A Machine Learning Approach for Ozone Forecasting in Pacific Northwest

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.

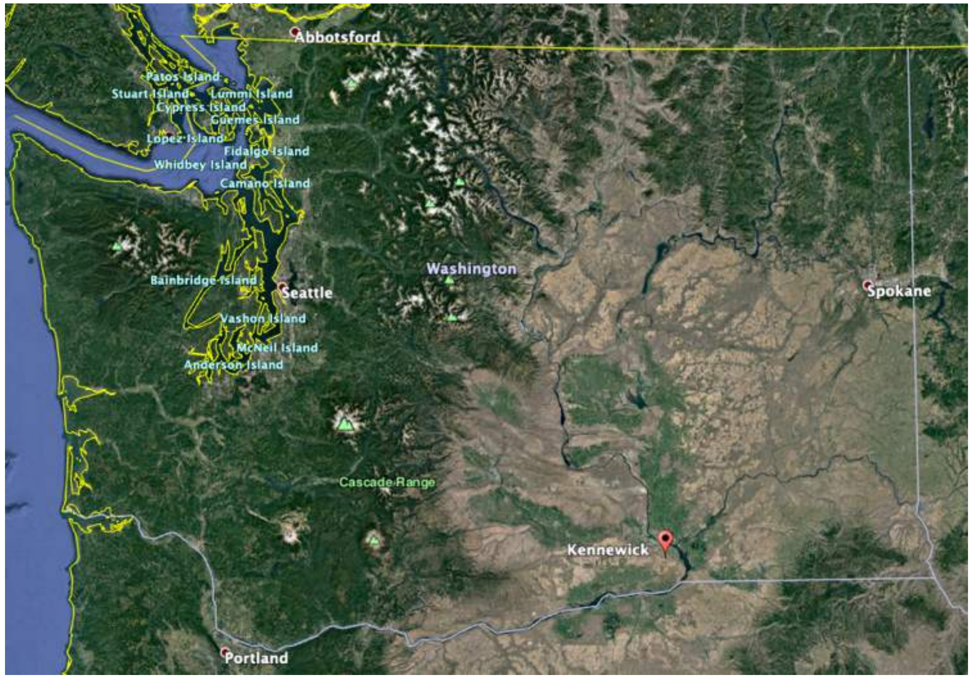
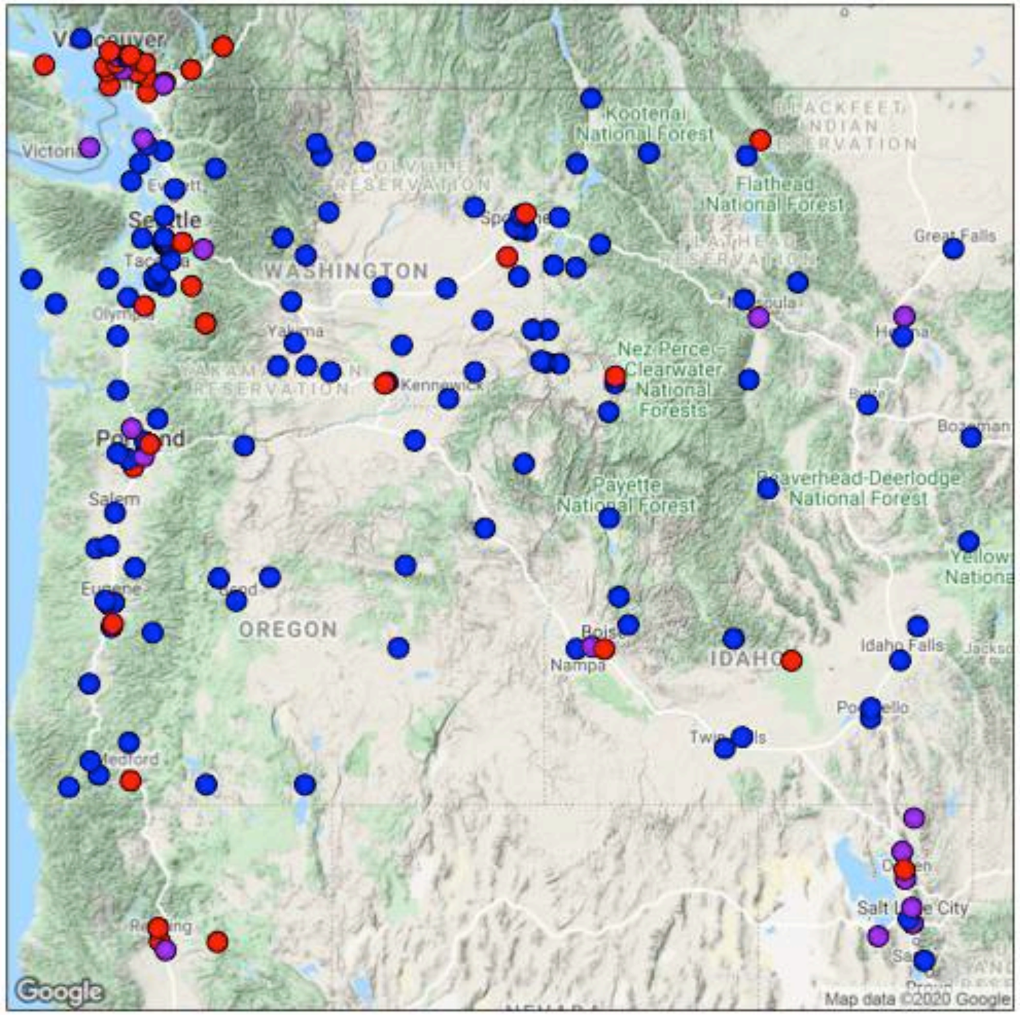


*Image from Google Earth

- 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. And then the ML models can be used for more sites in the Pacific Northwest (PNW).

