Machine Learning Applications for State Implementation Planning

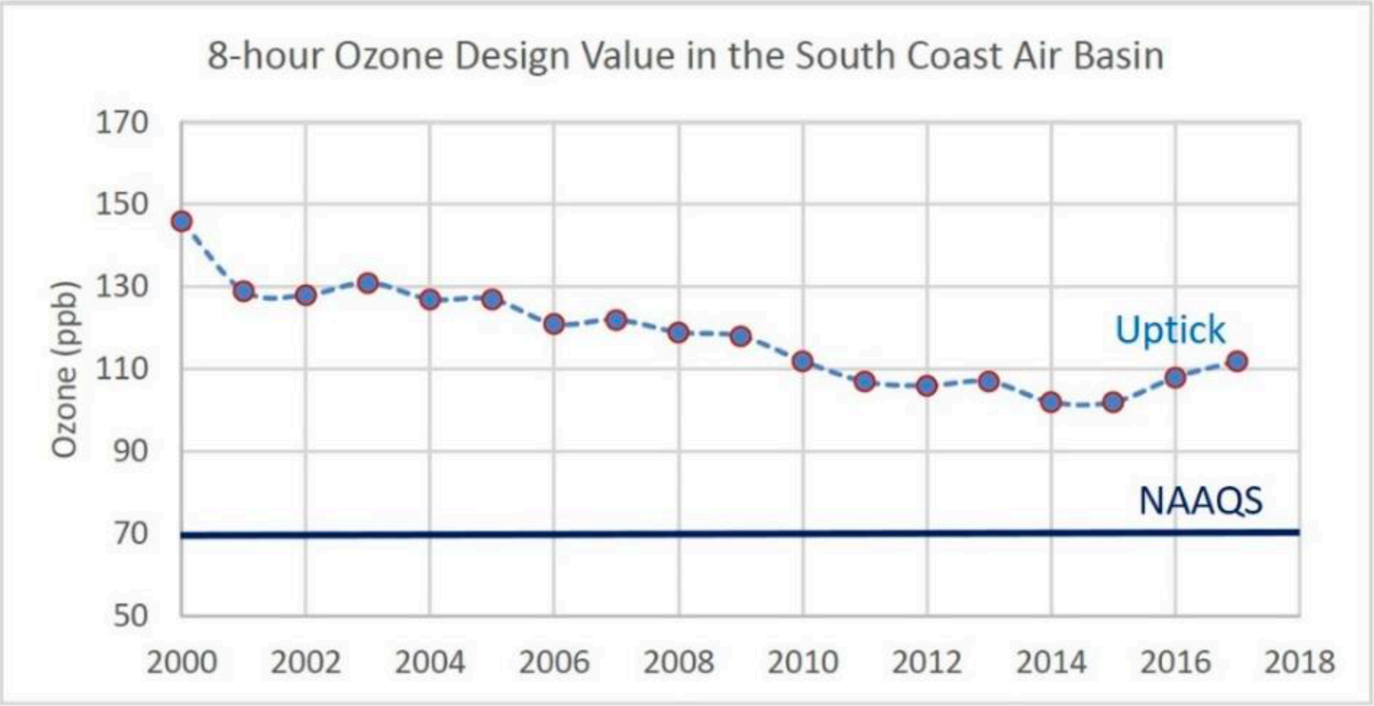
Khanh Do, Arash Kashfi Yeganeh, **Cesunica E. Ivey**

2020 CMAS Virtual Conference
October 26-30, 2020

Motivation: Recent Ozone Uptick in SoCAB

8-Hour ozone design value reached an inflection point in 2014

- 2017 = 112 ppb (Crestline)
- 2018 = 111 ppb (Crestline)
- 2019* = 108 ppb (Redlands)
- 2020* = 114 ppb (Redlands)



The federal 8-hour design value is based on the "three year average" of the annual 4th high

Motivation: Recent News

Unhealthy air pollution episodes strike Southern California despite reduced activity

L.A. records worst smog in 26 years

[**Smog,** from A1]
Compton's eight-hour reading was 115 ppb, its highest since monitoring began there in 2008.

Sunday's readings at the downtown L.A. air monitoring station, located on North Main Street in Chinatown, were so far above normal that they triggered a quality control check designed to prevent the release of erroneous data, air quality officials said.

The downtown L.A. readings did not initially appear online and were provided by the South Coast air district in response to questions from The Times about the missing data.

The figures were not reported immediately because the quality control check requires additional, manual validation if pollution readings exceed historic highs, South Coast AQMD spokeswoman Nahal Mogharabi said. If instruments are having problems, they "can show erroneously high levels and the quality control check prevents the automated release of high data that could be incorrect."

But it was no glitch.

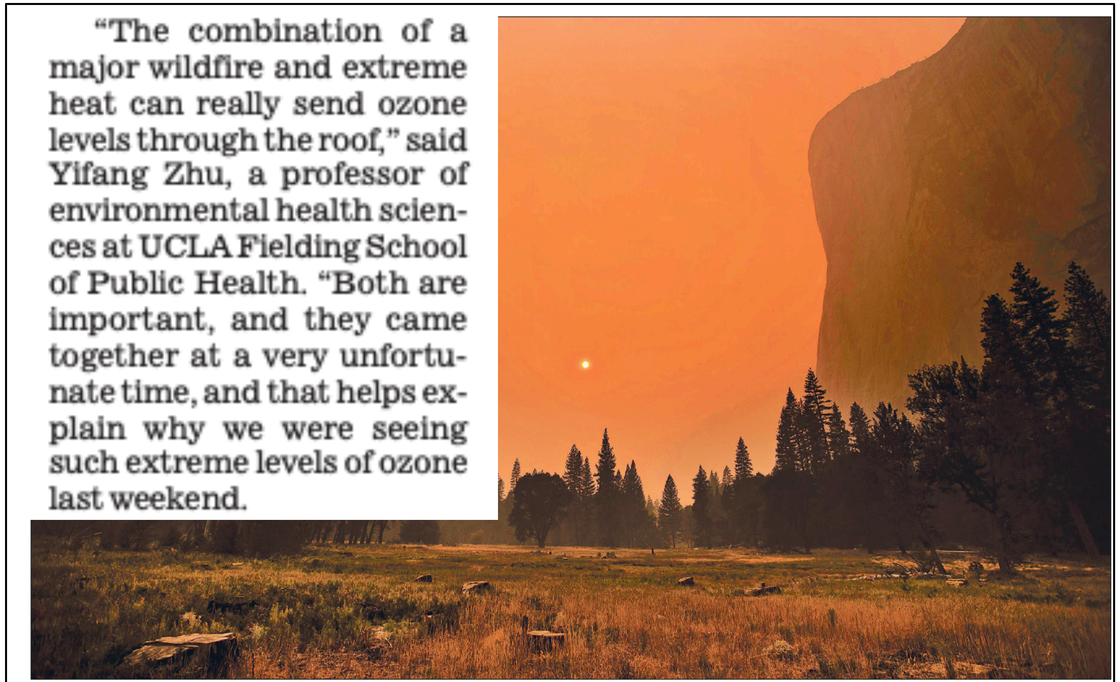
"The value for noon on Sunday has been reviewed and is preliminarily valid at 185 ppb," Mogharabi said.

Air quality officials said the high pollution readings were a result of intense heat combined with stagnant

ROBERT GAUTHIER Los Angeles Times

LA Times, September 11, 2020

"The combination of a major wildfire and extreme heat can really send ozone levels through the roof," said Yifang Zhu, a professor of environmental health sciences at UCLA Fielding School of Public Health. "Both are important, and they came together at a very unfortunate time, and that helps explain why we were seeing such extreme levels of ozone last weekend.



BRIAN VAN DER BRUG Los Angeles Times

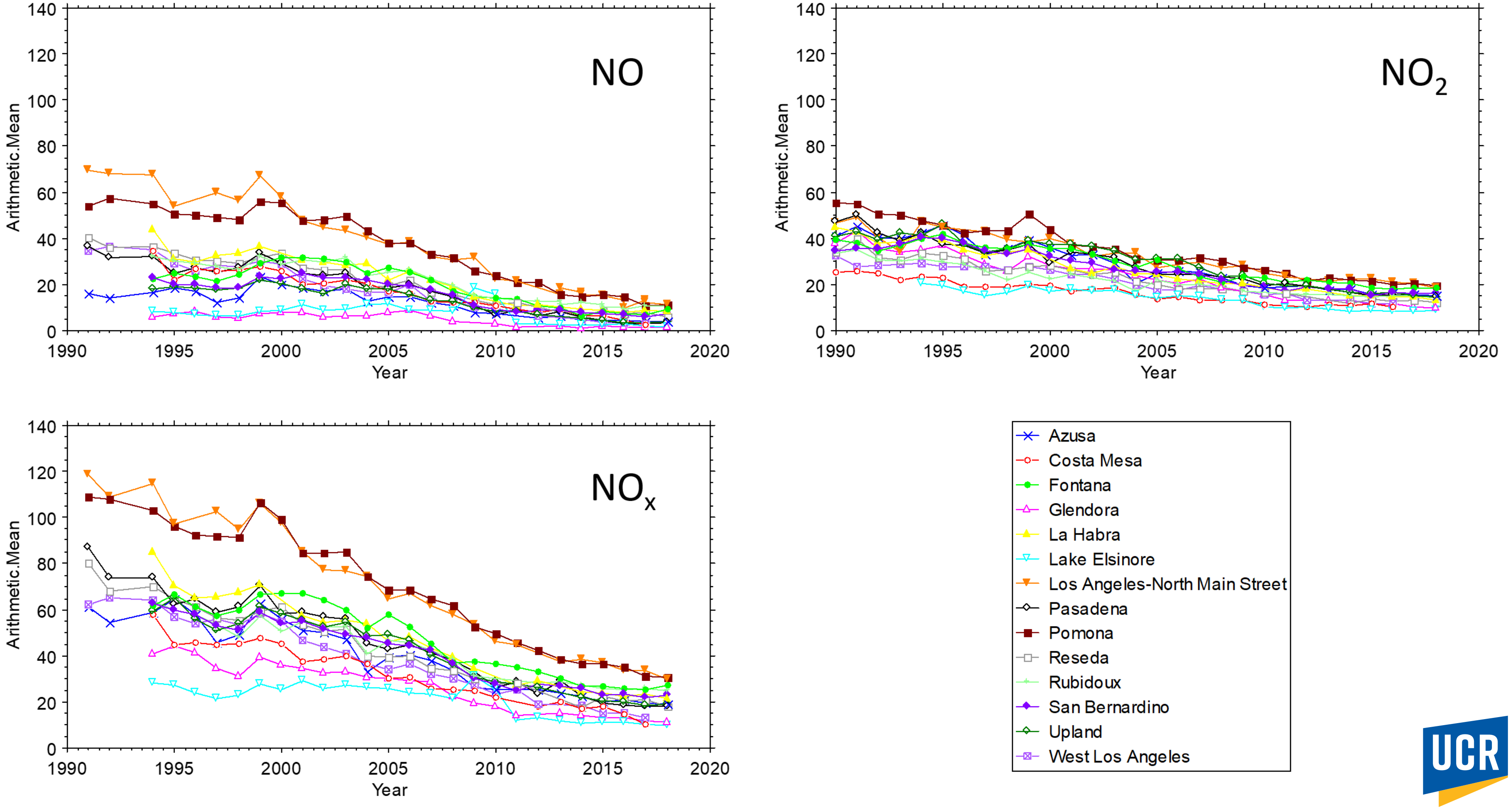
THICK SMOKE from multiple forest fires Saturday shrouds the iconic El Capitan rock formation and granite walls of the Yosemite Valley at Yosemite National Park.

