A Model Evaluation Framework for a Wild Fire Air Quality Forecast Model

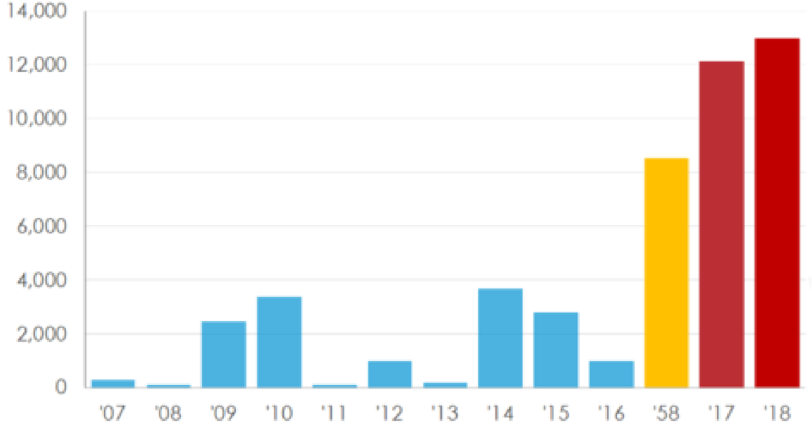
Rita So and Bruce Ainslie

Environment & Climate Change Canada

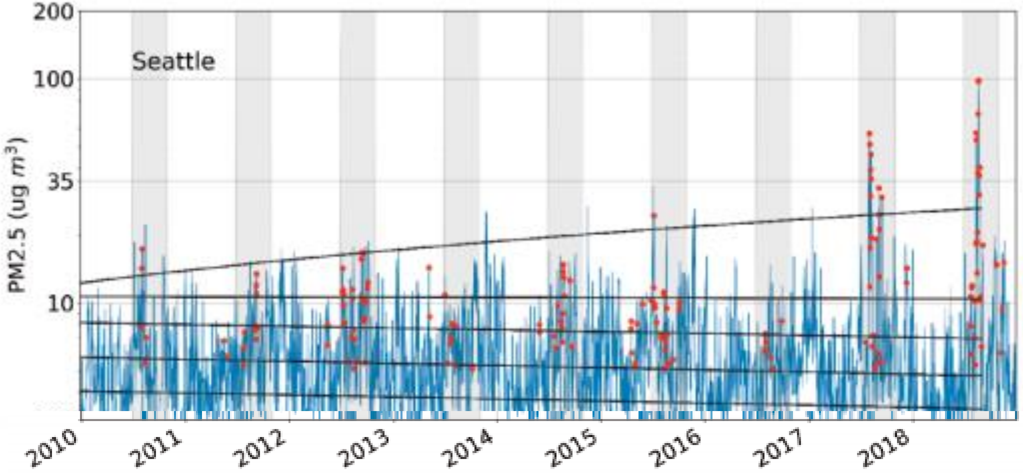
June 13th, 2019

NASA Aqua MODIS imagery on Aug 1, 2017.
Source: LANCE/EOSDIS MODIS Rapid Response Team

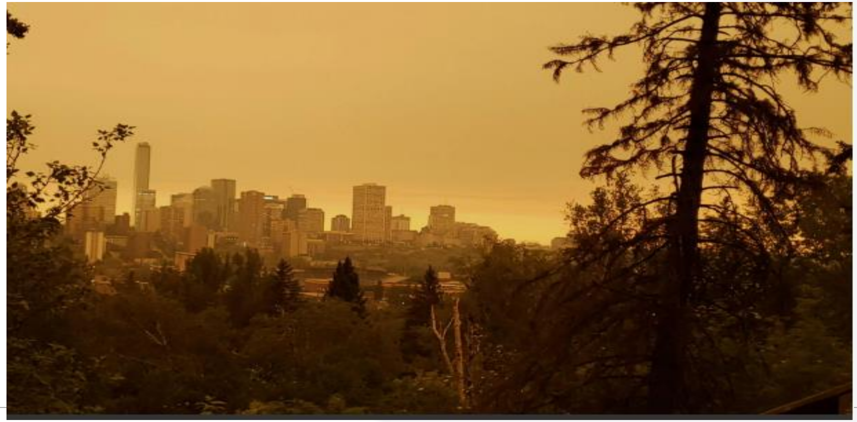
Motivation



Total Area Burned by Wildfire per Season (BC Wildfire)



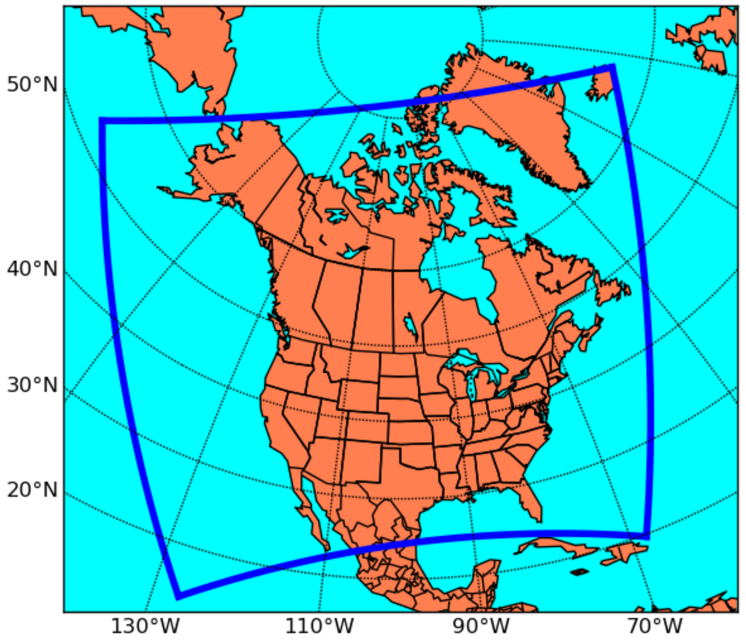
Timeseries of PM2.5 at Seattle Wa. (source: The Magazine for Environmental Managers • A&WMA • June 2019)



Downtown Edmonton on May 30, 2019 (source: Global News https://globalnews.ca/news/5333507/alberta-wildfire-edmonton-air-quality-statement/

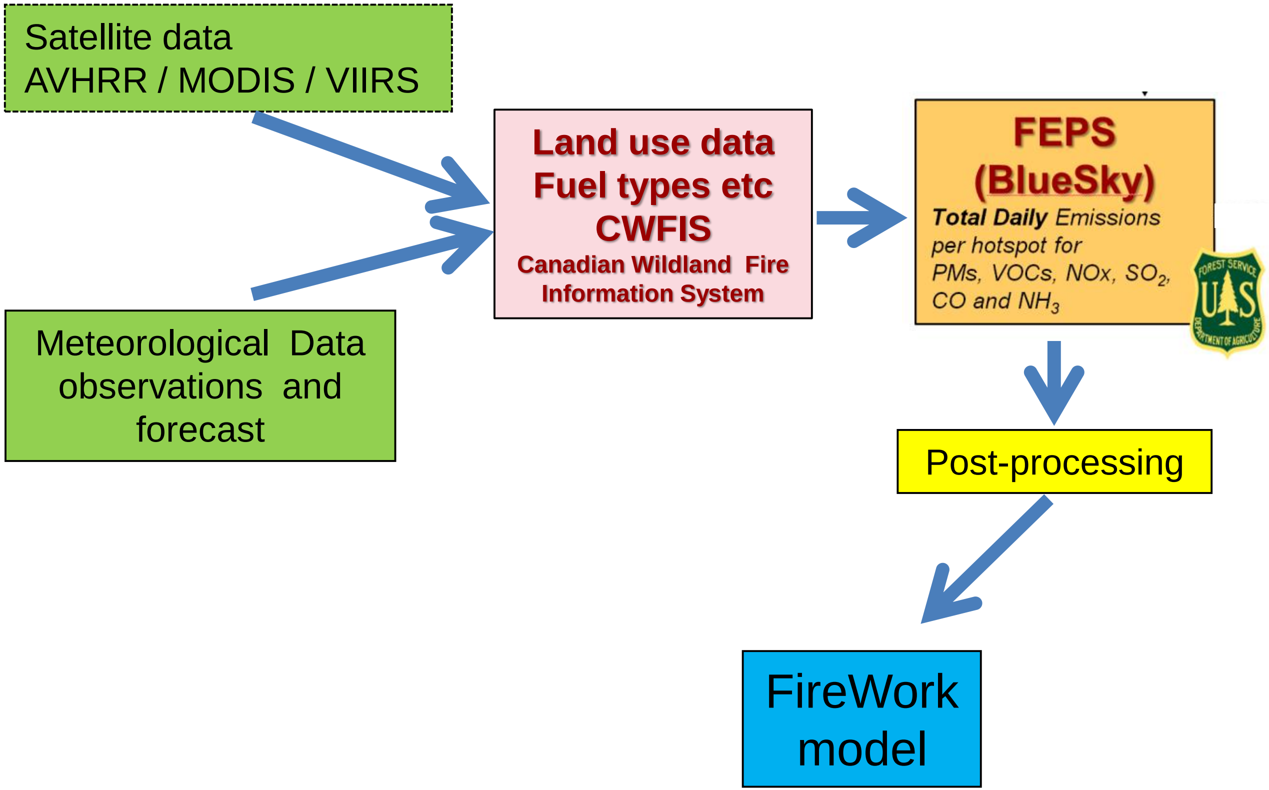
FireWork – ECCC's Air Quality Forecast System

