Ozone Forecasting Using Machine Learning Methods

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

[1]Laboratory for Atmospheric Research,
Civil and Environmental Engineering, Washington State University
[2]State University of New York at New Paltz
[3]Washington State Department of Ecology

Site and Objective

- Kennewick, WA lies 32 km (20 mi) north of Washington's southern border, near where the Columbia River meets the Snake River

- The **goal** is to predict high ozone events through the use of Machine Learning models

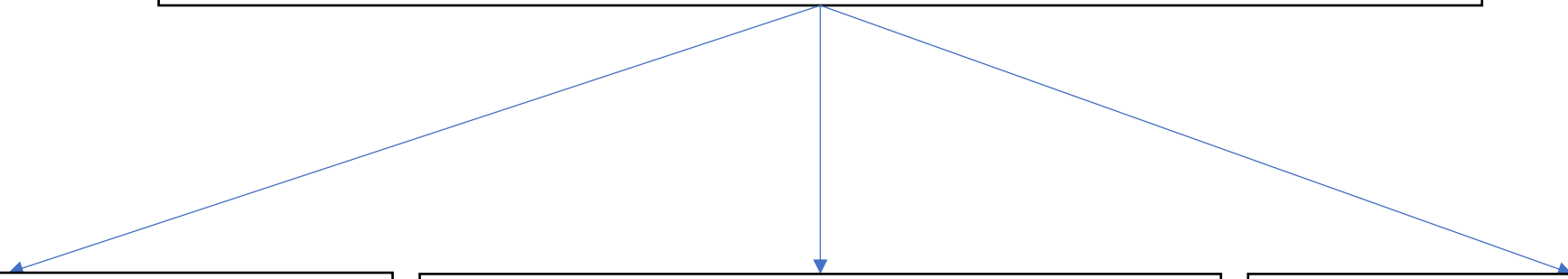
- We aim to provide better warning for days of poor air quality due to ozone

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 model**, **generalized additive model** and **random forest model**
- Multiple linear regression fits a line between multiple independent variables and one dependent variable
- Generalized additive model where the linear relation between dependent and independent variables is replaced with nonlinear smooth functions
- Random forest is the average result of many decision trees

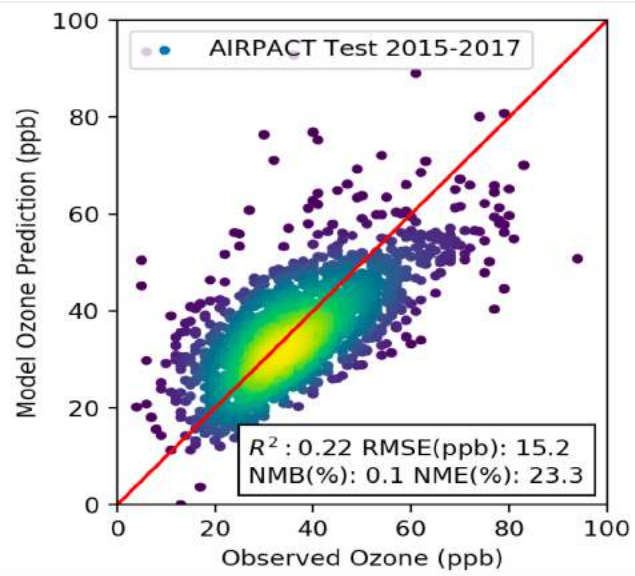
Machine Learning Scheme for the Kennewick Monitoring Site

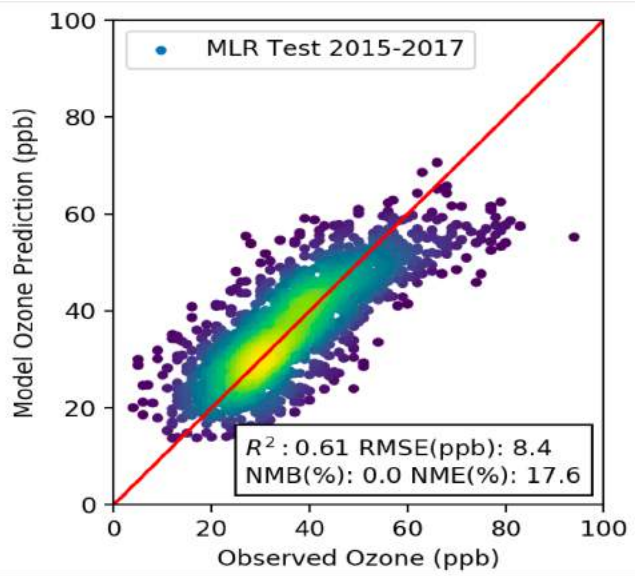
WRF met in Kennewick (PBL, P, Temp, U, V, RH)
+ month + weekday + hour
+ previous day's 8-hr avg. O_3

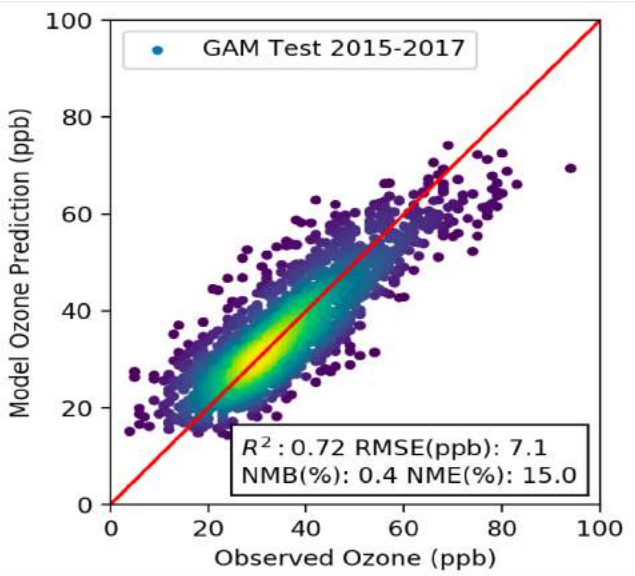


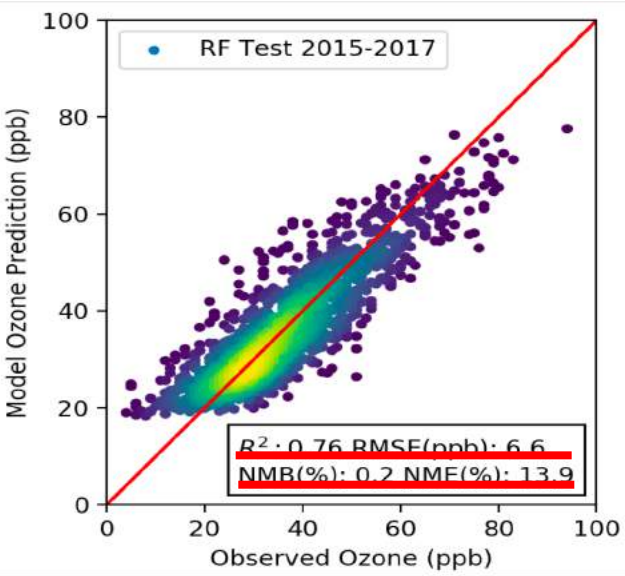
Multiple Linear Regression (MLR)	Generalized Additive Model (GAM)	Random Forest Model (RF)

