FireWork evaluation from a forecaster point of view

Bruce Ainslie & Rita So

Environment & Climate Change Canada

NW-AIRQUEST 2020 Annual Meeting

June 10, 2020

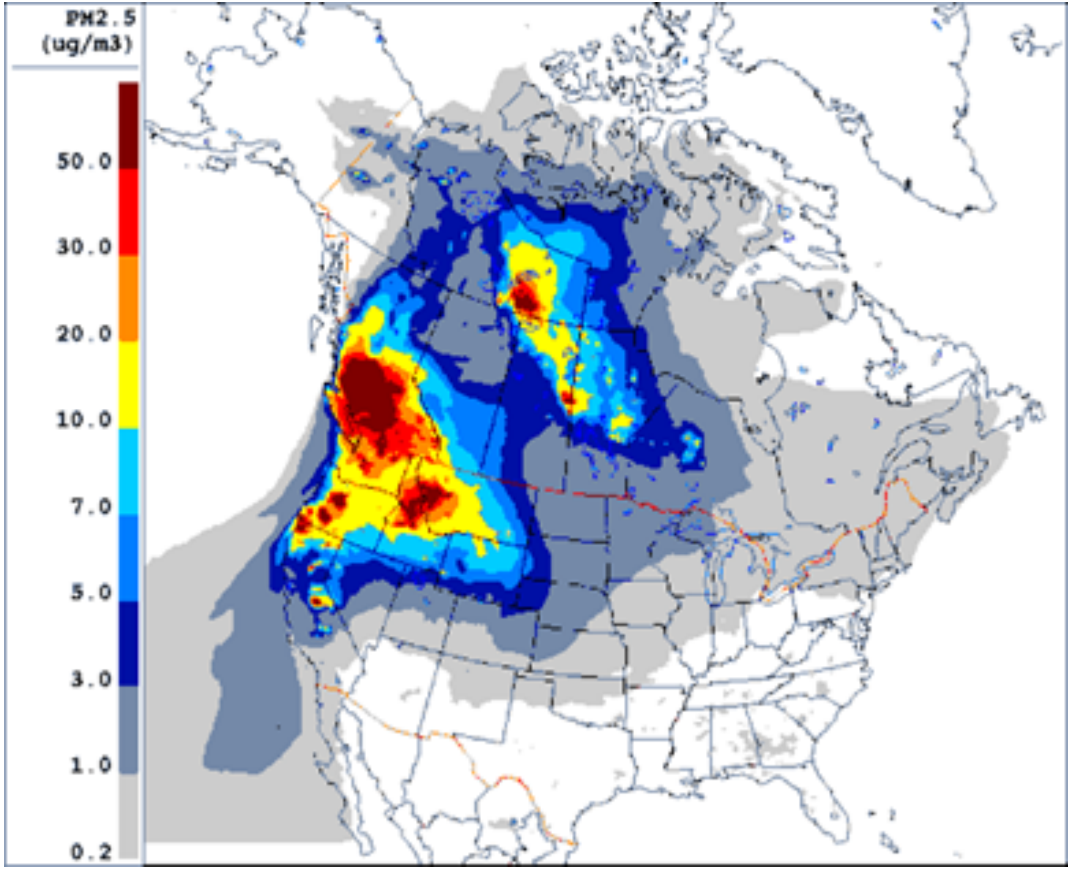
Vancouver Island
Vancouver

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

ECCC's FireWork System

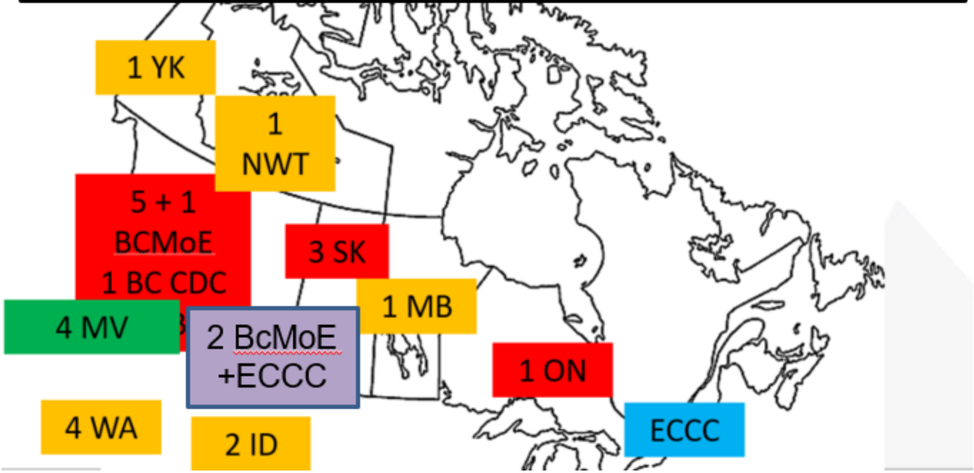
https://weather.gc.ca/firework

- Twice daily (00z/12z) during fire season (Apr- Oct)
- Incorporated into the ECCC Regional Air Quality Deterministic Prediction System
- NRT fire data from Canadian Wildland Fire Information System (based on NOAA/NASA satellite info.)
- Products:
  - PM2.5 & PM10 maps
  - AQ point forecasts
  - 24-hr accumulated PM2.5



How to evaluate performance?
How do we make use of forecast guidance?

I. Informal meeting with MV
II. https://www.surveymonkey.com/r/ZKXHJG8 (original)
III. Teleconference with ECCC
IV. https://www.surveymonkey.com/r/FL7KKT9 (updated)
V. Informal face to face meetings with forecasters



FireWork User Survey

FireWork User Survey - Takeaways

Survey shows emphatically **model guidance is useful**

Four key **perspectives**:

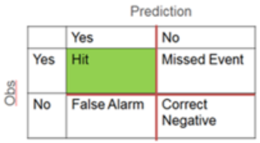
I. Jurisdiction-based, but never at a monitor

II. Event-based: **25 ug/m3**

III. Missing an event >> False alarm

IV. 1st day forecast (0 – 24 hrs) is the MOST IMPORTANT

+



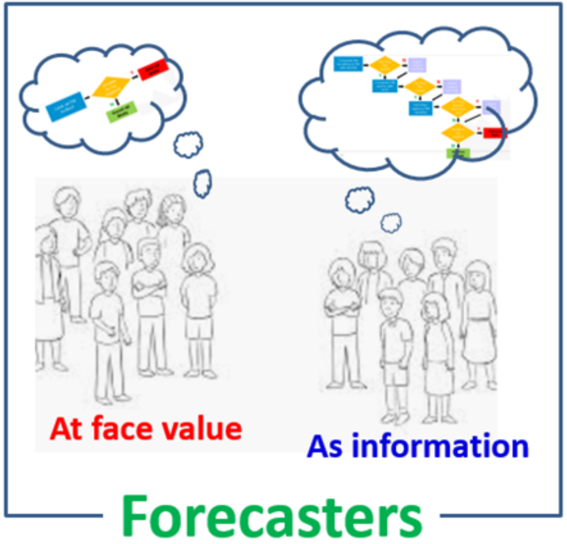
		Prediction	
		Yes	No
Obs	Yes	Hit	Missed Event
	No	False Alarm	Correct Negative

@

Two **types** of users:

