennewick Monitoring Site



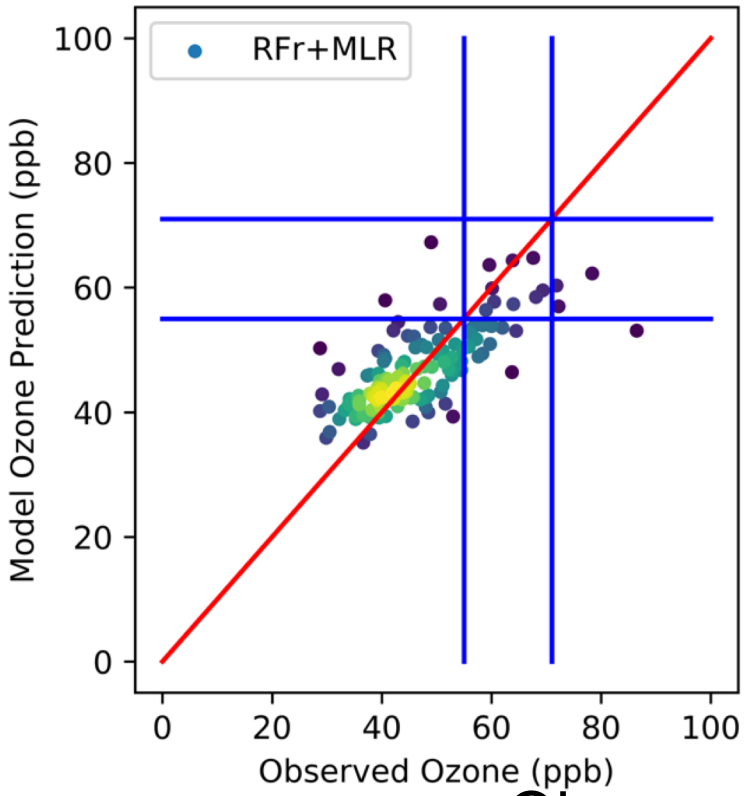
RF classifier + MLR Modeling Framework

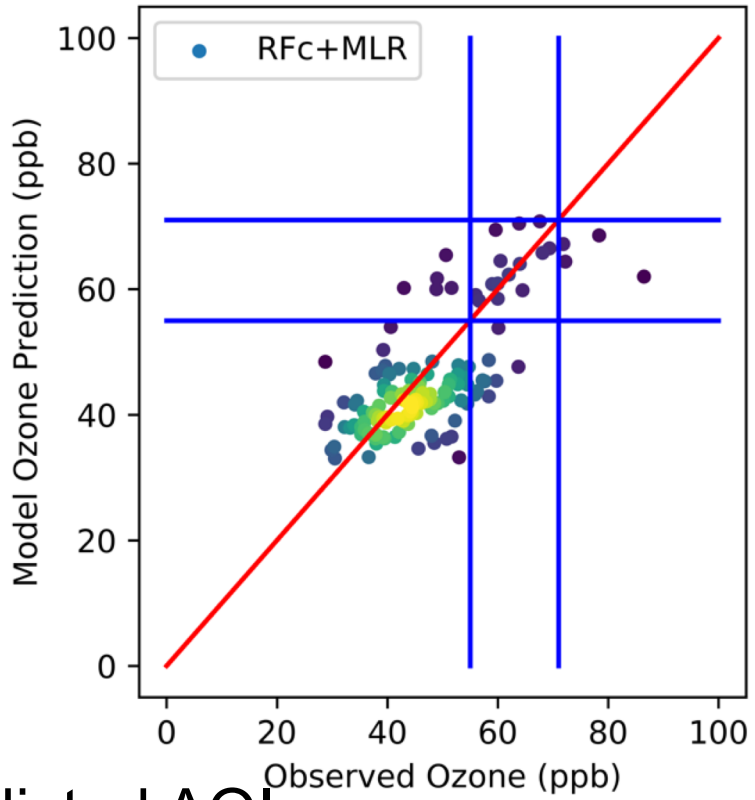
- RF approach does a good job on classification
- MLR approach shows better performance to predict high O_3 days