Machine learning prediction in PNW

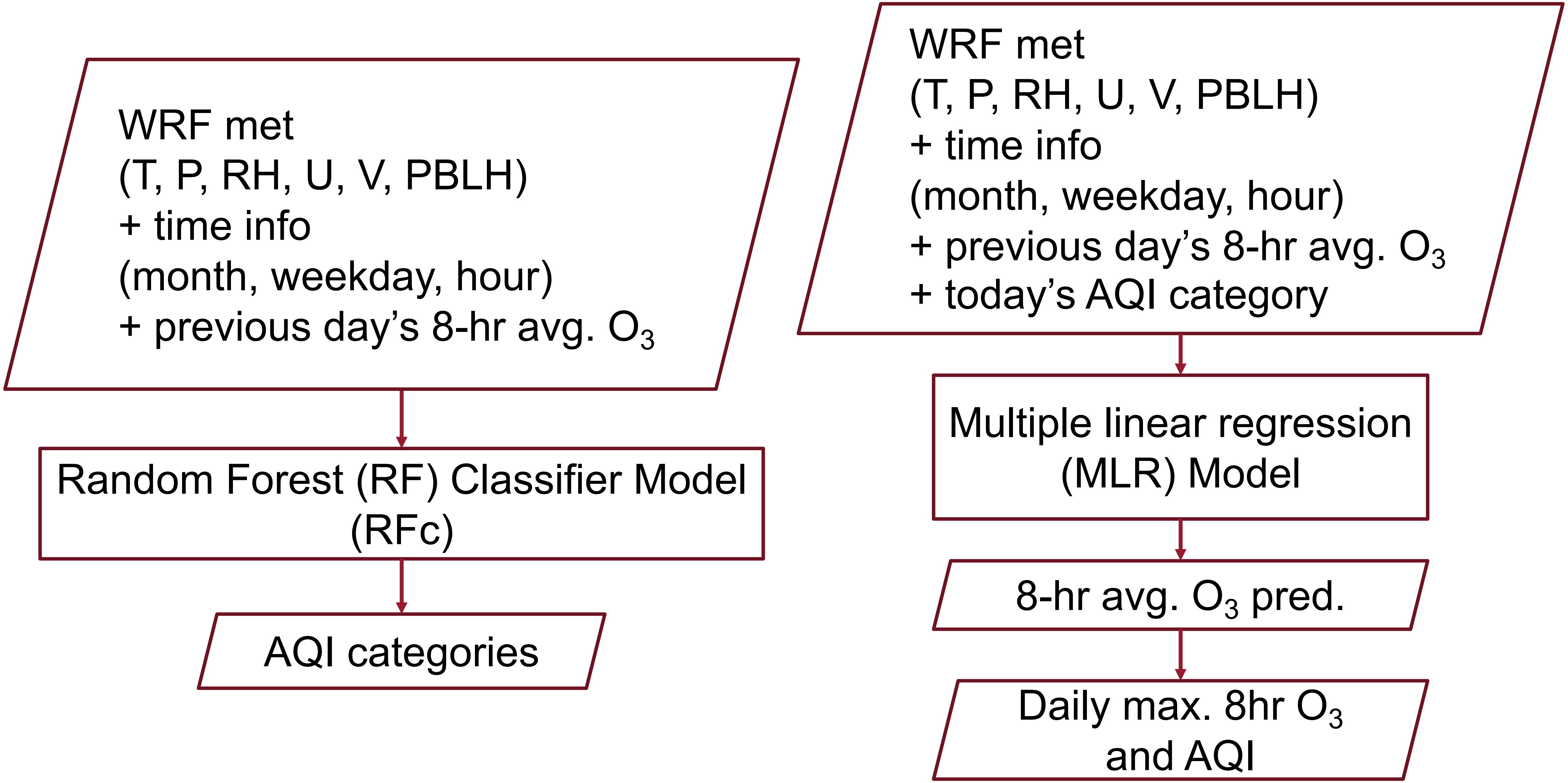


- O3
- PM2.5
- O3 and PM2.5

- There are 55 observation sites with O_3 observations in 2017-2019, 135 sites have PM2.5 observations, and 21 of them have both in the PNW.
- The ML models are trained individually in each site with archived WRF meteorology and observations.
- The analysis is based on the simulations in summer (June, July and August)

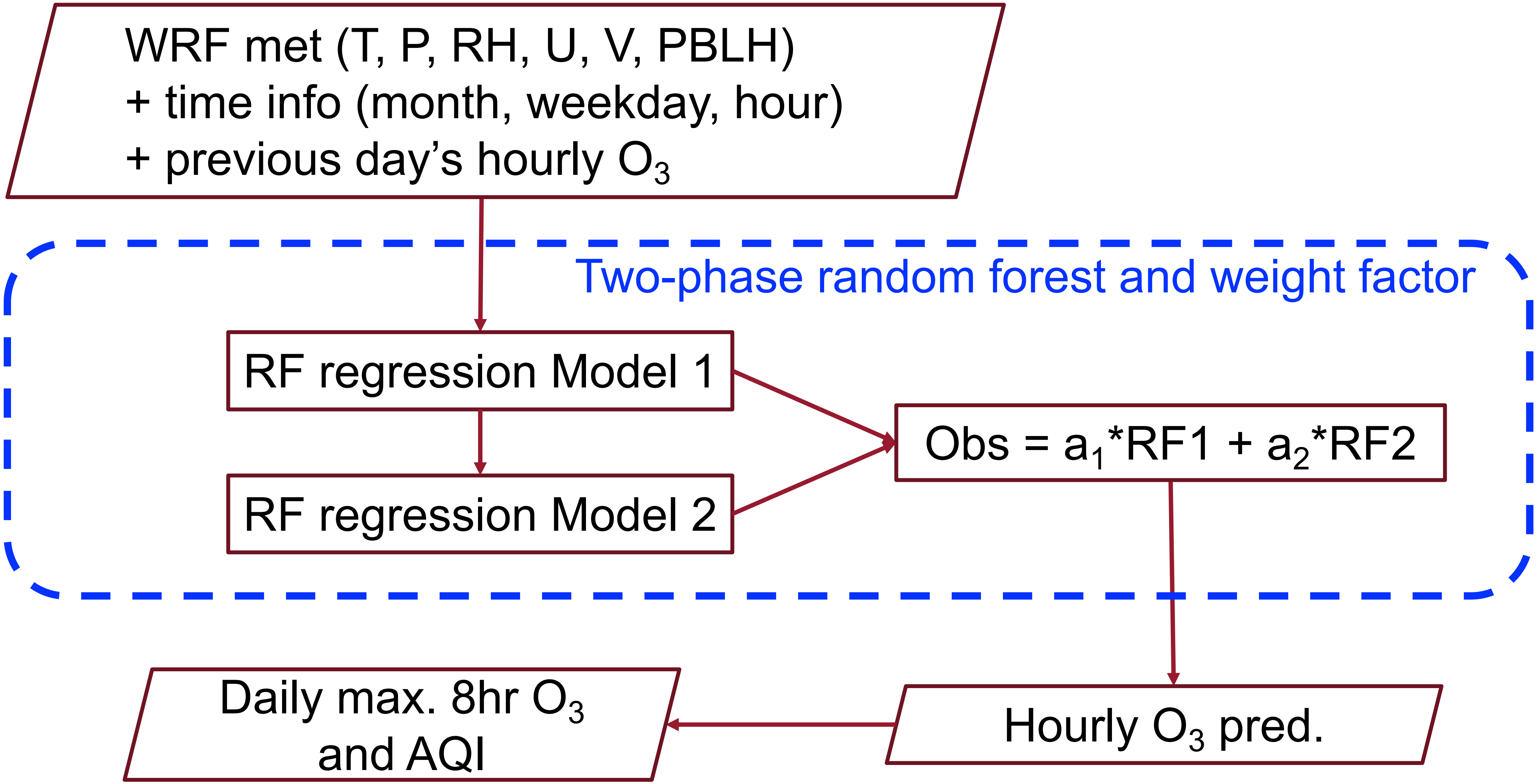
Machine Learning Model Framework 1: ML1

Combining Random Forest and Multiple Linear Regression methods



Machine Learning Model Framework 2: ML2

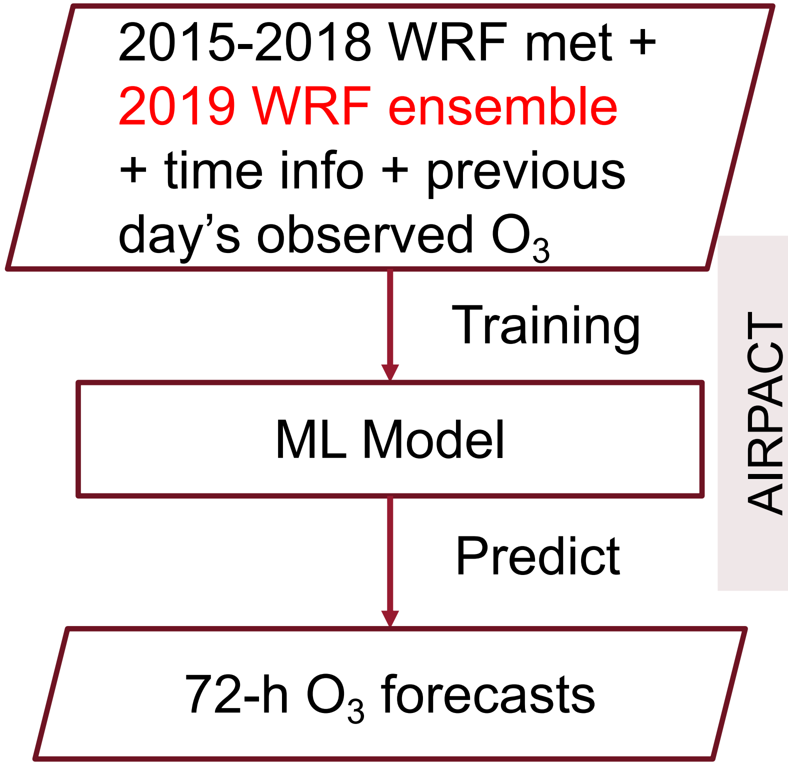
Two RF models weighted for optimal results



* 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.

Tri-Cities Ozone (ensemble mean) Forecast in 2019

- In 2019, there are 152 AQI1 days and 21 AQI2 days in Kennewick from April 6th to October 3rd.
- ML1 predicts the most hits and most false alarms
- ML2 reduces the false alarms significantly.