- 48hr forecast 2X daily (0z / 12z) from Apr to Oct at 10km horizontal grid spacing
- Near-real time fire data from Canadian Wildland Fire Information System (CWFIS)
- **FEPS***: operational since 2016 - 2018 (experimental mode 2014 - 2015)
- **CFFEPS**: current operational version (2019-)



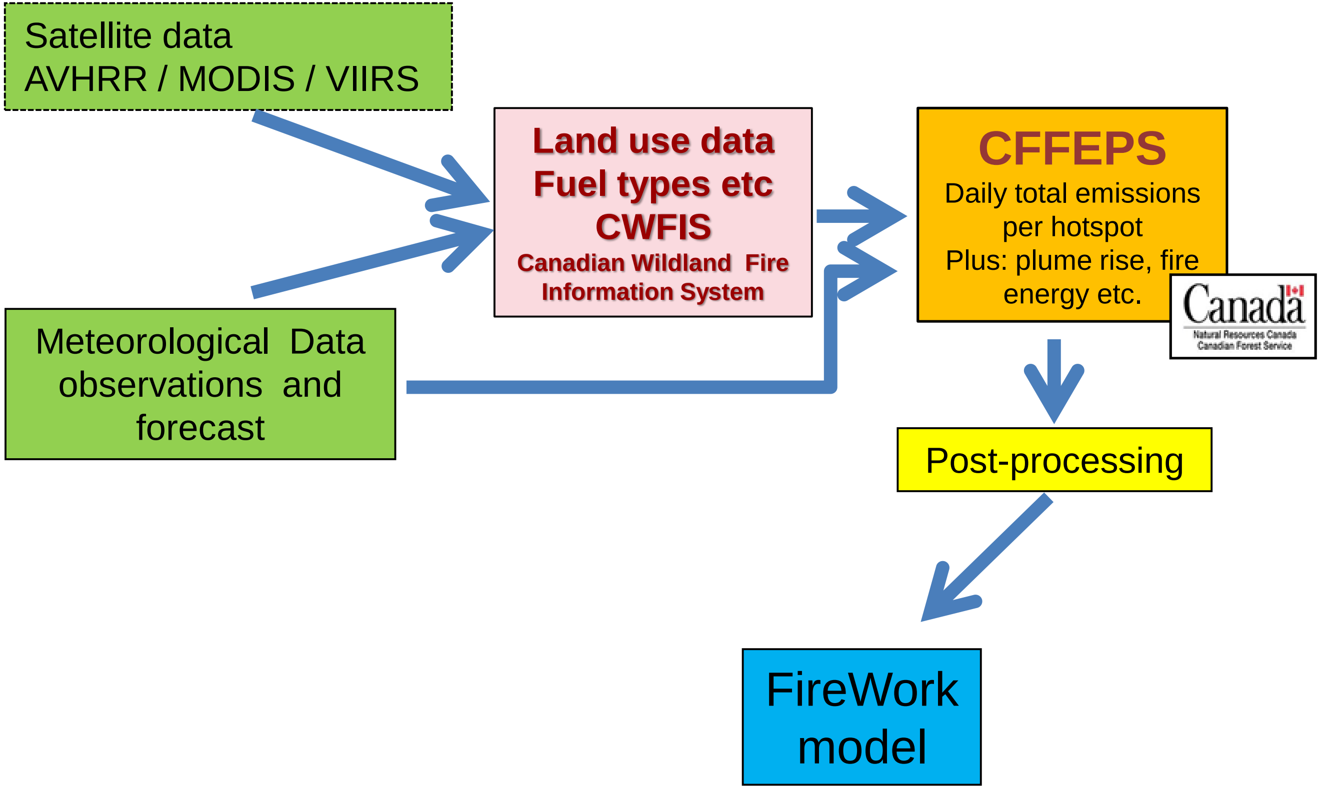
* I will sometimes loosely refer to FW-FEPS as FireWork or just FW

Old Operational System (FEPS + FireWork)



More details:
Chen, J. 2019: https://www.geosci-model-dev-discuss.net/gmd-2019-63/gmd-2019-63.pdf

Current operational System (CFFEPS + FireWork)



More details:
Chen, J. 2019: https://www.geosci-model-dev-discuss.net/gmd-2019-63/gmd-2019-63.pdf

Research Goal

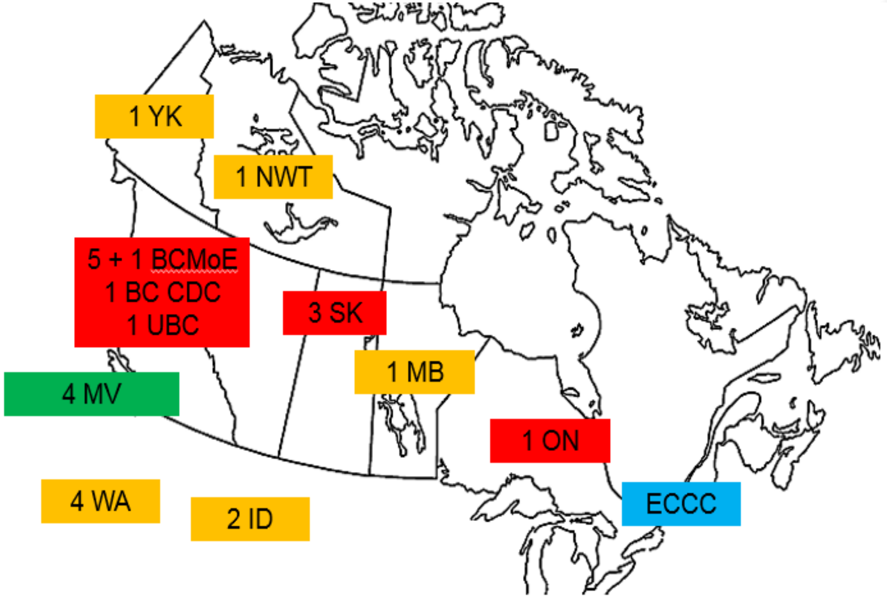
Tease out FW's inherent skill despite confounding effects of weather, fire location etc.

Utility

Inform model development group where FW needs improvement

FireWork User Survey

- Jurisdiction-based, but never at a monitor…
- Timing & Location >> Concentration & Footprint
- Missing an event* >> False alarm
- 1st day forecast (0 – 24 hrs) is the MOST IMPORTANT