Modeling Framework RFr+MLR vs. RFc+MLR

Model vs. Obs. Daily max. O_3



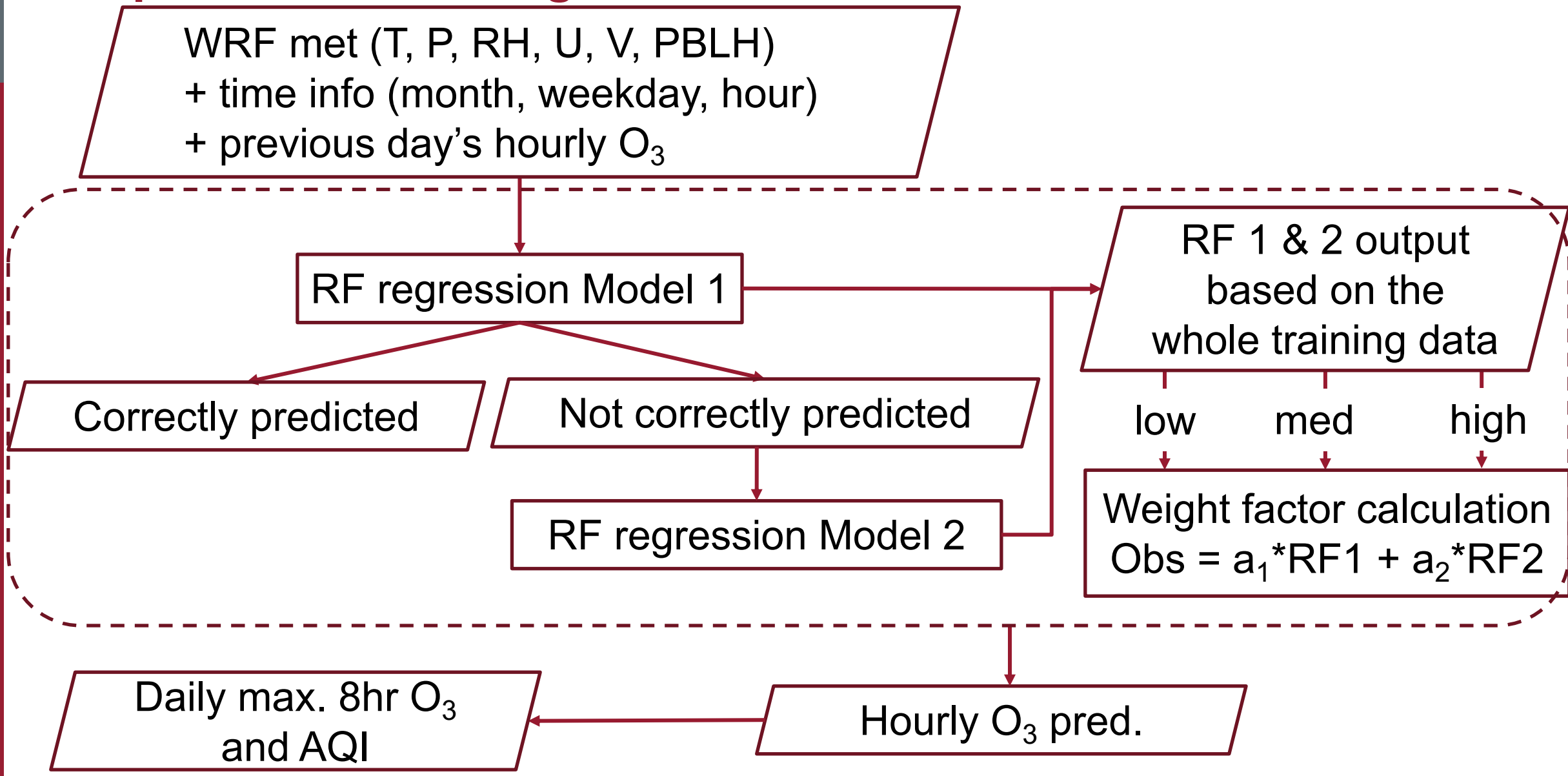


Observed AQI vs. predicted AQI

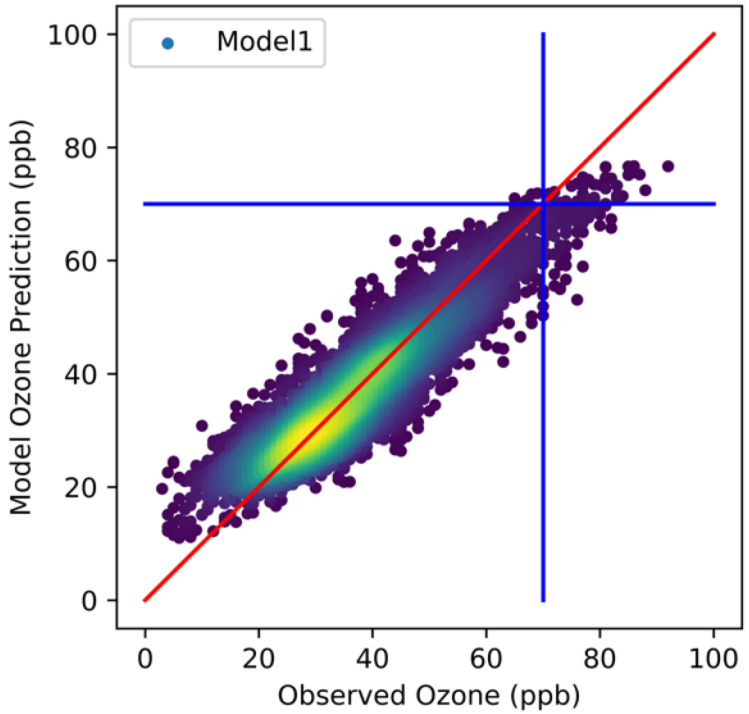
		Obs. AQI (days)			
		1	2	3	4
Model pred. AQI (days)	1	100	15	0	1
	2	3	9	3	0
	3	0	0	0	0
	4	0	0	0	0