AIRPACT		Observation AQI 1	Observation AQI 2
	AQI 1	142	12
	AQI 2	10	9

ML1		Observation AQI 1	Observation AQI 2
	AQI 1	117	4
	AQI 2	35	17

ML2		Observation AQI 1	Observation AQI 2
	AQI 1	146	16
	AQI 2	6	5

Forecast evaluation parameters

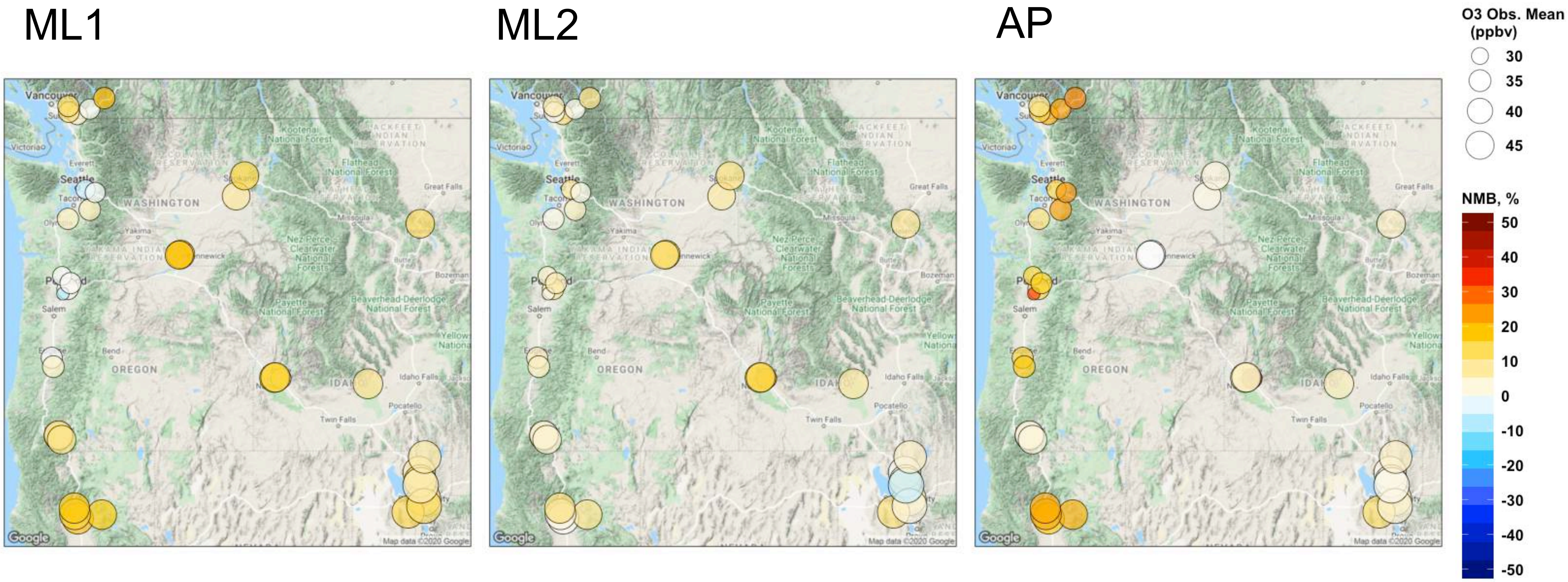
The Hanssen-Kuiper Skill Score (KSS)

- Also called Peirce Skill Score (PSS) or true skill statistic (TSS)
- How well does the model separate different categories?
- Range –1 to 1
- Perfect score = 1
- KSS = Hit rate – False alarm rate

Heidke Skill Score (HSS)

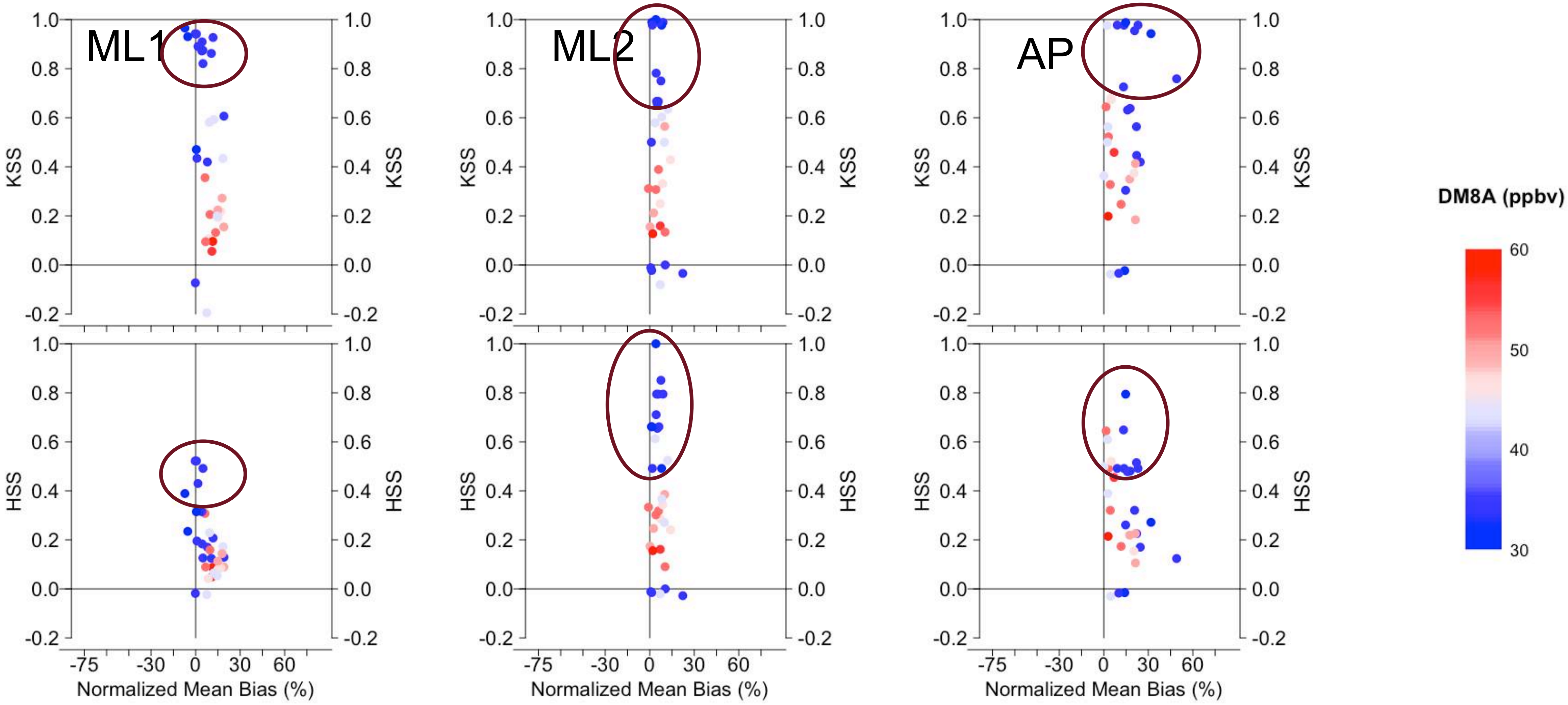
- What is the accuracy of the forecast in predicting the correct category, relative to that of random forecasts?
- Range –∞ to 1
- Perfect score = 1

2019 O_3 prediction in PNW



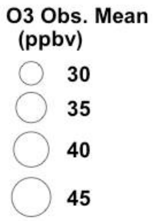
- 2017 and 2018 data is used to train the models, and 2019 data is used to evaluate them
- AIRPACT overpredicts DM8A O_3 of most sites along the coast. This does not happen for ML models
- The NMB is close between two ML models.