*Event = 24hr avg PM2.5 > 25 µg/m^3

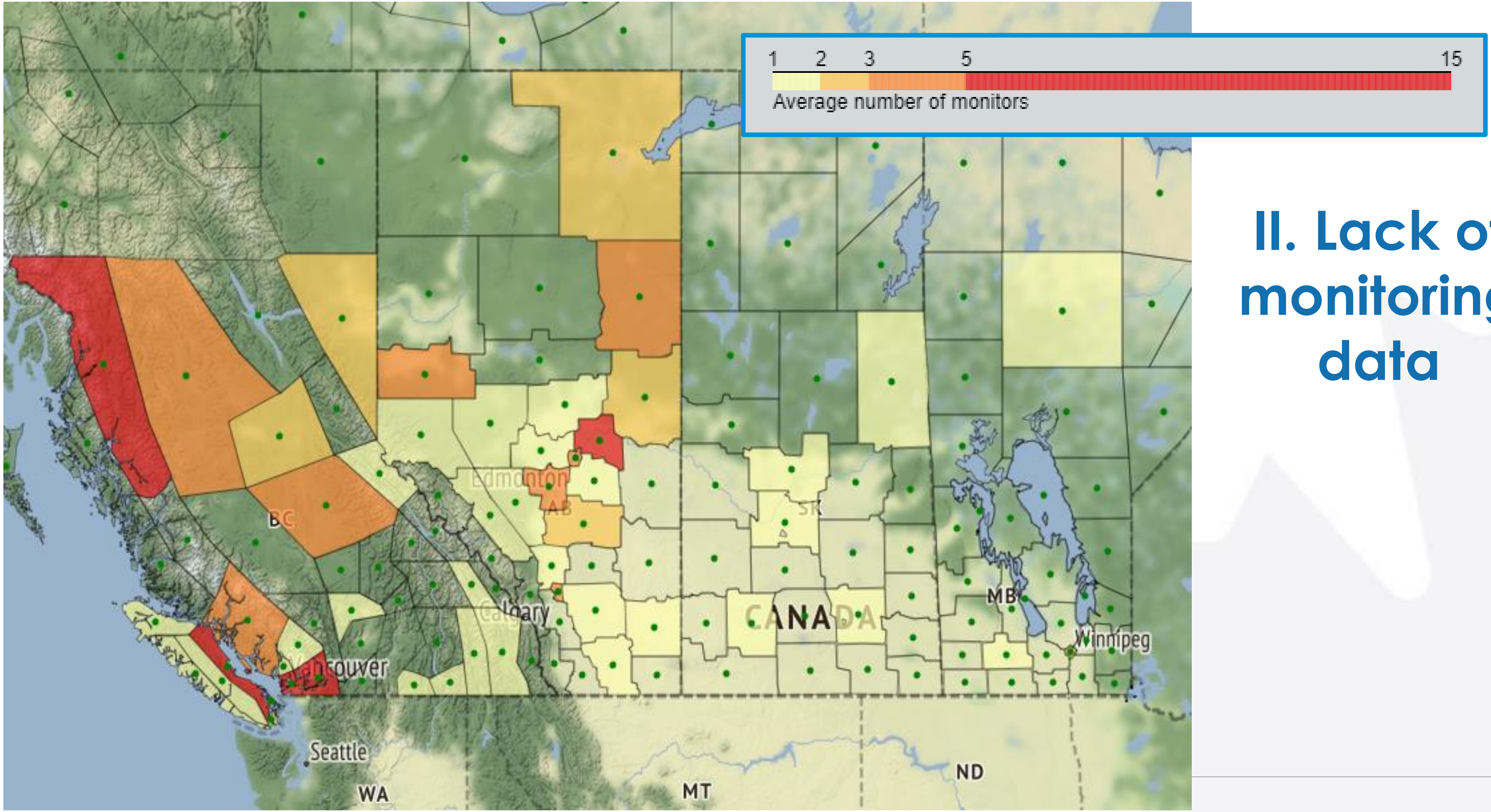
Four facets of FW model evaluation that have caused us a lot of grief

Four Horsemen of the Apocalypse by Viktor Vasnetsov

I. Forecast Quality versus Value

- A forecast has high **quality** if it predicts the observed conditions well according to some objective or subjective criteria. It has high **value** if it helps the user to make a better decision.

		Quality	
		Low	High
Value	Low	Ugh!	A forecast of clear skies over the Sahara Desert during the dry season
	High	"high resolution model predicts the development of smoke in a particular region, and smoke is indeed observed in the region but not in the particular spots suggested by the model. By most standard verification measures this forecast would have poor quality, yet it might be very valuable to the forecaster in issuing a public smoke forecast." WWRP/WGNE Joint Working Group on Forecast Verification Research	Everyone's dream



II. Lack of monitoring data

III. Pooled versus Unpooled data

- To get reliable verification statistics, a large number of forecast/obs pairs may be pooled over time and/or space. The larger the number of samples, the more reliable the verification results but …
- Pooling can mask variations in forecast performance when data are not homogeneous. Pooling all station-data will generally result in statistics that are biased towards the most commonly sampled regime:
  - in regions with higher station density (e.g. LFV)
  - days with no smoke impacts

IV. Class Imbalance – *the curse of correct negatives*

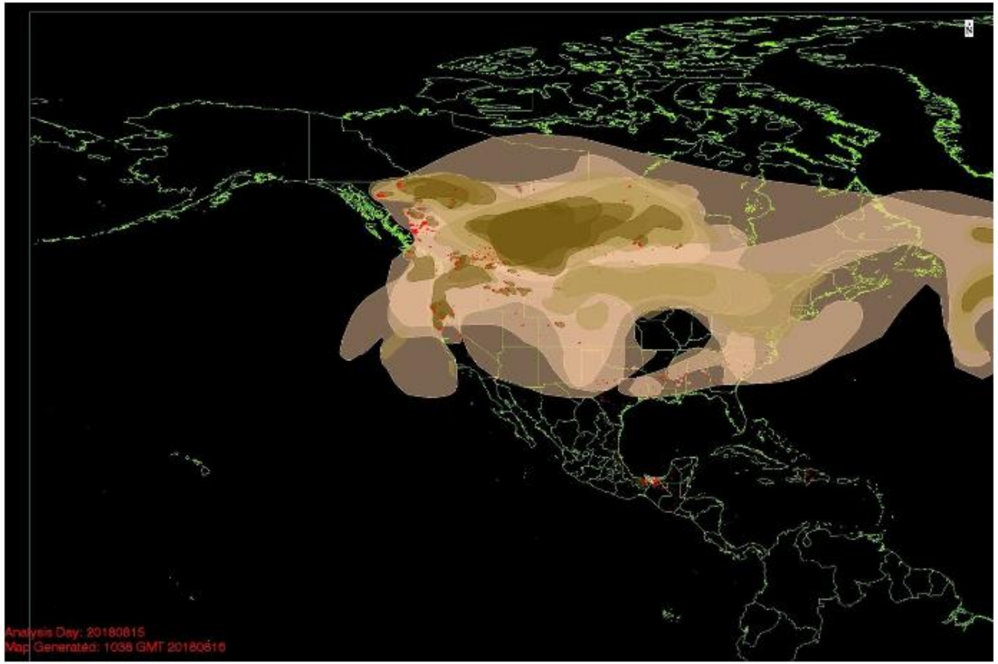
- In smoke forecasting, filtering the 'correct negative' cases is important since skill can be inundated by verification of the numerous and often trivial non-events.
- *Is the prediction of smoke-free conditions in Atlantic Canada from a BC wildfire skillful?*

Capital Weather Gang

Smoke from California's wildfires is reaching Washington and Baltimore

By Jason Samenow

August 16



Smoked analyzed across the United States on Wednesday. (NOAA's Hazard Mapping System via Joel Dreessen)

Road Map

Measure-based: During smoke-impacted days, how well does the model forecast ambient PM2.5 concentrations?

Contingency-based: How well does the model distinguish between smoke and non-smoke impacted days?

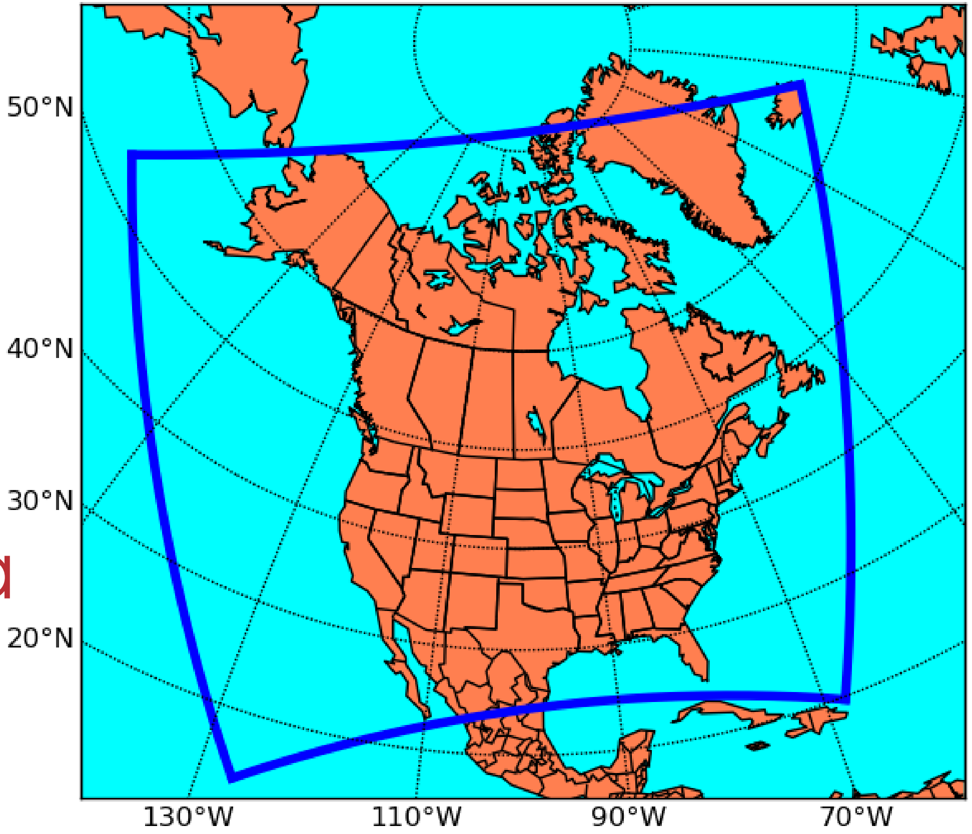
Monitor-level: How well does model agree at monitor locations?

Forecast Zone-level: How well does model agree when averaged over forecast zones?