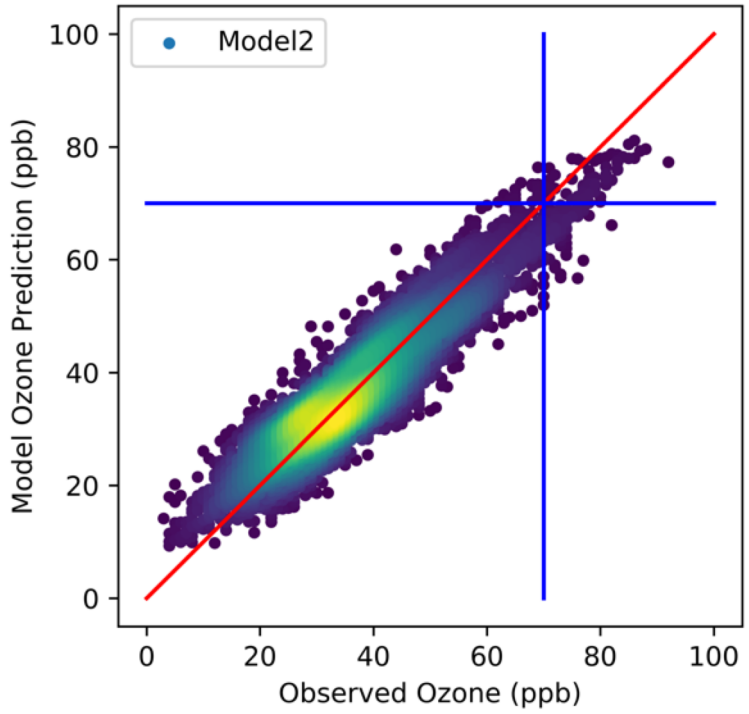
		Obs. AQI (days)			
		1	2	3	4
Model pred. AQI (days)	1	98	10	0	0
	2	5	13	3	1
	3	0	1	0	0
	4	0	0	0	0

Two-phase RF Modeling Framework*

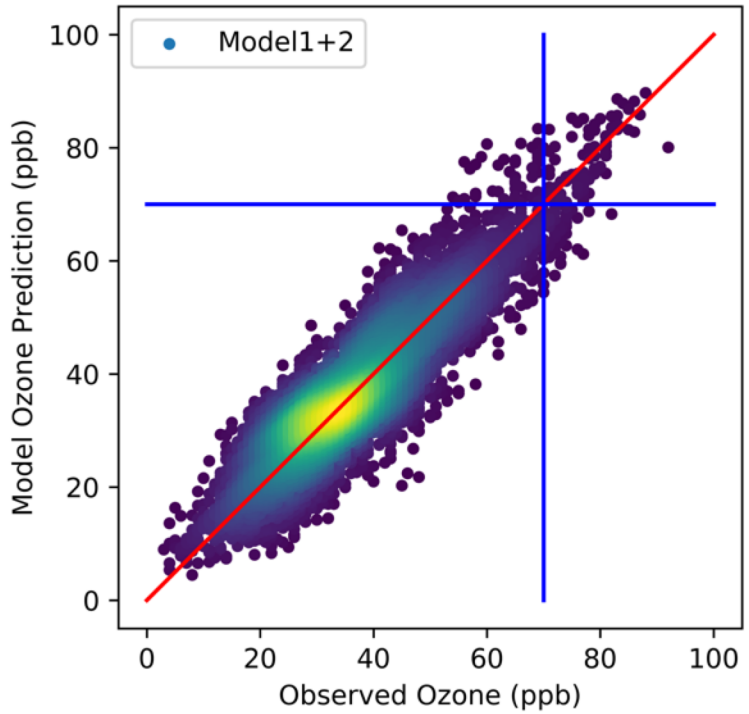


* Jiang, N., & Riley, M. L. (2015). Exploring the utility of the random forest method for forecasting ozone pollution in SYDNEY. *Journal of Environment Protection and Sustainable Development*, *1*(5), 245-254.





