Informing Climate Retrieval Development Using Data Mining

-or-

How to help scientists draw better lines

Lukas Mandrake

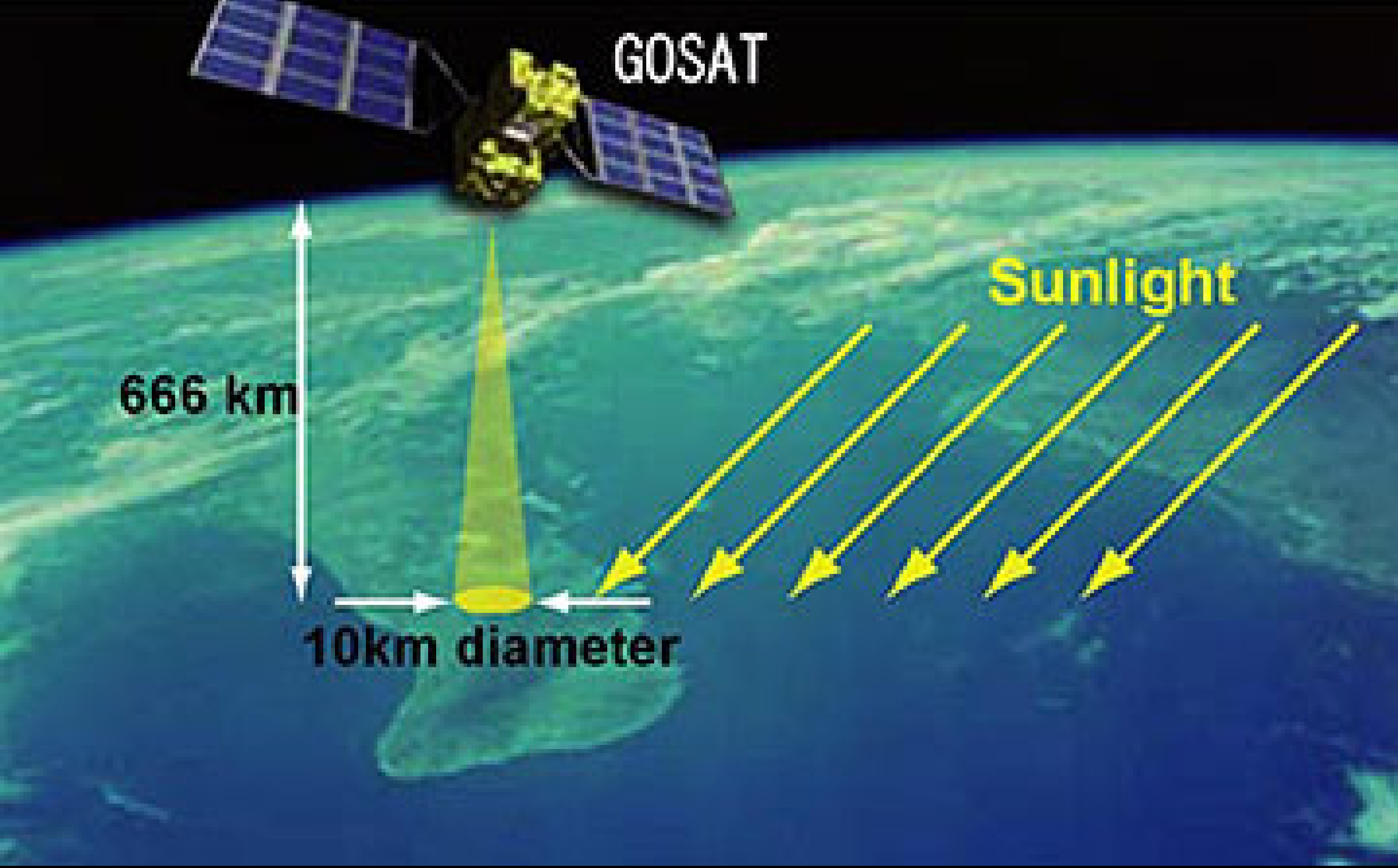
Jet Propulsion Laboratory, California Institute of Technology.

"Great things are done by a series of small things brought together"
– *Vincent Van Gogh*