		Spatial Scale	
		Monitor-level	**Forecast Zone-level**
Analysis	**Measure-based** (viia boxplots of model bias)	"Likely" smoke-impacted days & when both obs and forecast > 25 $\mu g/m^3$ (broken out by L, M, H)	Not ready
	Contingency-based (via Performance Diagram)	"Likely" smoke-impacted days (Broken out by L, M, H)	BIAS only

FEPS model output

- Experimental 2014 - 2015
- Operational 2016 - 2018
- 10km horizontal grid spacing
- Grab daily first 24hrs from 12Z forecast over MJJAS summer months

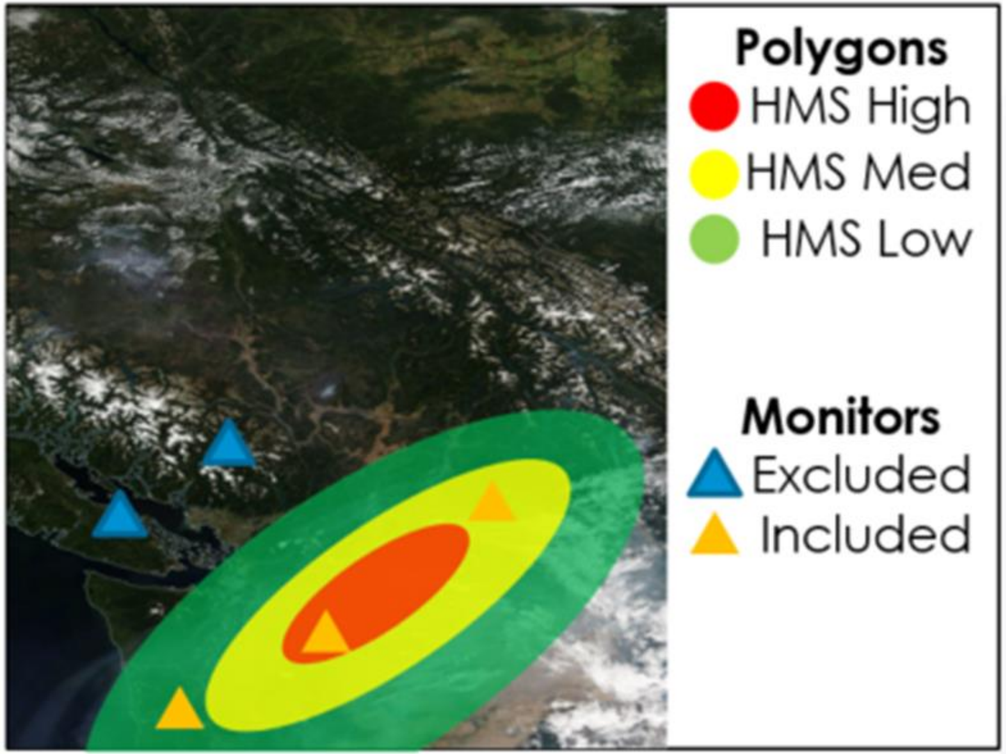


NAPS/BcMoE/ADE data

- Daily 24-hr obs of PM 2.5
- ~ 120 Monitors over Western Canada only

Data Pooling: HMS-filter

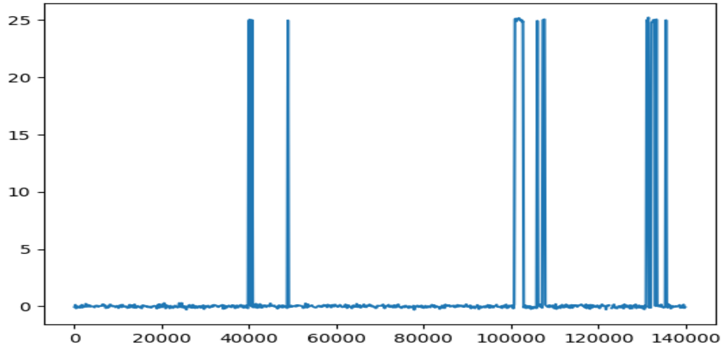
NOAA's Hazard Mapping System (HMS) Polygons

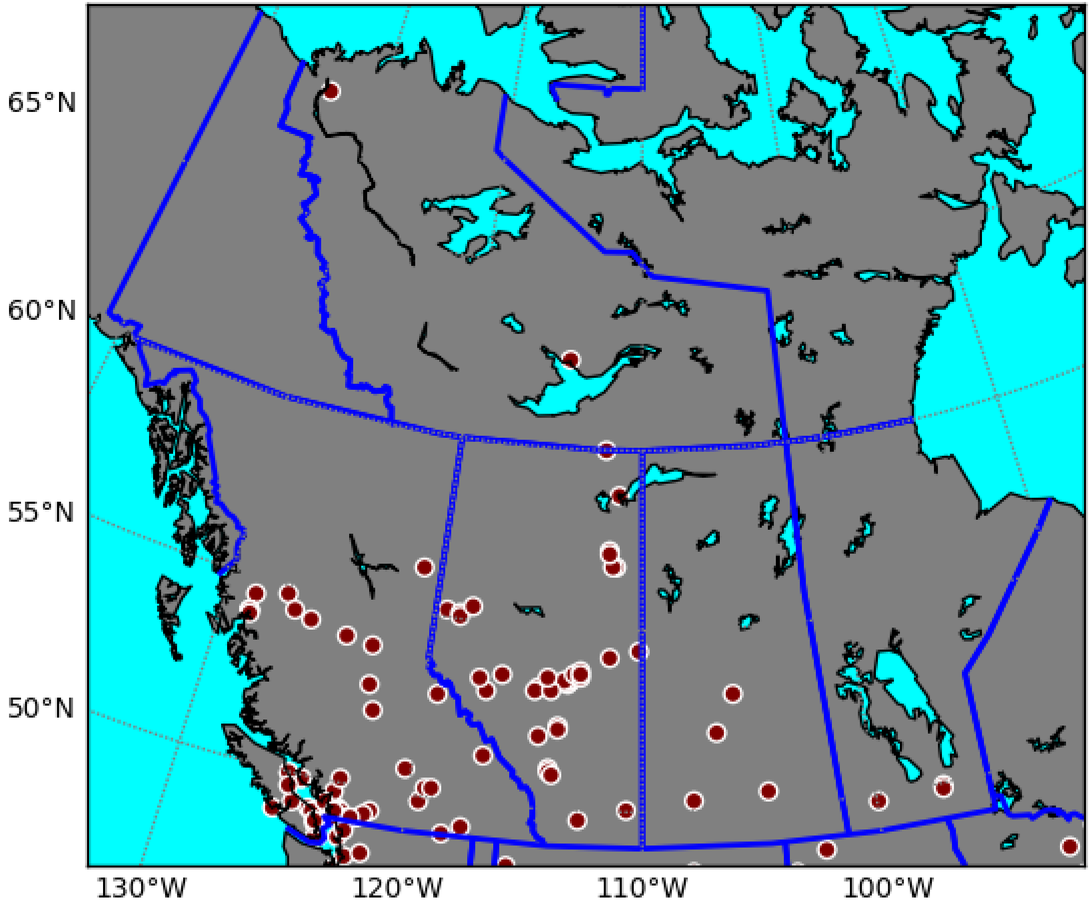


Data Pooling: PCA-filter

Principal Component Analysis

- Pairwise comparison of modified obs time series





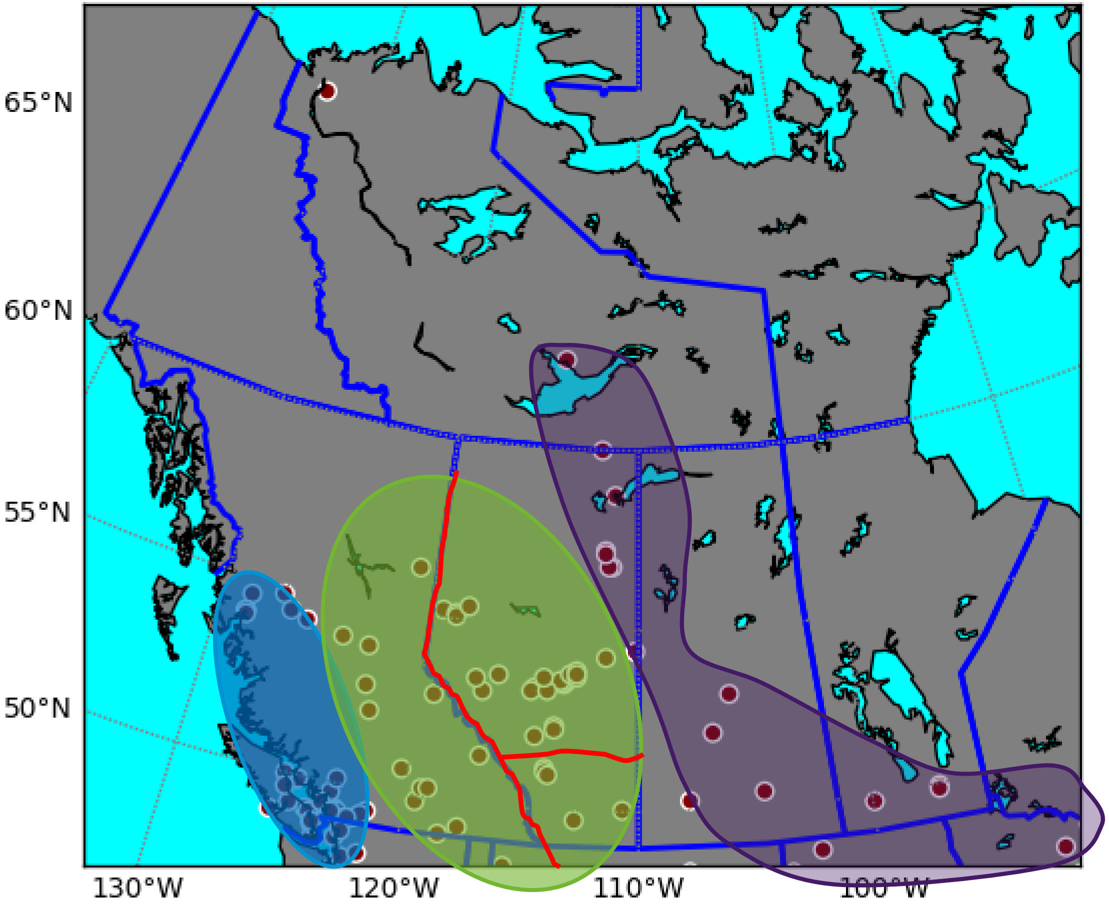
DATA ➜ PCA ➜ NETWORKS ➜ BOXPLOTS ➜ MixedEffects ➜ FINDINGS