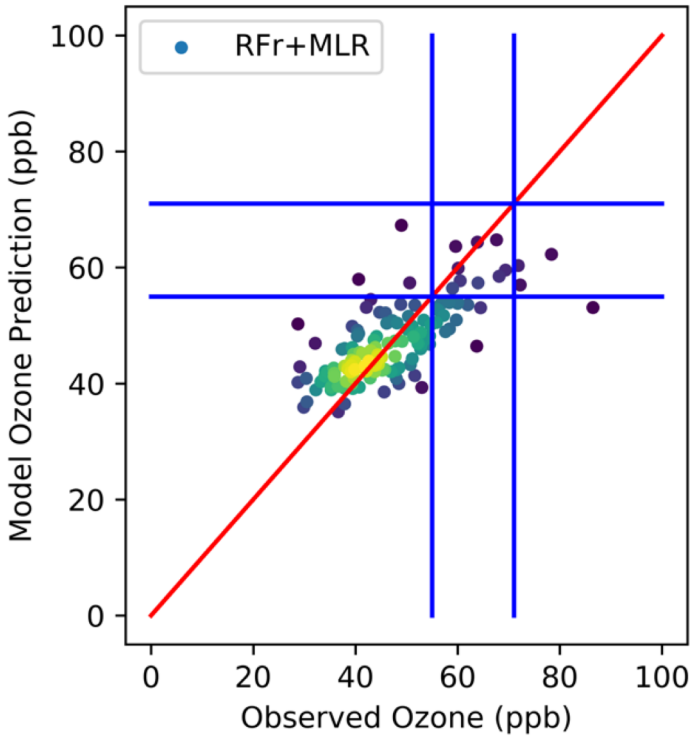
Weight Factor	RF 1	RF 2
Low (0 – 40 ppb)	-0.9	1.9
Medium (40 – 60 ppb)	-1.4	2.4
High (>60 ppb)	-1.1	2.1

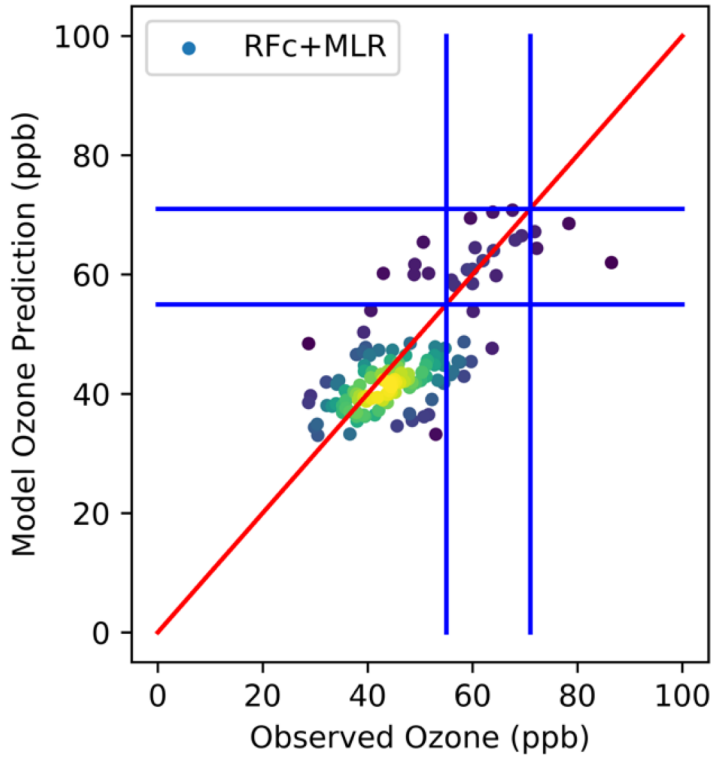


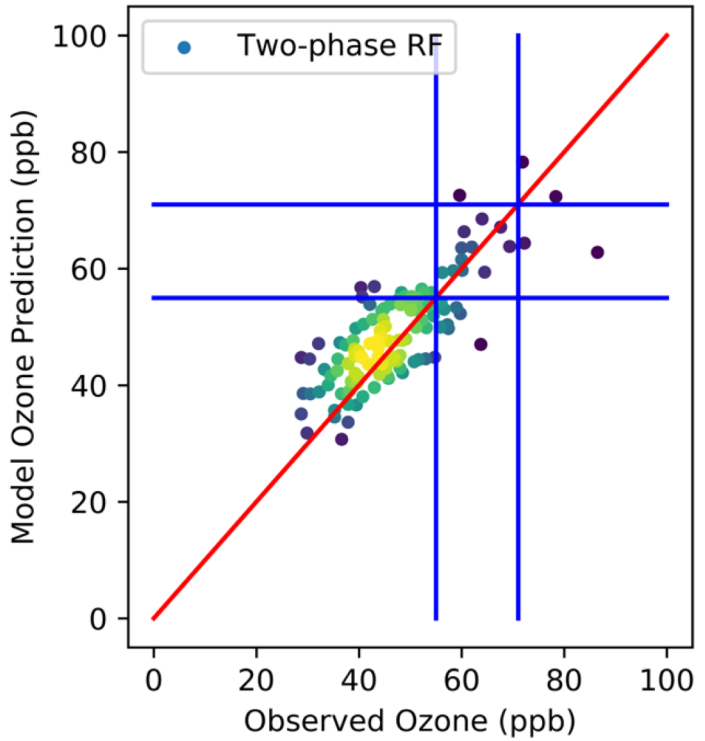
	RF 1	RF 2	Two-phase RF
NMB_high	-7.6%	-6.1%	-2.8%