Test 2015-2017

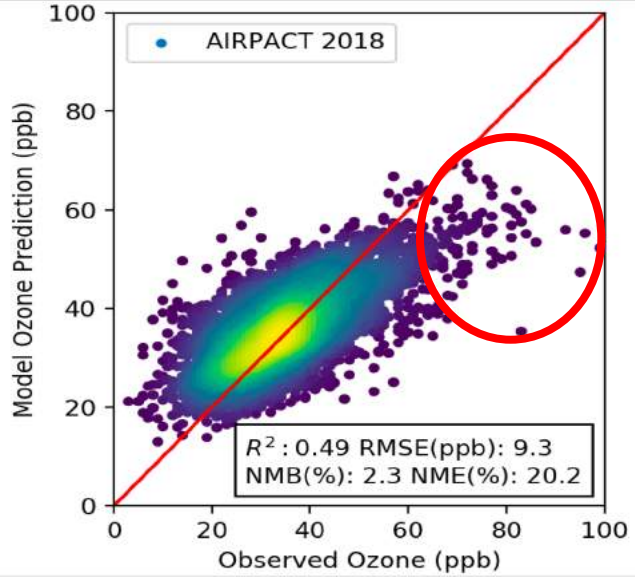


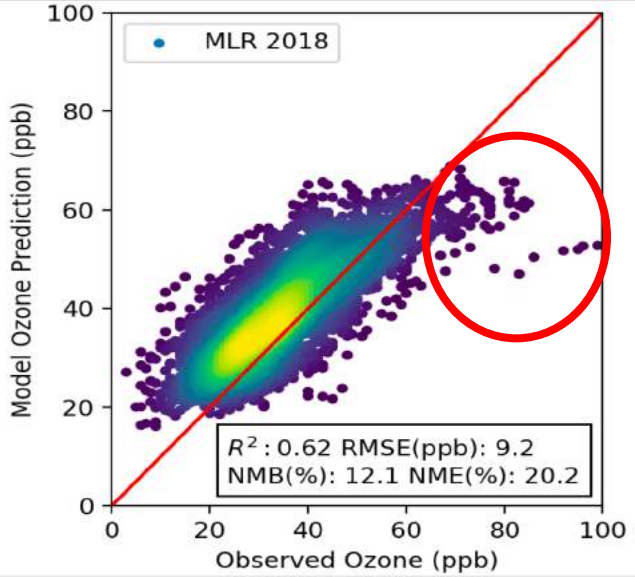


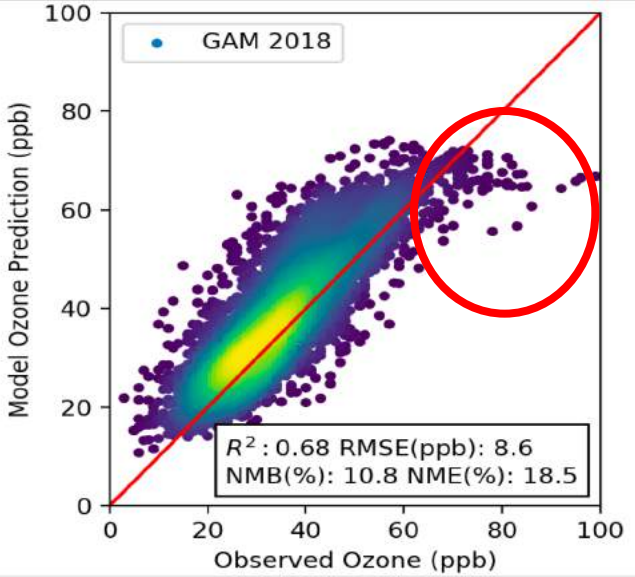


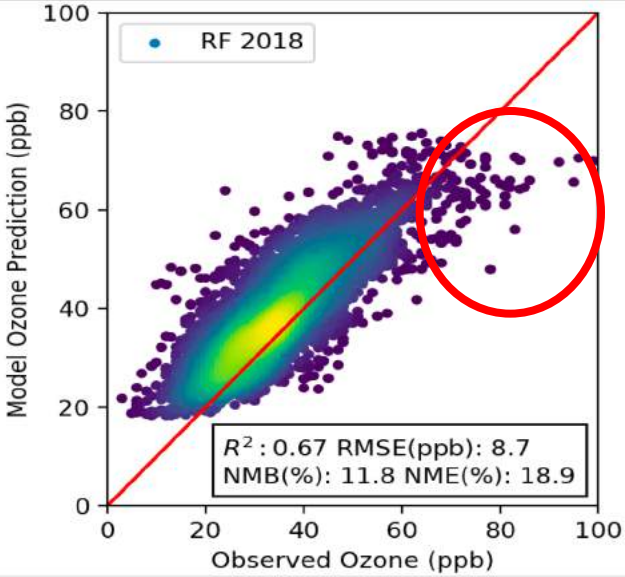


Evaluate 2018

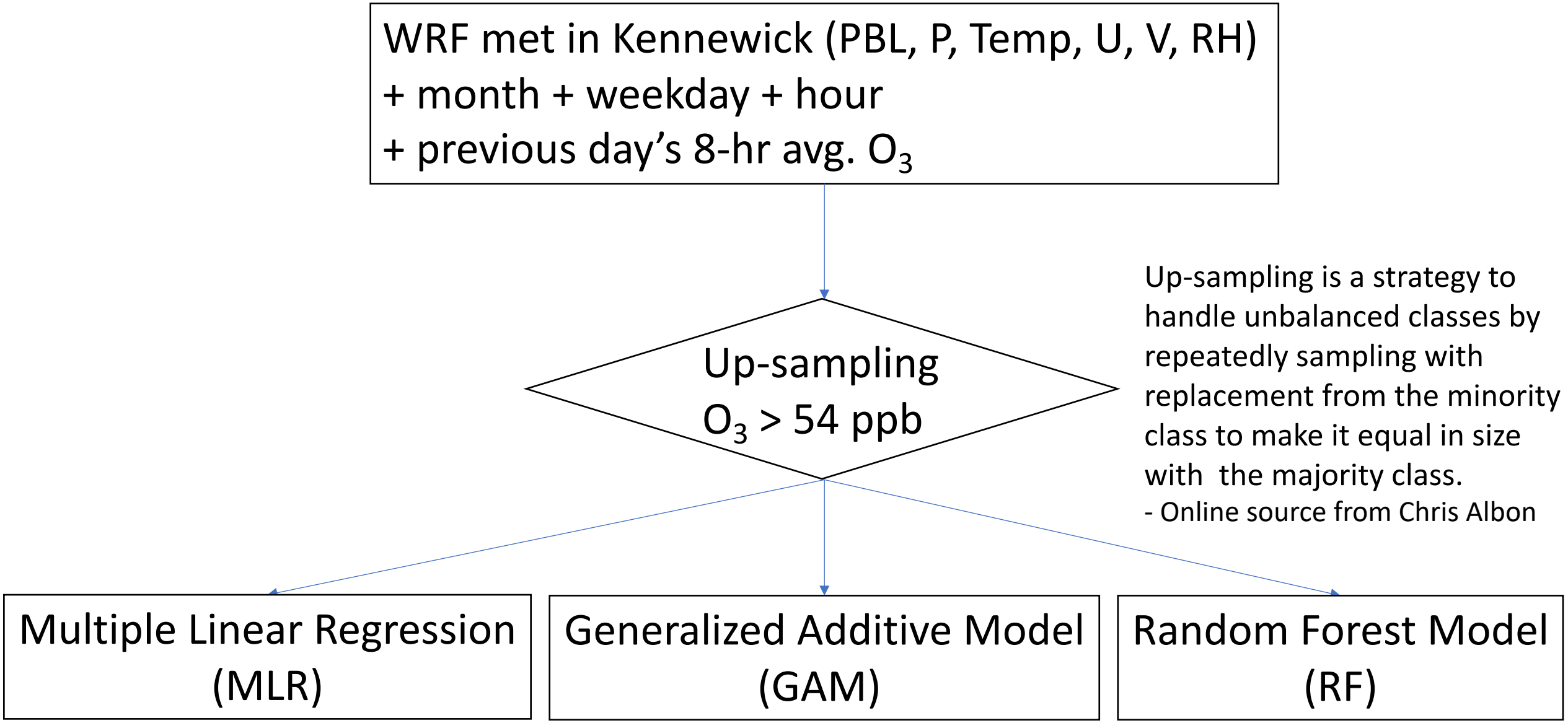








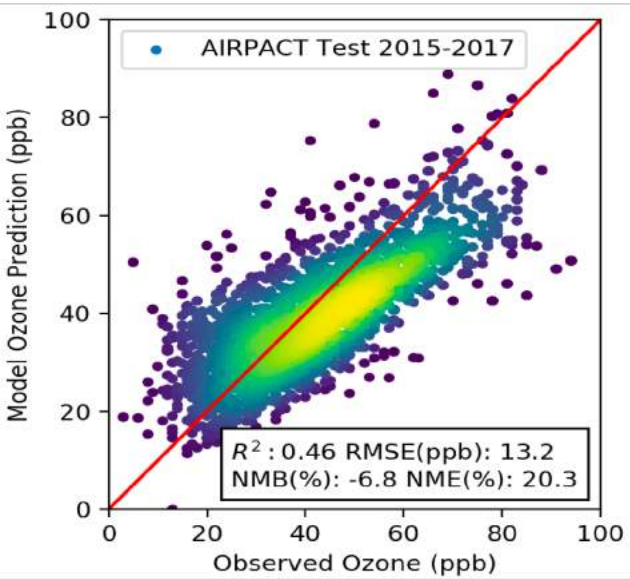
All models underestimate peak O_3, but perform well for lower O_3 levels