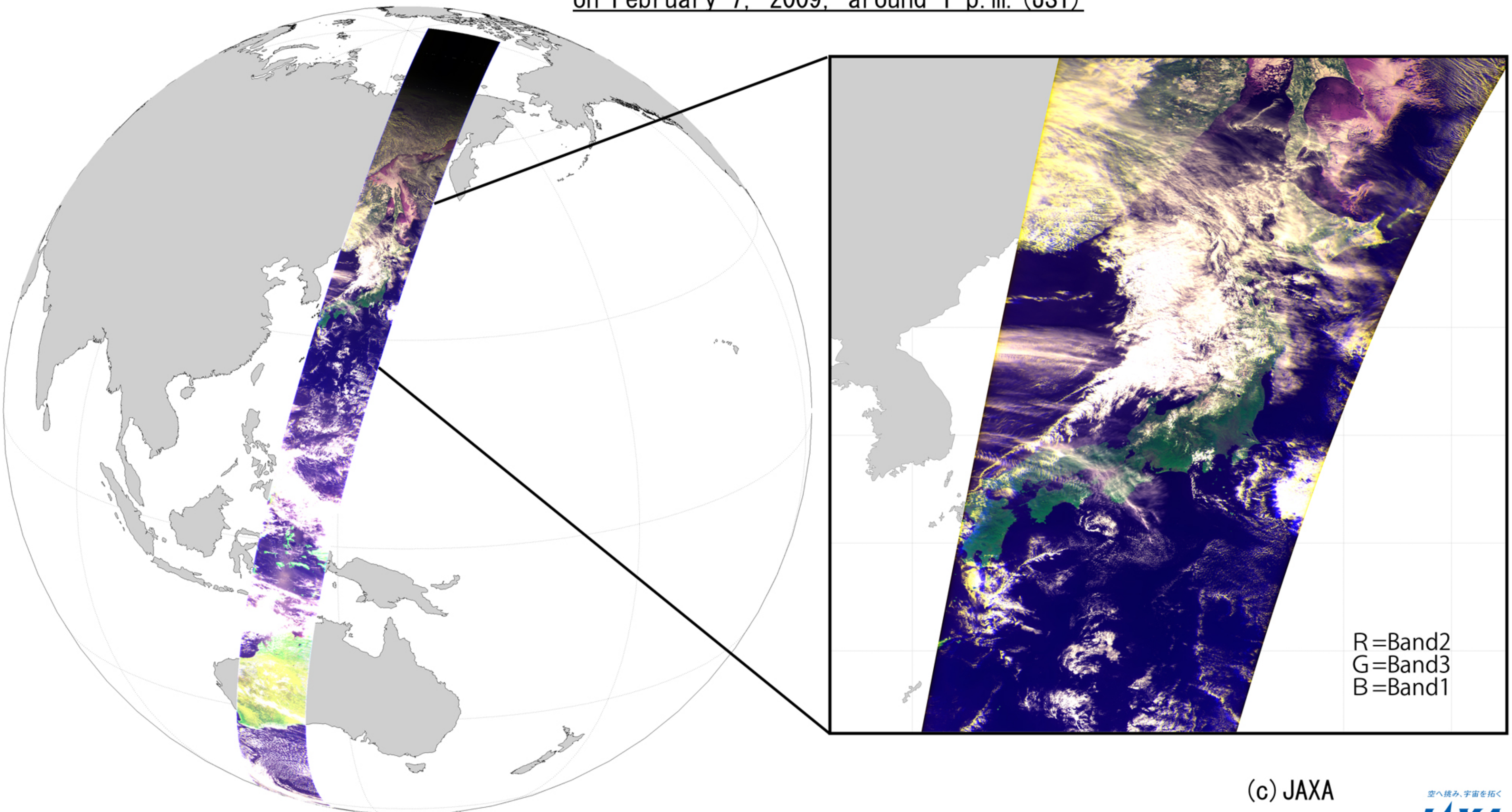
Data Source: GOSAT

(Mostly want to get CO_2)

Looming…


GOSAT
Sunlight
666 km
10km diameter

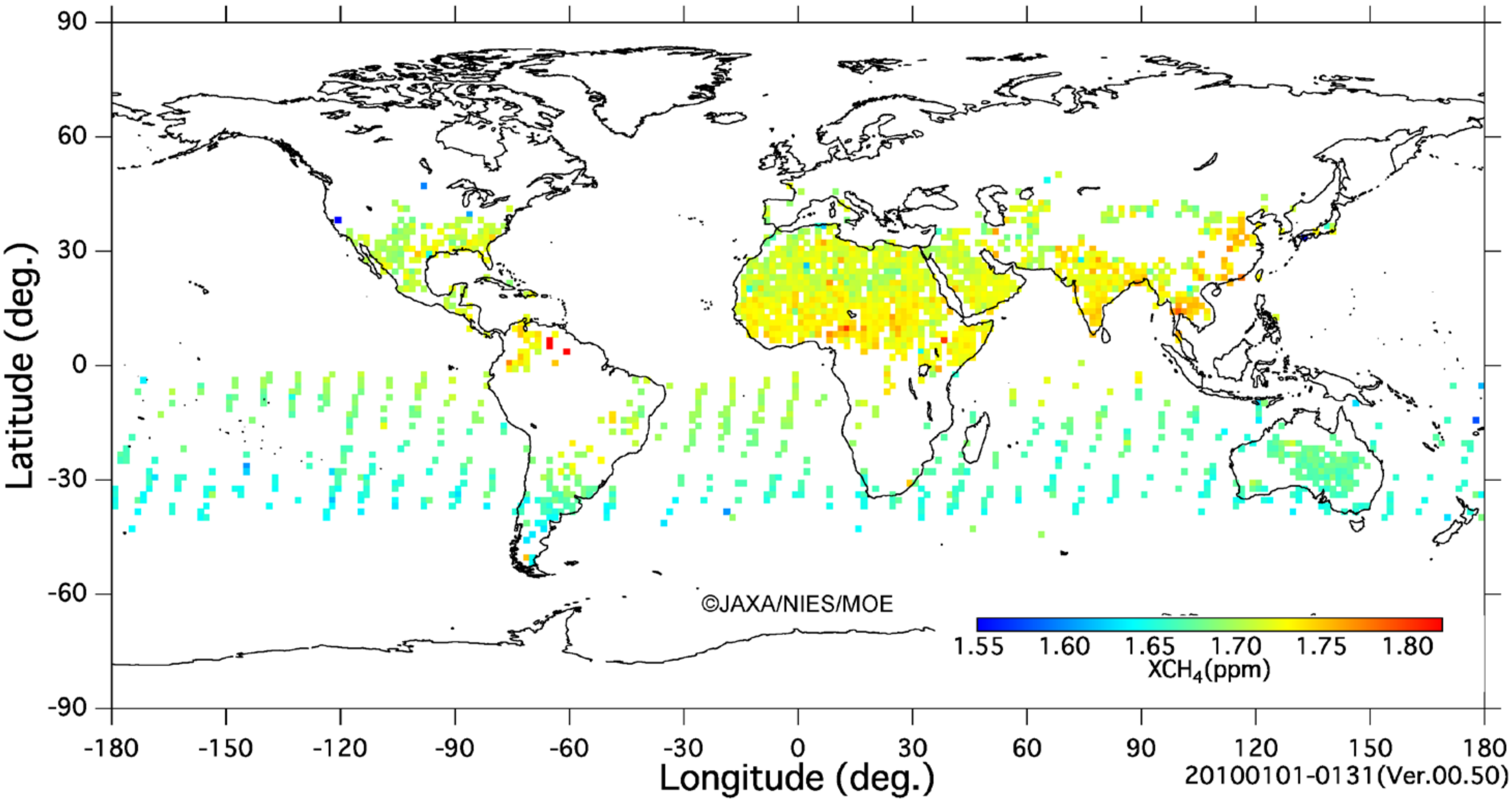
“IBUKI” Cloud and Aerosol Imager (TANSO-CAI) “FIRST LIGHT” False Color-composite Image on February 7, 2009, around 1 p.m.(JST)