Modeling Frameworks Evaluation

Model vs. Obs. Daily max. O_3







Observed AQI vs. predicted AQI

		Obs. AQI (days)			
		1	2	3	4
Model pred. AQI (days)	1	100	15	0	1
	2	3	9	3	0
	3	0	0	0	0
	4	0	0	0	0

		Obs. AQI (days)			
		1	2	3	4
Model pred. AQI (days)	1	98	10	0	0
	2	5	13	3	1
	3	0	1	0	0
	4	0	0	0	0

		Obs. AQI (days)			
		1	2	3	4
Model pred. AQI (days)	1	85	9	0	0
	2	11	11	1	1
	3	0	1	2	0
	4	0	0	0	0

Modeling Frameworks Evaluation

Model prediction accuracy (days)

	Obs low	Obs high
Model low	127	4
Model high	0	0

	Obs low	Obs high
Model low	125	4
Model high	2	0

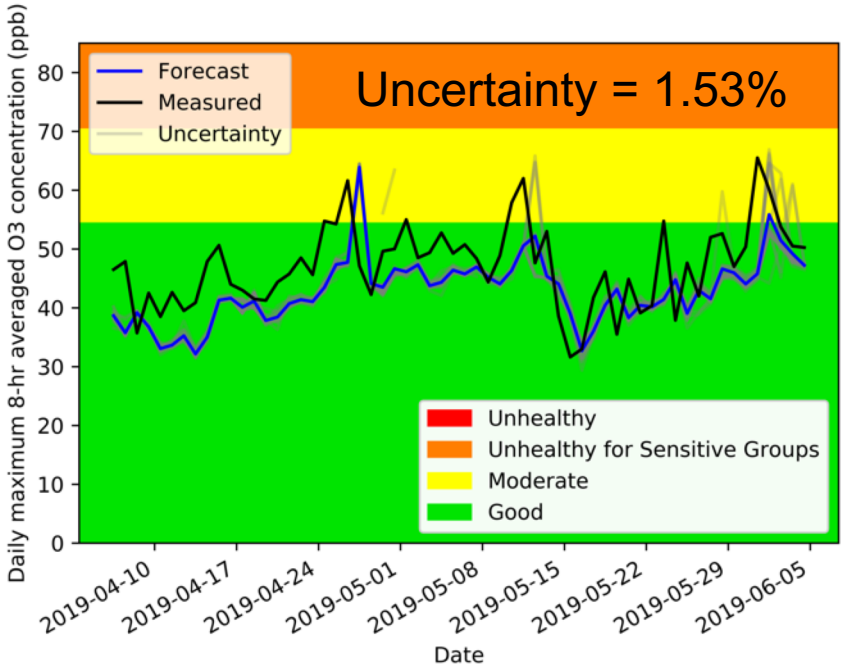
	Obs low	Obs high
Model low	116	2
Model high	1	2