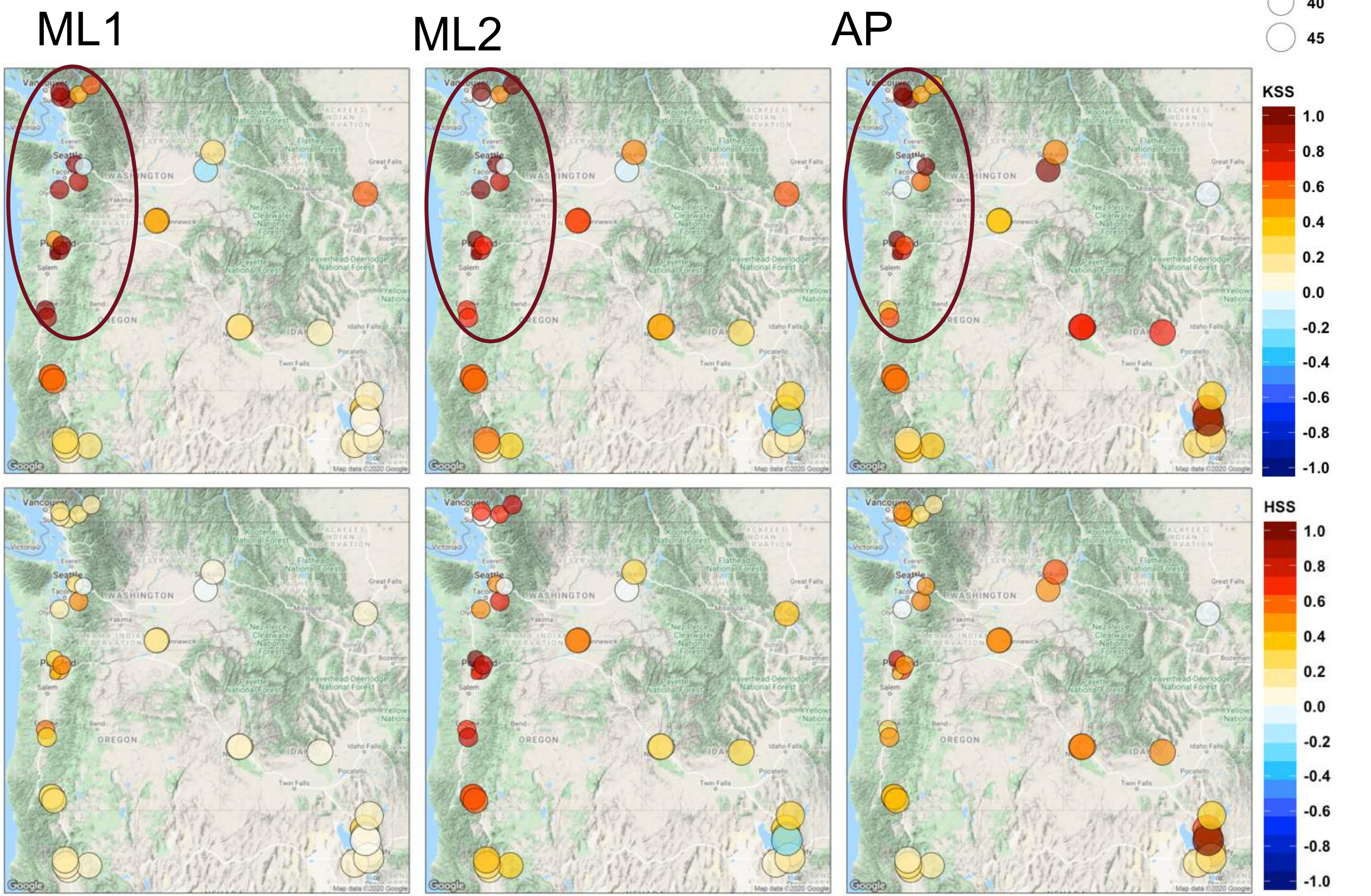
2019 O_3 prediction in PNW



- The ranges of NMB from ML models are narrow. AIRPACT overpredicts O_3, especially for low O_3 sites, which can be due to the NO titration at night.
- KSS and HSS are higher for low O_3 sites. It means the hit rate is higher than false alarm rate in these sites.

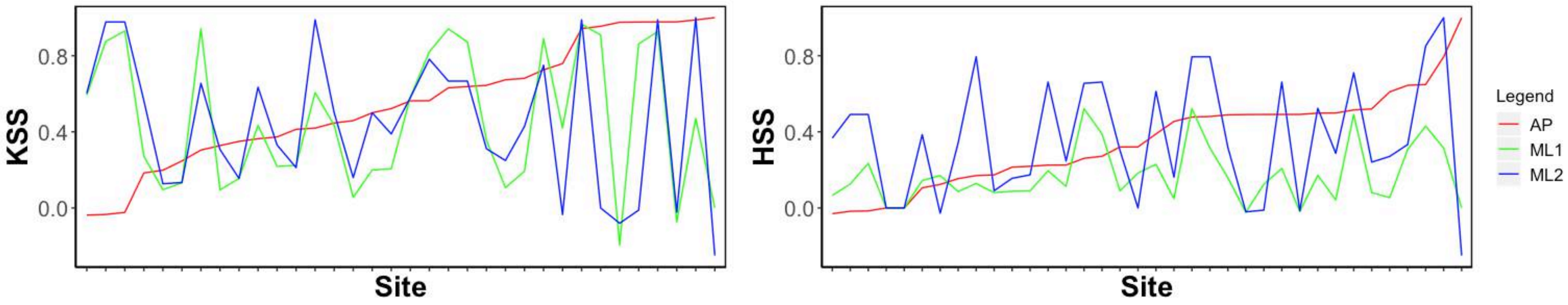
2019 O_3 prediction in PNW





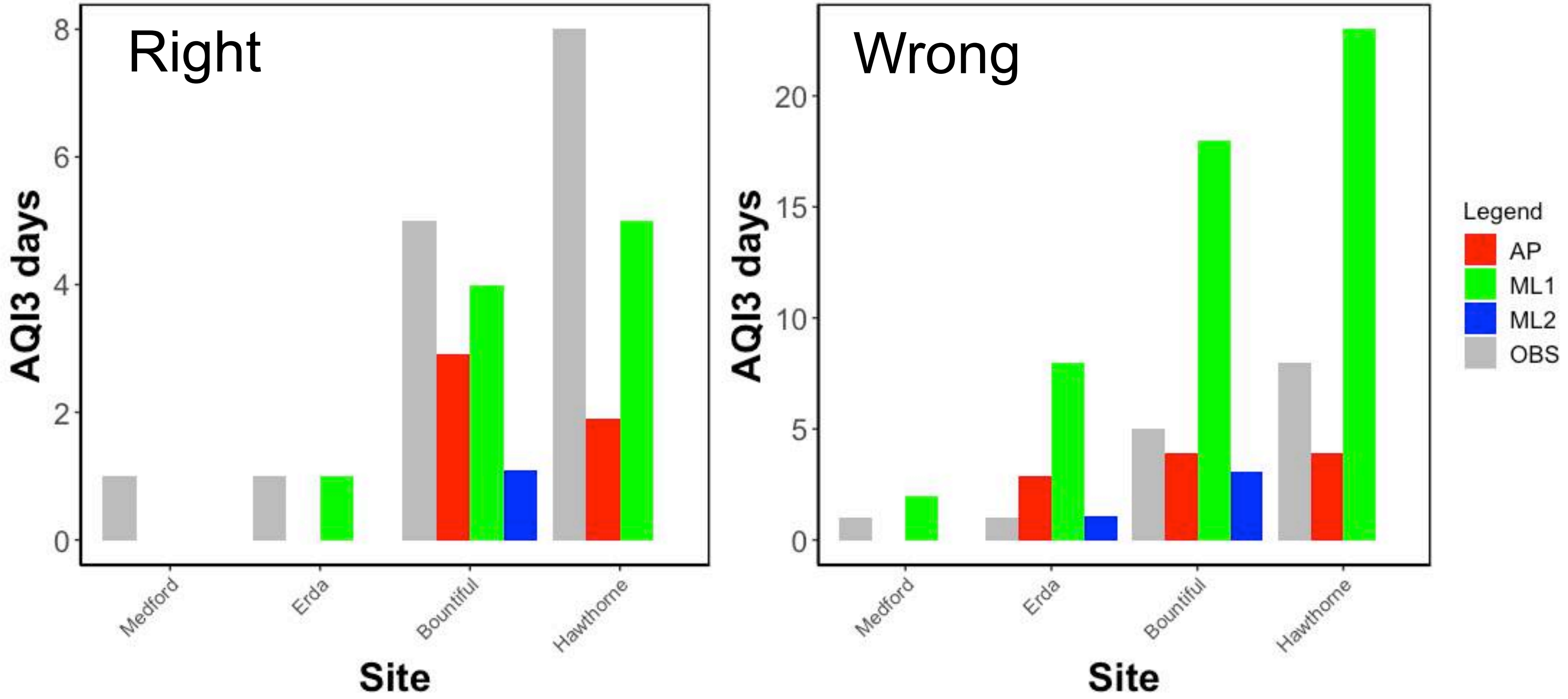
- The DM8A O_3 is relatively lower near Vancouver, Seattle, Portland and Eugene. KSS at these sites is higher.
- For HSS, ML1 and AIRPACT do not perform well in big cities near the western coast. ML2 shows higher HSS at these sites.