Climate apocalypse has struck California

LA Times, September 13, 2020

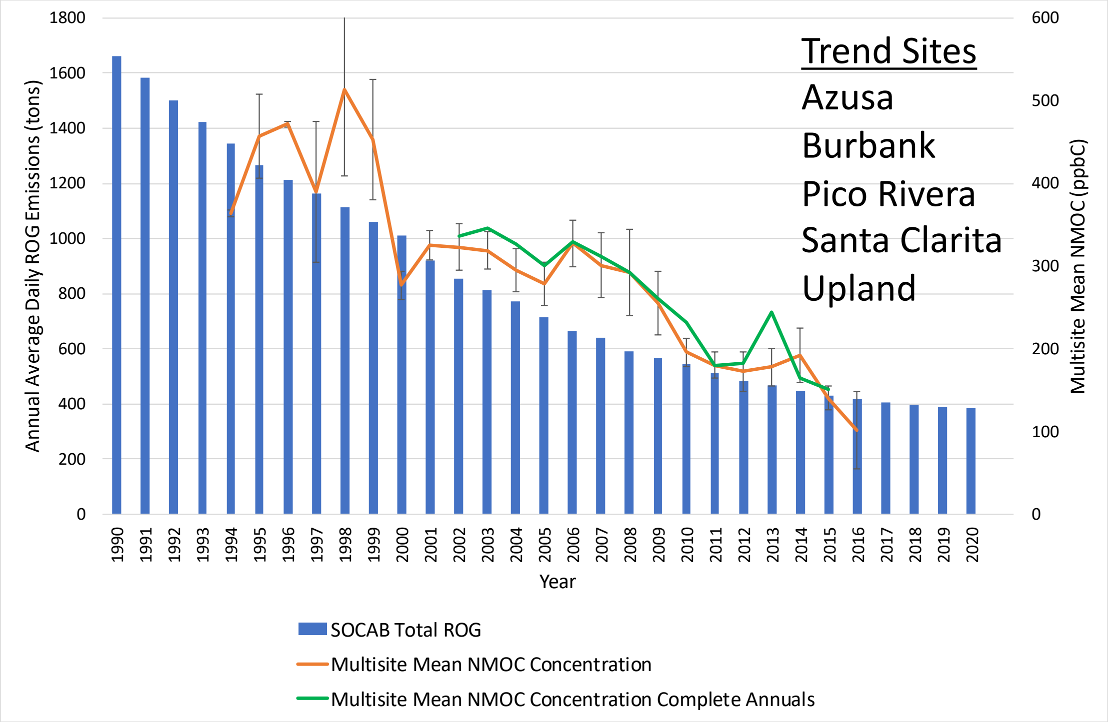
Decreasing Levels of Nitrogen Oxides

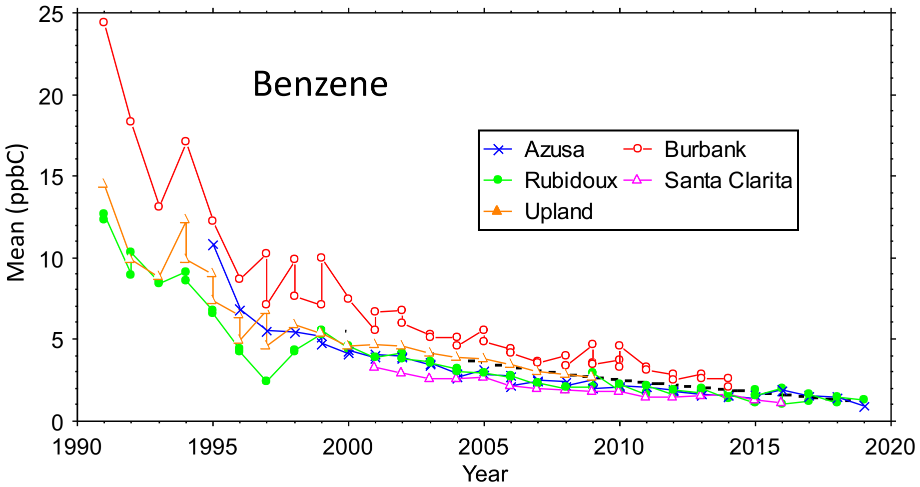
Declining Mean Annual NO, NO_2, and NO_x at All Trend Sites (N > 25 Years)