I. Take forecast "at face value"

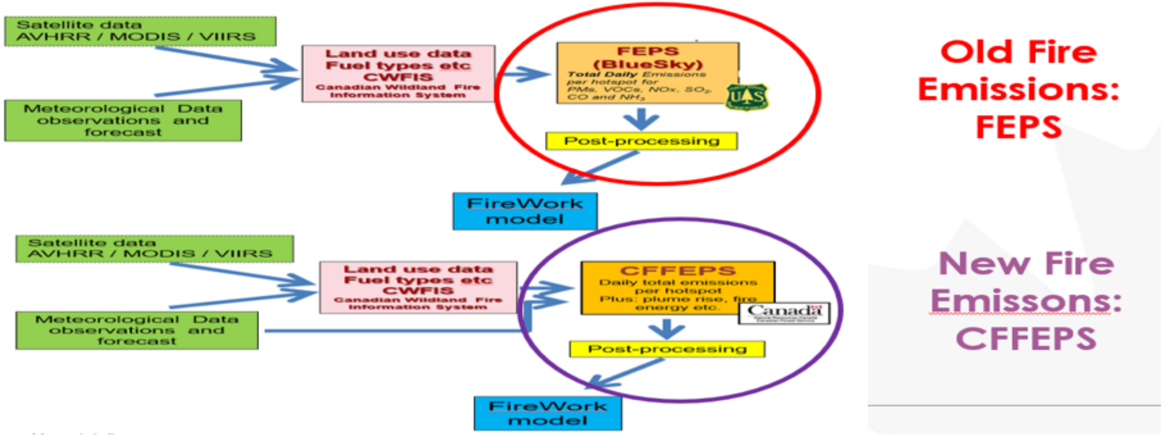
II. Take forecast "as information"



Model Evaluation Framework

Use survey results to design evaluation framework and test ECCC's old **(FEPS)** and new (CFFEPS) operational wild fire smoke models over 2016-2018 fires seasons

Meteorology (GEM) + Fire Emissions → FireWork



Satellite data AVHRR / MODIS / VIIRS

Meteorological Data observations and forecast

Land use data Fuel types etc CWFIS Canadian Wildland Fire Information System

FEPS (BlueSky) *Total Daily Emissions per hotspot for PMs, VOCs, NOx, SO_2, CO and NH_3*

Post-processing

FireWork model

Old Fire Emissions: FEPS

Satellite data AVHRR / MODIS / VIIRS

Meteorological Data observations and forecast

Land use data Fuel types etc CWFIS Canadian Wildland Fire Information System

CFFEPS Daily total emissions per hotspot Plus: plume rise, fire energy etc.

Canada

Post-processing

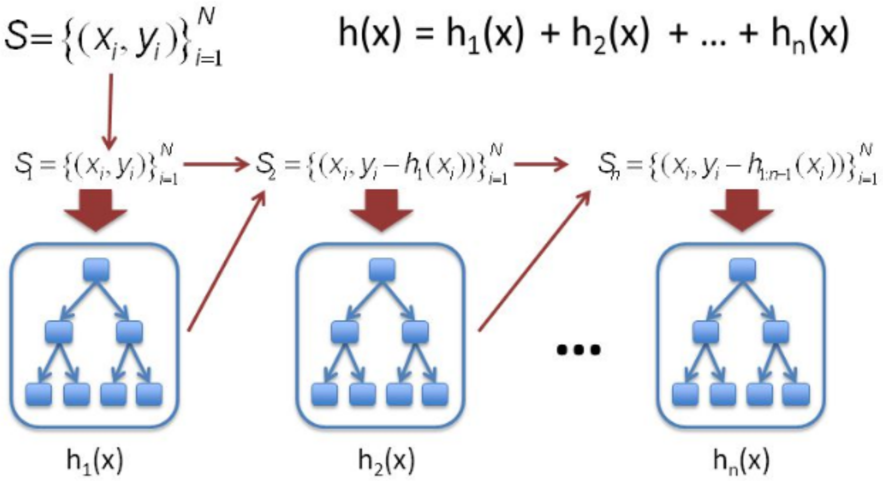
FireWork model

New Fire Emissons: CFFEPS

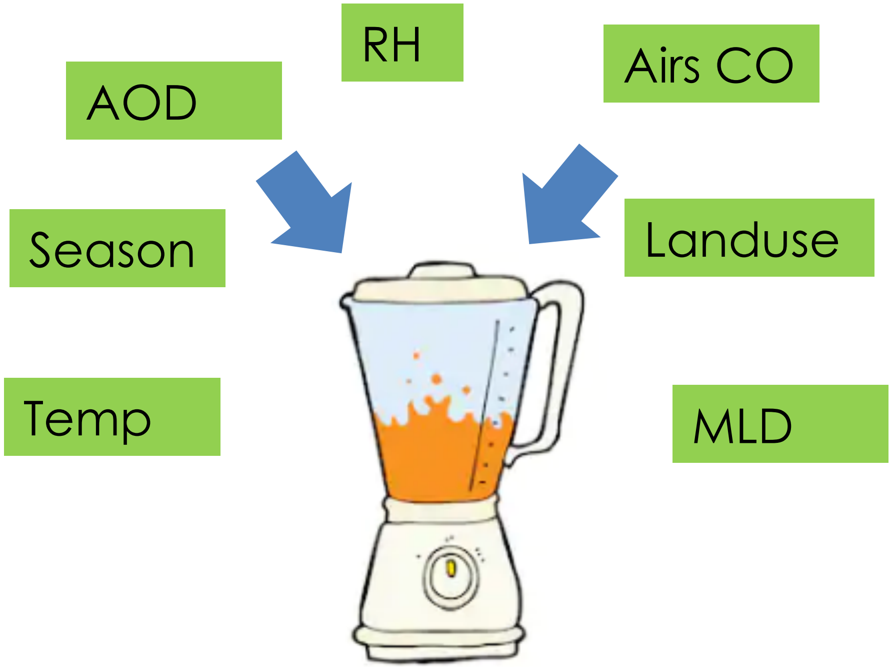
Gridded Spatial Observation dataset

Detour: Extreme Gradient Boosting (xGBM)

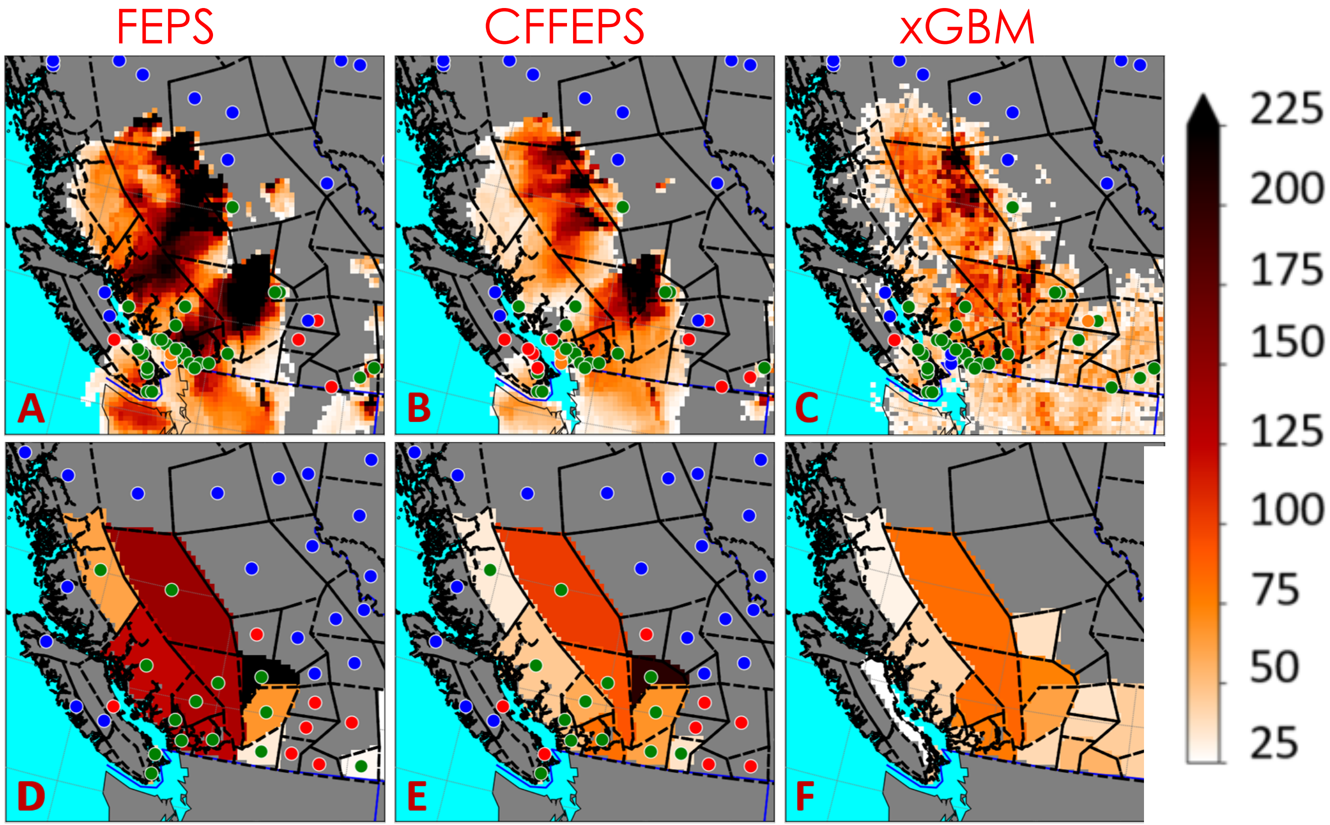
- Use machine learning to downscale satellite data any many other dataset to construct daily 24-hr PM2.5
- GBM predicts in the form of an ensemble of simple models
- It builds model in a stage-wise fashion