2019 O_3 prediction in PNW



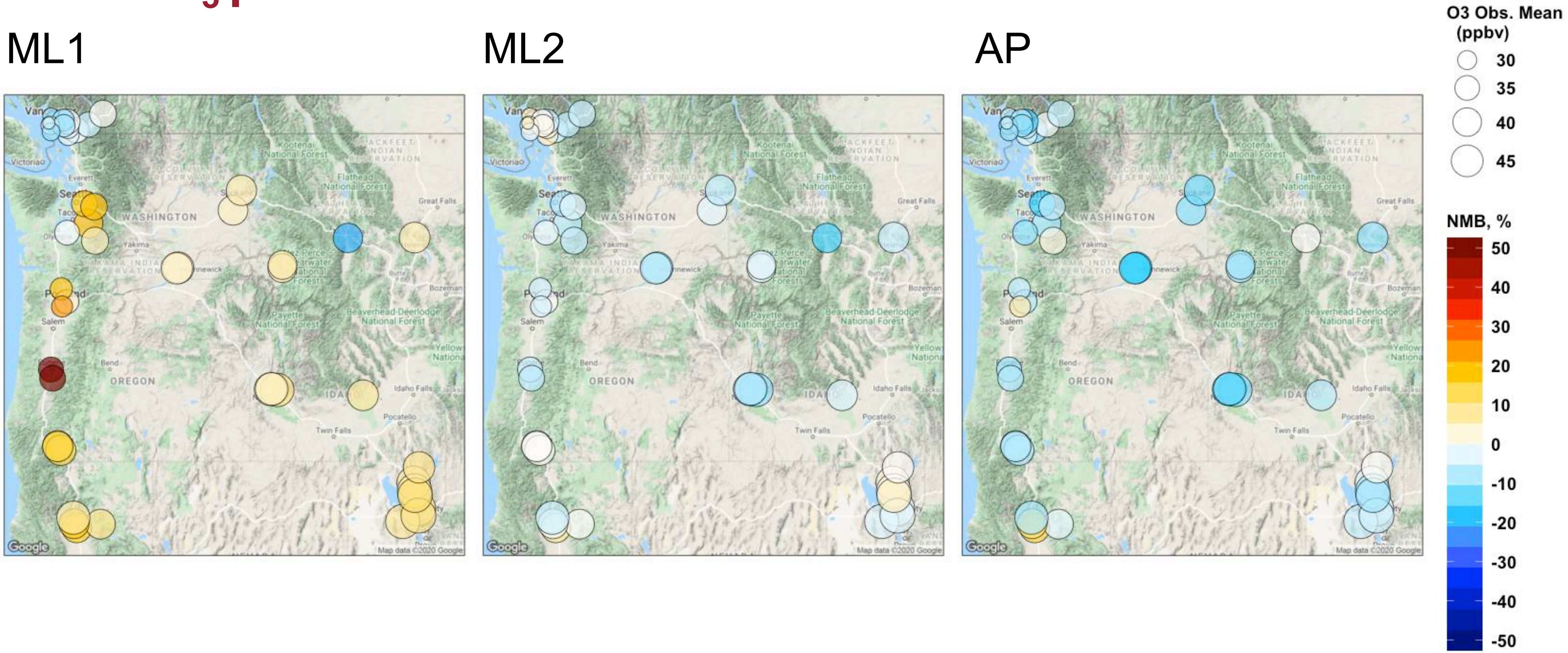
- HSS and KSS show that the model performance varies at these sites
- HSS and KSS from ML models do not follow the trend of AIRPACT
- ML1 and ML2 shows close KSS at most sites
- ML2 shows higher HSS at most sites

2019 O_3 prediction in PNW



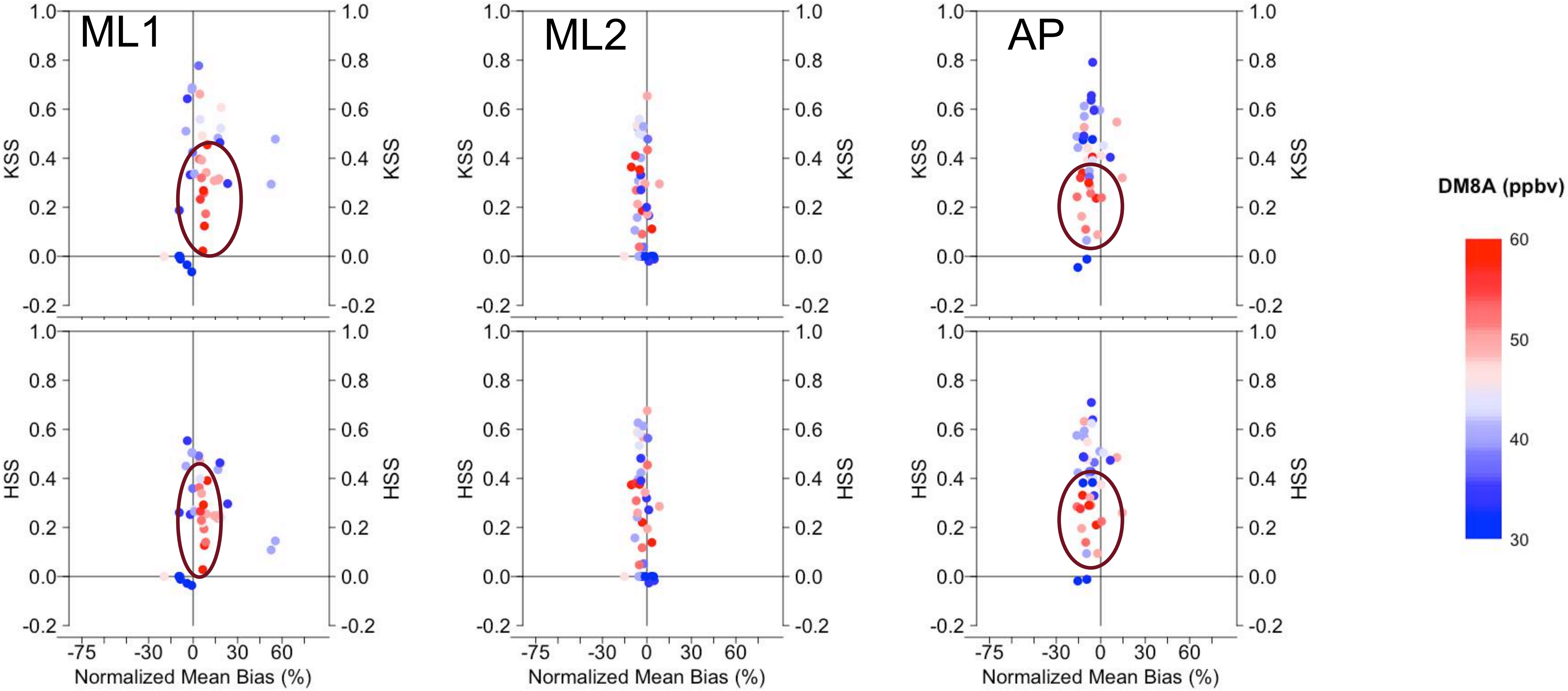
- There are only four sites where AQI3 days occurred. These figures show the right predicted (hit) and wrong predicted (false alarm) AQI3 days at these sites.
- ML1 captures the most hit, but also the most false alarms.