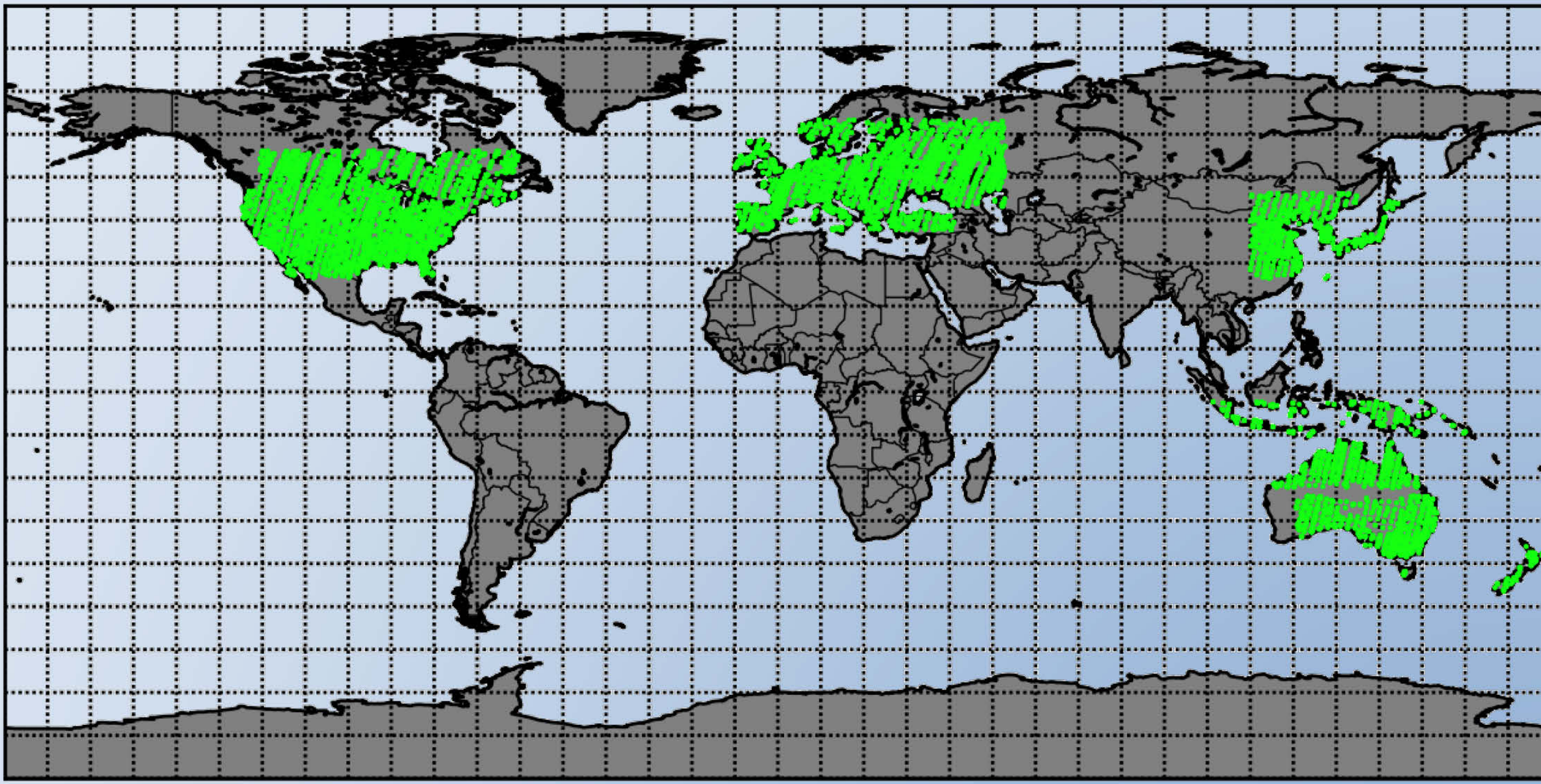
(c)JAXA

GOSAT Measurements = Soundings

Up to every ~4 seconds



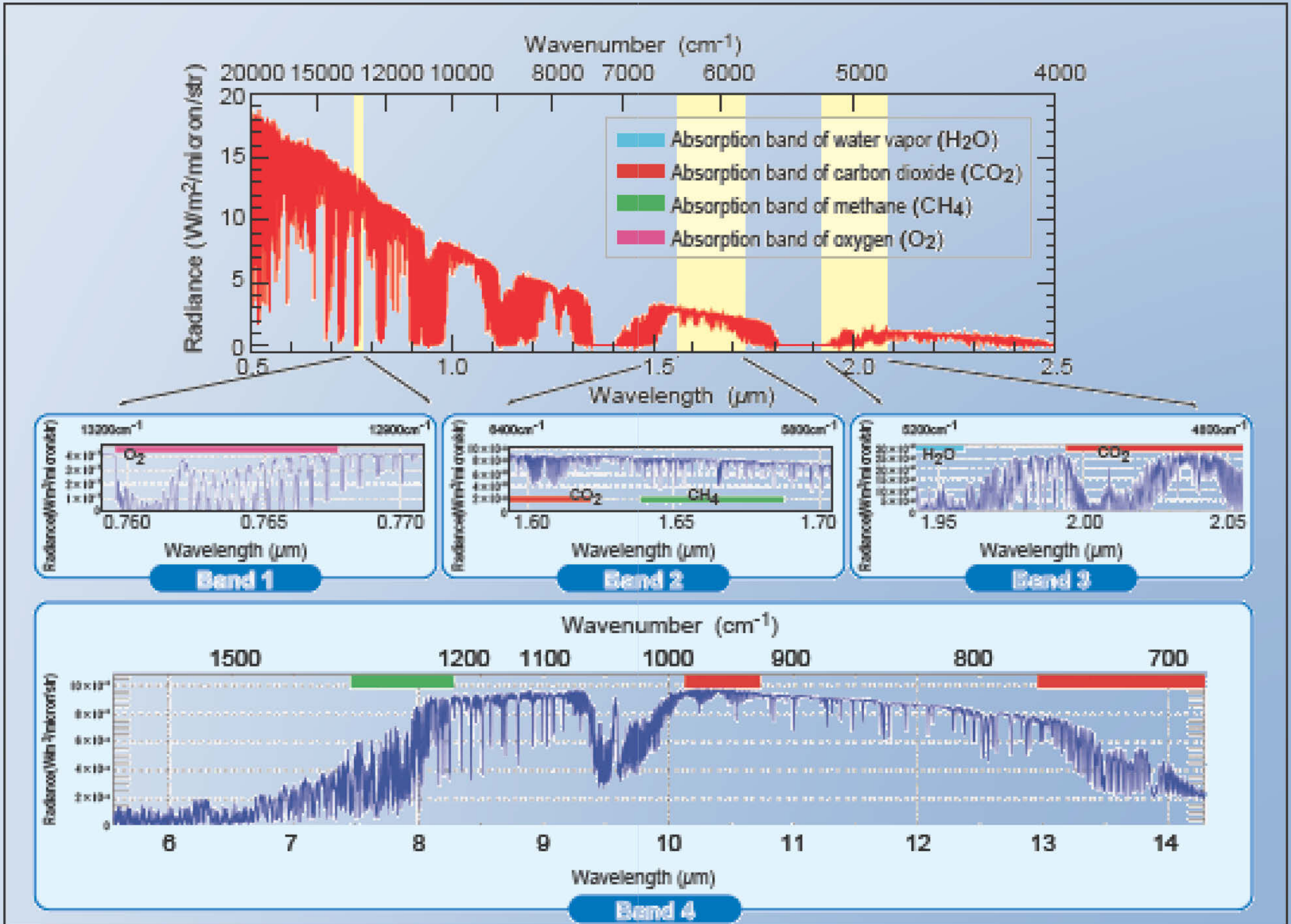
Places where you have "truth"



Around 62k observations over 2 years

What is a raw "sounding?"