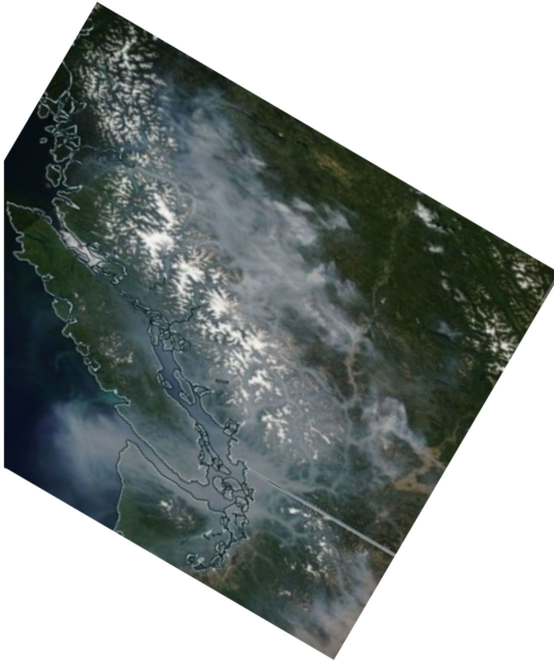


$S=\{(x_i, y_i)\}_{i=1}^{N}$
$h(x) = h_1(x) + h_2(x) + \ldots + h_n(x)$
$S_1=\{(x_i, y_i)\}_{i=1}^{N}$
$S_2=\{(x_i, y_i-h_1(x_i))\}_{i=1}^{N}$
$S_n=\{(x_i, y_i-h_{1:n-1}(x_i))\}_{i=1}^{N}$
...
$h_1(x)$
$h_2(x)$
$h_n(x)$



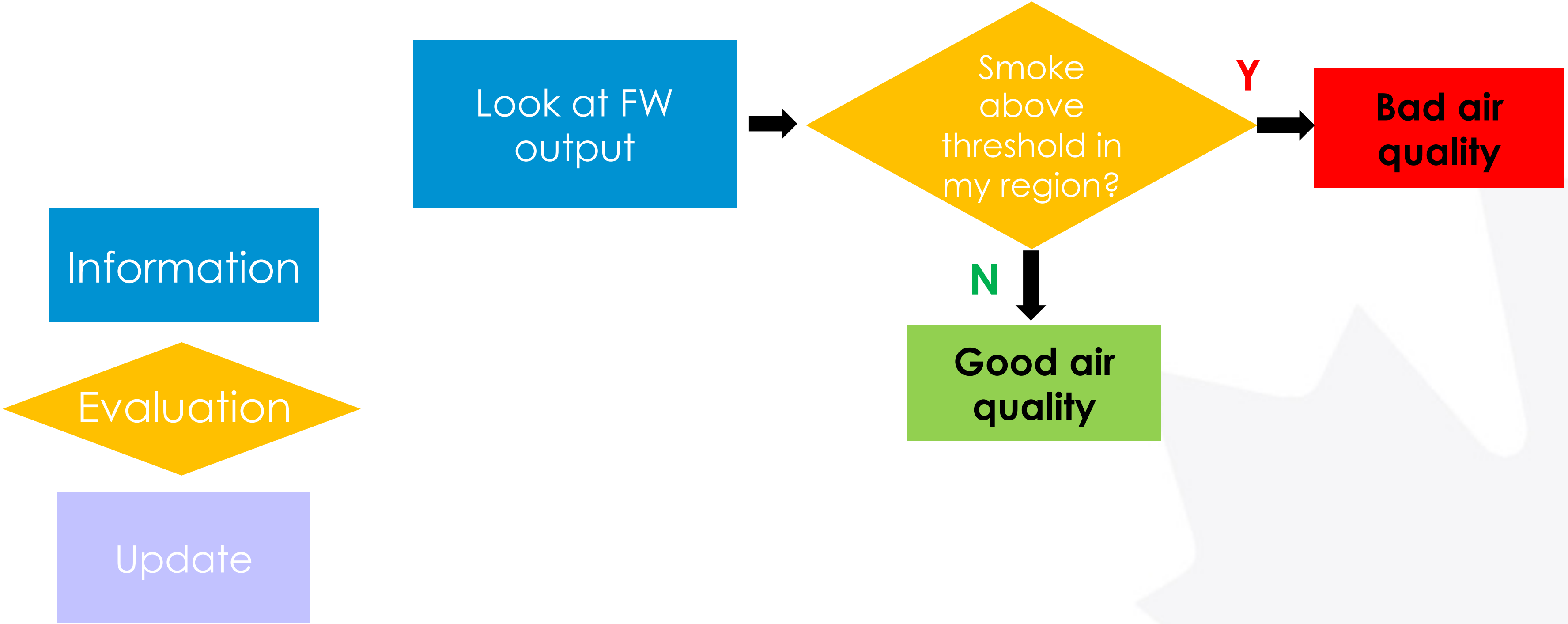
Detour: Spatial Footprints August 2nd 2017





Hit ●	False Alarm ●
Missed ●	Cor. Neg ●

Model Evaluation I:
“At face value” decision making process



Look at FW output

Smoke above threshold in my region?

Y

Bad air quality

N

Good air quality

Information

Evaluation

Update

At face value event-based verification: contingency table

Event-based + at face value → use standard 2x2 contingency metrics

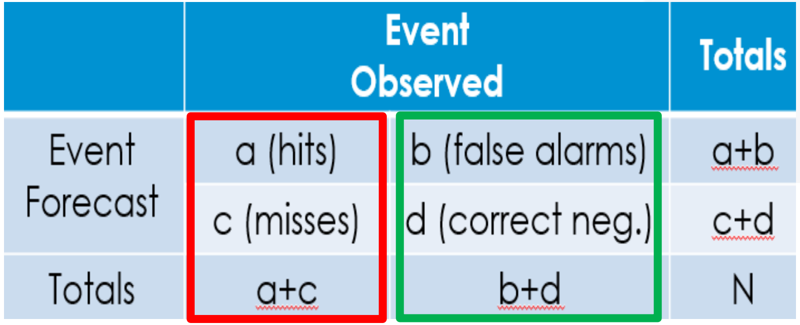
Would like metric(s) to be:

1. Independent of **forecast situation** (e.g. separate ***forecast system*** from ***forecast situation***)
2. Punish MISS more than FA

→ Unfortunately, treated misses differently than false alarms is **value** not a **quality**