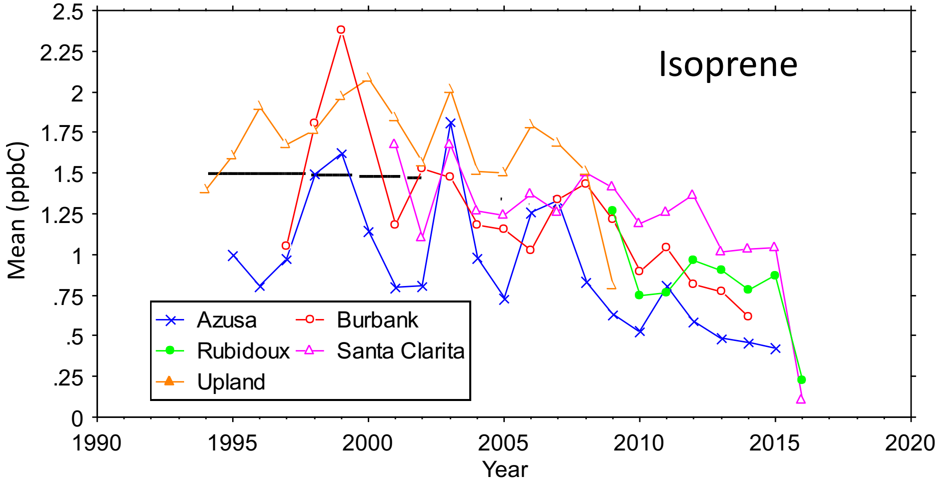


Decreasing Levels of VOCs

Average ambient NMOC tracked ROG emissions



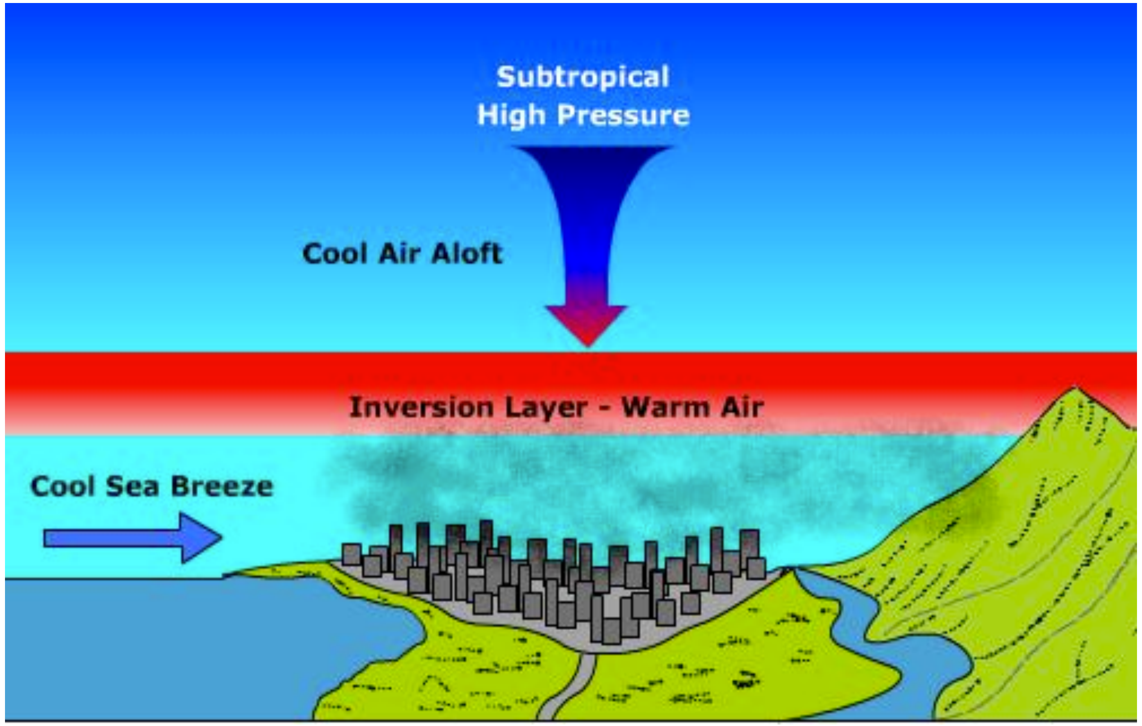


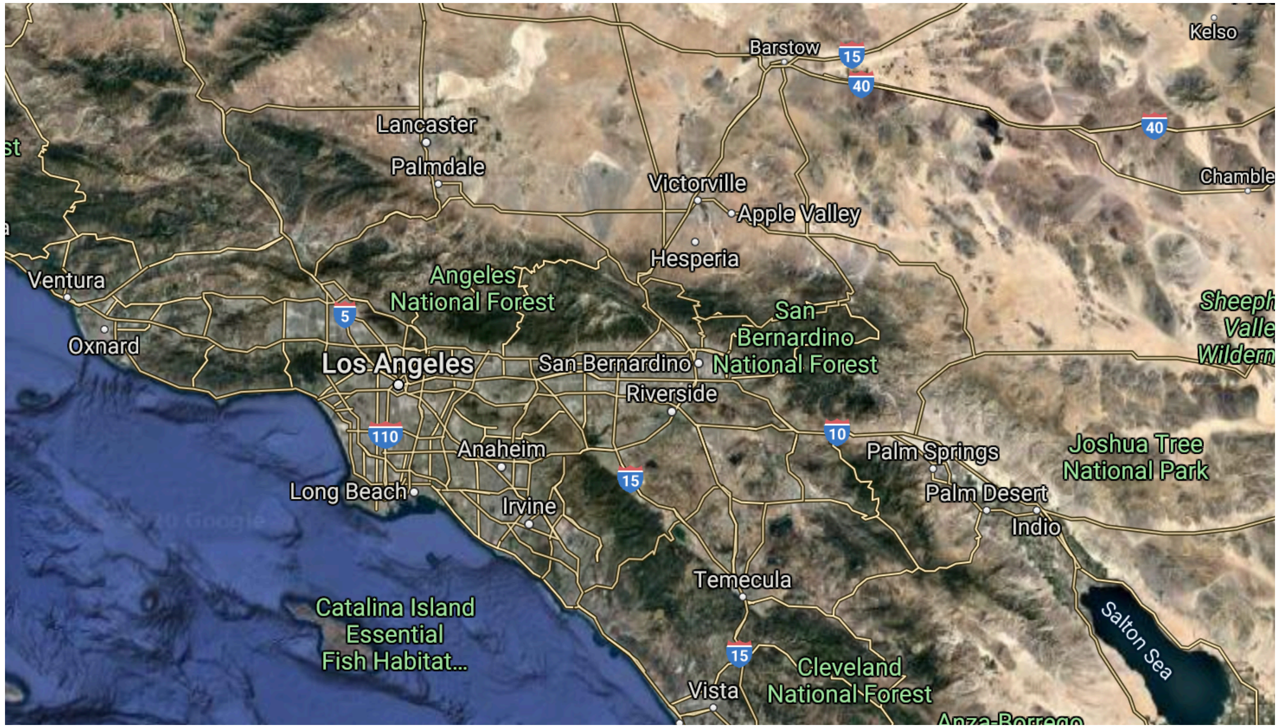


Implication: ambient data corroborate emission inventory trends but lack of data post-2015 hinder interpetation of 2016-2018 period

South Coast Air Basin

Unique meteorology, emissions, chemistry, and topography

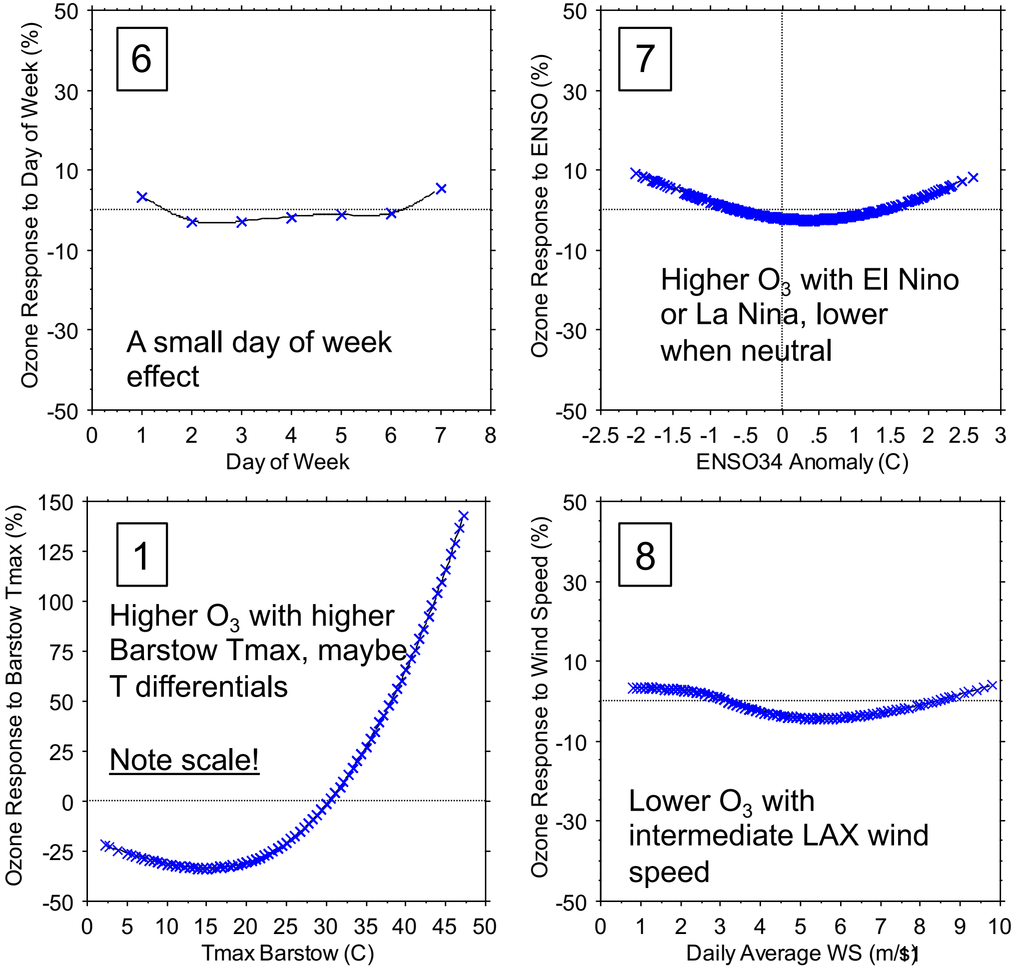




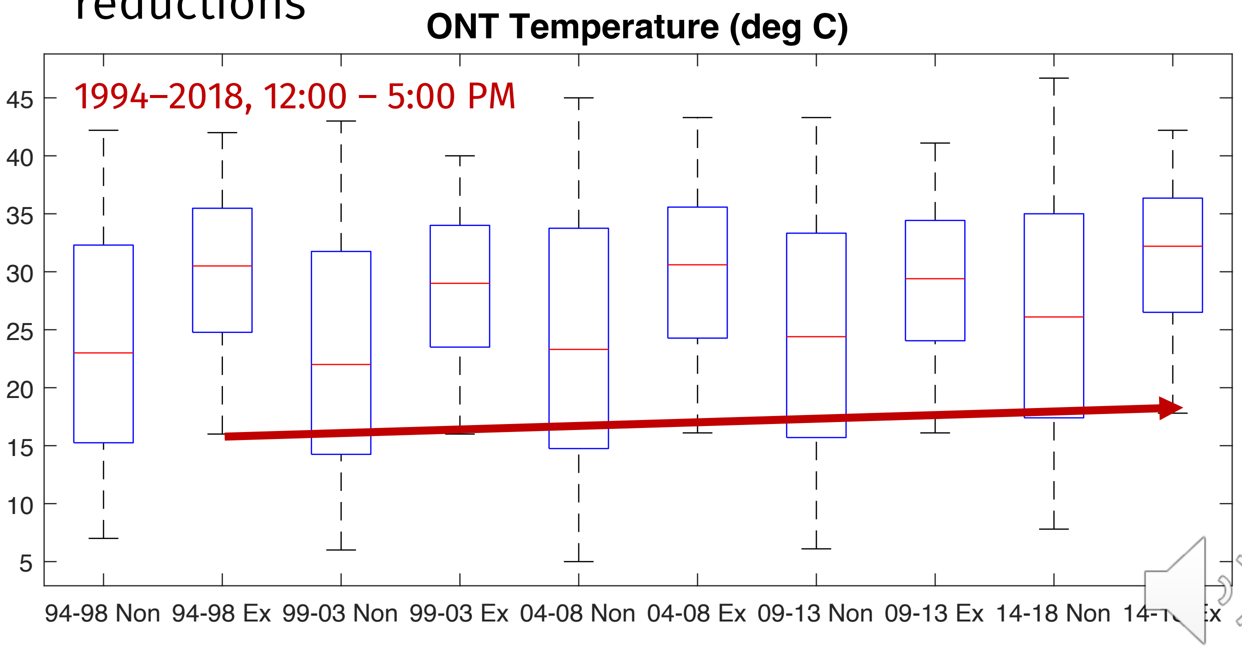
On-shore winds prevail, transporting precursor emissions from west to east in the Basin. High pressure systems induce temperature inversions, with warmer air atop a cooler surface layer.

Ozone Response to Meteorology

GAM Finds Key Weather Influences on Crestline O3



- High ozone concentrations are associated with high temperatures
- Positive trend in temperatures over the last 25 years
- Expected increases in ozone design value due to climate change, in the absence of major precursor emission reductions