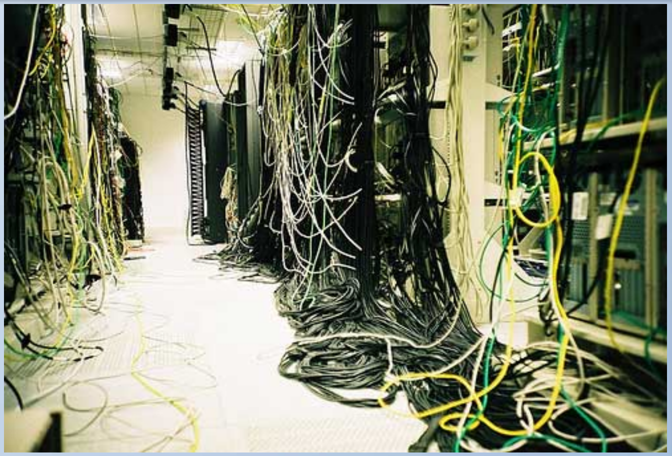
An infrared spectrum



How do you convert spectra to CO2 value?

Spectrum
1000's of λ

Models and Bayesian optimization



TAKES AROUND 8 MINUTES

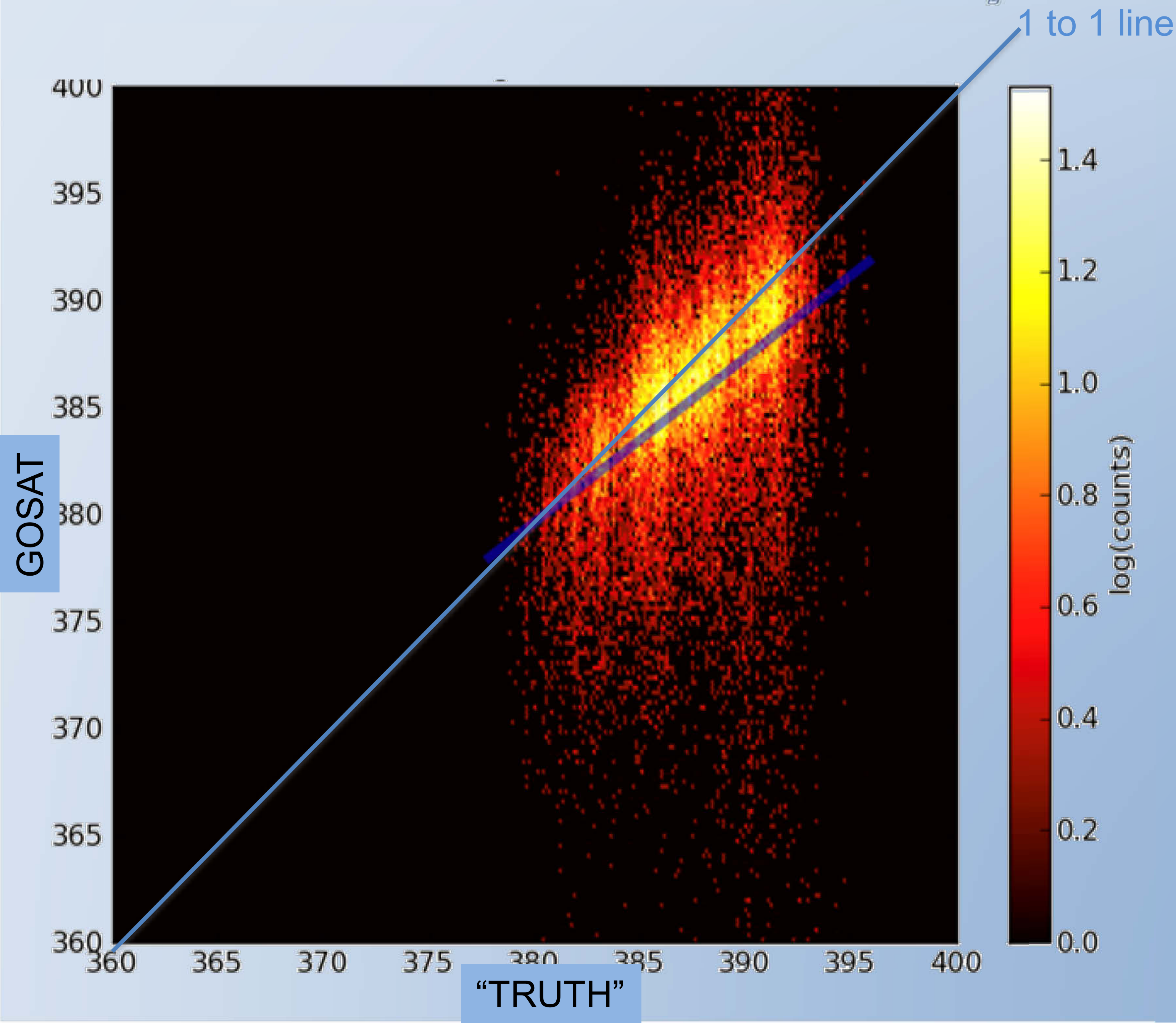
A single CO_2 value

381.5 ppm

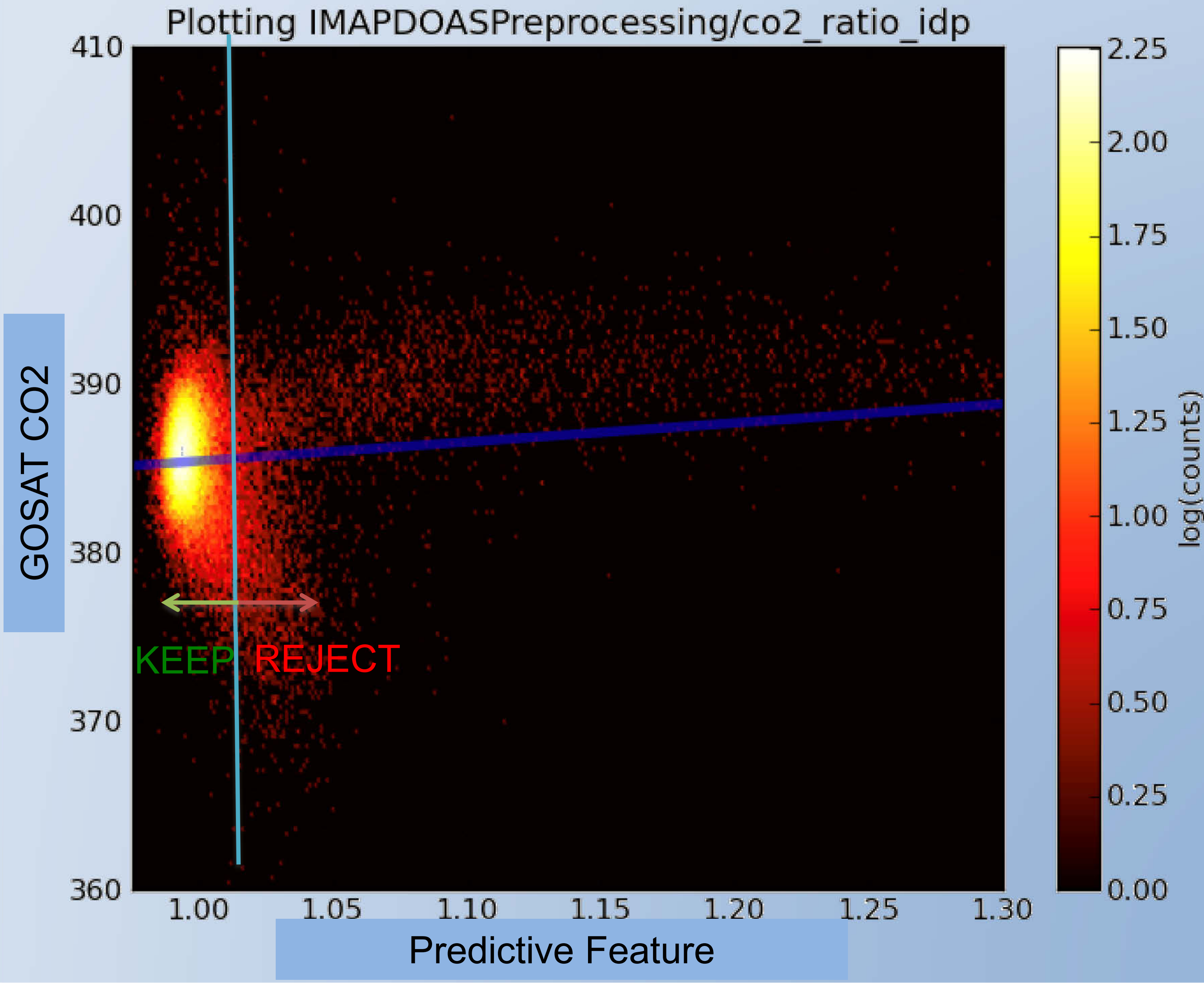
We're just practicing on GOSAT… for the new OCO-2 mission, we have data every 0.05 seconds or so. That's a lot of computation. Processing is expensive.