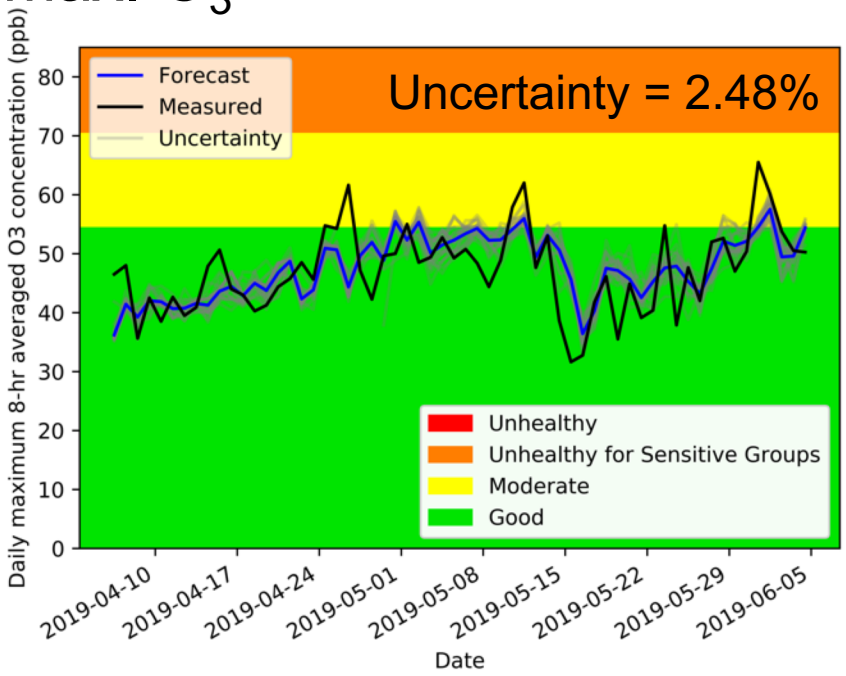
Failure rates between three modeling frameworks

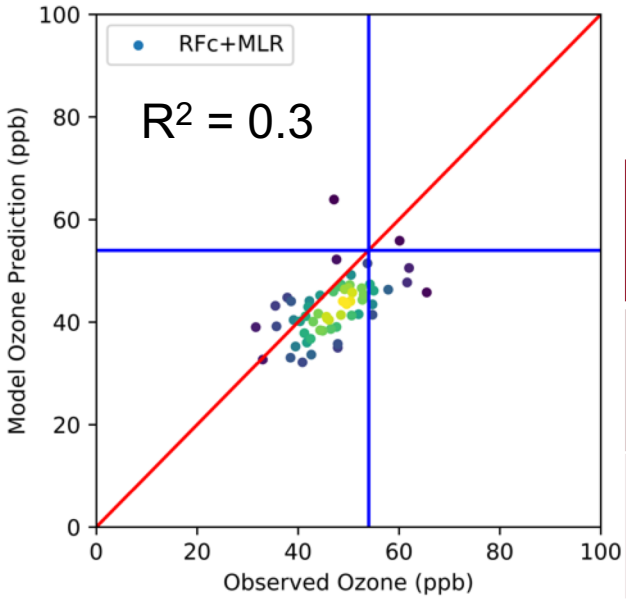
	RFr+MLR	RFc+MLR	Two-phase RF
False alarm rate	0	1.6%	0.8%
Failed to raise alarm	100%	100%	50%

Tri-Cites Ozone Forecast (RFc+MLR vs. Two-phase RF)