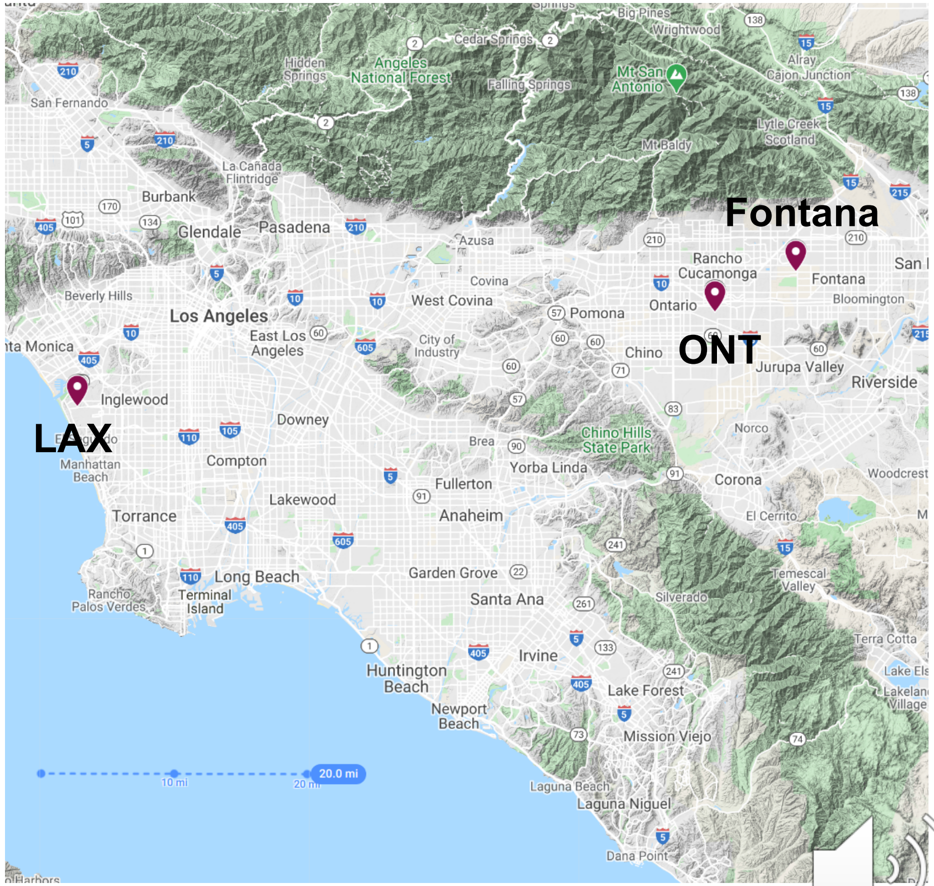


Ivey, Russell, and Blanchard, Ozone Meteorology Study: Quarter 2 Report, 2020

Machine Learning Methods

Building a data-driven, multi-feature framework for environmental predictions

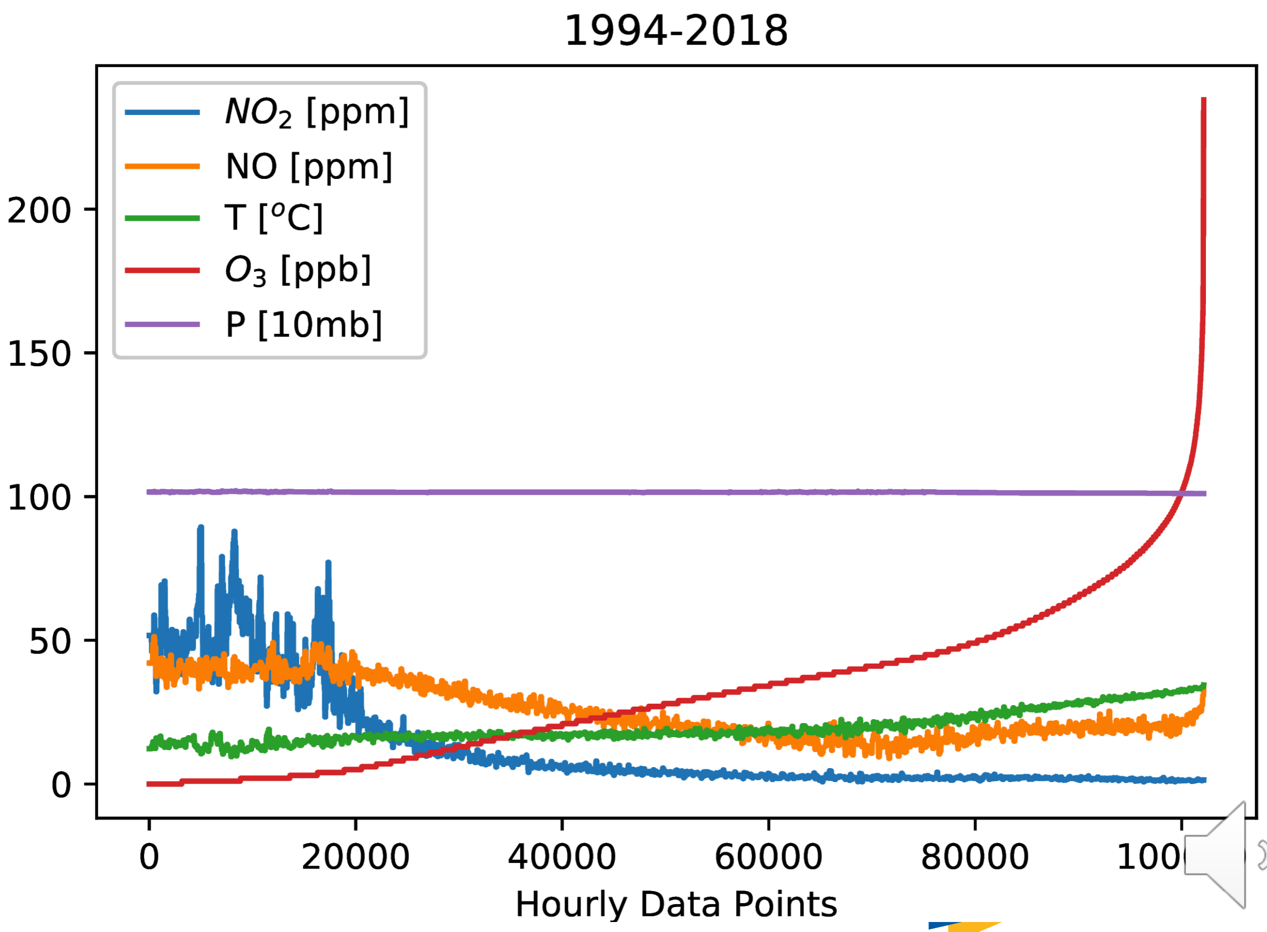
- Analysis period 1994 to 2018, 12:00–5:00 PM (5 hours)
- Surface meteorological data are obtained for LAX and ONT International Airports
  - Temperature, relative humidity, pressure, wind speed, wind direction, visibility, dew point temperature
- Fontana air quality data are retrieved from Air Quality System (AQS) data mart
  - NO, NO_2, O_3
- 80% for training and 20% for model testing and validation



Time-Independent Trends

Historical air quality trends from 1994-2018 were sorted based on the increment of ozone

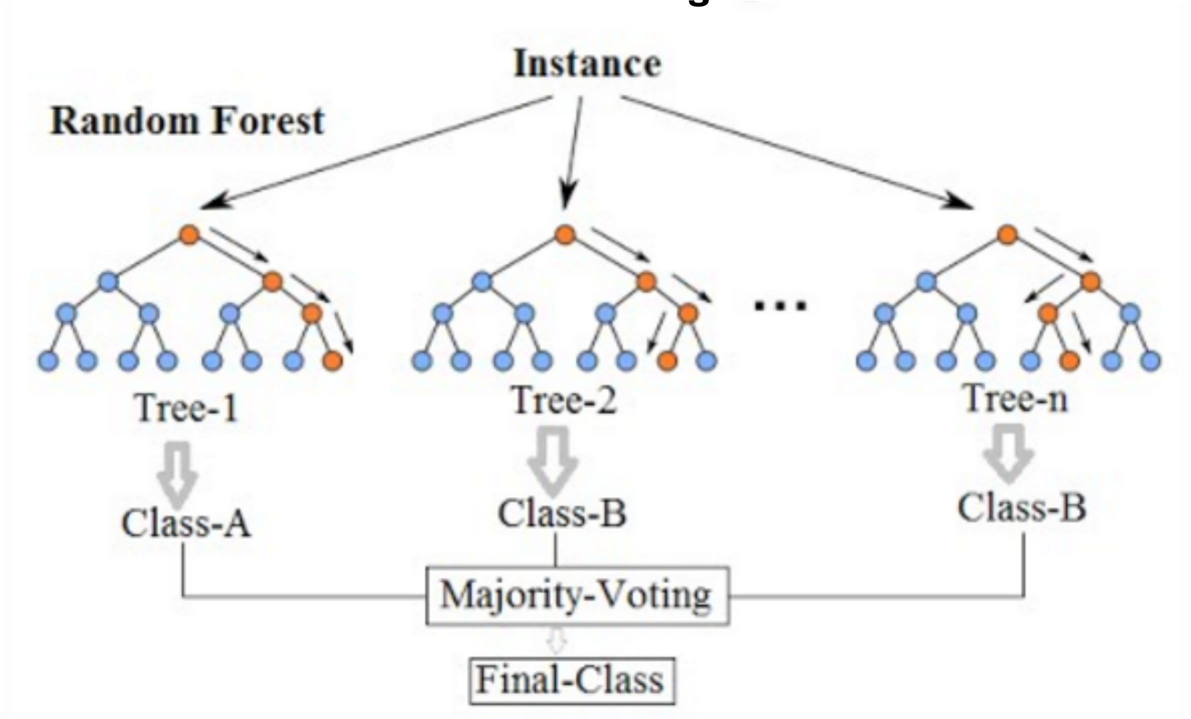
- 12:00–5:00 PM meteorology data sorted by time-independent, ascending ozone concentrations
  - Clear inverse relationship with NO_2
  - Clear direct relationship with temperature
  - No relationship with surface pressure
- Interesting NO relationship
  - Inflection point for NO
  - NO minimum at ~10 ppb ozone
  - Slightly higher NO concentrations, low NO_2, and high temperatures signify special regime



Machine Learning Algorithms

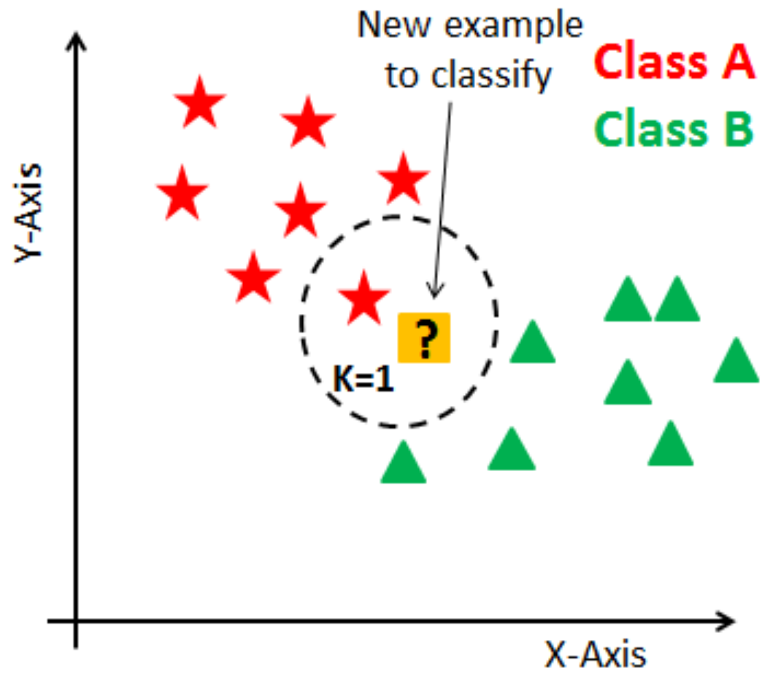
Which algorithm builds the most appropriate model for our system?