Problem: Disagreement

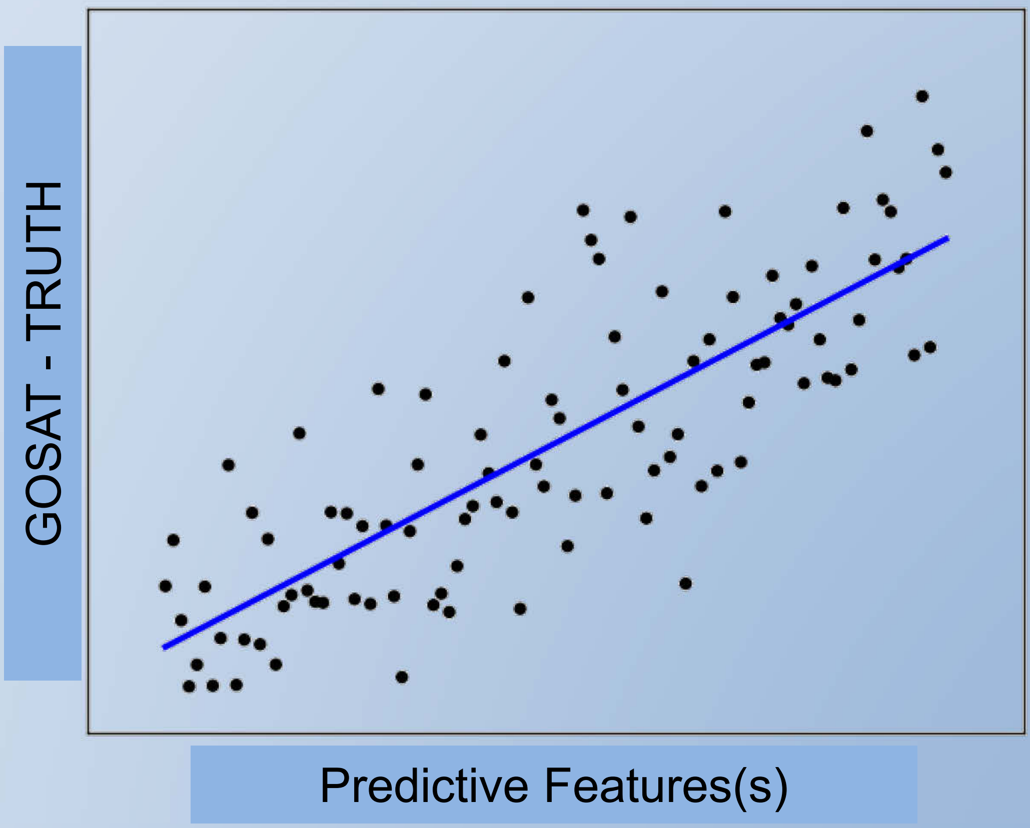
Retrieval doesn't always work



What do the scientists do? Filter fist...



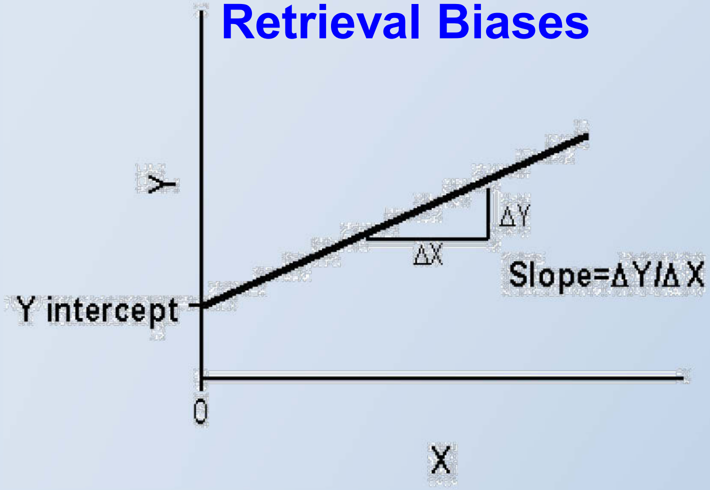
... then fit out discrepancies.



Then say: "I hope that got rid of all the bad data and biases…"

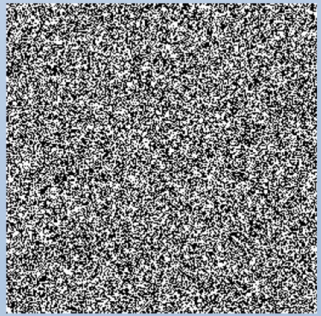
Mental Model of Data

Predictable Physics & Retrieval Biases

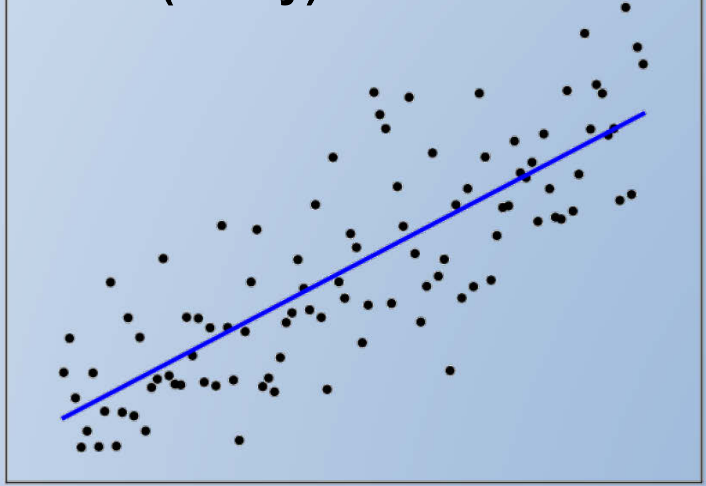


+

Noise (aerosols, unmodelled complexities, measurement error)