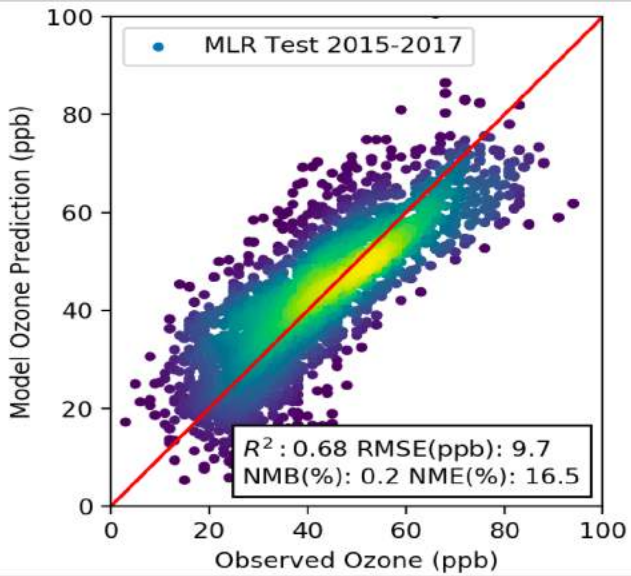


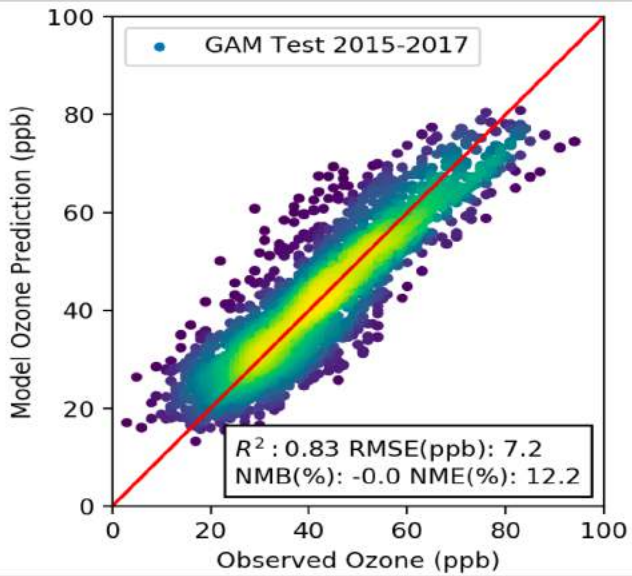
Up-sampling is a strategy to handle unbalanced classes by repeatedly sampling with replacement from the minority class to make it equal in size with the majority class.

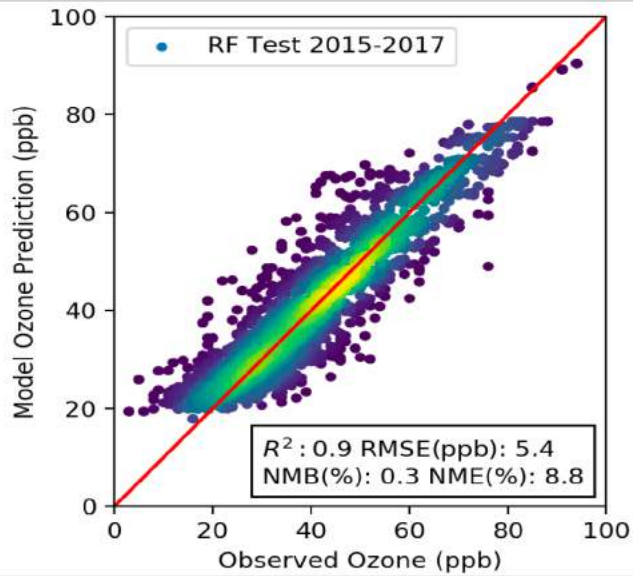
- Online source from Chris Albon

Test 2015-2017 (Up-sampling)

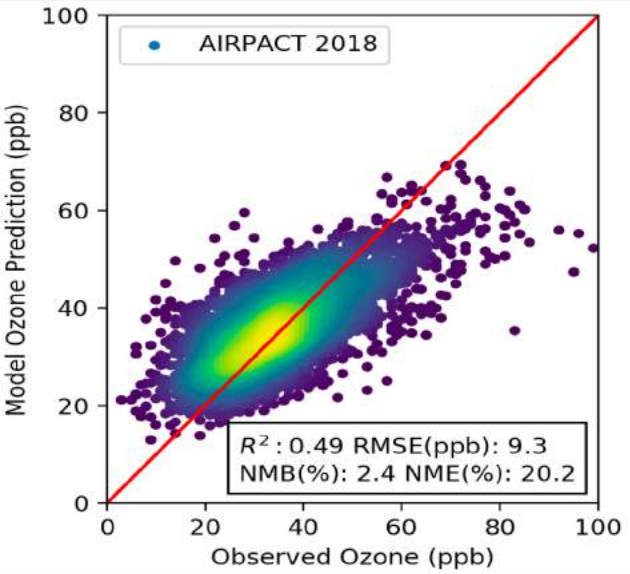


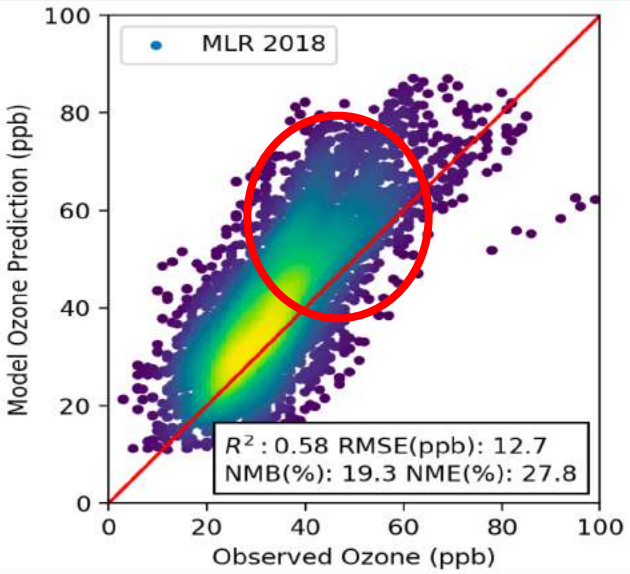


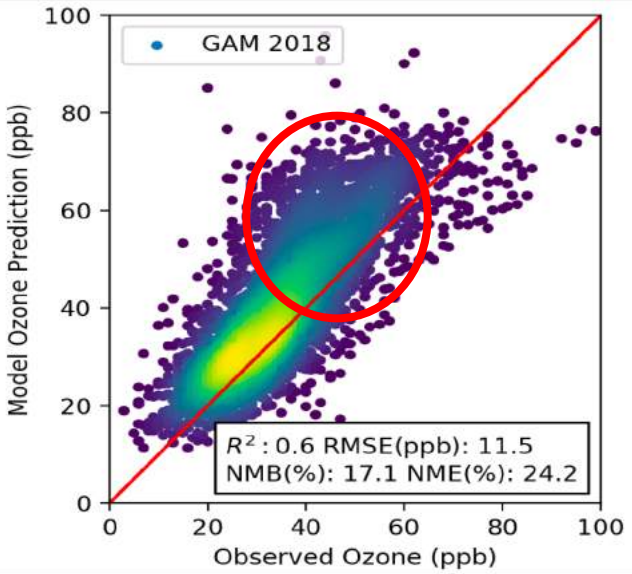


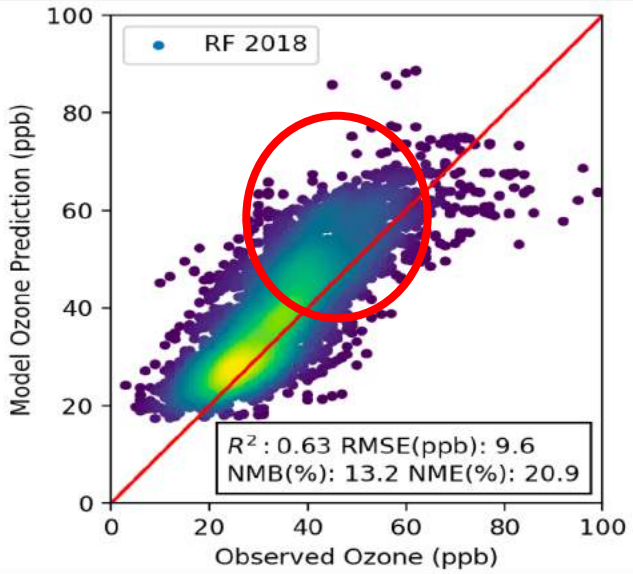


Evaluate 2018 (Up-sampling)