Random Forest Regression



https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

K-Nearest Neighbors



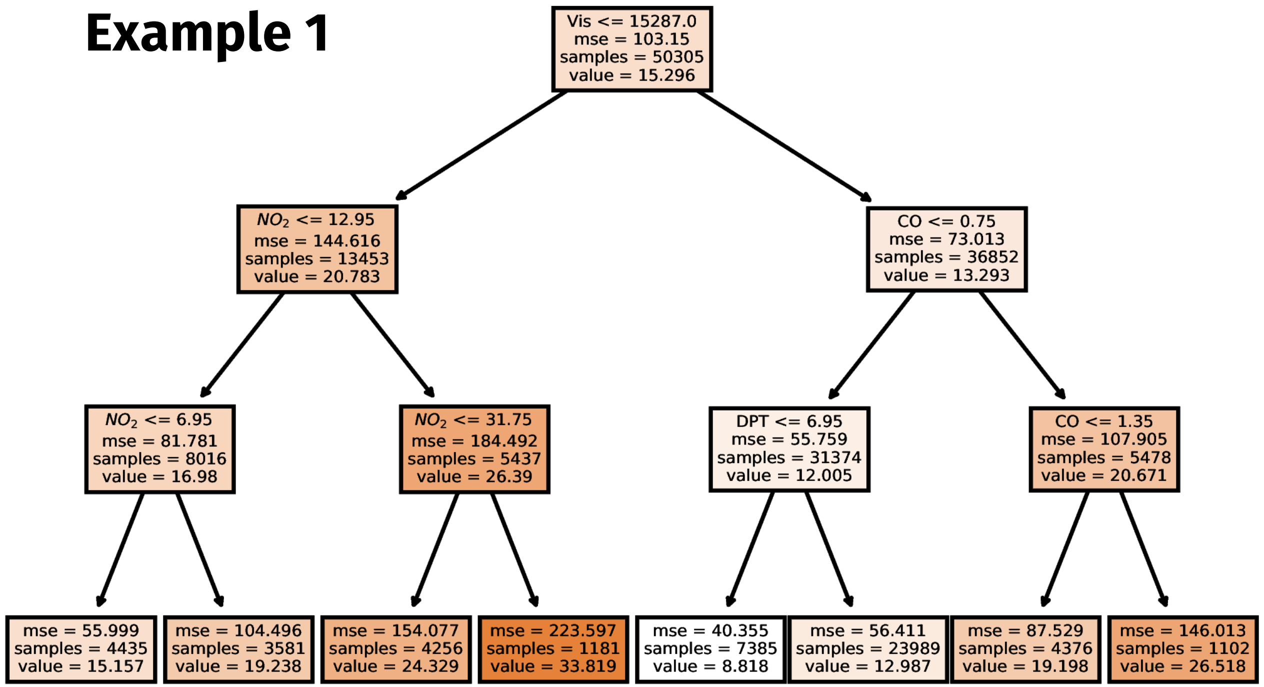
https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

RFR uses decision trees that are constructed by placing the most informative features at the root of the tree. K-NN selects the class for which the predicted value has the most features in common

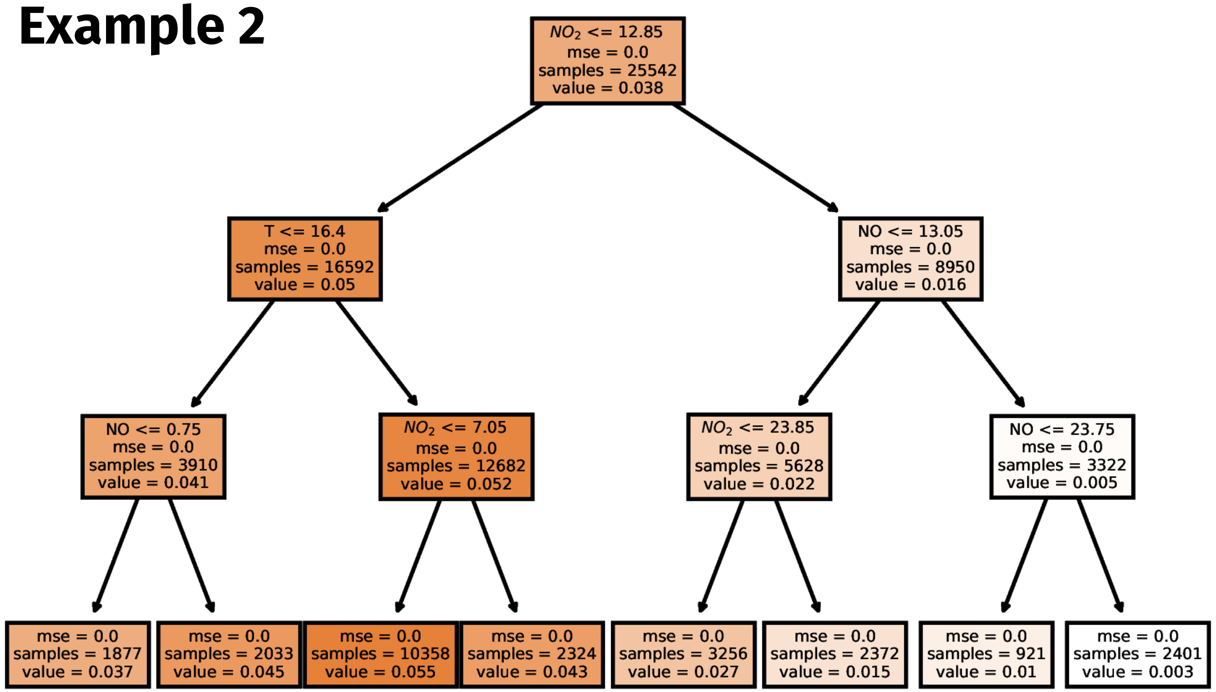
Machine Learning Algorithms

Example of three-branch RFR trees

Example 1



Example 2



RFR uses decision trees that are constructed by placing the most informative features at the root of the tree. K-NN selects the class for which the predicted value has the most features in common

Random Forest: Ozone Prediction

RFR also has satisfactory performance

RFR shows good performance

- Correlational factor > 0.86
- Mean error < 7%
- Underestimated: 1994-1998,2004-2008, 2009-2013
- Overestimated: 1999-2003, 2014-2018

Hyperparameters:
n_estimators = 16
max_features = auto
max_depth = none
min_samples_split = 2
min_samples_leaf = 1
min_weight_fraction_leaf = 0
max_leaf_nodes = none
n_jobs = 8

Year	CC	MB	NMB	FB	FAE	M	O
1994-1998	0.88	-0.004	-0.069	-0.039	0.234	0.055	0.059
1999-2003	0.882	0	0.002	0.028	0.235	0.048	0.048
2004-2008	0.884	-0.002	-0.041	-0.025	0.182	0.052	0.054
2009-2013	0.884	-0.002	-0.029	-0.024	0.15	0.055	0.057
2014-2018	0.869	0	0.003	0.017	0.143	0.058	0.058

Random Forest: Feature Sensitivity

Wilcoxon rank sum tests determine significance of input features

Actual one variable-LAX					
p-values	94-98	99-03	04-08	09-13	14-18
no drop	0.36	0.12	0.00	0.00	0.37
drop visibility	0.44	0.33	0.00	0.00	0.91
drop no2	0.92	0.24	0.00	0.00	0.82
drop RH	0.27	0.16	0.00	0.00	0.48
drop windSp	0.07	0.14	0.00	0.00	0.59
drop windDir	0.38	0.44	0.00	0.00	0.66
drop dewT	0.59	0.14	0.00	0.00	0.35
drop no	0.00	0.81	0.42	0.00	0.36
drop temperature	0.45	0.10	0.00	0.00	0.44
drop pressure	0.90	0.14	0.00	0.00	0.16

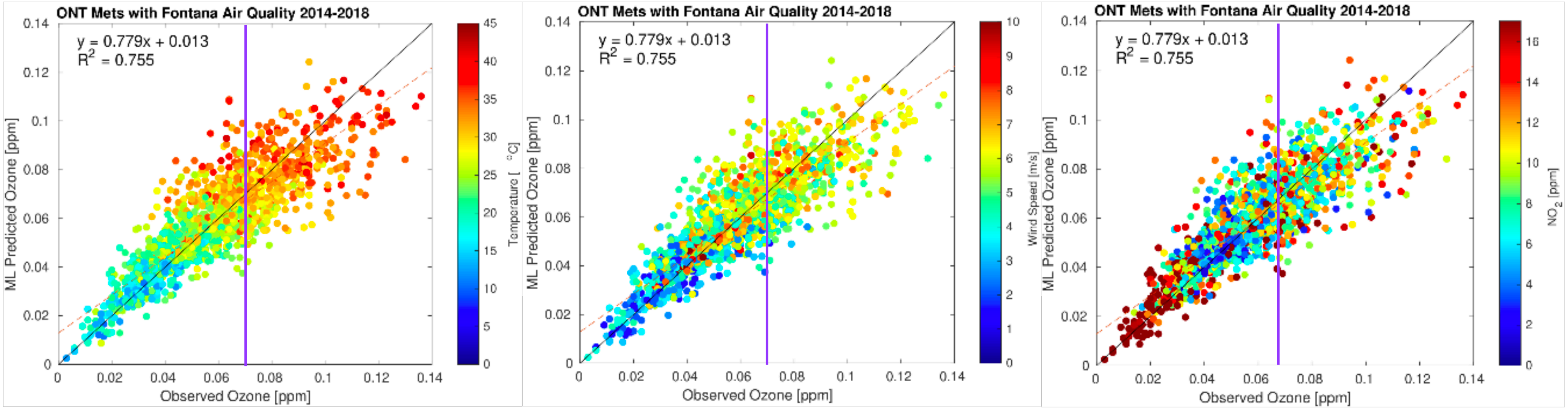
Actual one variable-ONT					
p-values	94-98	99-03	04-08	09-13	14-18
no drop	0.17	0.30	0.10	0.10	0.25
drop visibility	0.01	0.75	0.02	0.07	0.29
drop no2	0.13	0.14	0.00	0.00	0.73
drop RH	0.17	0.32	0.09	0.12	0.20
drop windSp	0.06	0.92	0.00	0.06	0.57
drop windDir	0.17	0.26	0.05	0.07	0.28
drop dewT	0.17	0.28	0.09	0.20	0.20
drop no	0.00	0.49	0.11	0.11	0.09
drop temperature	0.19	0.31	0.08	0.16	0.17
drop pressure	0.17	0.30	0.08	0.16	0.27