=



(Noisy) Retrievals

Conclusion: Filter's job is to remove random outliers, fit the remaining physics
Bias removal can "correct" lingering retrieval issues
Expectation: Filtering noise first will improve bias fit

Past Work

Example Filter 1

Filter	Filter criterion
Retain data with good spectral □ts	reduced_chi_squared_o2_fph $< 1.2 + 0.088 \times (f_{year} - 2009.26)$ reduced_chi_squared_strong_co2_fph $< 1.2 + 0.040 \times (f_{year} - 2009.26)$ reduced_chi_squared_weak_co2_fph $< 1.2 + 0.064 \times (f_{year} - 2009.26)$
Retain data with well-retrieved surface elevation	$\lvert(\Delta P) - \overline{\Delta P}\rvert < 5$ hPa (ΔP = surface_pressure_fph − surface_pressure_apriori_fph)
Retain scenes without extreme aerosol optical depth values	0.05 < retrieved_aerosol_aod_by_type < 0.15 (use the □rst of the 5 rows of the matrix)
Retain data with no diverging steps	diverging_steps = 0
Retain scenes with no cloud	cloud_□ag = 0
Retain data that converge	outcome_□ag = 1 or 2
Retain data with 'H' gain only	gain_□ag = 'H '
Retain no glint data	glint_□ag = 0
Retain scenes without cloud over ice	2.4×albedo_o2_fph −1.13×albedo_strong_co2_fph < 1
Retain scenes unless with nonzero X_{CO_2} uncertainties	xco2_uncert ≠ 0

Parameter	Mean value	Coef□cients		
		Assumption 1	Assumption 2	Assumption 3
blended_albedo	0.3	6.5 ± 0.4	6.3 ± 0.4	6.2 ± 0.4
ΔP	0.59 hPa	-0.15 ± 0.01	-0.14 ± 0.01	-0.16 ± 0.01
airmass	2.6	-1.3 ± 0.4	-1.3 ± 0.4	-1.5 ± 0.4
signal_o2	3.7×10^{-7} W cm^{-2} sr^{-1} (cm^{-1})$^{-1}$	-0.47 ± 0.08	-0.45 ± 0.08	-0.47 ± 0.08

Fit Params: dP_{surf}, airmass, signal O_2, albedo

Example Filter 2

Table 3. Advanced screening criteria for the L2 in the v2.10 data.

Variable	Allowed Range		
	Ocean Glint	Land H	Land M
RetrievalResults/outcome_flag	1 or 2	1 or 2	1 or 2
RetrievalResults/aerosol_total_aod	< 0.3	< 0.5	< 0.3
RetrievalResults/aerosol_water_aod	< 0.15	< 0.15	< 0.15
RetrievalResults/diverging_steps	<= 2	<= 2	<= 2
SoundingGeometry/sounding_altitude_stddev	< 200	< 200	< 200
IMAPDOASPreprocessing/co2_ratio_idp	0.985 to 1.005	0.985 to 1.005	0.985 to 1.005
IMAPDOASPreprocessing/h2o_ratio_idp	0.96 to 1.10	0.96 to 1.10	0.96 to 1.08
ABandCloudScreen/dp_cld • 10^{-2} [hPa]	-8 to 8	−8 to 8	−10 to 10
RetrievalResults/aerosol_ice_aod	< 0.02	< 0.02	0.004 to 0.04
SpectralParameters/reduced_chi_squared_o2_fph	< 1.5	< 1.25	< 1.5
SpectralParameters/reduced_chi_squared_weak_co2_fph	< 1.8	< 1.6	< 1.6
SpectralParameters/reduced_chi_squared_strong_co2_fph	< 2.0	< 2.0	< 2.0
RetrievalResults/xco2_uncert • 10^6 [ppm]	< 1.5	< 2.0	< 2.0
RetrievalResults/albedo_slope_strong_co2	> 1.2e-5	> -1e-4	> -1e-4
ΔP_s [hPa]	-10 to 10	-10 to 7	-12 to 2
RetrievalResults/albedo_slope_o2	< 4e-6	< 0	
RetrievalResults/albedo_slope_weak_co2	-7e-6 to -5e-7		
RetrievalResults/temperature_offset_fph	> -1		
RetrievalResults/zero_level_offset_o2	< 5.0		
RetrievalResults/albedo_strong_co2_fph	> 0.01		
Blended Albedo		< 0.8	< 0.8
SpectralParameters/signal_o2_fph • 10^7		1.5 to 6.5	

Lots of filter inputs, only segregates data into good/bad

$$X'_{CO2} = X_{CO2} + 0.19 \cdot (\Delta P_S + 1.0 \text{ hPa}) + 7.0 \cdot (A_{SCO2} - 0.20) + 1.2$$

Fit Params: XCO_2, dP_{surf}, Albedo SCO_2

Original Task

1) Can we make a data-driven filter to guide users to avoid "bad" GOSAT retrievals?

2) Can we "fit out" disagreement? If so, what are the features and coefficients?

3) Can we figure out when/why co2 differs so strongly?

4) Oh, and don't alienate the climate scientists…

Hope: A more principled, automated way to create these filters and fits

Available Features

240 potentially predictive features including:

- observation geometry
- solar geometry
- preprocessor outputs
- physical environment like humidity, temperature, pressure
- retrievals of peripheral gasses like methane
- signal quality and strength
- etc…

Filter Fit

How do filtration and bias correction relate?