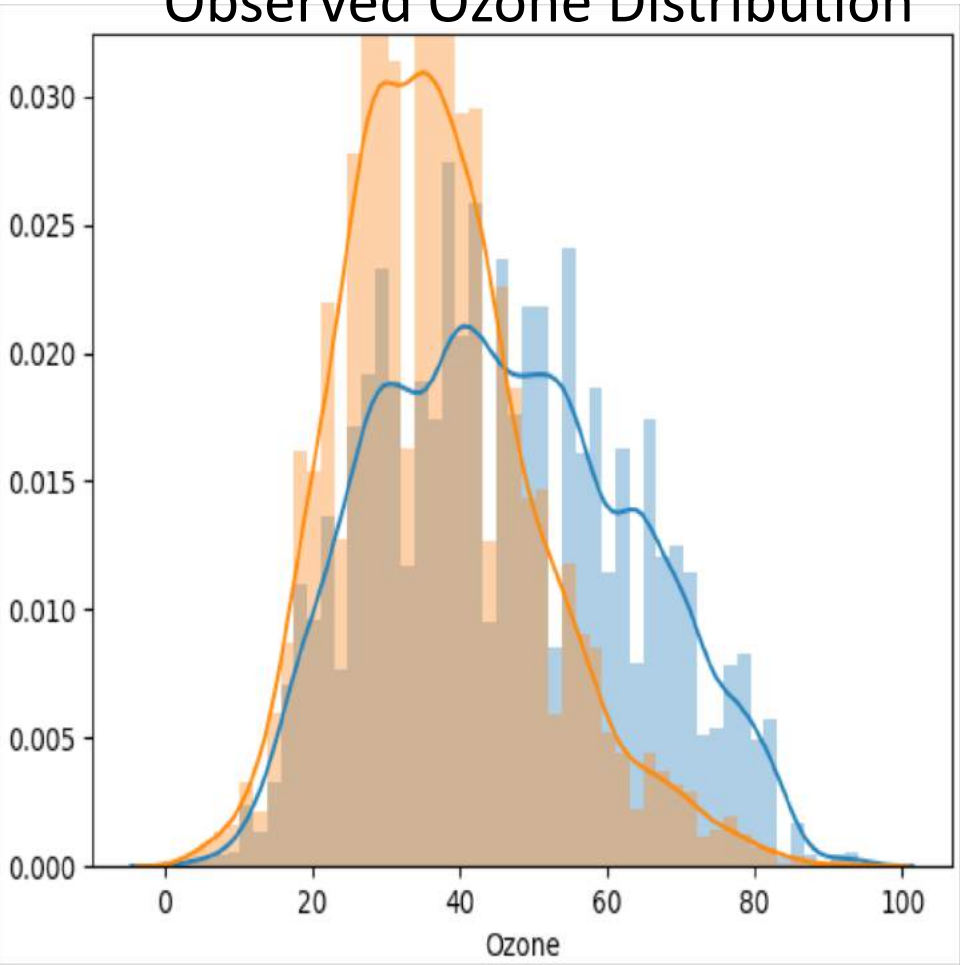




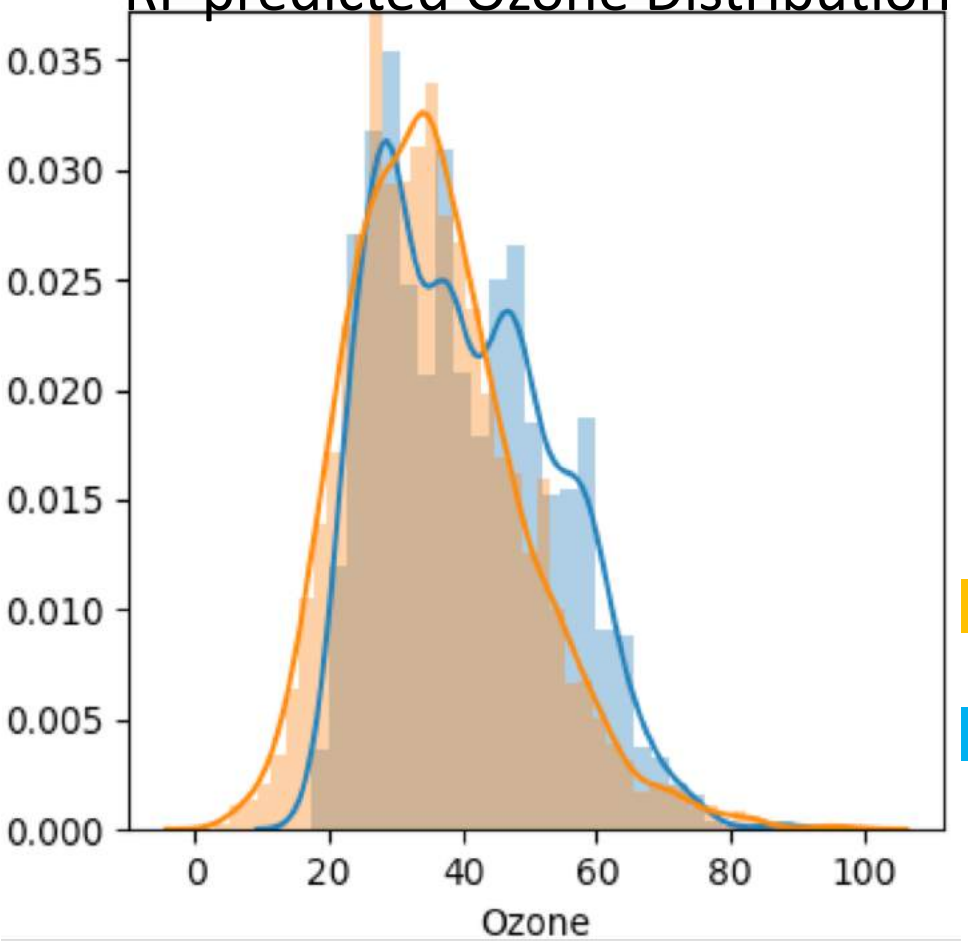
Up-sampling does not solve the high O_3 prediction problem.