2017 O_3 prediction in PNW



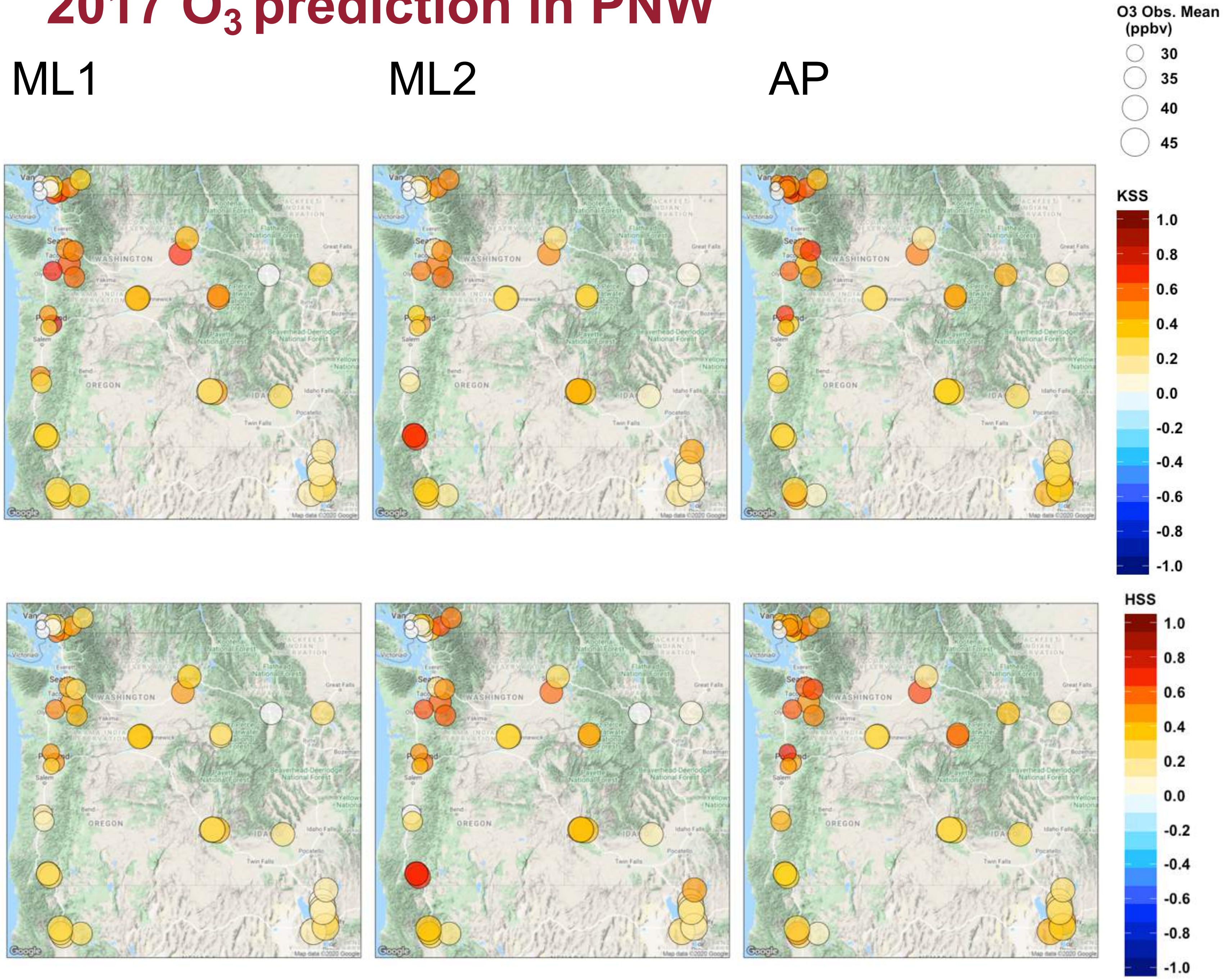
- There are not many high O_3 days in 2019. So, here 2018 and 2019 data is used to train the models and 2017 for evaluation
- ML1 overpredicts most sites. ML2 and AIRPACT underpredicts most sites.

2017 O_3 prediction in PNW



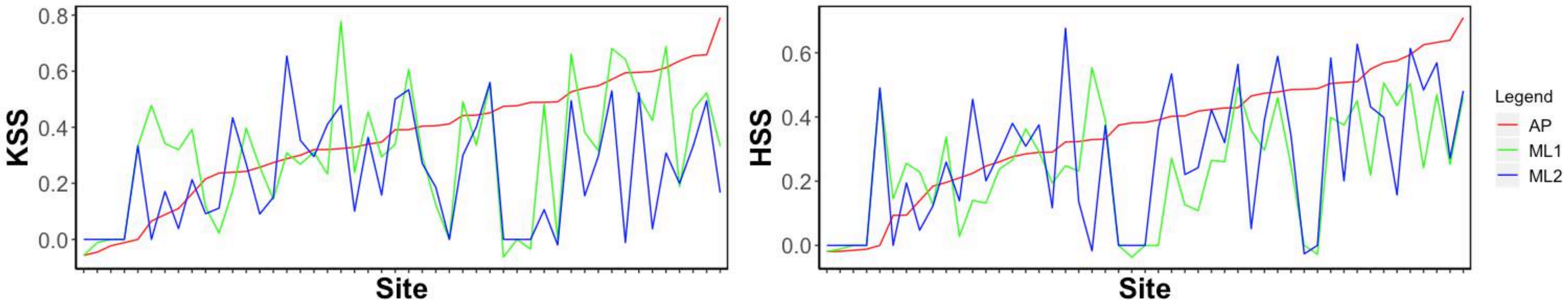
- The NMB range of ML2 is very narrow.
- ML1 overpredicts most sites. AIRPACT underpredicts most sites.
- ML1 and AIRPACT shows lower KSS and HSS at the high O_3 sites.

2017 O_3 prediction in PNW



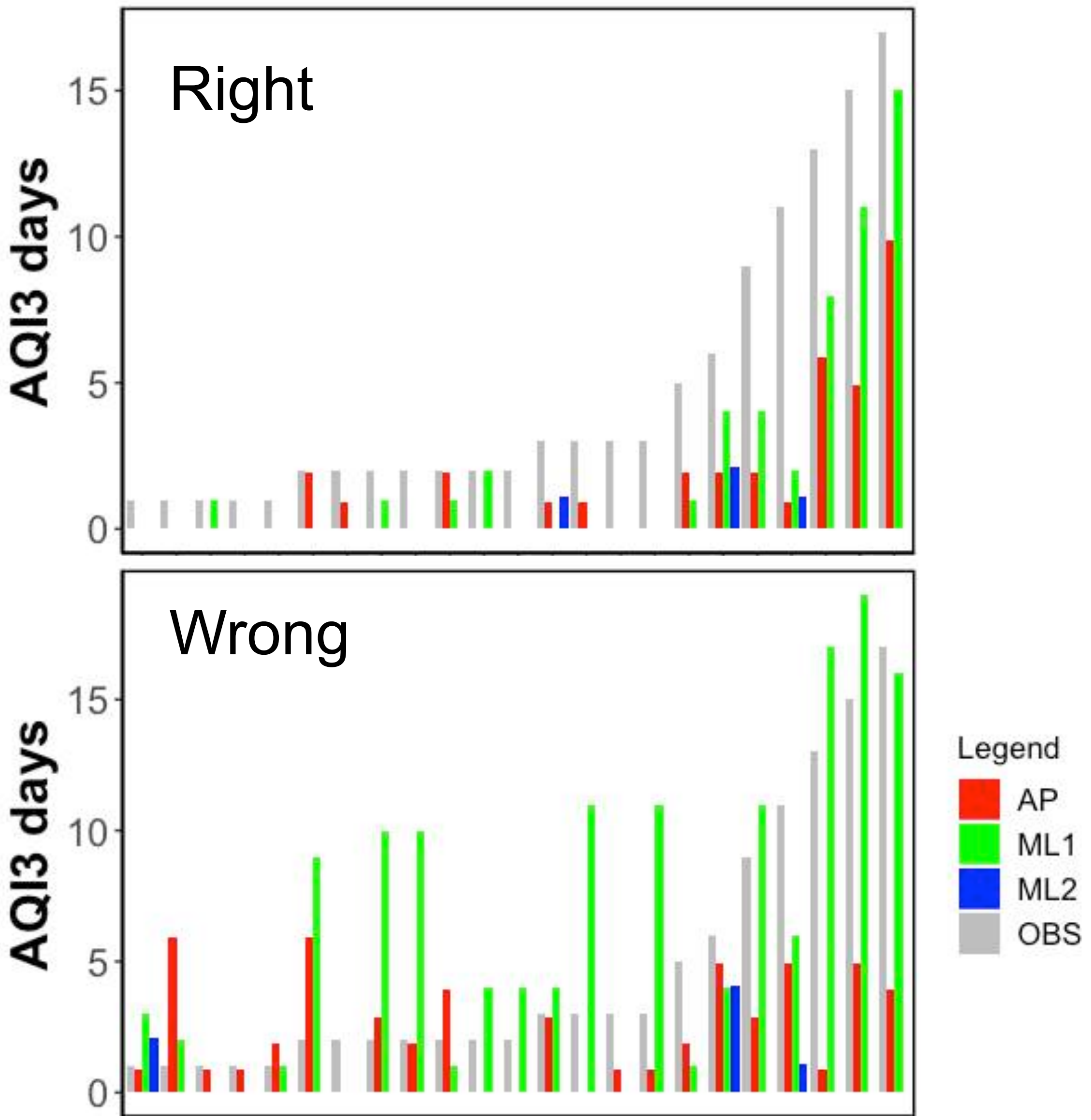
- HSS and KSS is close among three models at most sites

2017 O_3 prediction in PNW



- ML1 shows higher KSS at most sites.
- HSS is very close between two ML models.