Method 1 : Things Tried

Tried LDA (Fischer), SVM's, Decision Trees…

Got some results, some features of interest

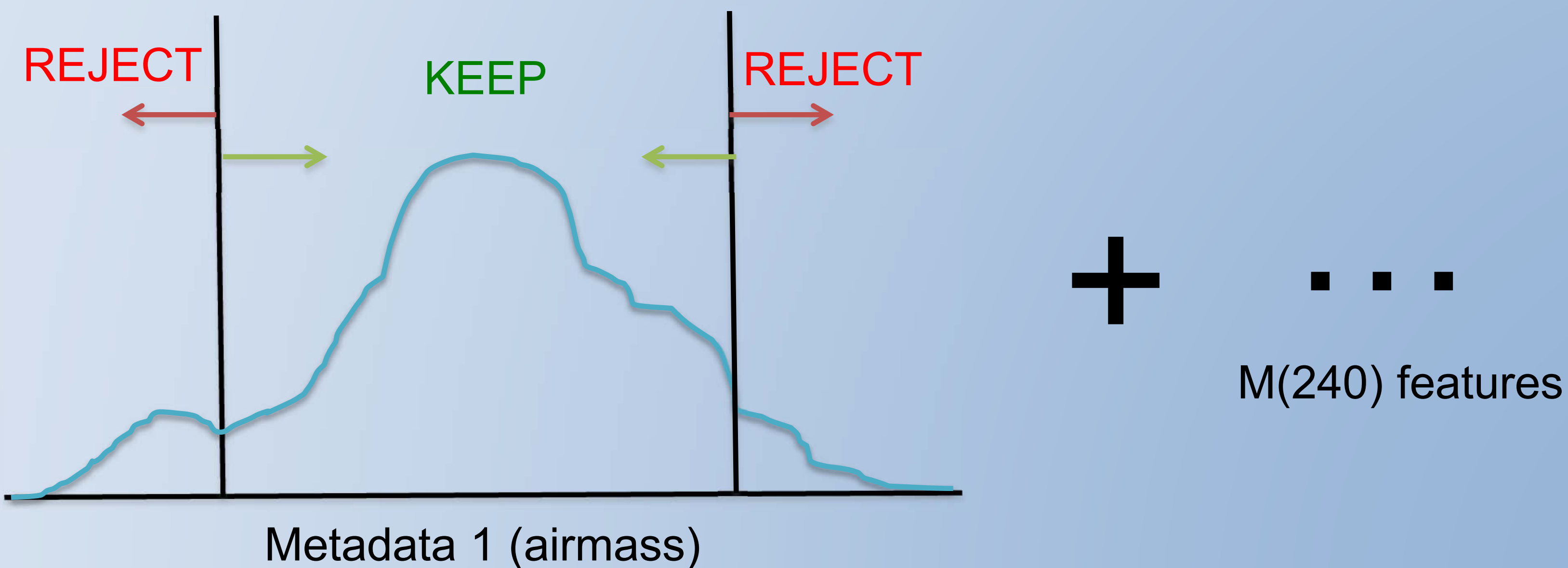
Each time when reporting in the larger mission sphere, you get the same response:

Nodding heads
"Very interesting"
Nothing changes…

It's quite hard to fight complex results against decades of training.

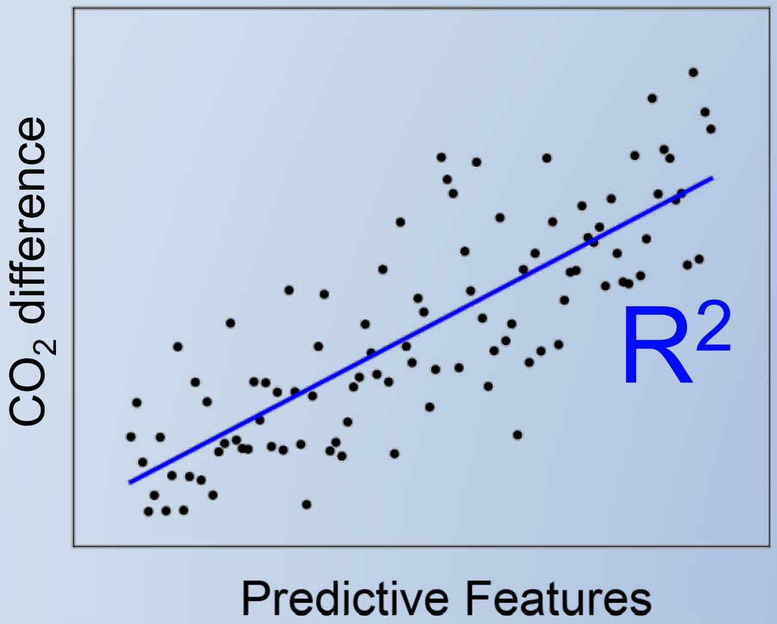
Method 2: Automated scientist

- Genetic algorithm for threshold filters
- Minimize CO_2 error vs. data accepted
- Examine features that dominate