“Networks”

Stations with similar loadings in each of the first 3 eigenvectors are grouped in “networks”

Three networks emerge:

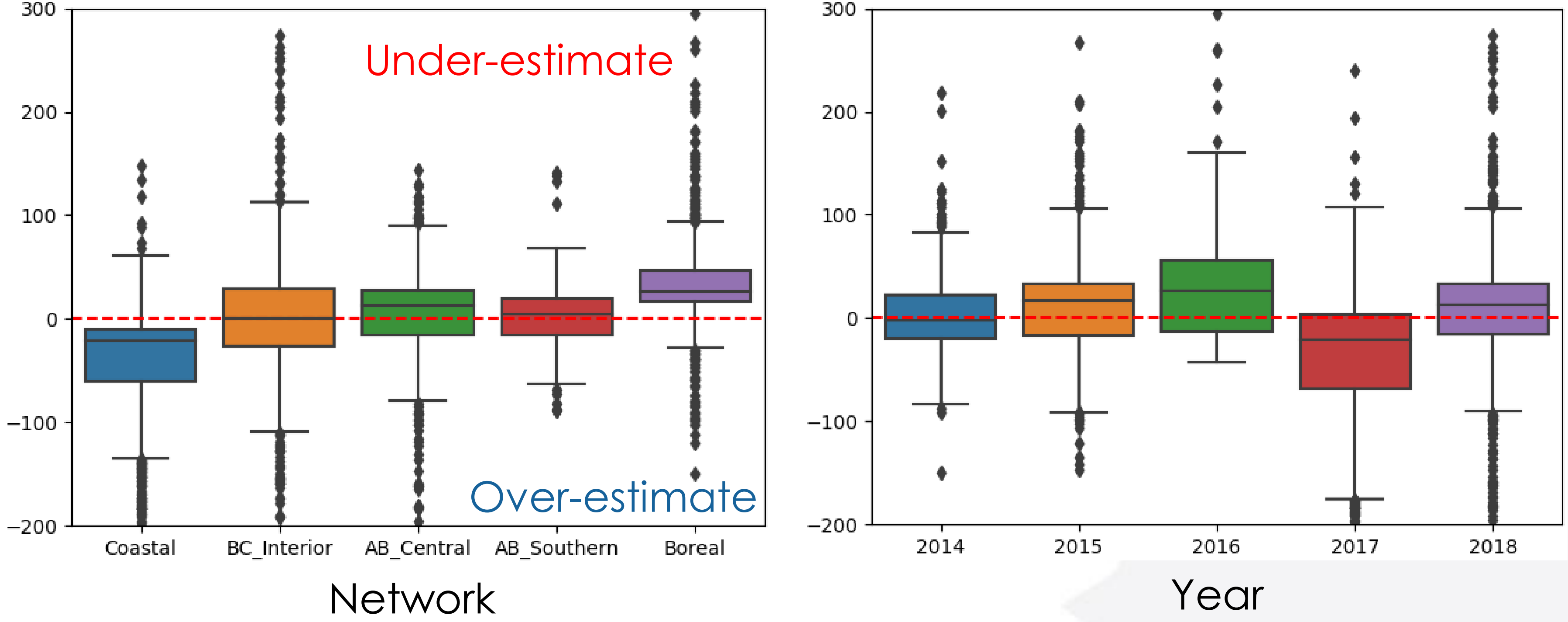
1. Coastal
2. Interior
    a. Int. BC
    b. S. AB
    c. Central AB
3. Boreal



DATA ➔ PCA ➔ NETWORKS ➔ BOXPLOTS ➔ MixedEffects ➔ FINDINGS

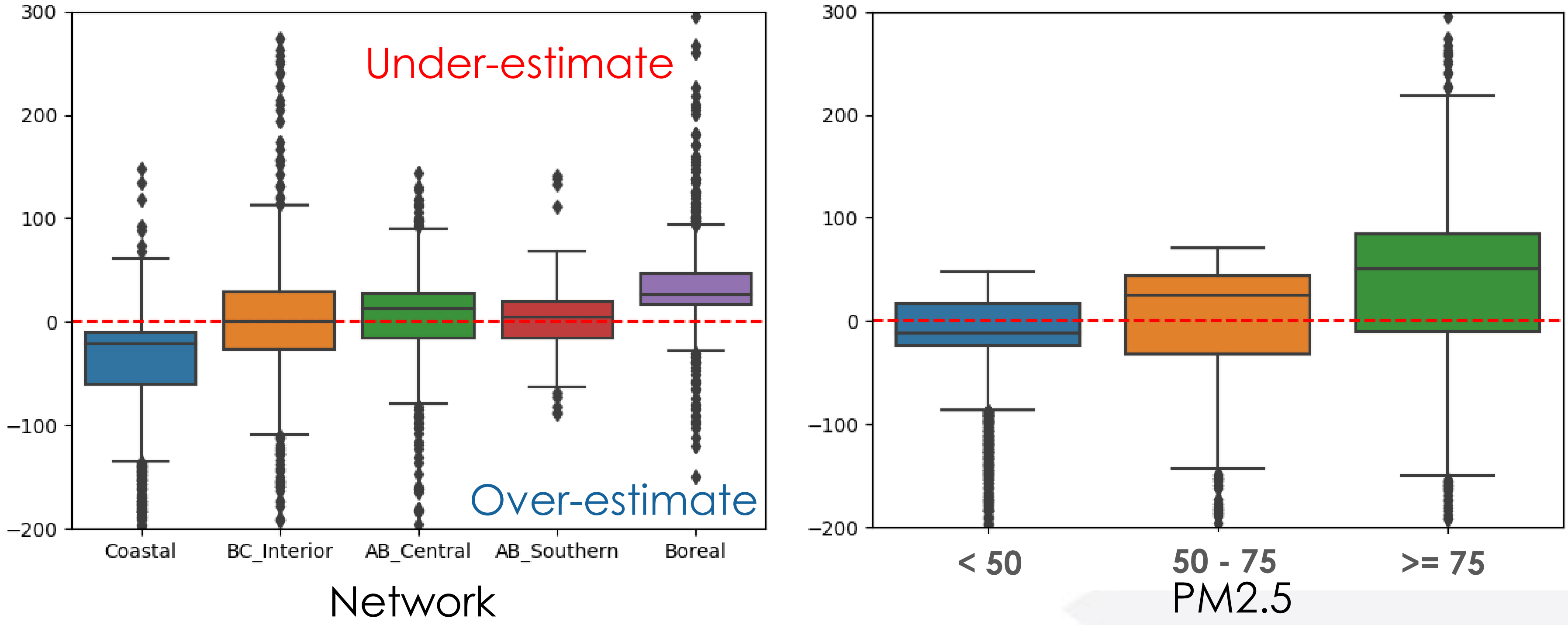
Measure-based model performance at the monitor level: Network-effect and Year-effect

24-hr avg Model Error (Obs – Mod; µg/m³)



Measure-based model performance at the monitor level: Network-effect and PM2.5-effect

24-hr avg Model Error (Obs – Mod; μg/m³)