Hourly Ozone Distribution

Observed Ozone Distribution



RF predicted Ozone Distribution

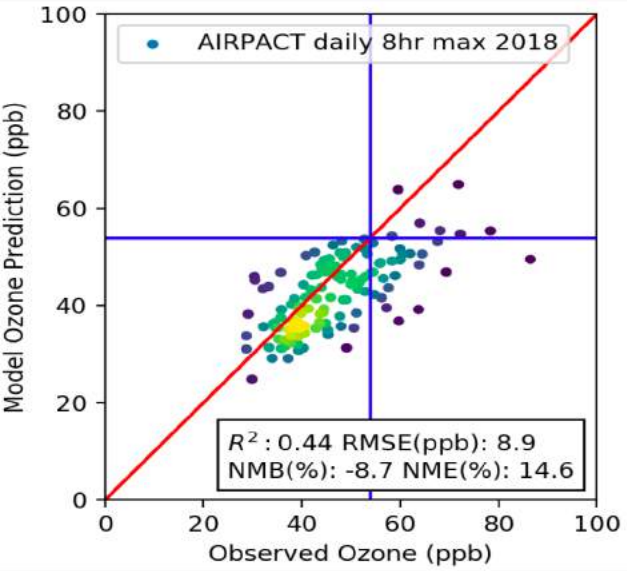


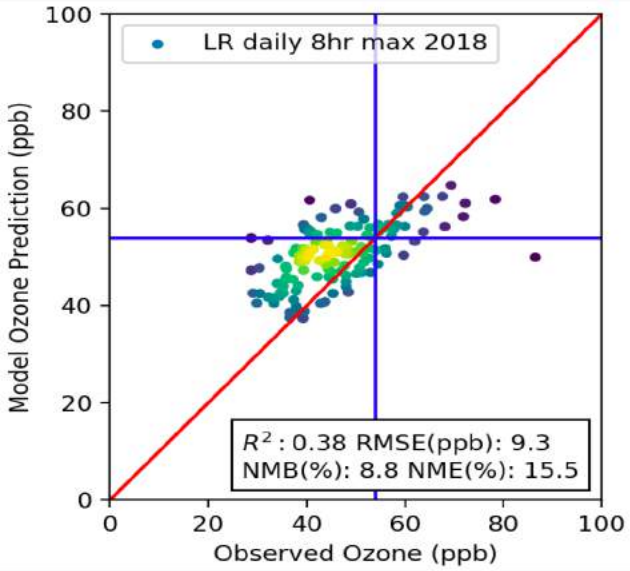
Original O_3

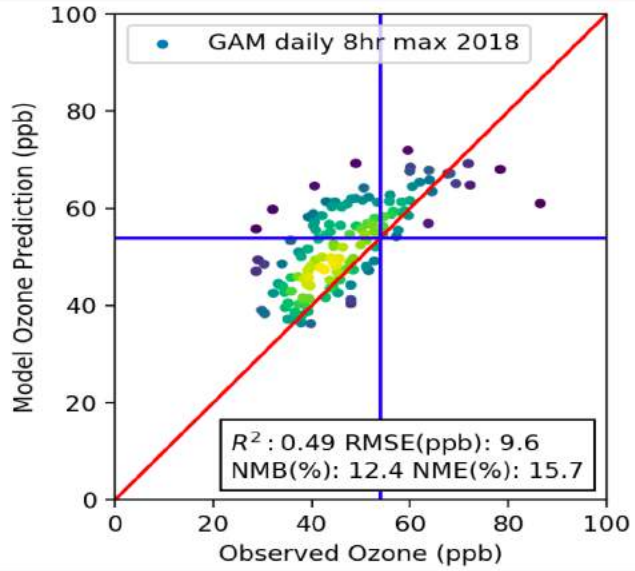
Resampled O_3

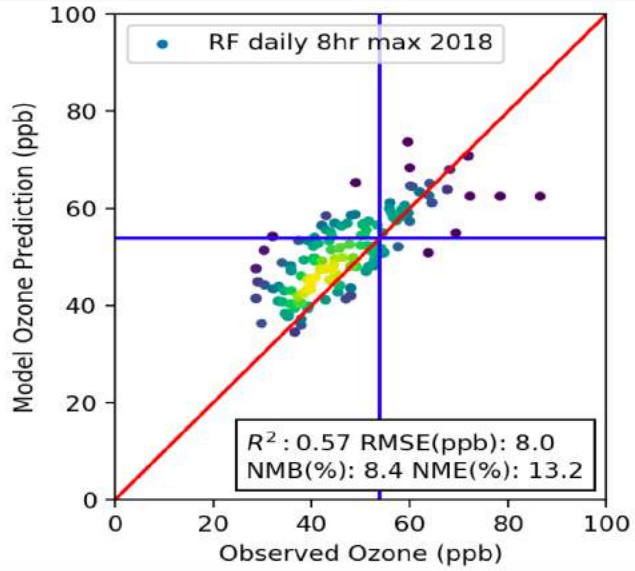
Model predicted ozone distribution does not show much difference as the observations after up-sampling.

Daily maximum 8-hour O_3 concentration

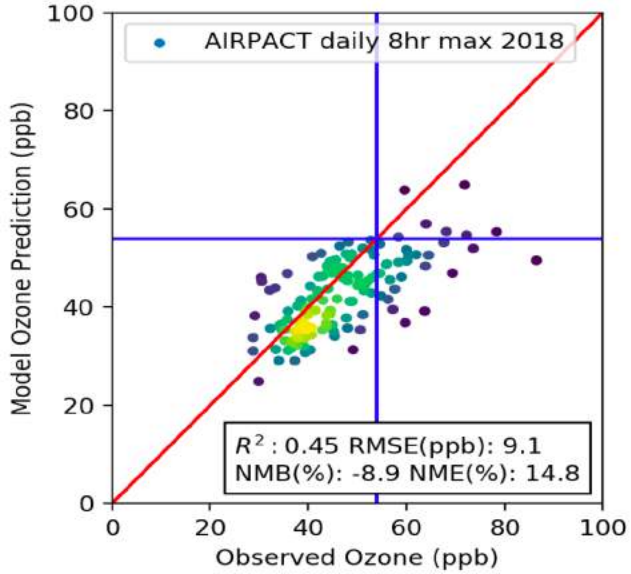


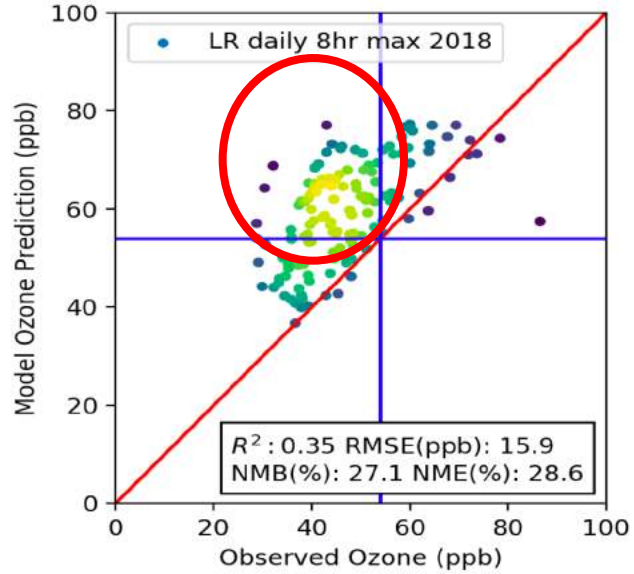


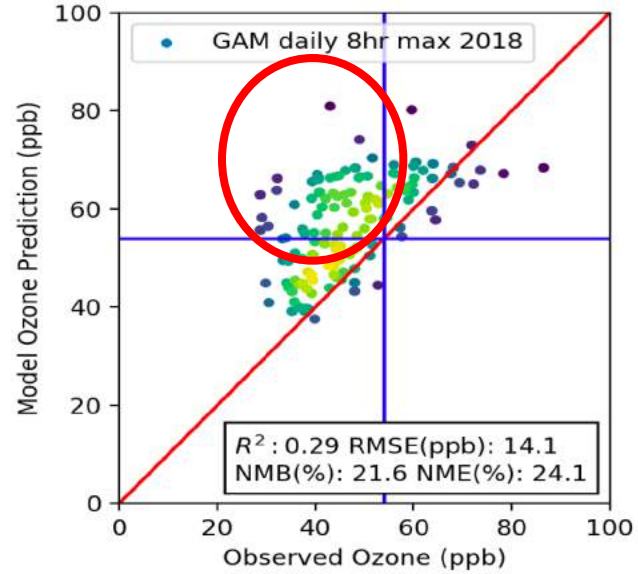


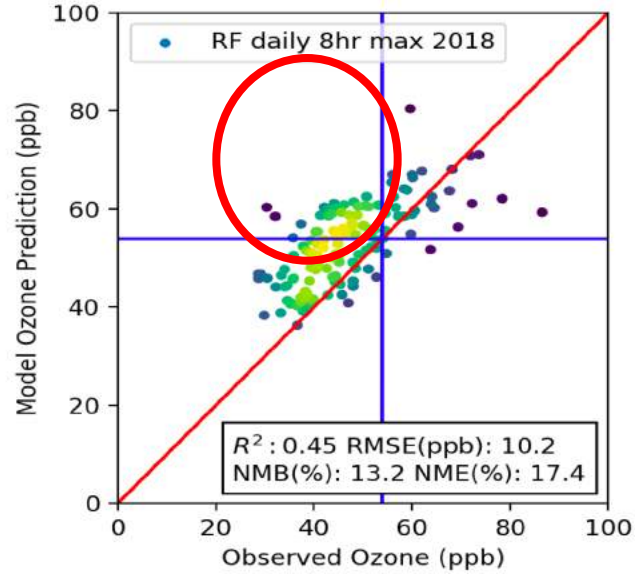


Daily maximum 8-hour O_3 concentration (Up-sampling)