RFR uses decision trees that are constructed by placing the most informative features at the root of the tree. K-NN selects the class for which the predicted value has the most features in common

Random Forest: Meteorology Trends

In a test of ozone exceedance prediction, correlation coefficient was ≥ 0.86 for all time periods

Ozone predictions with ONT meteorology



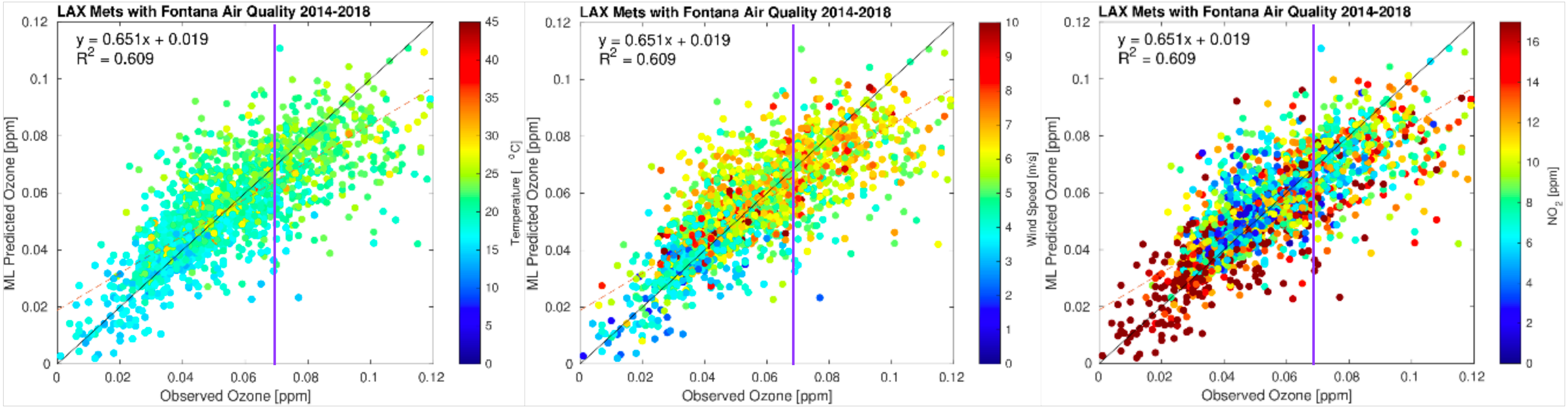
Temperature (0-45 ℃) | Wind Speed (0-10 m/s) | NO_2 (0-17 ppb)

During peak ozone hours (12:00 – 5:00 PM), high ozone is associated with high temperatures, moderate wind speeds, and lower NO_2

Random Forest: Meteorology Trends

A higher degree of scatter is observed using upwind meteorological predictors

Ozone predictions with LAX meteorology



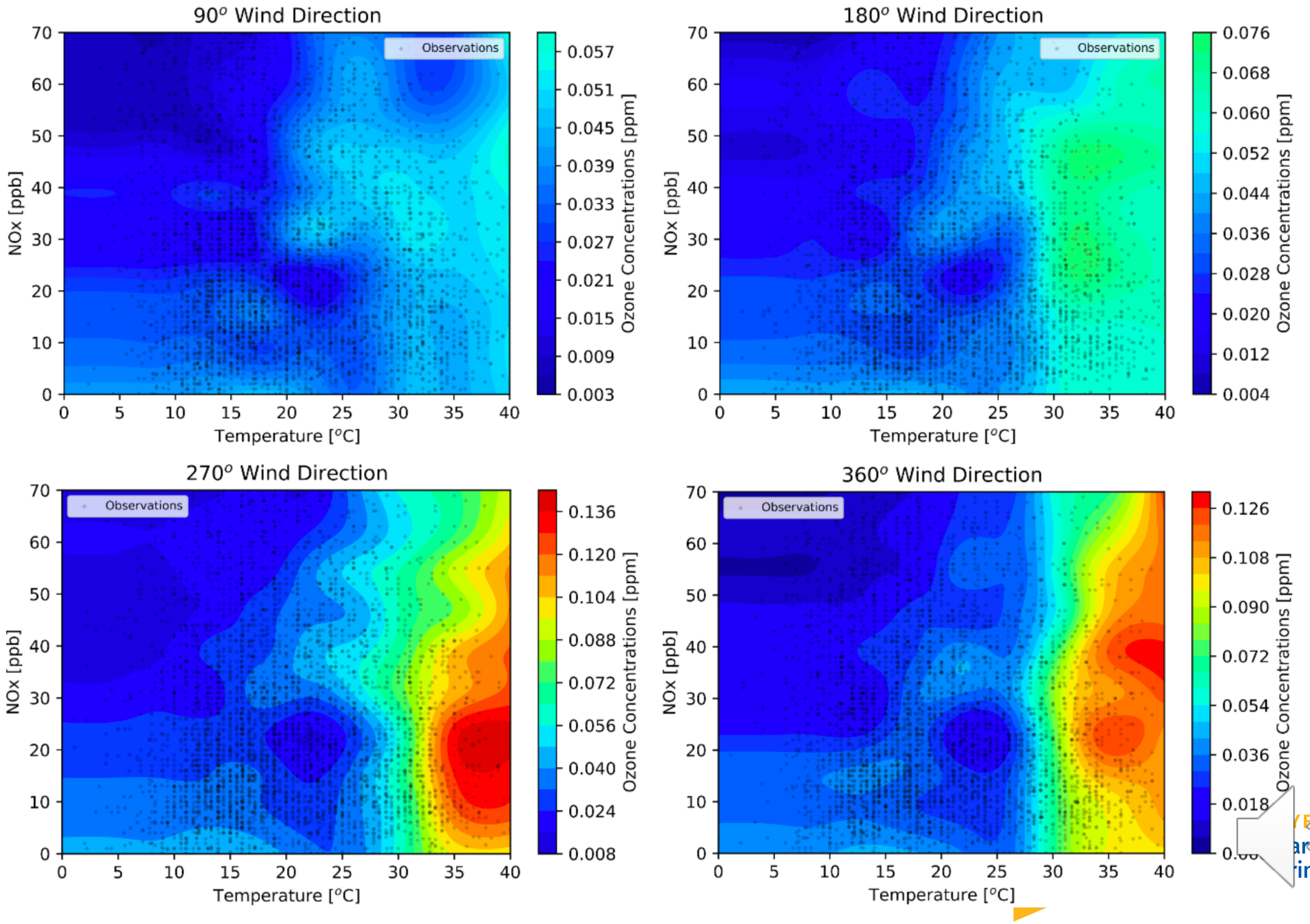
Temperature (0-45 ℃) Wind Speed (0-10 m/s) NO_2 (0-17 ppb)

During peak ozone hours (12:00 – 5:00 PM), high ozone is associated with high temperatures, moderate wind speeds, and lower NO_2

Random Forest: Meteorology Trends

Developing heat maps relates meteorological variables at various values

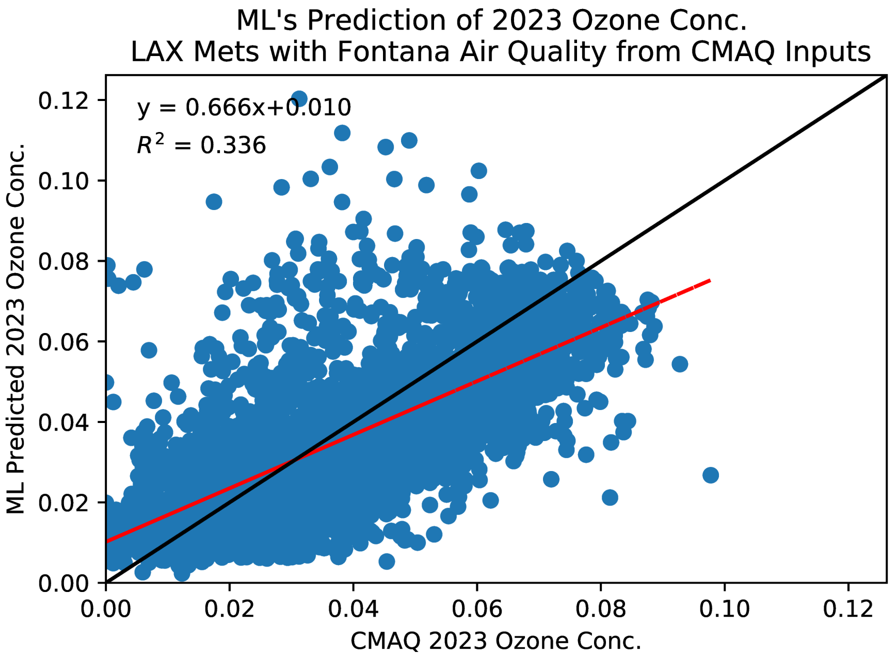
- Contour plots generated by random forest regression model
- Trained on ONT meteorology and Fontana air quality
- Constant:
  - wind speed (8 m/s)
  - visibility (16000 m)
- Adaptive:
  - pressure
  - humidity
- Four discrete wind direction levels (90, 180, 270, 360).