Peirce Skill Score

- Large number of ways of scoring contingency table
- Table can be expressed in terms of 3 parameters:
  1. **H: hit rate a/(a+c)**
  2. **F: false alarm rate b/(b+d)**
  3. **s: base rate (a+c)/N**



	Event Observed		Totals
Event Forecast	a (hits)	b (false alarms)	a+b
	c (misses)	d (correct neg.)	c+d
Totals	a+c	b+d	N

We have chosen **PSS** = **H** – **F** since:

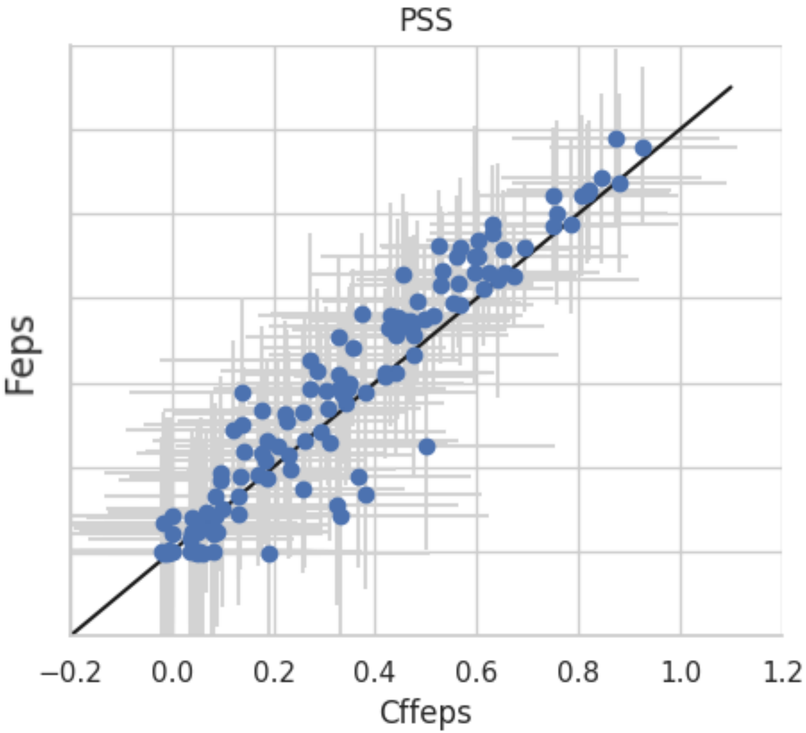
1. independent of base rate
2. Provides a link between forecast quality and value

PSS: CFFEPS vs FEPS

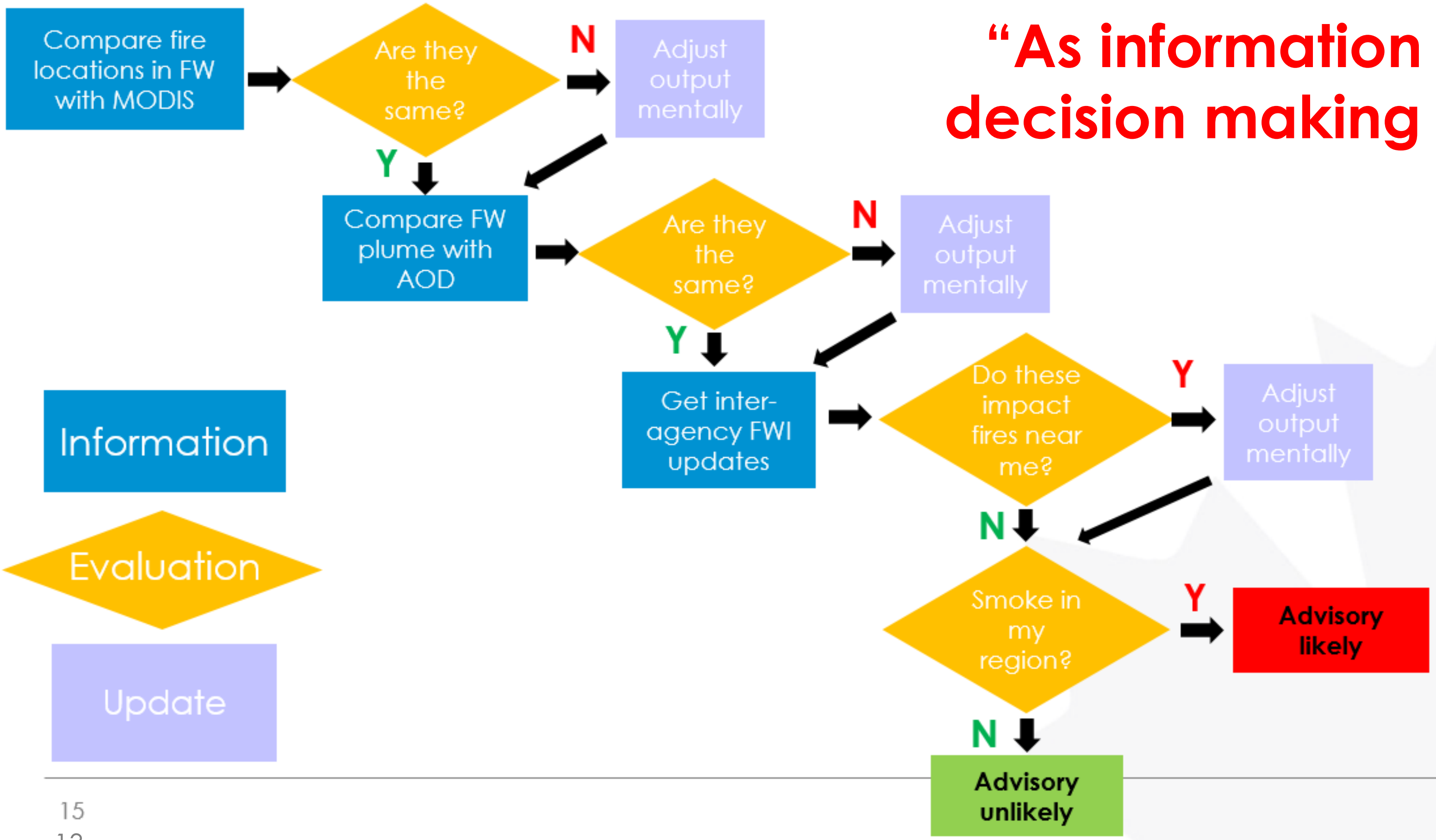
by MSC forecast zone

Wilcoxon signed-rank test suggests the two estimates for MSC Zone PSS scores do not come from the same population (p<0.01).