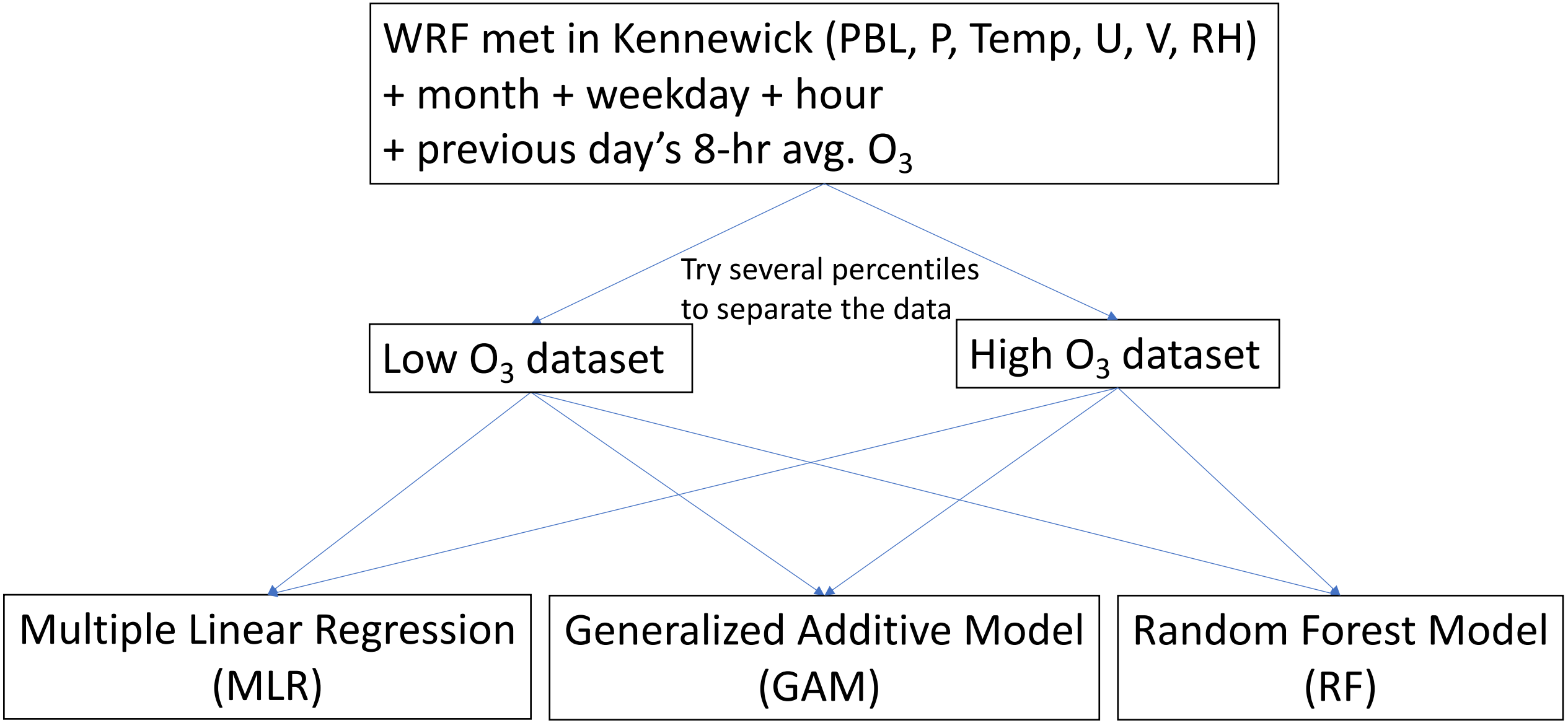








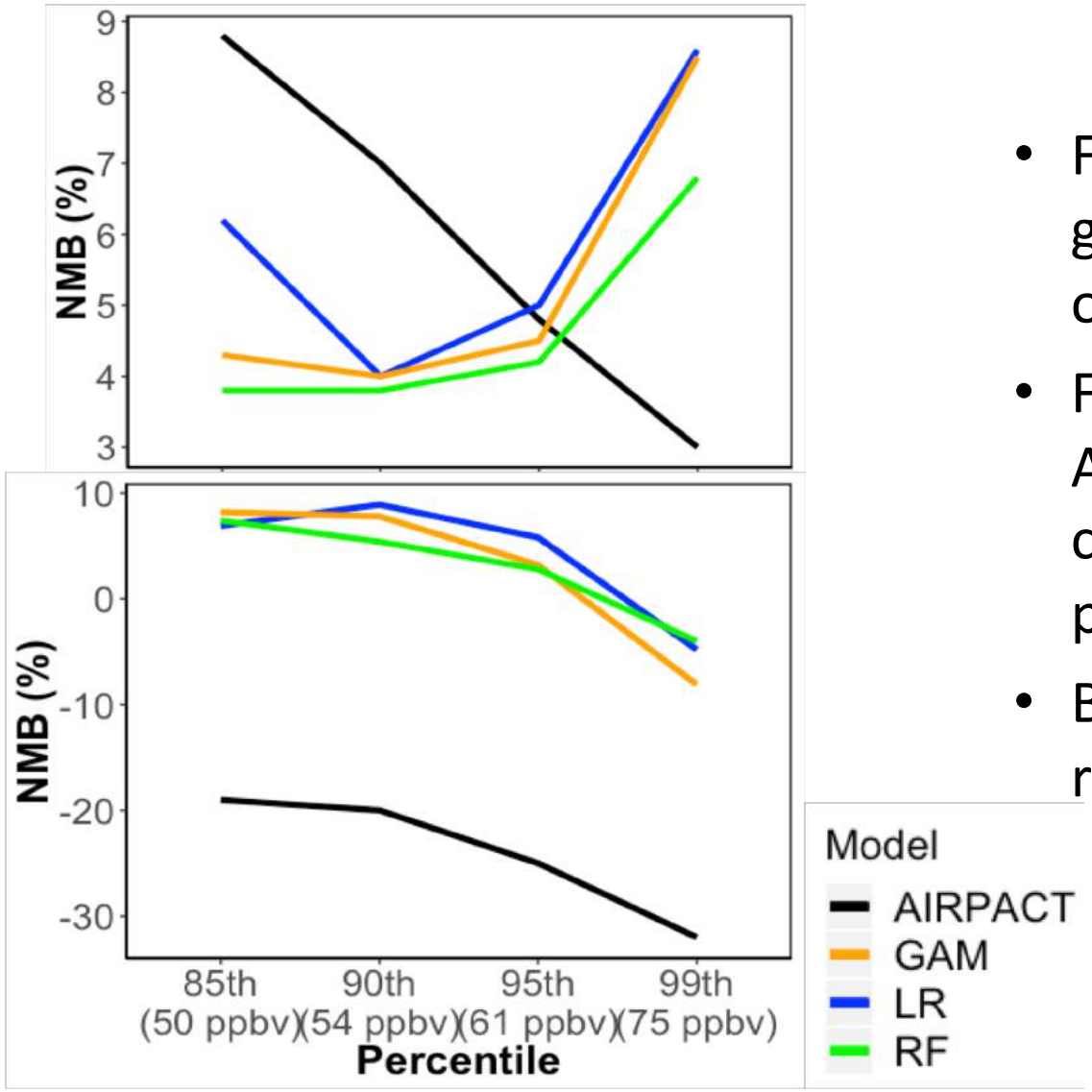
For daily maximum 8-hour O_3, up-sampling gives worse performance compared with the base model. This may be because up-sampling mostly predicts higher O_3 concentrations between 40 and 60 ppbv.