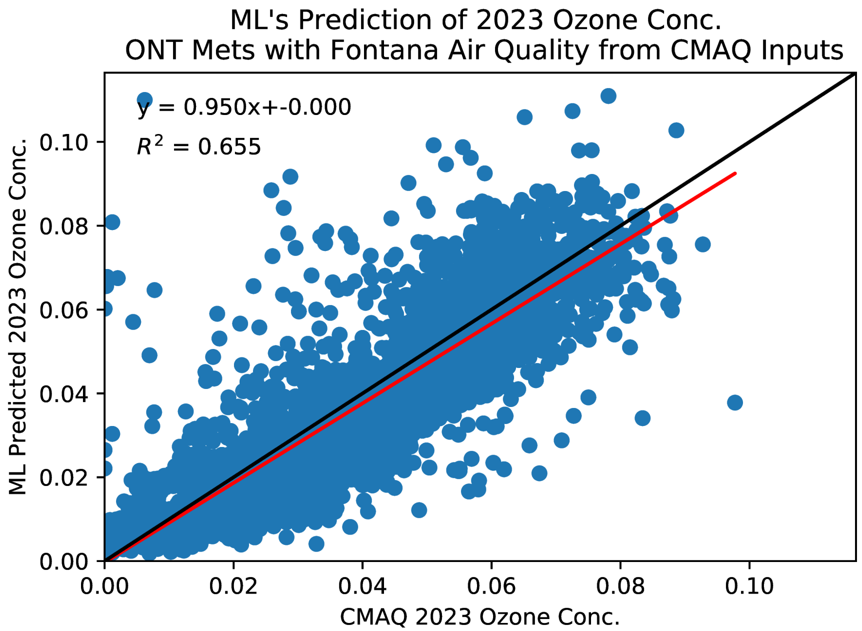


Random Forest: 2023 Predictions

Developing heat maps relates meteorological variables at various values

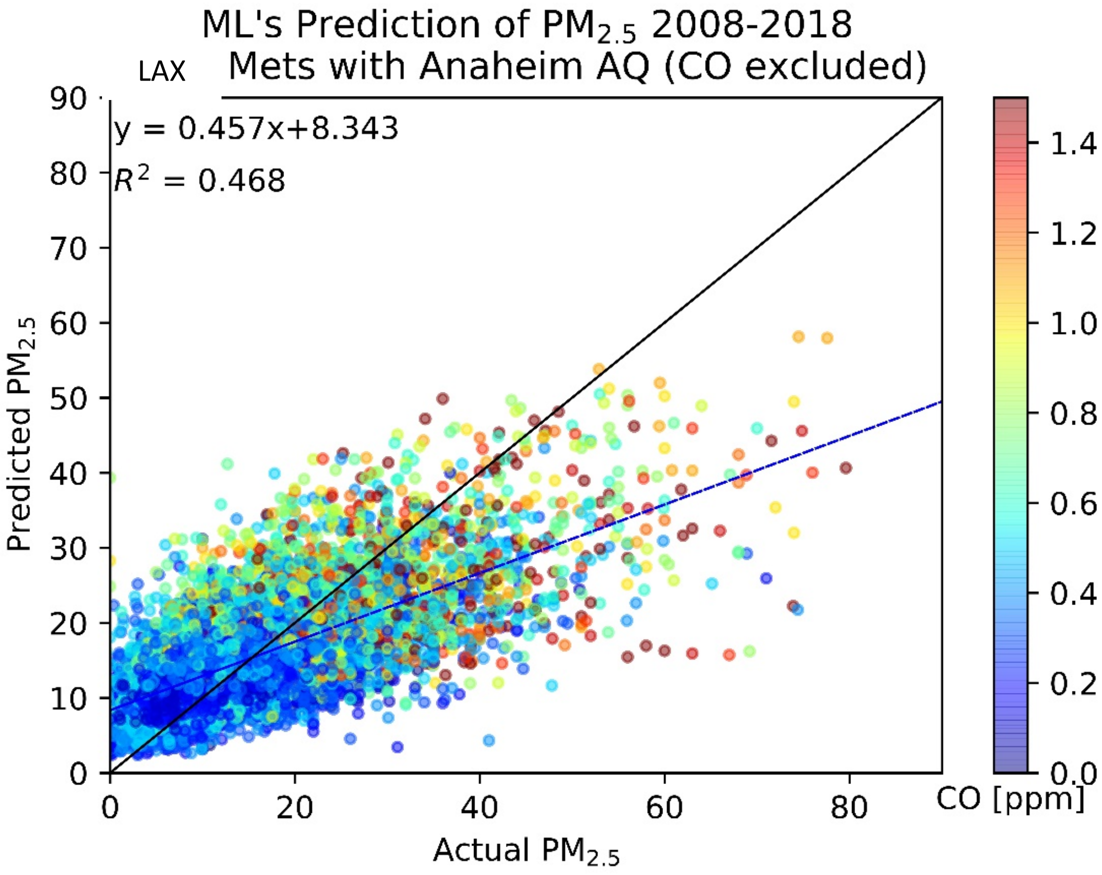
- RFR has good correlation with CMAQ simulation
- R^2 (0.66), intercept is 0, and slope of 0.95
- RFR model gives good prediction when trained with ONT meteorology to predict Fontana ozone
- Ontario has similar meteorology to Fontana

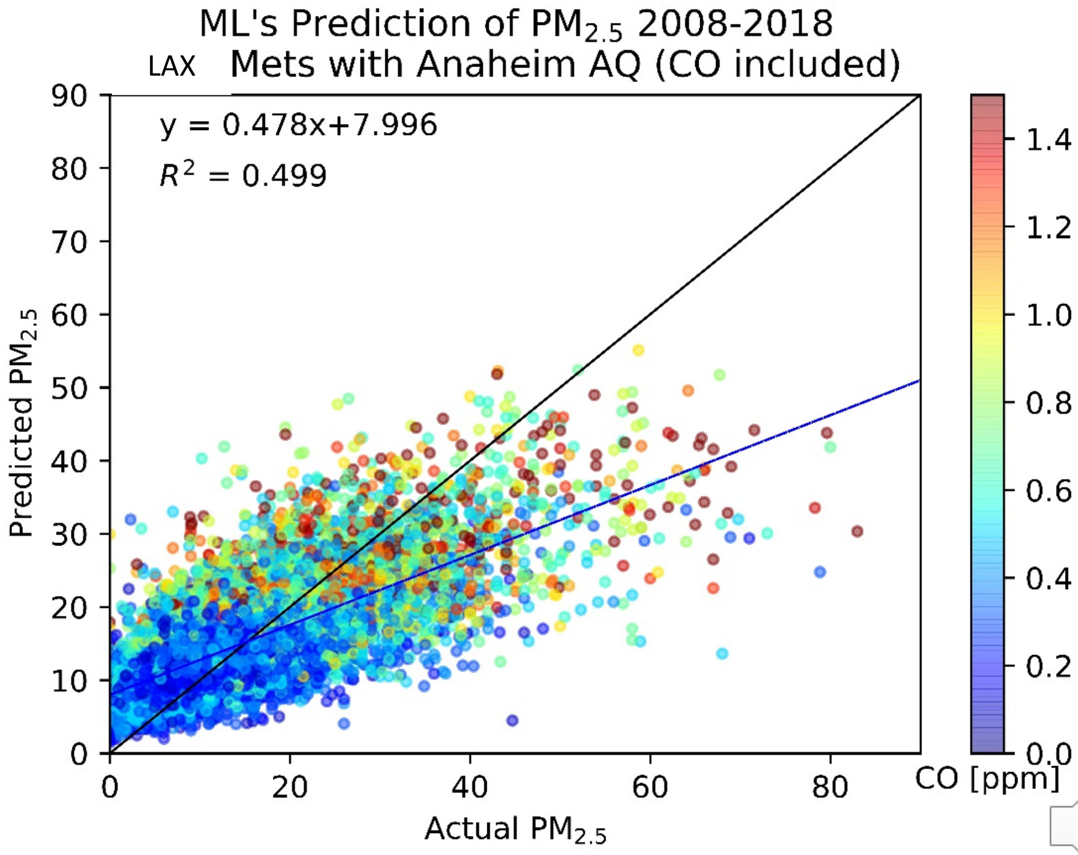




$PM_{2.5}$ Prediction

Meteorology inputs are not as effective for PM predictions





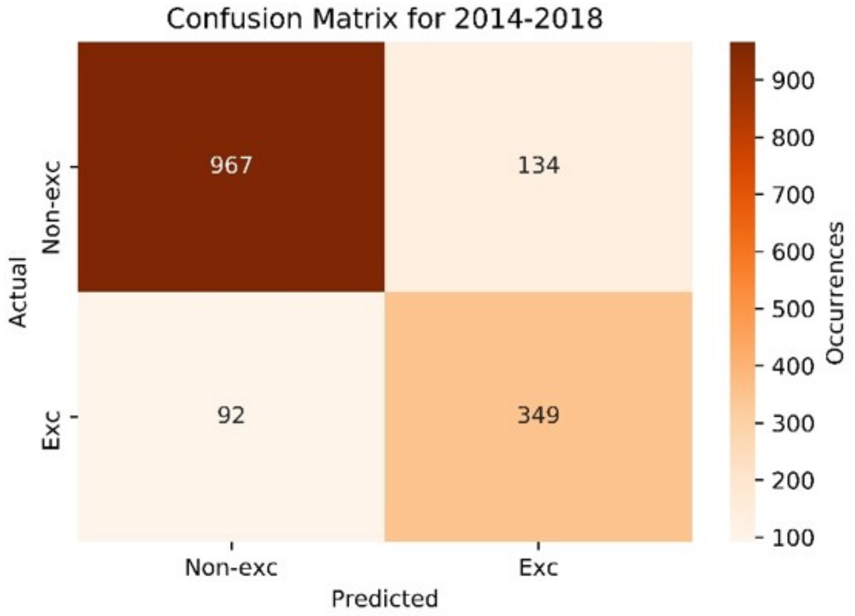
K-Nearest Neighbor: Ozone Prediction

KNN gives highest accuracy for binary classification

- Good probability of detection (PoD) in the later years
- Accuracy ranges from 84% to 87%
- High scores on confusion matrix