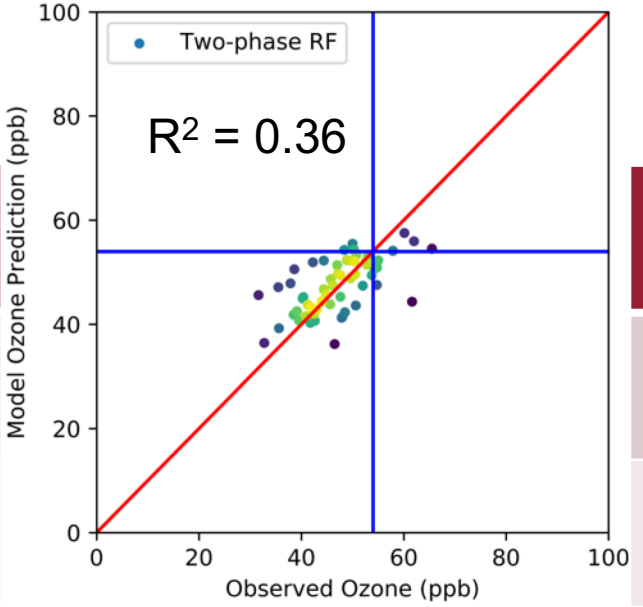
Time series of daily max. O_3







	Obs low	Obs high
Model low	50	8
Model high	1	1



	Obs low	Obs high
Model low	46	5
Model high	3	4

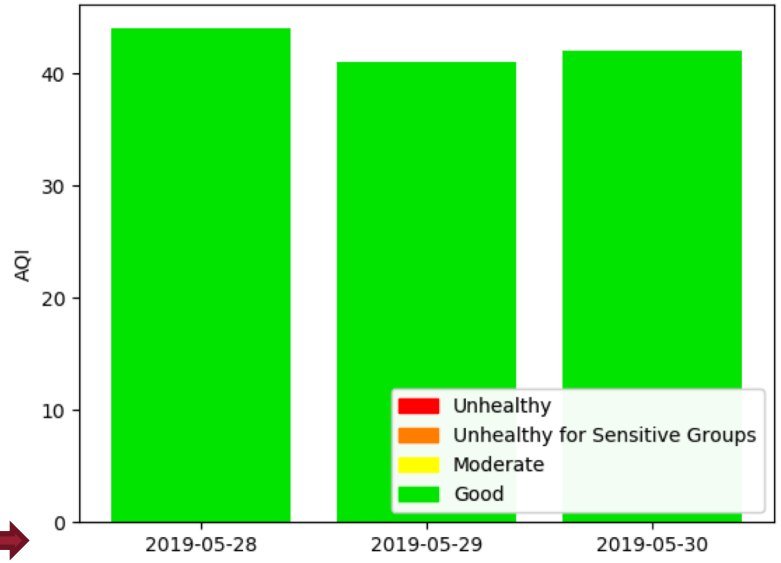
Tri-Cites Ozone Forecast

72-h ensemble UW-WRF forecast parameters for Kennewick

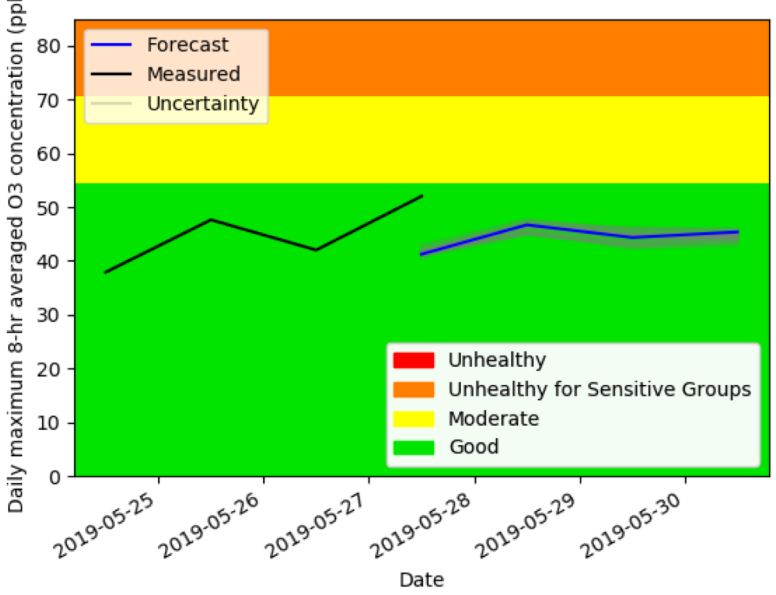
Previous day's measured ozone for Kennewick