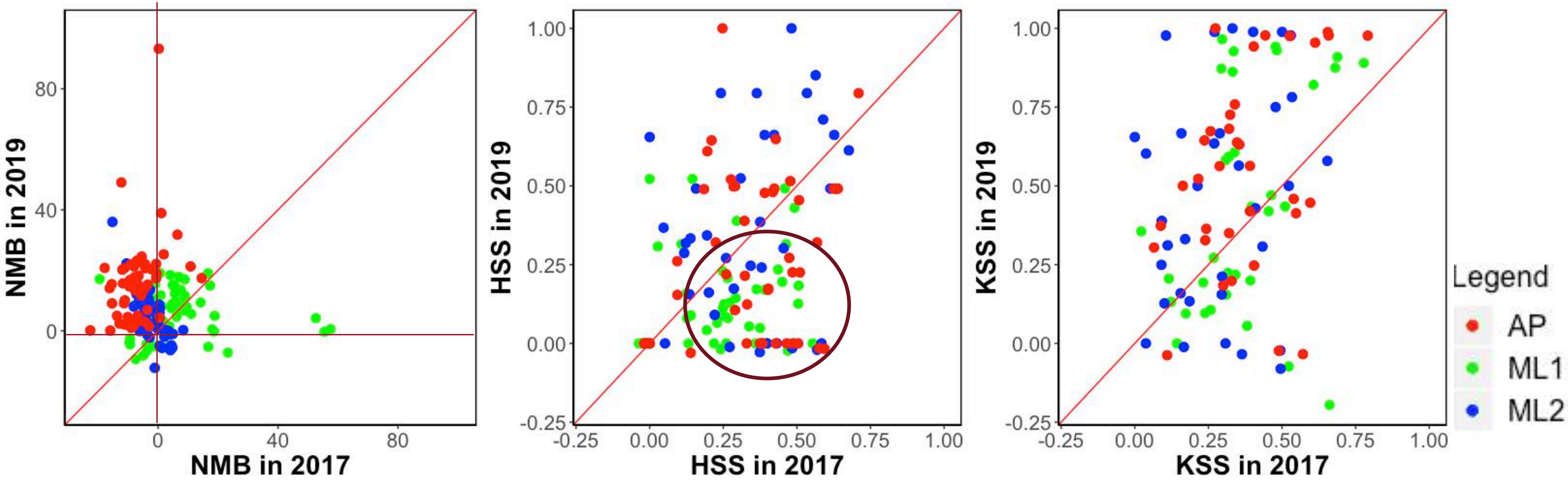
2017 O_3 prediction in PNW



- ML1 captures the most hit, but also the most false alarms.
- ML2 can reduce the false alarms, but miss many high O_3 days

Model performance in yearly variation

Scatter plots to compare NMB, HSS, KSS in 2017 and 2019.



- Three models tend to overpredict at most sites in 2019
- ML1 show higher HSS in 2017 than 2019
- No clear trend for KSS

Summary

- All models overpredict O_3 in 2019. ML1 overpredicts O_3 in 2017, ML2 and AIRPACT underpredict in 2017. The NMB range of ML2 is narrowest, so it could provide more accurate O_3 prediction for low O_3 days at most sites.
- HSS and KSS do not show a significant difference among three models, and they vary among sites and years.
- ML1 can capture more high O_3 days than ML2 and AIRPACT, but more false alarms than them. If ML1 can be improved to reduce the false alarm rate, it should be a good tool for O_3 forecasts.
- 2017 has more high O_3 days than 2019, so the ML model performance differs between them. This also can be due to the different training dataset.
- The similar ML models will be used to predict PM2.5 concentrations. And the cross validation method will be used to evaluate the models.

Thank you!

	1	2	3	hss	kss
1	100	0	0	na	na
2	0	0	0		
3	0	0	0		

	1	2	3	hss	kss
1	99	0	0	1	1
2	0	1	0		
3	0	0	0		

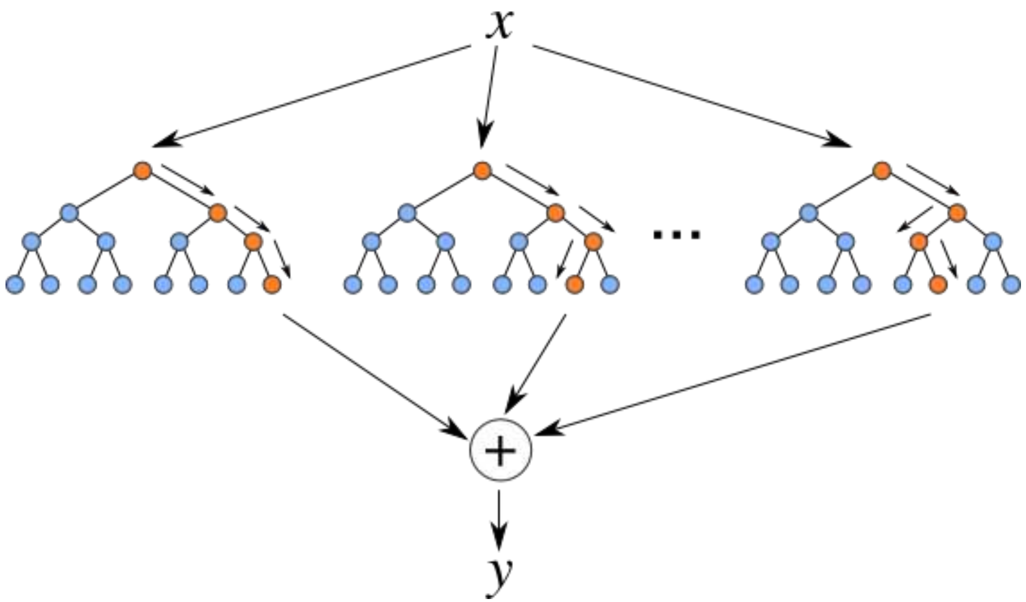
	1	2	3	hss	kss
1	99	1	0	0	0
2	0	0	0		
3	0	0	0		

	1	2	3	hss	kss
1	97	0	1	0.8	0.67
2	0	2	0		
3	0	0	0		

	1	2	3	hss	kss
1	97	0	0	0.83	0.83
2	0	2	1		
3	0	0	0		

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

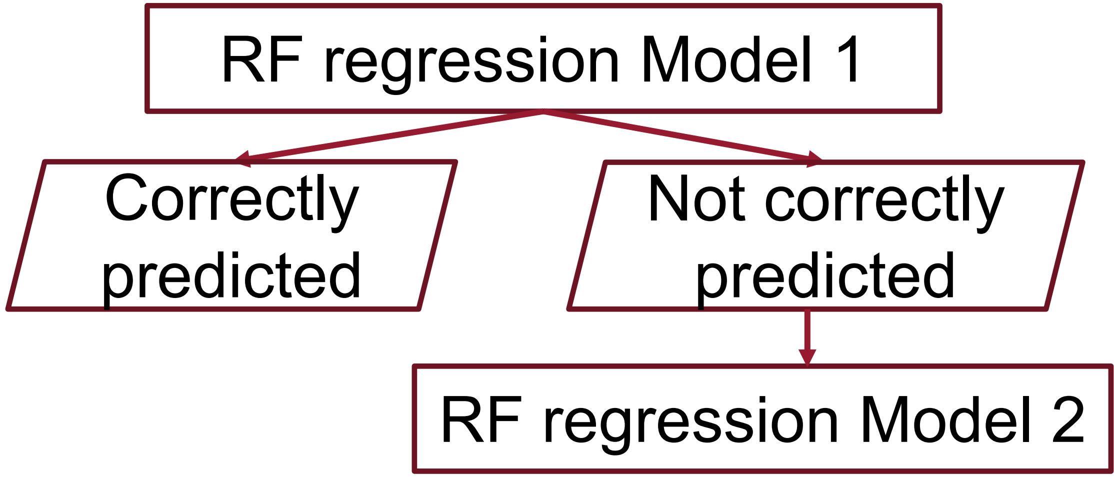
Multiple linear regression (MLR)