Performance for Ozone > 70 ppb

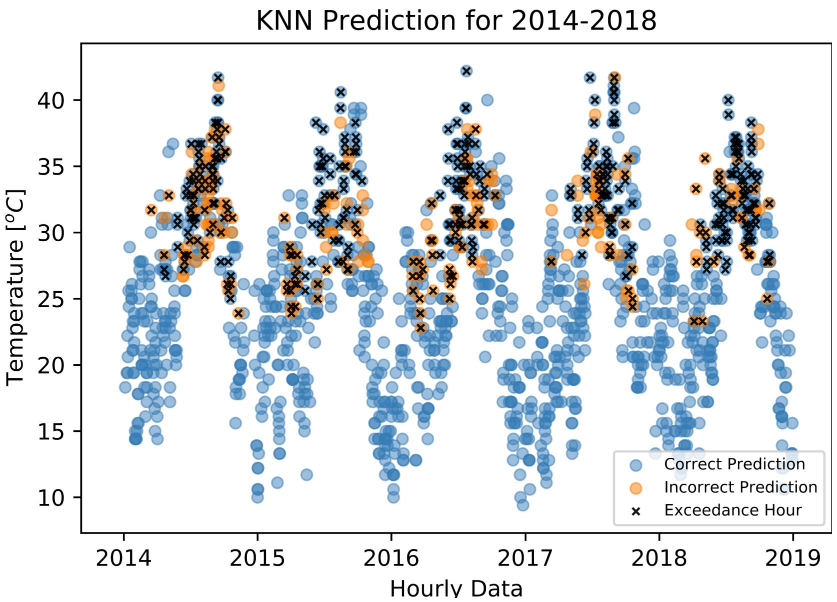
Year	Probability of Detection	Accuracy	Failure to Predict
1994-1998	0.58	0.84	0.42
1999-2003	0.69	0.87	0.31
2004-2008	0.69	0.86	0.31
2009-2013	0.74	0.87	0.26
2014-2018	0.79	0.85	0.21



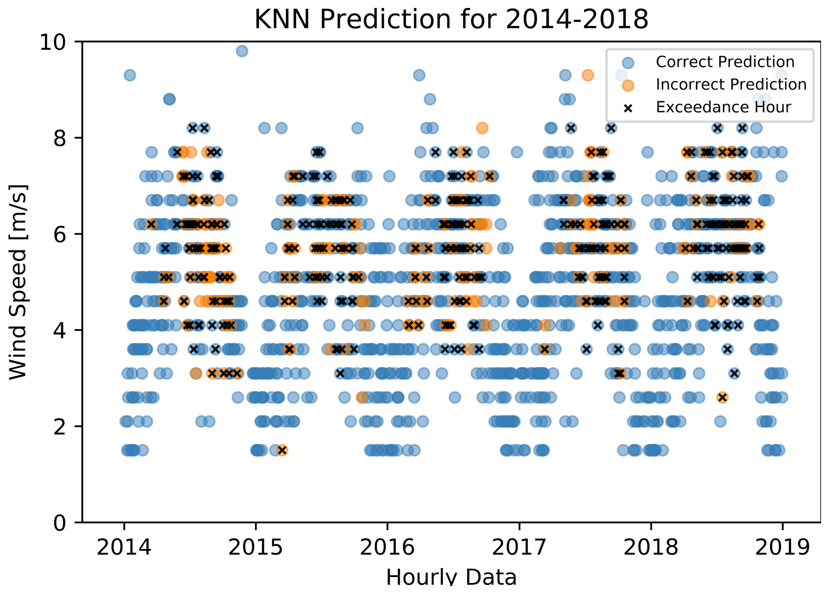
K-Nearest Neighbor: Meteorology Trends

Exceedance prediction accuracy was 85% based on a binary classification and 70 ppb threshold (N = 64)

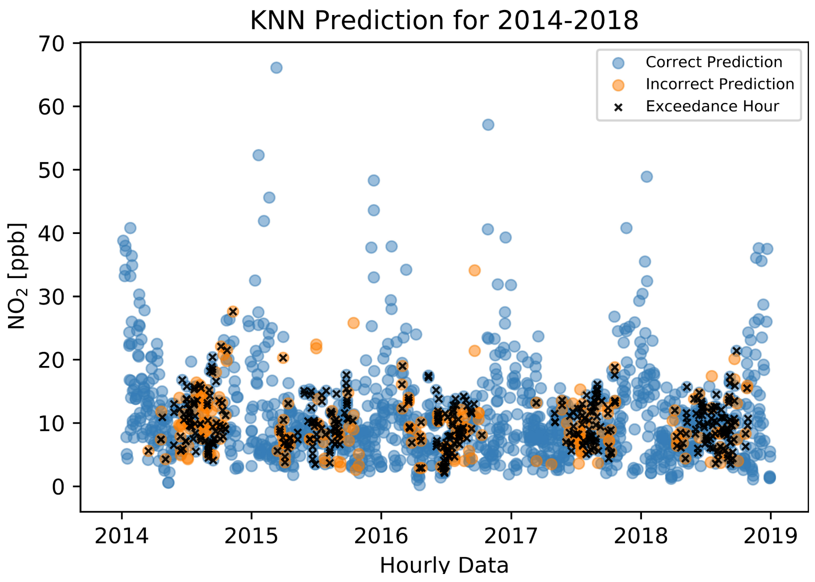
Ozone predictions with ONT meteorology



Temperature (0-45 °C)



Wind Speed (0-10 m/s)



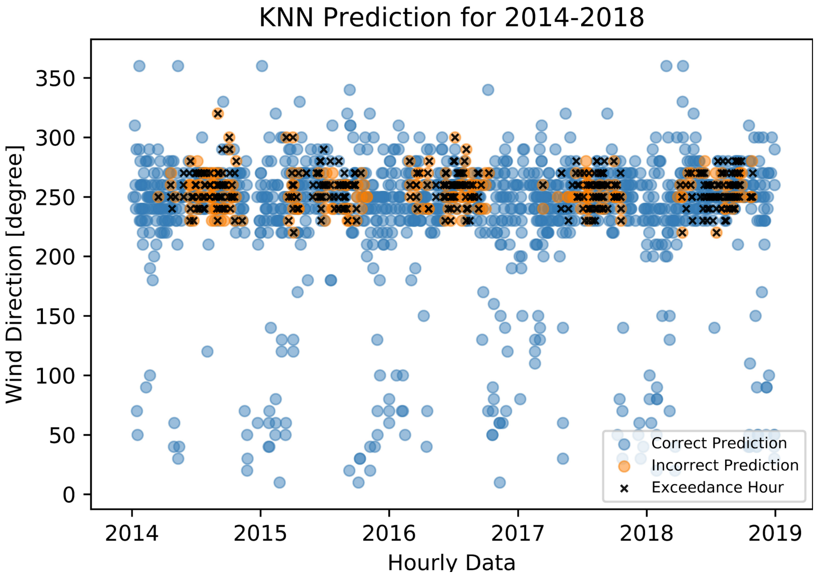
NO_2 (0-70 ppb)

Examination of temporal prediction patterns reveals a tendency to miss classify ozone exceedances compared to non-exceedances

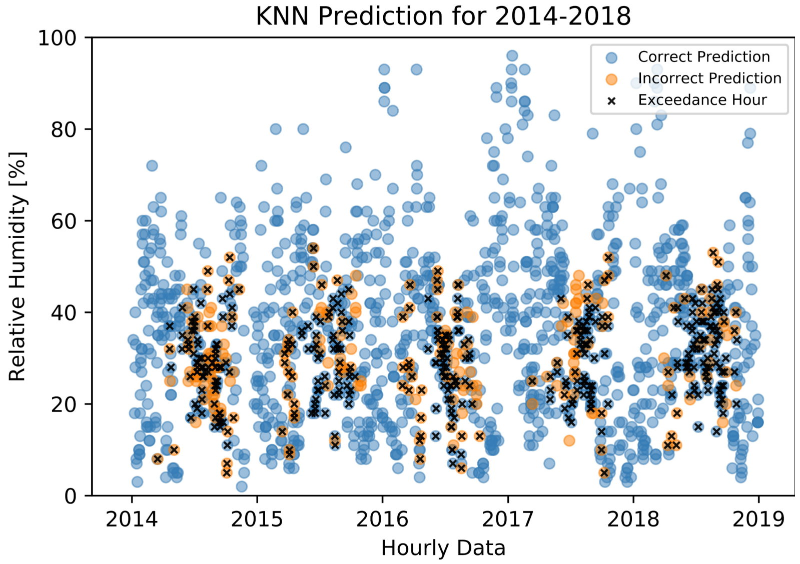
K-Nearest Neighbor: Meteorology Trends

Exceedance prediction accuracy was 85% based on a binary classification and 70 ppb threshold (N = 64)

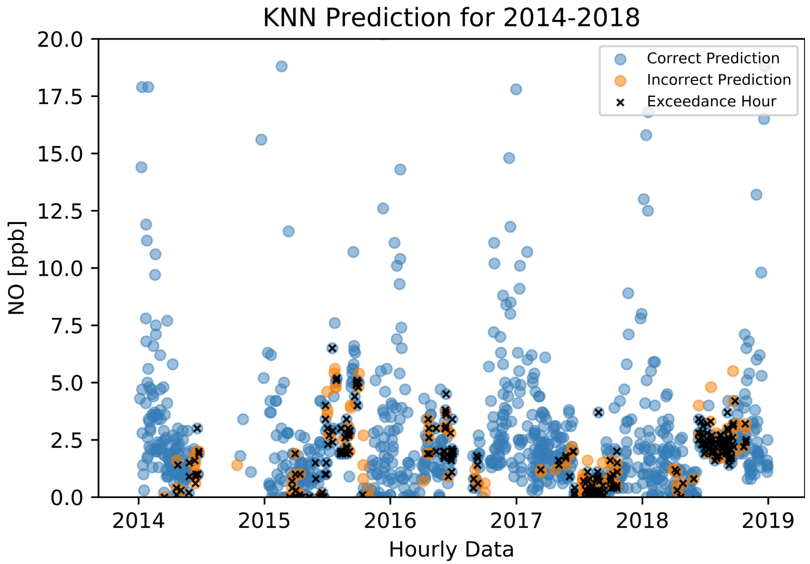
Ozone predictions with ONT meteorology



Wind Direction (0-360 °C)



Relative Humidity (0-100 %)



NO (0-20 ppb)

Examination of temporal prediction patterns reveals a tendency to miss classify ozone exceedances compared to non-exceedances

Summary

Takeaway messages

- Rising temperatures and more frequent heat waves will lead to higher ozone, especially in the absence of major precursor emission reductions
- There are ongoing efforts to test the sensitivity of algorithm/model performance with respect to hyperparameter choice
- Machine learning models will be evaluated for policy relevant performance (e.g., 8-hour ozone, 24-hour $PM_{2.5}$)

Acknowledgments

Sponsors and Partners

Funding
South Coast Air Quality Management District

Study Collaborators
Ted Russell, Charles Blanchard, Ziqi Gao

Contact Information
Cesunica Ivey, Ph.D.
Assistant Professor of Chem. and Env. Engineering
cesunica@ucr.edu