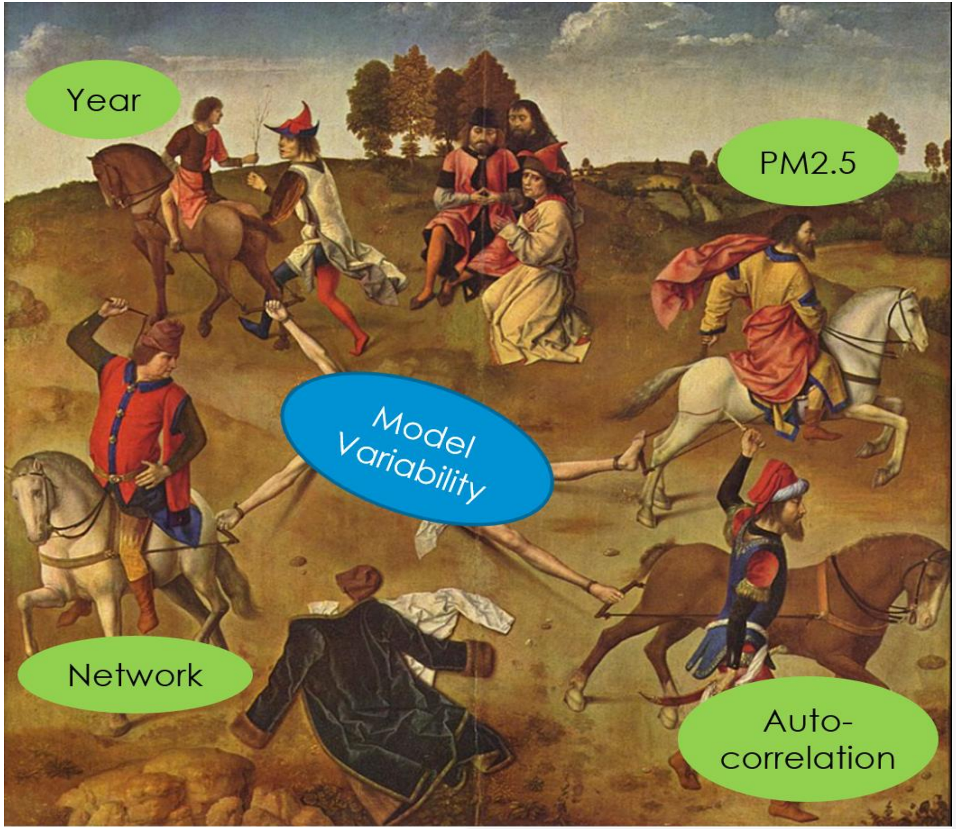


Statistical Analysis

Is there really a Year-effect?

Is the Network-effect just a manifestation of PM2.5-effect?

How to find out? → Torture the data till it confesses!



General Additive Model:

Model Error ~ Year + Network + s(PM2.5) + ARMA(2,0)

Findings:

- PM2.5 is a statistically significant factor in explaining model error across networks and year
- Year is not a significant predictor, **with the exception of 2017**
- Network is not significant factor

Implications:

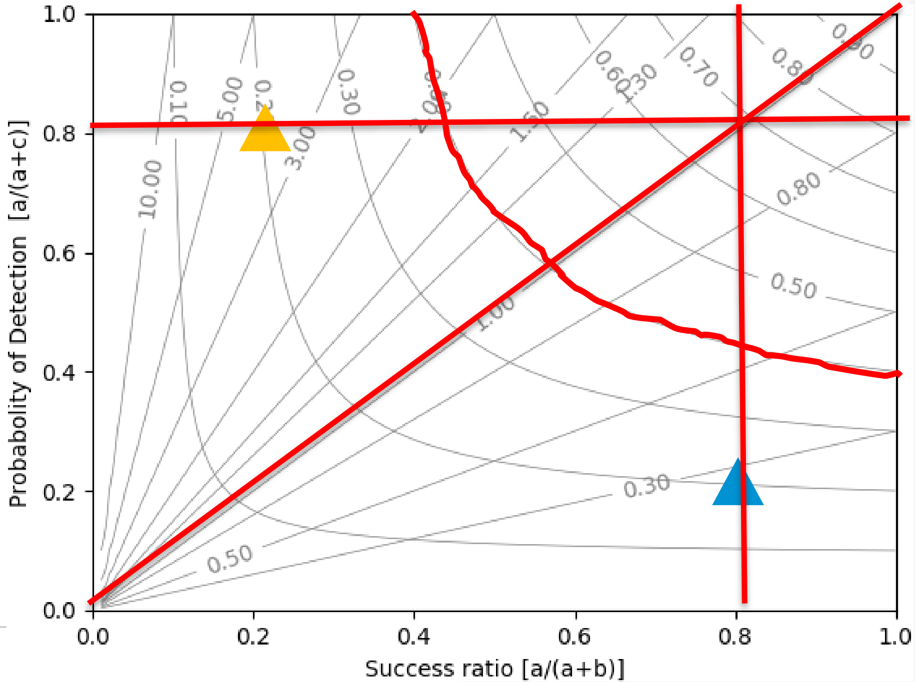
- Model bias changes with obs (or modeled) PM2.5 concentration.
- PM2.5 acting as a surrogate for proximity to fire?
- Significance of 2017 points to unaccounted for factor

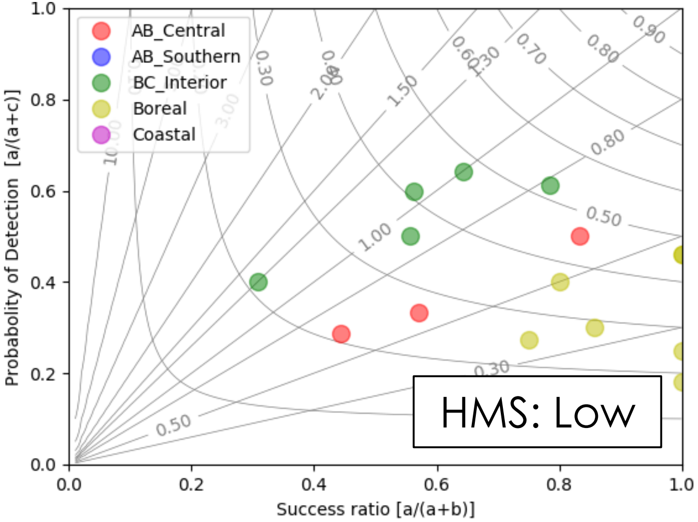
Contingency-based performance at the monitor level

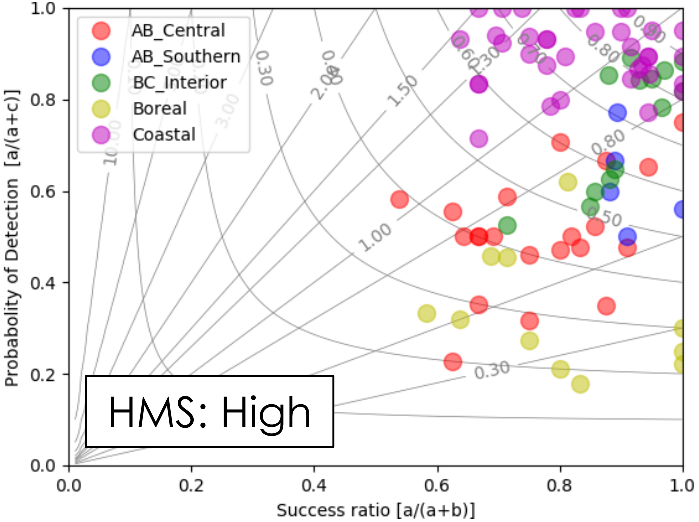
Observation

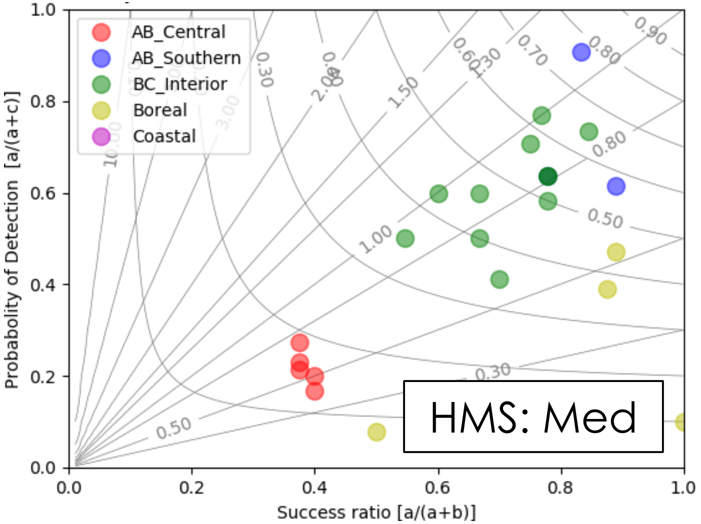
Model	Yes	No
Yes	Hit (a)	False Alarm (b)
No	Missed (c)	Cor. Neg (d)

- Performance Diagram (Roebber, 2009)
- Probability of Detection (PoD)
- Success Ratio (SR)
- Bias (under/over forecast)
- CSI - Accuracy when correct negatives are removed









Findings:

- Network groupings are more distinct under LOW and MED conditions
- All network seems to have improved perf. during HIGH smoky days
- More variability within network under HIGH HMS days

Road Map- II – Forecast zone based

User survey → evaluate model over MSC zones

To re-do the analysis at the FX-zone level, gridded 'obs' are needed

But most MSC zones have *no* monitors!