➢ Feps provided more 'value' than Cffeps

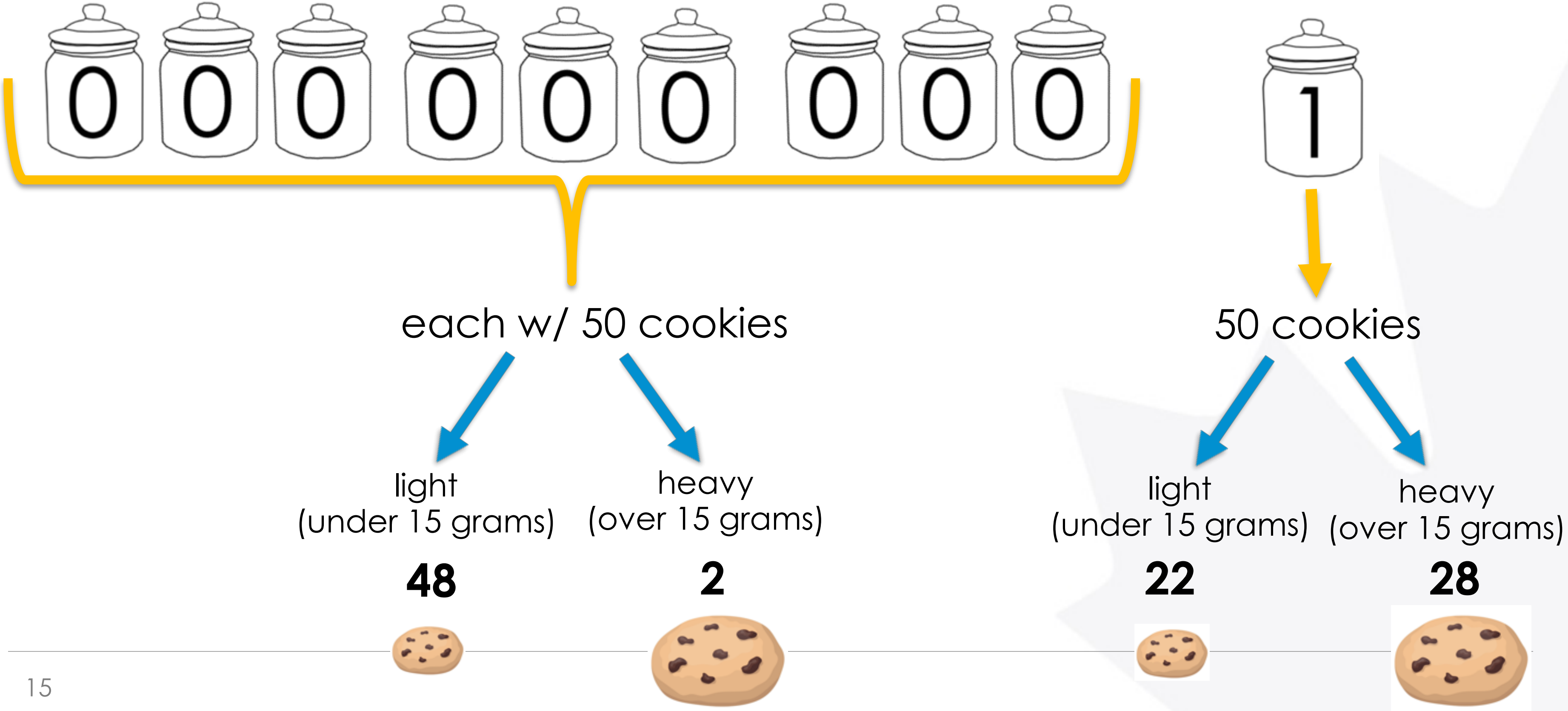


Model Evaluation II: "As information source" decision making process



Bayesian Refresher

Detour: Which cookie jar?



each w/ 50 cookies

light
(under 15 grams)

48

heavy
(over 15 grams)

2

50 cookies

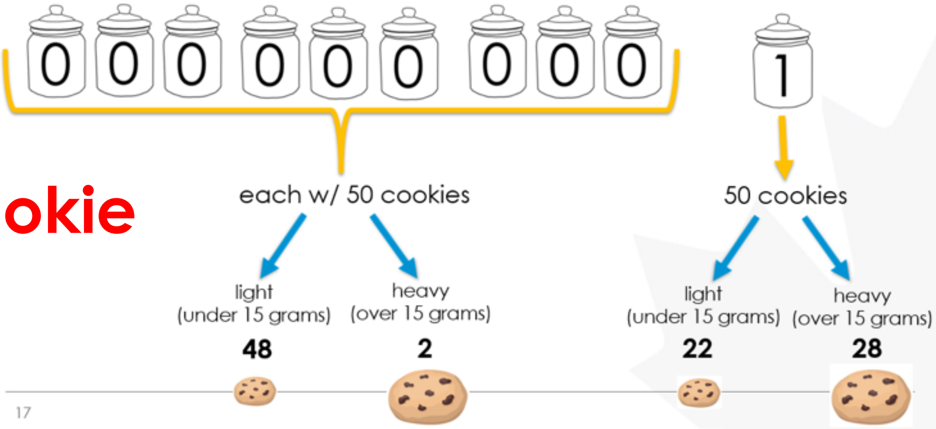
light
(under 15 grams)

22

heavy
(over 15 grams)

28

A single cookie is taken from one of the ten jars.



What is the probability the cookie was drawn from jar 1?

$$p(J=1) = 0.1$$

Now, if I tell you the drawn cookie weighted over 15 grams –

how does this information help refine your guess?

$$\mathbf{p(J=1 \mid heavy)} = p(heavy \mid J=1)/p(heavy) \quad x \quad p(J=1)$$

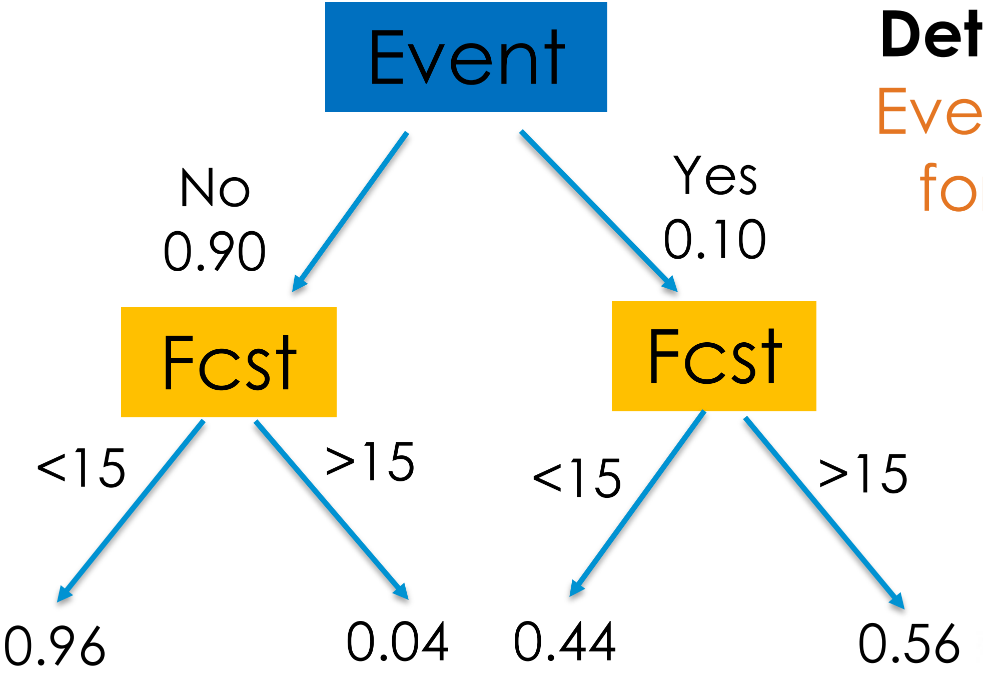
$$= (28/50)/(0.1*28/50 + 0.9*2/50) \quad x \quad 0.1$$

$$\mathbf{0.61 = 6.1 \ x \ 0.1}$$

Updated guess

New information

Initial guess



Detour: Chance of an Event today given the forecast PM2.5 was greater than 15 ug/m3?

p(**Event** | **Fcst>15**) = p(Fcst>15 | Event)/p(Fcst>15)x BaseRate
= 6.1 x 0.1
= **0.61**