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

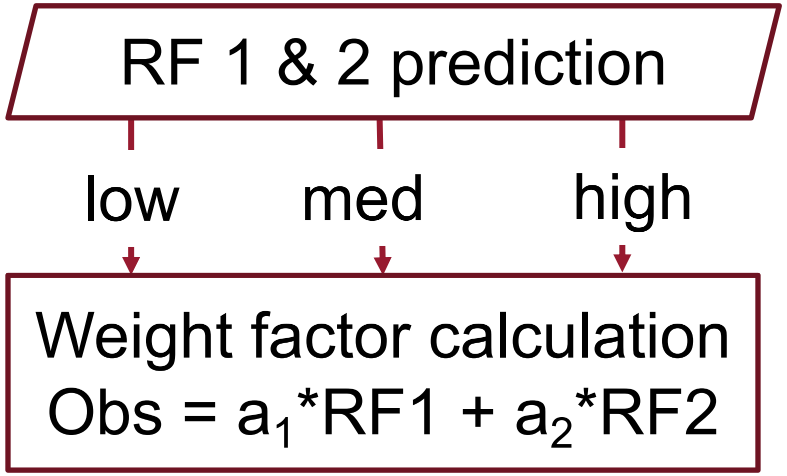
- MLR approach is used to predict the 8-h average O_3, which shows good performance to predict high O_3 days.

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.

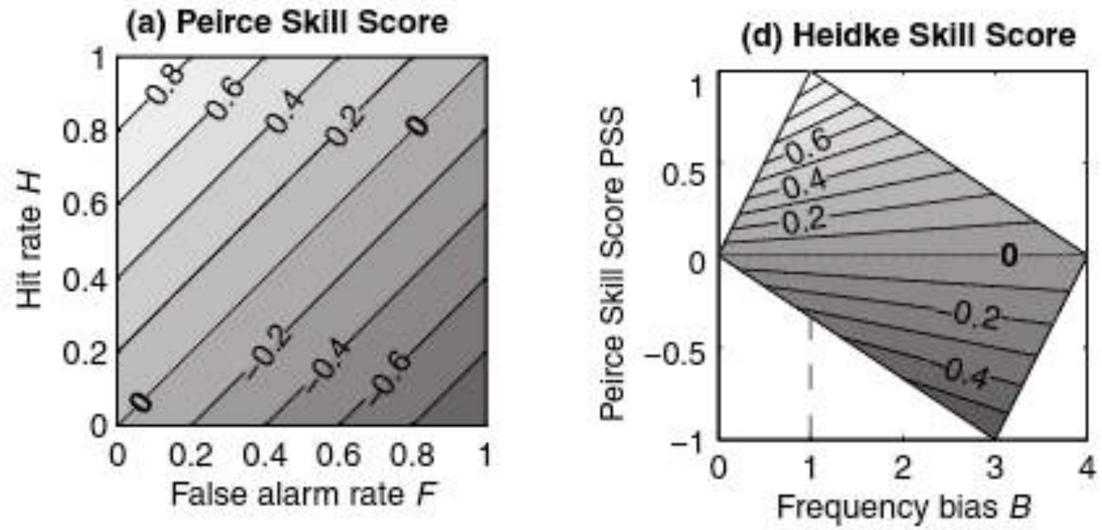


Forecast evaluation parameters

Table 3.1 Schematic contingency table for deterministic forecasts of a sequence of n binary events. The numbers of observations/forecasts in each category are represented by a, b, c and d

	Event observed		
Event forecast	Yes	No	Total
Yes	a (Hits)	b (False alarms)	$a+b$
No	c (Misses)	d (Correct rejections)	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=n$

Hit rate, H = a/(a+c)
False alarm rate, F = b/(b+d)
Frequency bias = forecast rate r / base rate s = (a+b) / (a+c)



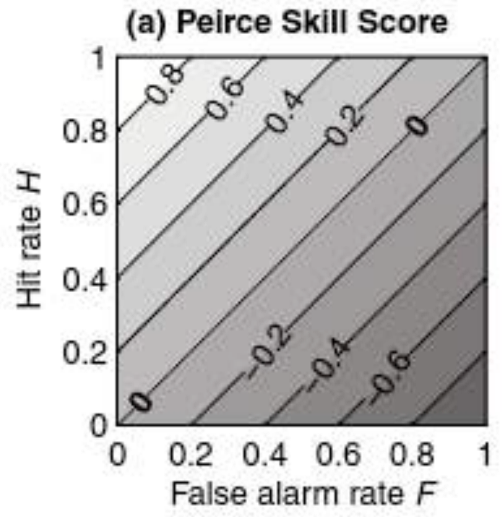
Both PSS and HSS are truly equitable, awarding random and constant forecasts an expected score of zero. They are equal for unbiased forecasts, and when s = 1/2 they are equal for all forecasts. They therefore differ only in the way they treat biased forecasts for s ≠ 1/2. Figure 3.3d shows that when s < 1/2, isopleths of HSS are further apart than isopleths of PSS for forecasting systems that overpredict occurrence, but closer together for forecasting systems that underpredict. Therefore, for systems with positive skill, PSS will treat overpredicting systems more generously than HSS and underpredicting systems more harshly. The opposite is true when s > 1/2 . Being truly equitable and difficult to hedge, both measures are more robust indicators of skill than the previous ones discussed in this section. In terms of properties listed in Table 3.4, the only difference is that HSS is transpose symmetric while PSS is base-rate independent.

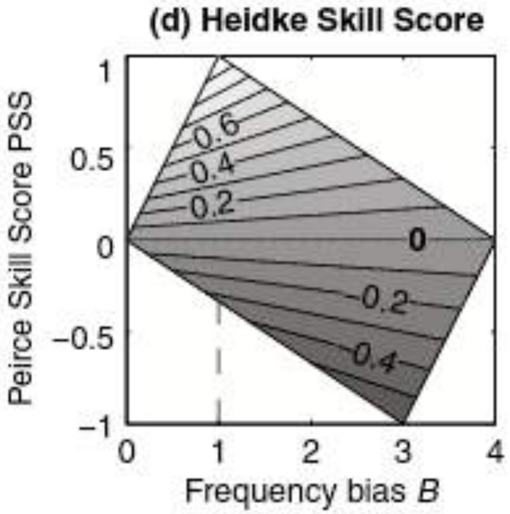
Forecast evaluation parameters

Table 3.1 Schematic contingency table for deterministic forecasts of a sequence of n binary events. The numbers of observations/forecasts in each category are represented by a, b, c and d

	Event observed		
Event forecast	Yes	No	Total
Yes	a (Hits)	b (False alarms)	$a+b$
No	c (Misses)	d (Correct rejections)	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=n$

Hit rate, H = a/(a+c)
False alarm rate, F = b/(b+d)
Frequency bias = forecast rate r / base rate s = (a+b) / (a+c)





Random forecast
$a_r = (a+b)(a+c)/n$
$d_r = (b+d)(c+d)/n$

$S = (x-x_r)/(x_p-x_r)$

x=a+d or x=PC=(a+d)/n

$$\text{HSS} = \frac{a+d-a_r-d_r}{n-a_r-d_r}$$

$$\text{PSS} = \frac{ad-bc}{(b+d)(a+c)} = H - F$$