That's okay –lets use satellite data and machine learning

Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level $PM_{2.5}$ ☆

Yongming Xu [a], Hung Chak Ho [b,*], Man Sing Wong [b,c], Chengbin Deng [d], Yuan Shi [e], Ta-Chien Chan [f], Anders Knudby [g]

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1029/2018JD028573

Satellite-Based Daily $PM_{2.5}$ Estimates During Fire Seasons in Colorado

Guannan Geng[1], Nancy L. Murray[2], Daniel Tong[3,4,5], Joshua S. Fu[6,7], Xuefei Hu[1], Pius Lee[3], Xia Meng[1], Howard H. Chang[2], and Yang Liu[1]

Key Points:
- A Bayesian ensemble model was used to predict the daily $PM_{2.5}$

Environmental Science & Technology

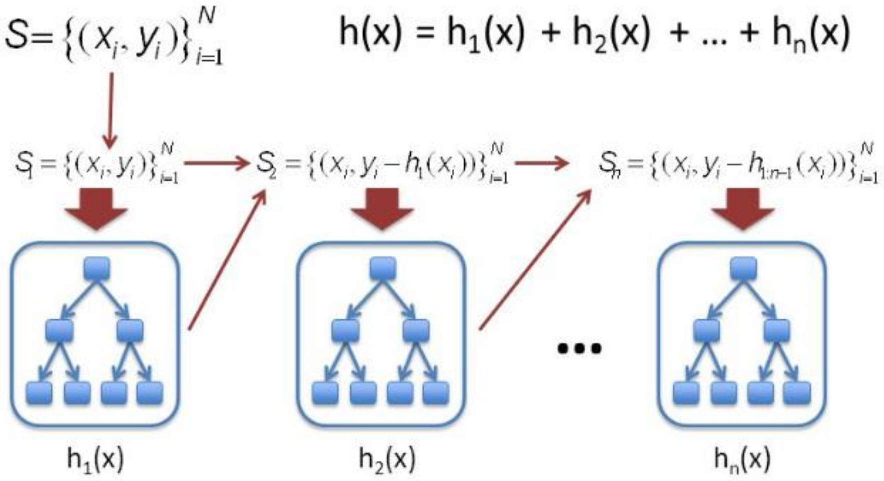
Article

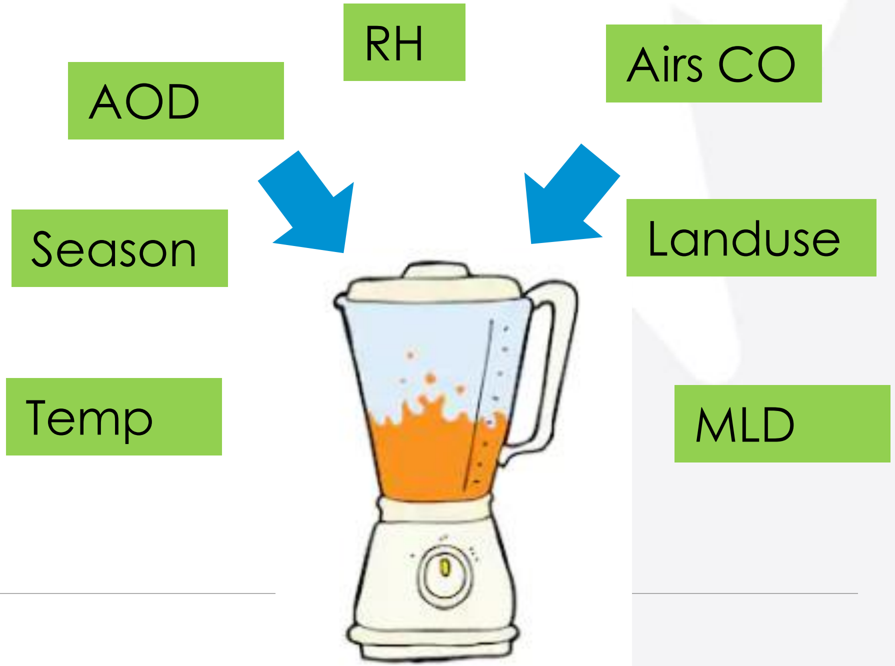
pubs.acs.org/est

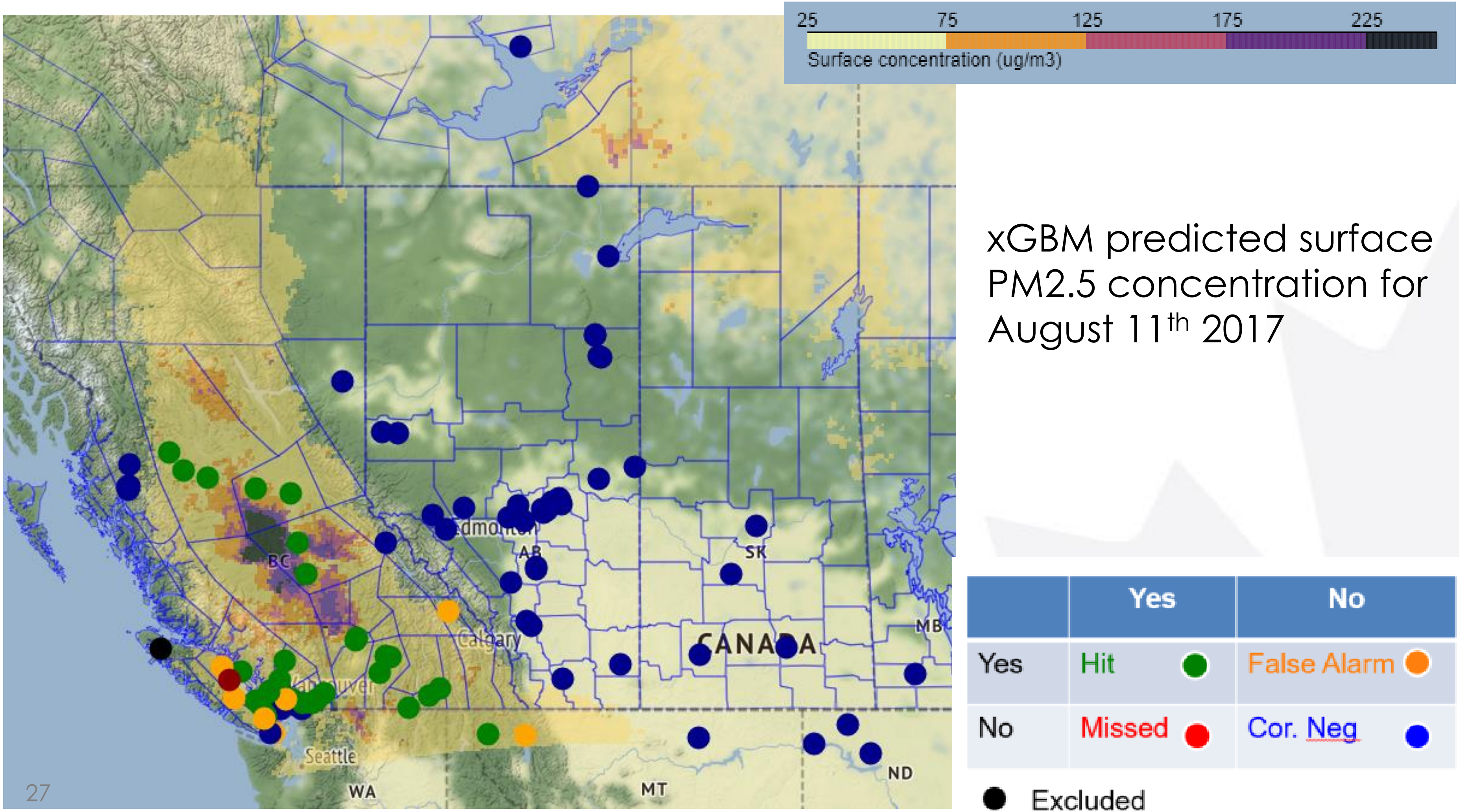
Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning

Gradient Boosting Machine

- Use machine learning to downscale satellite data construct daily 24-hr PM2.5
- GBM predictics in the form of an ensemble of simple models
- It builds model in a stage-wise fashion