→ ML modeling framework →

AQI forecasts

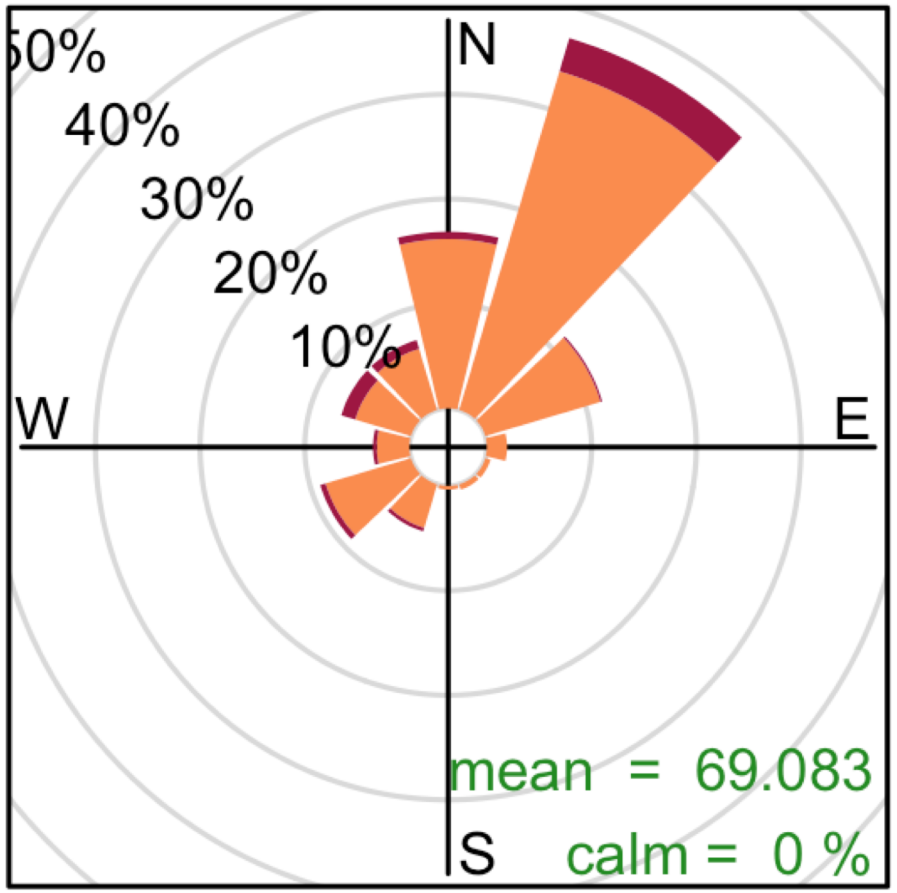


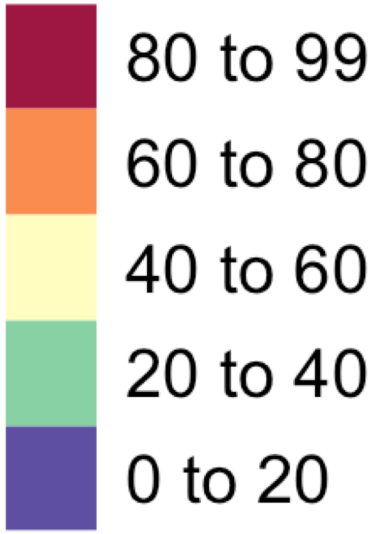
Daily max. O_3 forecasts



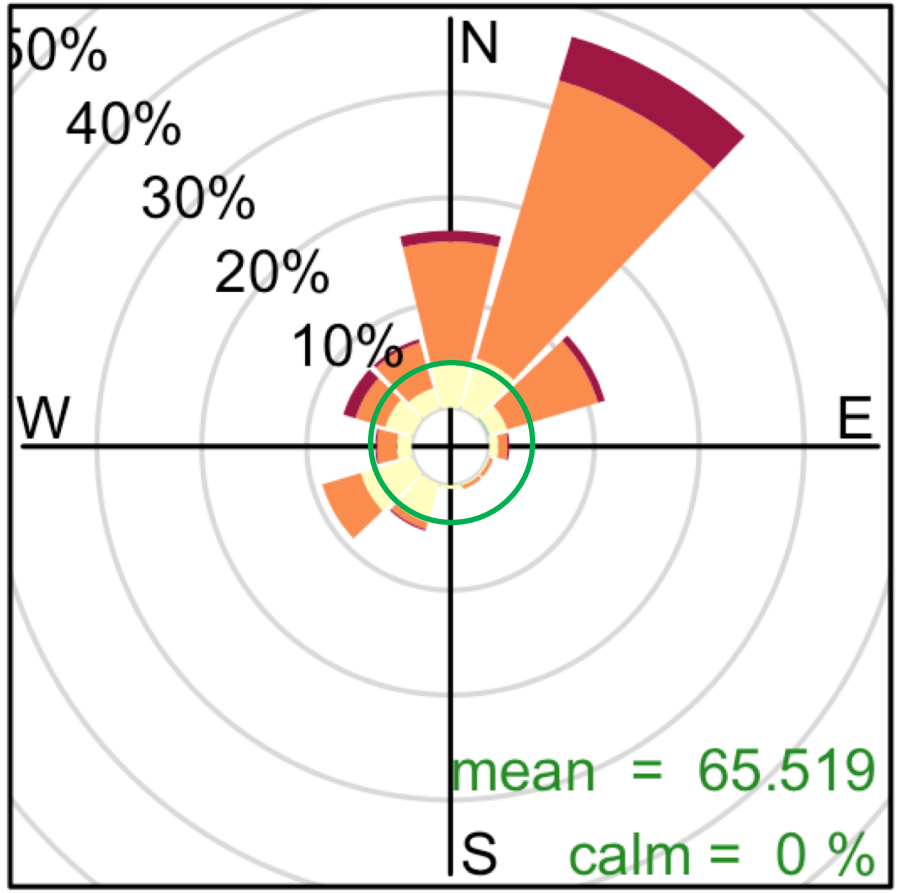
The pollution rose (hourly O_3 > 60 ppb)

Observation





Model



The pollution rose showing light NNE winds tend to coincide with high O_3

Imbalanced data problem

Year	Simulated days	AQI days				High AQI ratio (>2)
		1	2	3	4	
2015	94	68	22	4	0	4%
2016	133	113	18	2	0	2%
2017	100	62	31	7	0	7%
2018	138	108	25	4	1	4%
Total	465	351	96	17	1	4%

Summary

- The RFr+MLR model underestimates the O_3 mixing ratios and cannot capture the high O_3 days
- The model with RF classifier and MLR does not improve the high ozone day prediction compared with RFr+MLR
- The model with 2-phase RF regression can capture half high O_3 days in 2018 and perform well in 2019
- The imbalanced dataset makes it difficult to predict high O_3 days
- The light NNE winds tend to coincide with high O_3. In future, this will be considered in the model

Thank you!