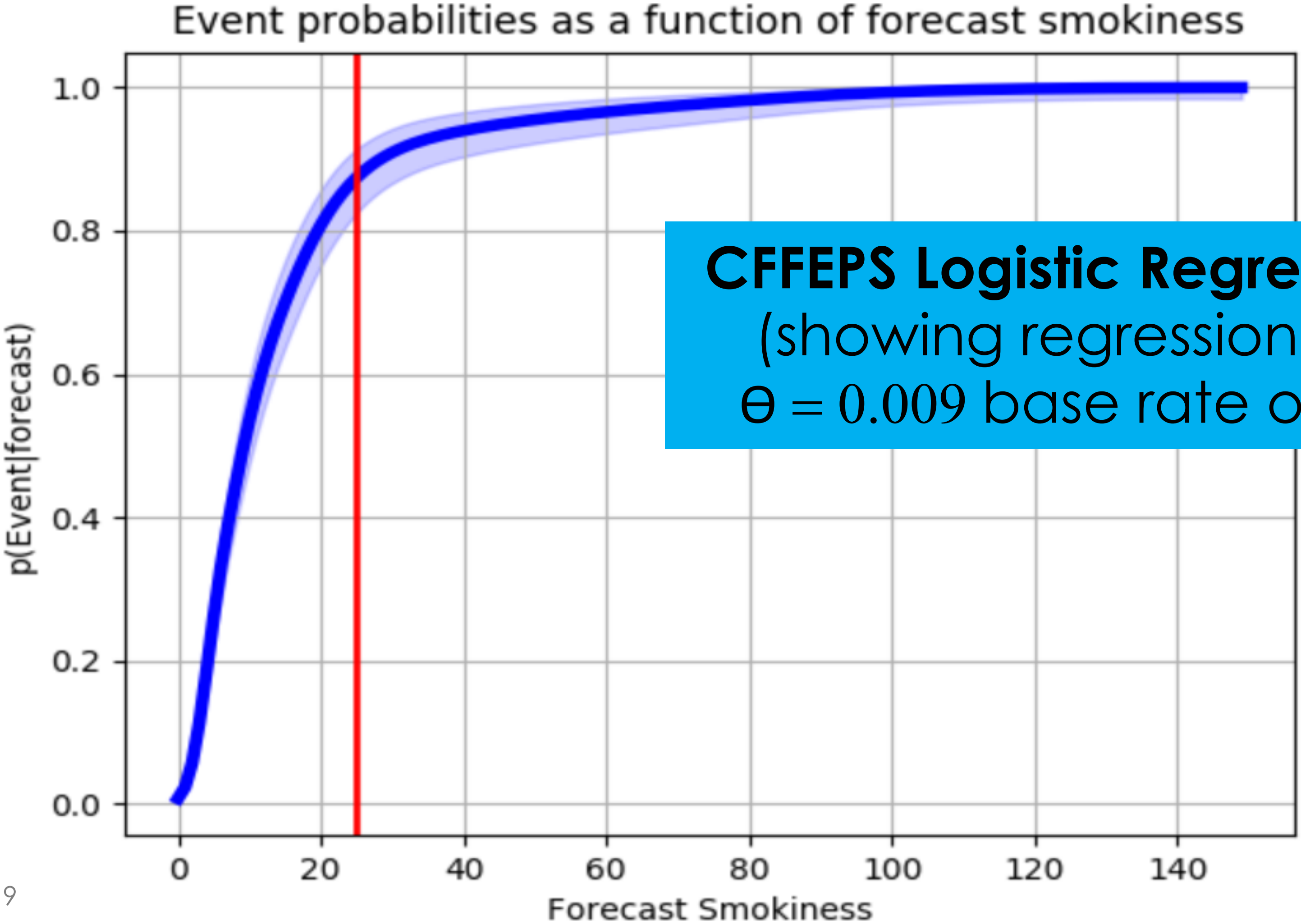
As Information source: convert forecast → p(Event)

Binomial (Logistic) Regression

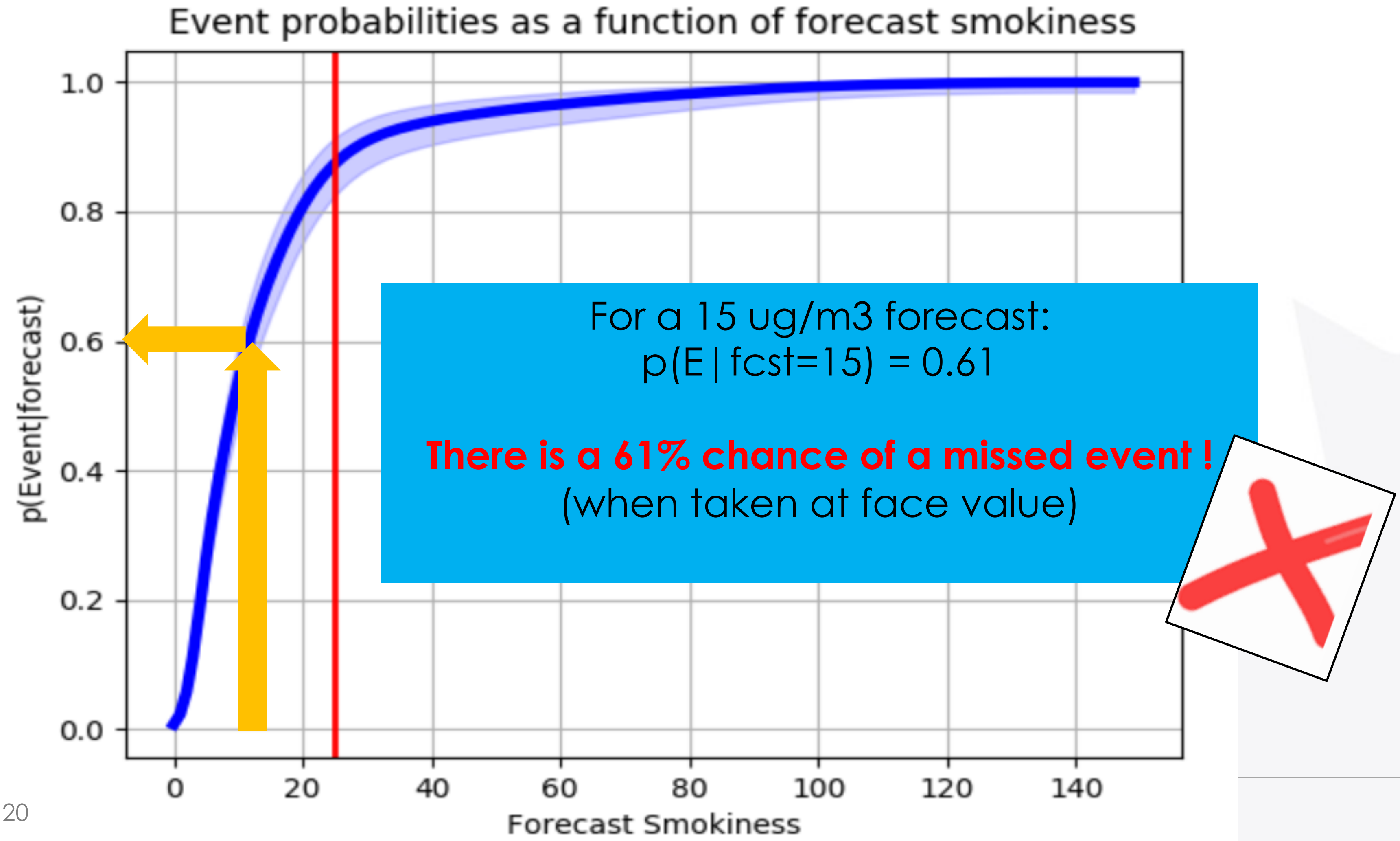
$$P(Event|fcst)\sim\frac{1}{1+exp\{-1[bo+b1*conc+b2*\Theta+b3*met]\}}$$