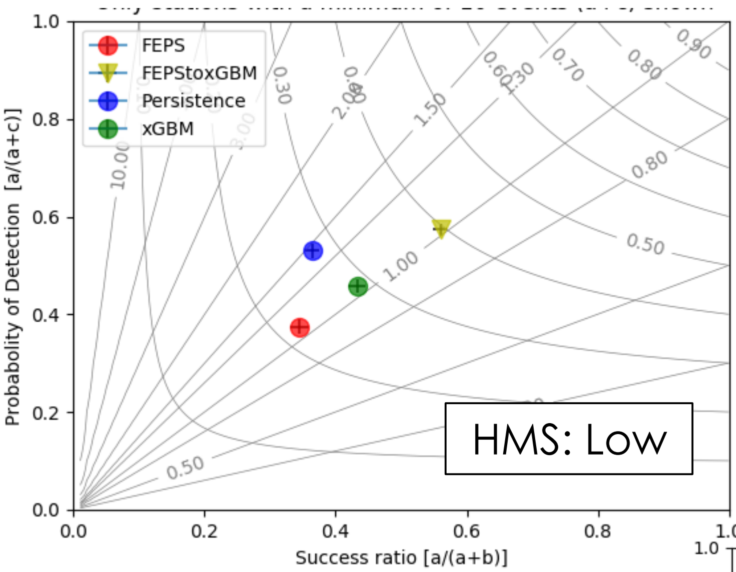


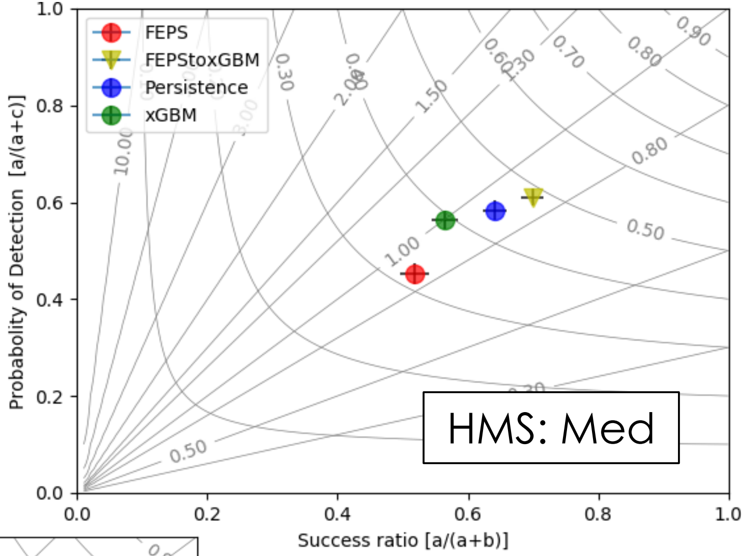


xGBM predicted surface PM2.5 concentration for August 11th 2017

	Yes	No
Yes	Hit	False Alarm
No	Missed	Cor. Neg

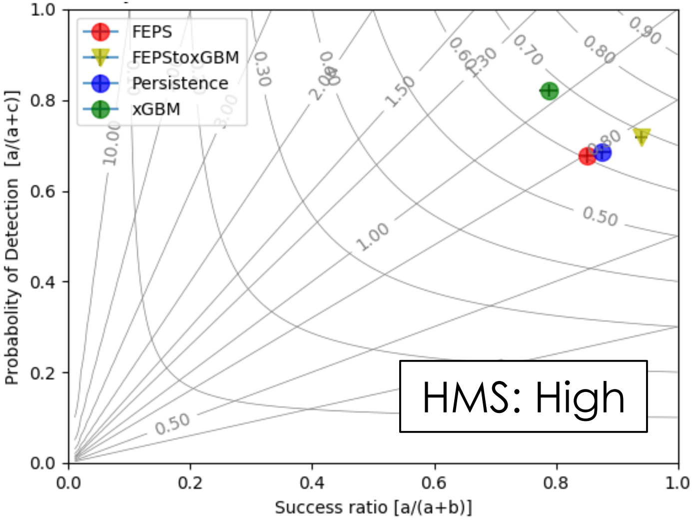
Excluded

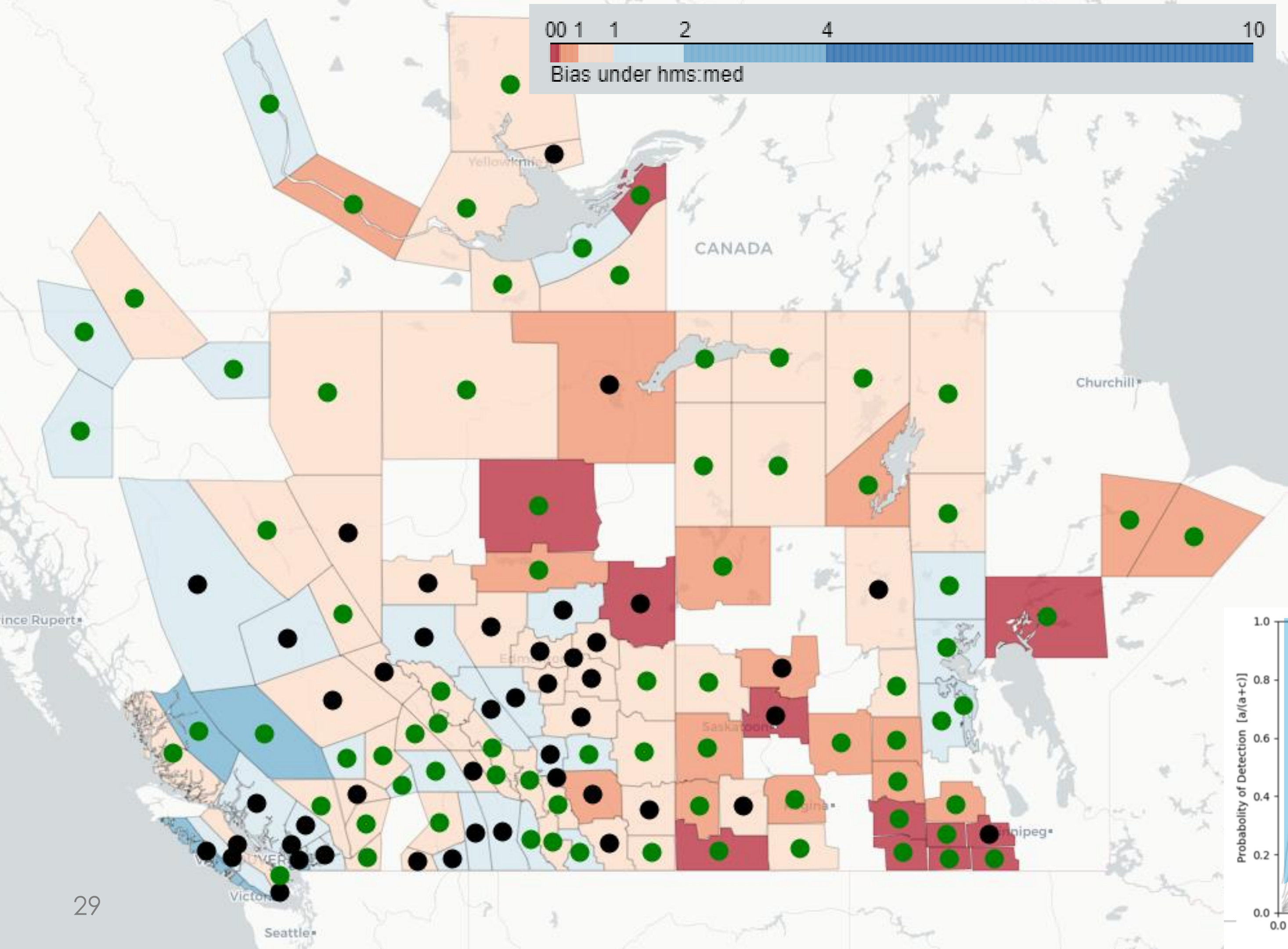




Observation

Model	Yes	No
Yes	Hit (a)	False Alarm (b)
No	Missed (c)	Cor. Neg (d)

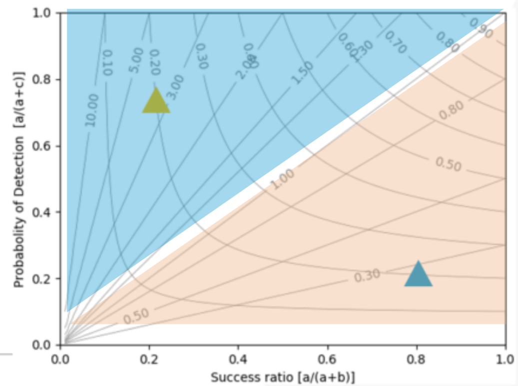




Spatial distribution of Contingency-Bias based under HMS Medium conditions between 2014-2018

Unmonitored

Monitored



Conclusion

Developing framework to analyze model performance from a wild fire smoke forecast model that:

- looks at performance at the monitor- and forecast region-level
- Is designed from the forecasters perspective
- Takes into account importance of events
- Is aware of class imbalance

Future Work

- Improve xGBM to allow comparisons with contingency SR, PoD and CSI as well as zone-averaged rmse, bias, etc.
- Develop ways of measuring forecast “value”
- Examine spatio-temporal relationship between predicted PM2.5 and estimated fire emissions.
- Apply methods to current operational model