WRF met in Kennewick (PBL, P, Temp, U, V, RH)
+ month + weekday + hour
+ previous day's 8-hr avg. O_3
Try several percentiles
to separate the data
Low O_3 dataset
High O_3 dataset
Multiple Linear Regression (MLR)
Generalized Additive Model (GAM)
Random Forest Model (RF)

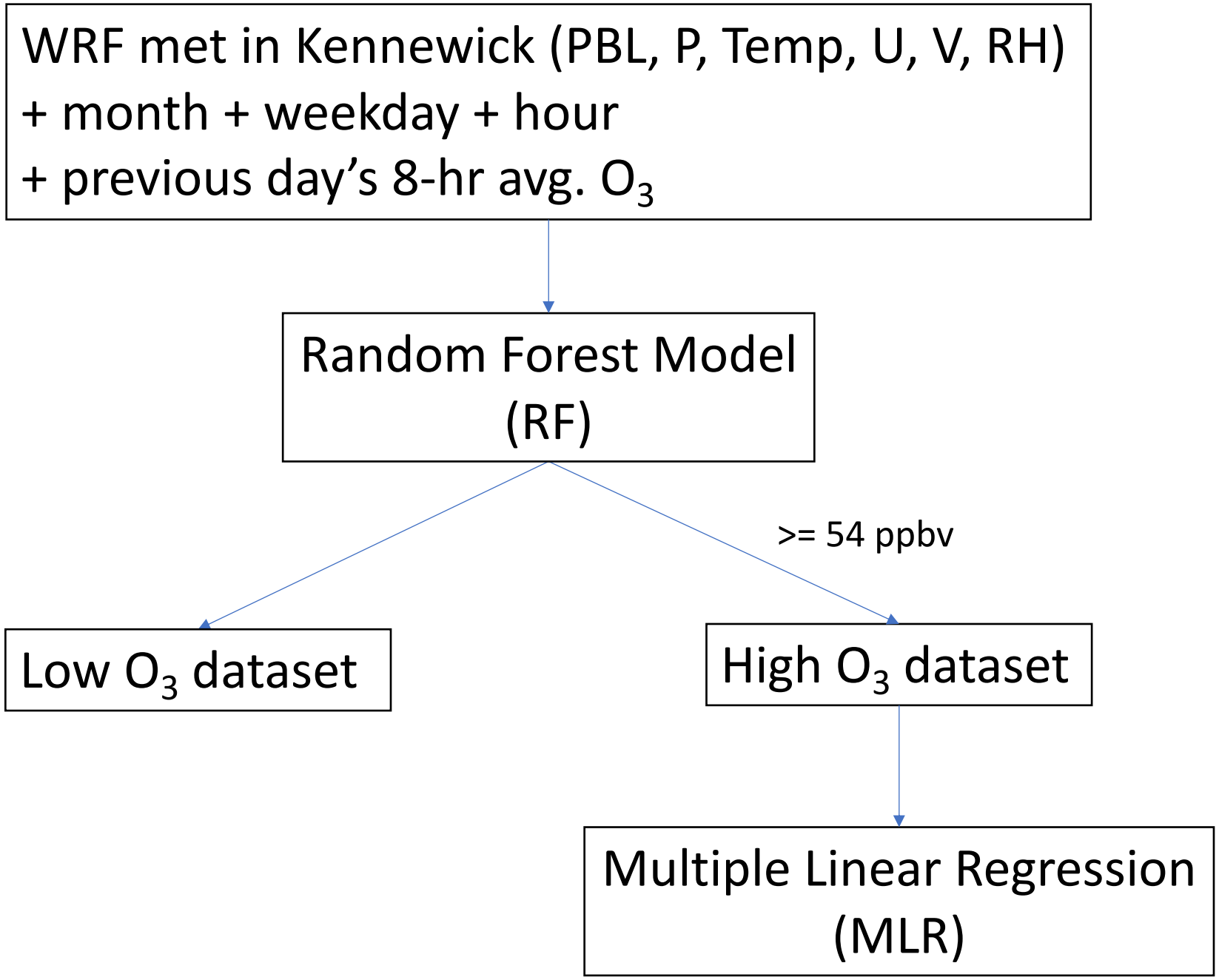
Model performance for different thresholds

Lower than the threshold

Higher than the threshold



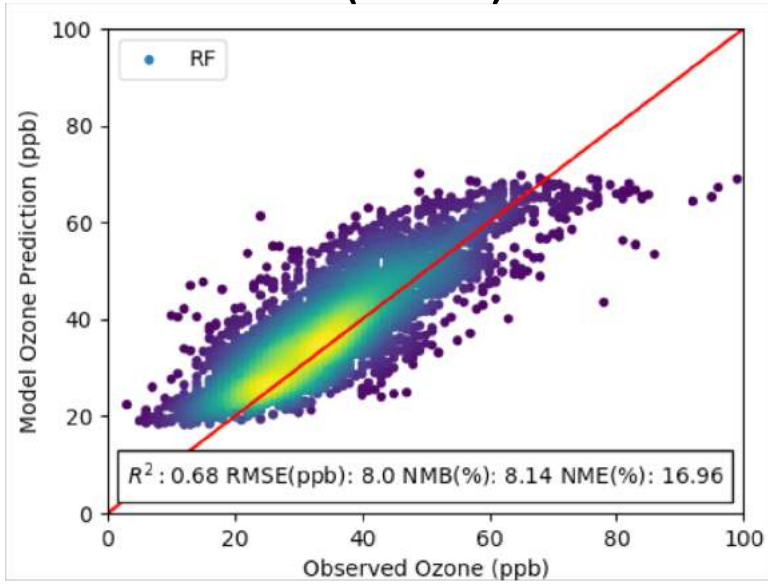
- For the data lower than the percentile, generally RF shows lower NMB than other models.
- For the data higher than the percentile, AIPRACT largely underestimates the O_3 concentrations, and the concentrations predicted by LR are higher than RF.
- Based on the NMB, 90th percentile is a reasonable point to separate the data.

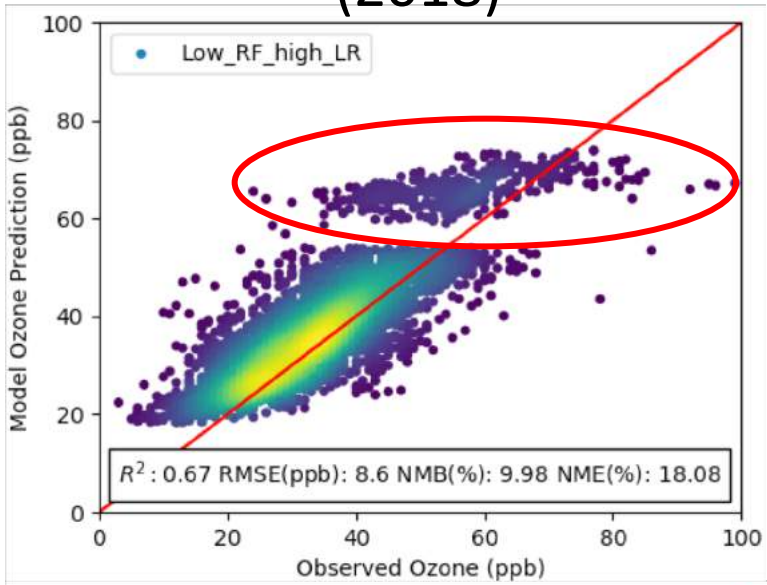

WRF met in Kennewick (PBL, P, Temp, U, V, RH)
+ month + weekday + hour
+ previous day's 8-hr avg. O_3
Random Forest Model (RF)
>= 54 ppbv
Low O_3 dataset
High O_3 dataset
Multiple Linear Regression (MLR)

Random Forest (2018)