Our regression using GAM

$$\log\left(\frac{P}{1-P}\right)\sim f(conc)+g(\Theta)+j(NCEP)$$

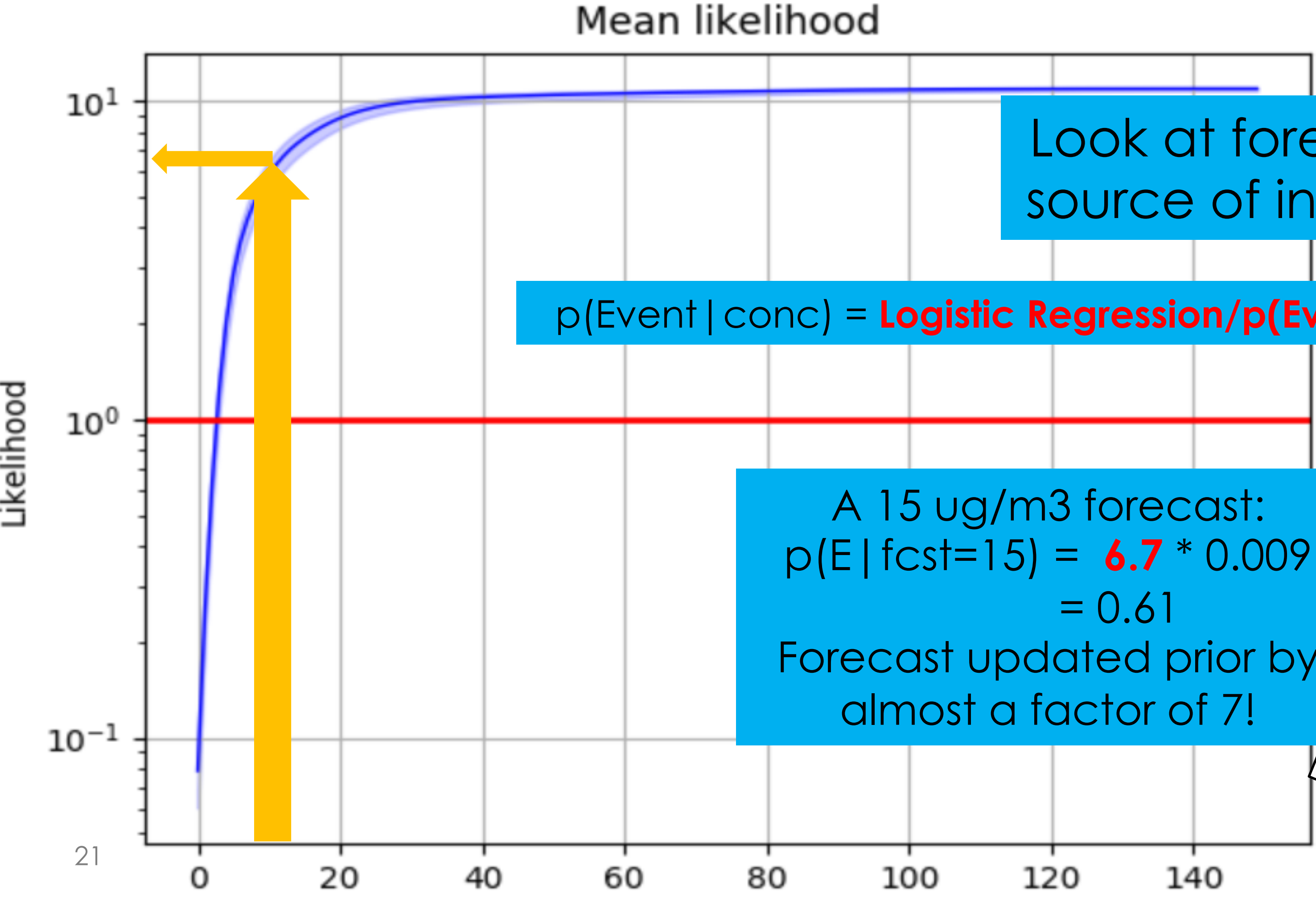


CFFEPS Logistic Regression
(showing regression for $\theta = 0.009$ base rate only)



For a 15 ug/m3 forecast:
p(E|fcst=15) = 0.61

There is a 61% chance of a missed event !
(when taken at face value)



Look at forecast as a source of information

p(Event|conc) = **Logistic Regression/p(Event)** * p(Event)

A 15 ug/m3 forecast:
p(E|fcst=15) = **6.7** * 0.009
= 0.61
Forecast updated prior by almost a factor of 7!

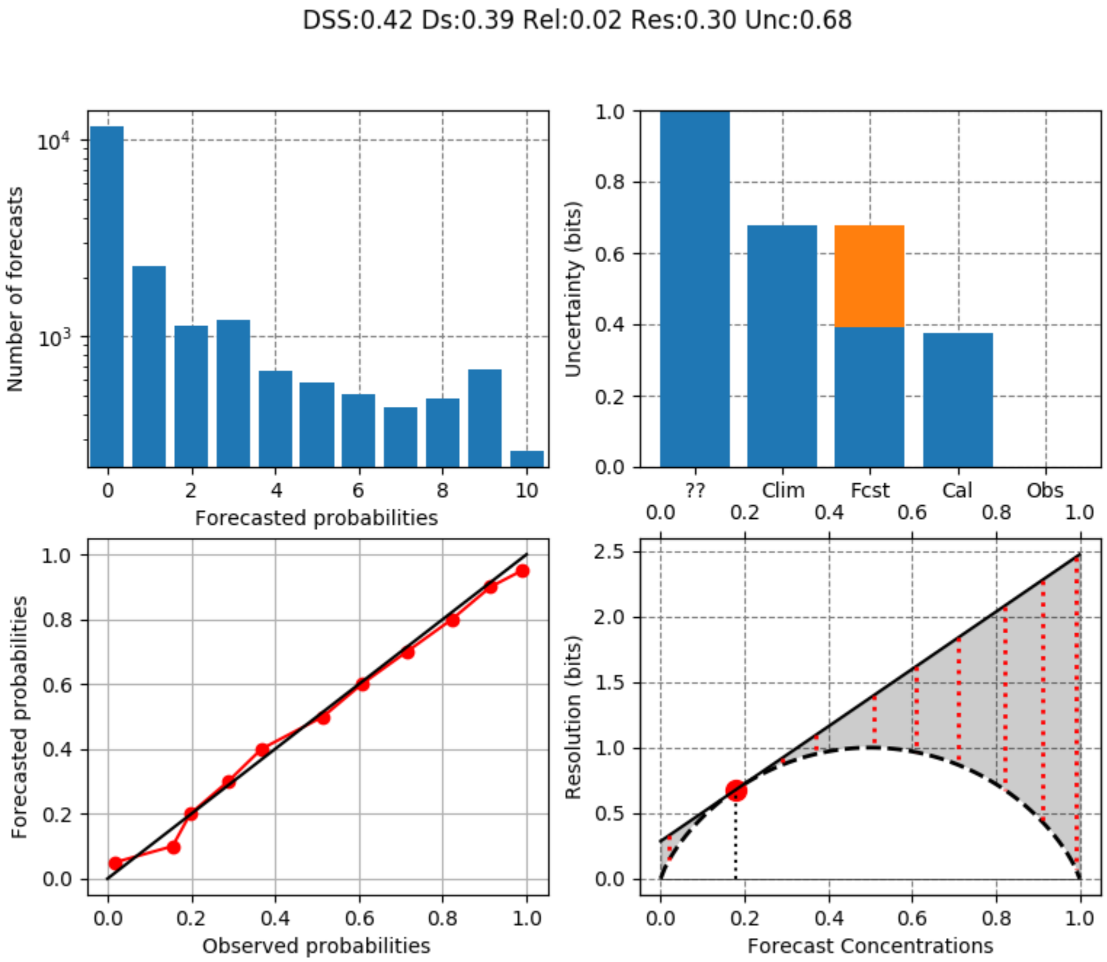
✓

Question:
How to tell which model supplies more information?

1. Ranked Mutual Information Skill Score (RMIS)
2. Divergence Skill Score (DSS)
3. Ignorance Skill Score (IGN)

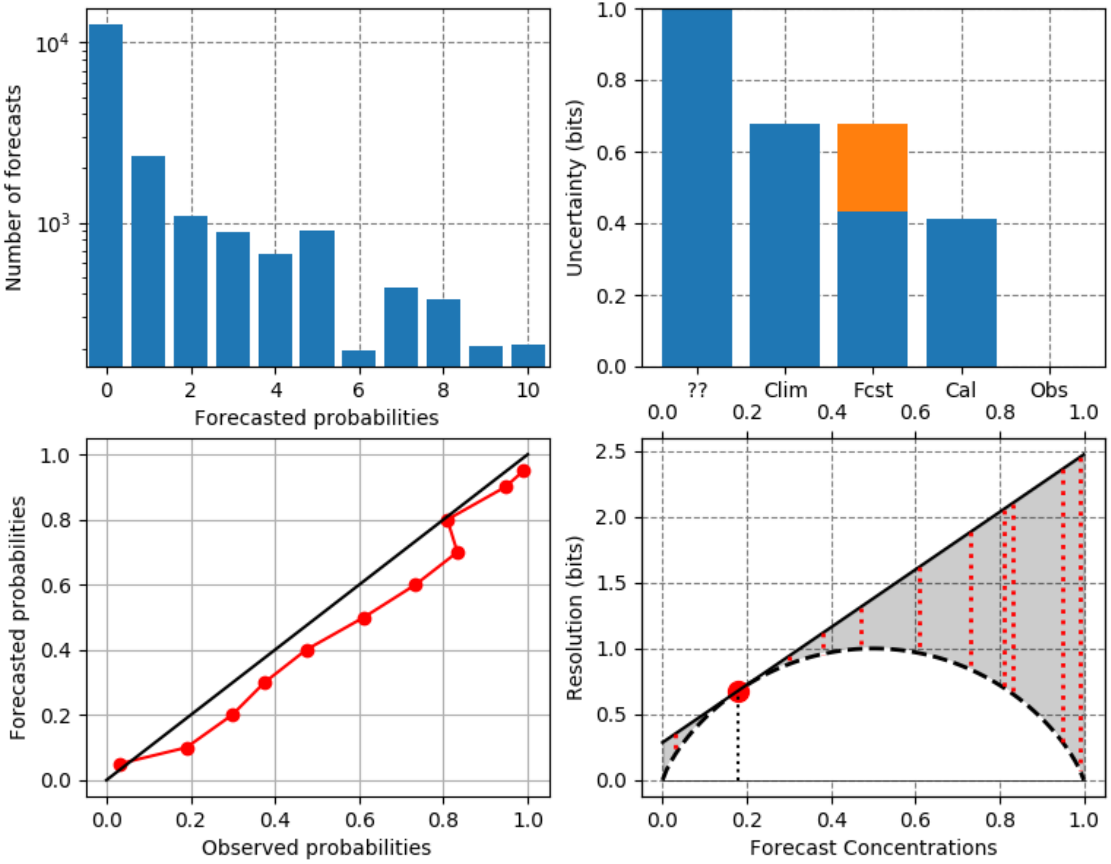
Measures information in forecast conveyed to user

→ how much more you know about the future after seeing the forecast, but starting from base rate

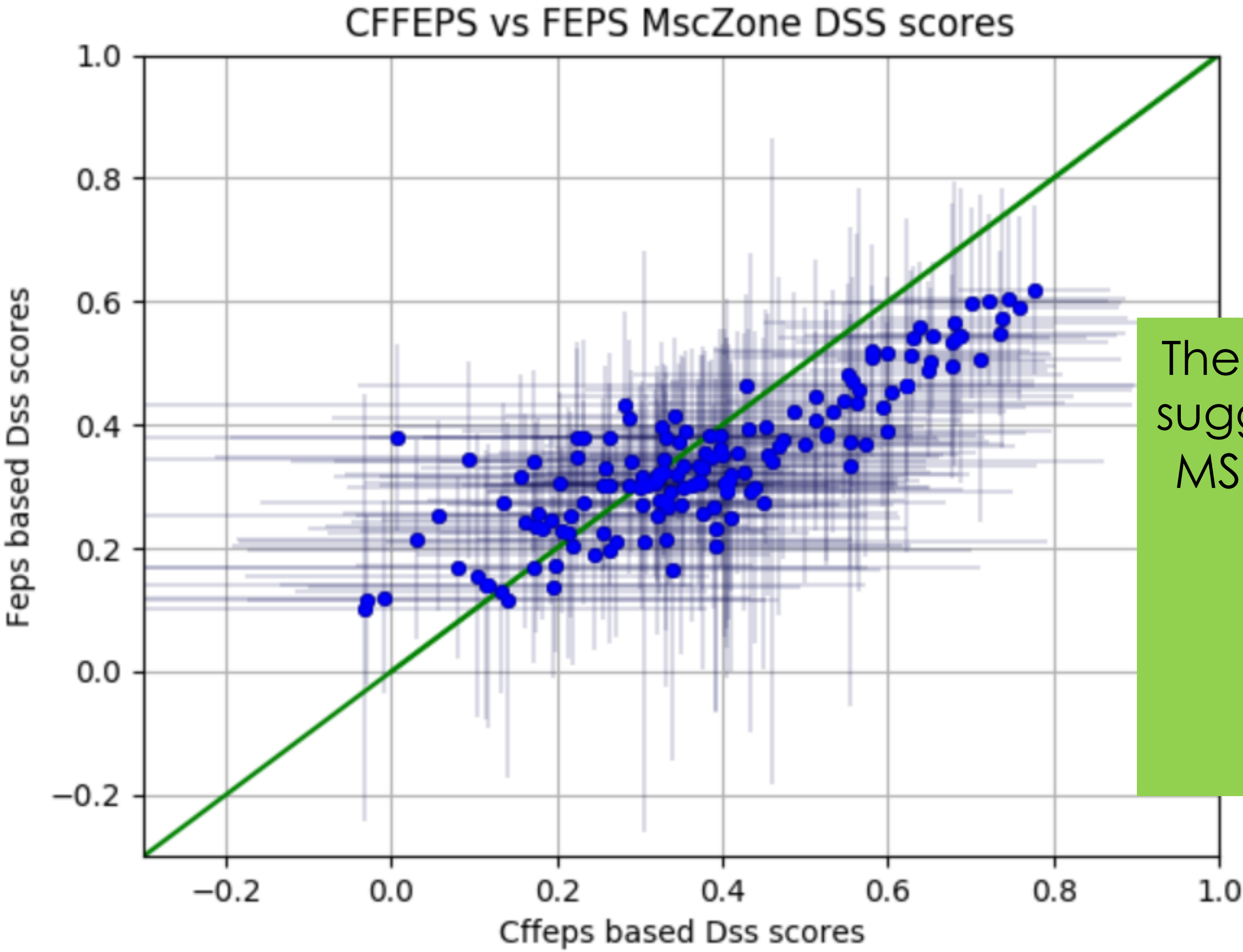

DSS:0.42 Ds:0.39 Rel:0.02 Res:0.30 Unc:0.68
Number of forecasts
10^4
10^3
0
2
4
6
8
10
Forecasted probabilities
Uncertainty (bits)
1.0
0.8
0.6
0.4
0.2
0.0
??
Clim
Fcst
Cal
Obs
Forecasted probabilities
Observed probabilities
0.0
0.2
0.4
0.6
0.8
1.0
Resolution (bits)
2.5
2.0
1.5
1.0
0.5
0.0
Forecast Concentrations

CFFEPS

DSS:0.36 Ds:0.43 Rel:0.02 Res:0.26 Unc:0.68



FEPS



The Wilcoxon signed-rank test suggests the two estimates for MSC Zone DSS scores do not come from the same population ($p<0.01$).

- ➢ CFFEPS provides more information to forecaster than FEPS

Conclusions

- Guided by survey: 2 interpretations of model guidance, both focussed on events
- Evaluation at the forecast zone-scale and over all of Western Canada via xGBM
- "At face value" analysis via PSS score: FEPS a little more skillful/valuable than CFFEPS
- "As information" analysis via DSS: CFFEPS a little more informative than FEPS

Did we miss something in how guidance is used?