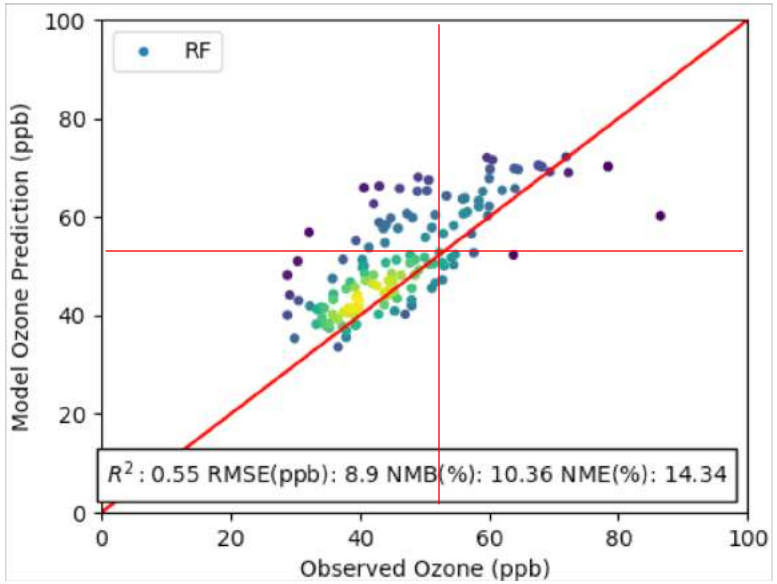
Random Forest + Multiple Linear Regression (2018)

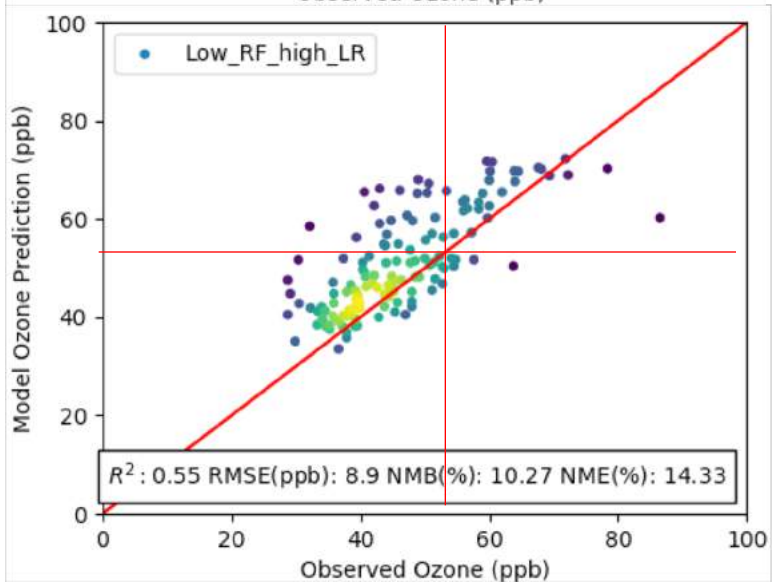
Hourly O_3





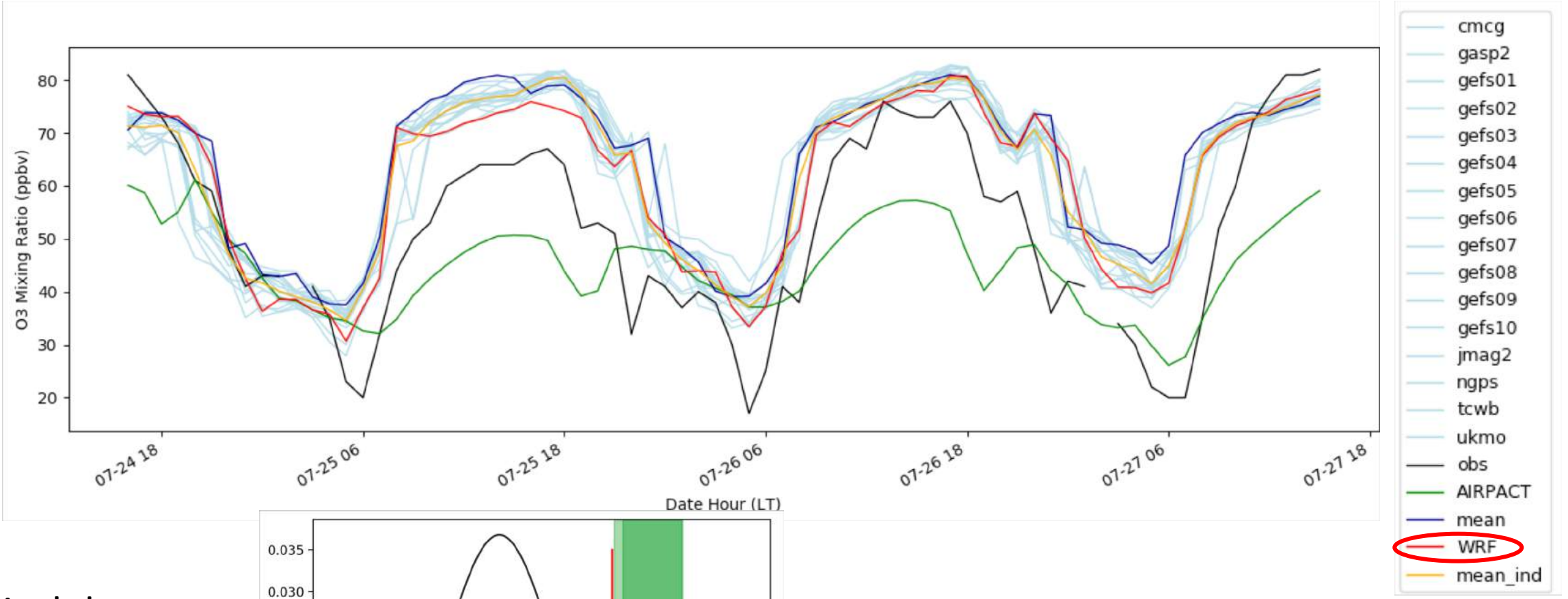
Daily maximum 8-hour O_3



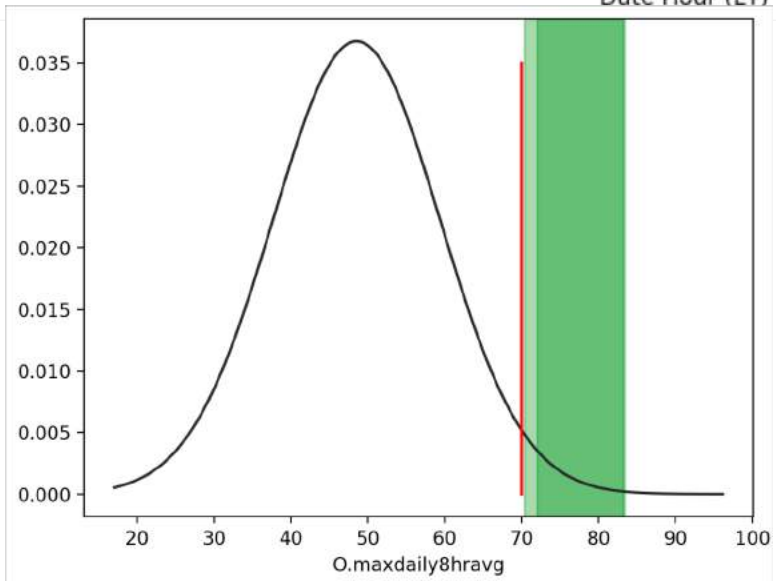


The hybrid approach leads to higher hourly O_3 prediction, but does not improve the daily maximum 8-hour O_3 prediction significantly.

3-day O3 forecasting by Random Forest and Multiple Linear Regression



Historical data distribution and forecasted daily maximum 8-hour O_3



Daily maximum 8-hour O_3

2018-07-25 74 ppb (90% above 70 ppb)

2018-07-26 77 ppb (90% above 70 ppb)

Summary

- WRF meteorology, time information and previous day's ozone concentrations are used to train machine learning models
  - None of the models forecast peak ozone concentrations correctly.
  - The Random Forest model give the best performance.
- Up-sampling and a data separation approach are used to improve the model predictions.
  - Up-sampling does not improve peak ozone predictions.
  - After data separation, LR predicts higher ozone concentrations than RF.
- Currently, the best approach is to use the random forest model for ozone levels less than a threshold (54 ppb, 90^{th} percentile) and a multiple linear regression model for ozone levels greater than the threshold.
- Next steps
  - Construct a website with time series and alarm for 3 day ozone forecasting.
  - Further development and testing of machine learning methods

Thank you!

Q&A