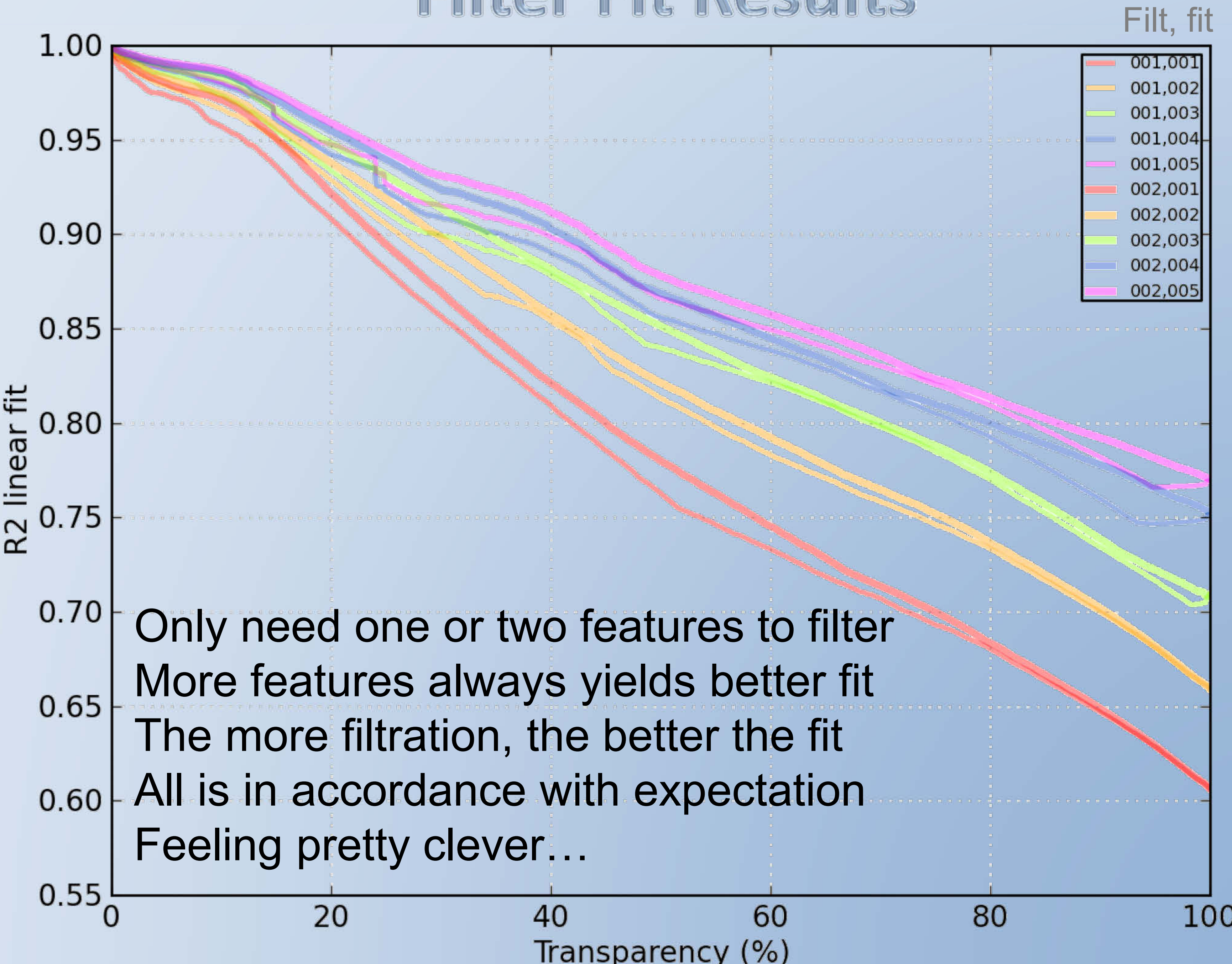
Method 2: Automated scientist

Once we have our filters, judge them based on how well we "fit out" remaining discrepancy. Metric: R^2



Works! Gets you beautiful, compact graphs that make intuitive sense…

Filter Fit Results



Only need one or two features to filter
More features always yields better fit
The more filtration, the better the fit
All is in accordance with expectation
Feeling pretty clever…

Filter Failure

Only one problem:

Filters are throwing out known good regions first

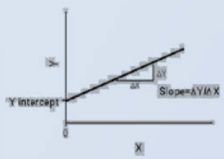
I made a contaminated sounding selector!

Now why would my code find that solution?

All I told it to do was improve the fit…

(New) Mental Model of Data

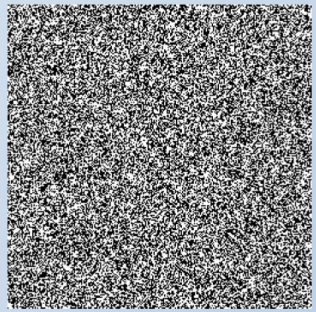
Predictable Physics & Retrieval Biases



(airmass, etc.)

+

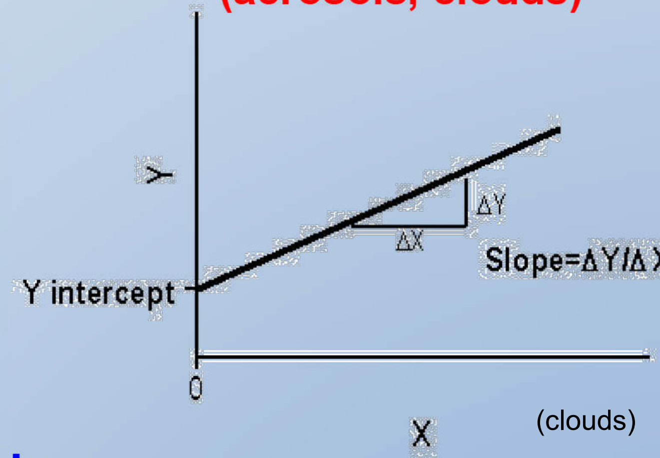
Actual sounding variability



(unmodeled physics, surface topology, etc.)

+

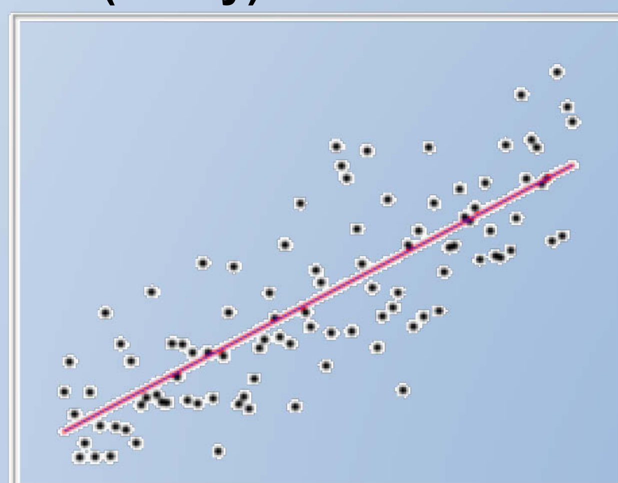
Systematic Contamination (aerosols, clouds)



(clouds)

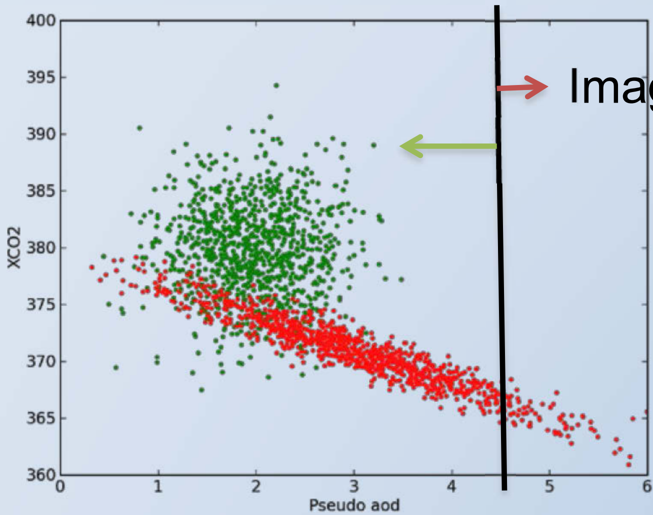
=

(Noisy) Retrievals



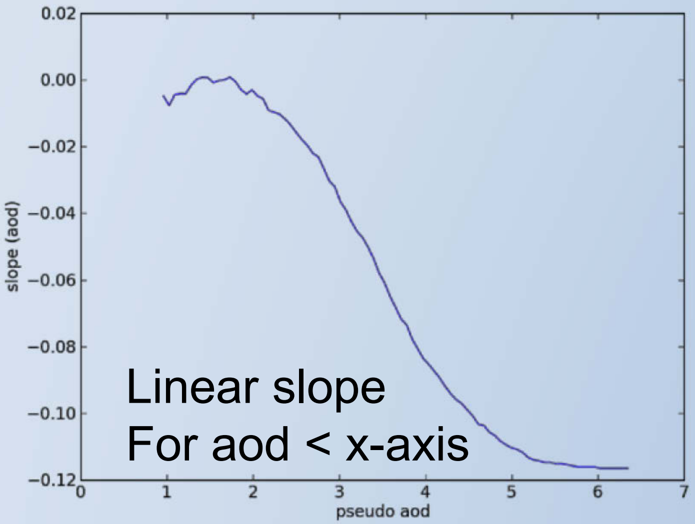
Conclusion: We have a much more difficult task to discover any underlying XCO2 bias beneath a much larger systematic bias due to aerosol/cloud contamination

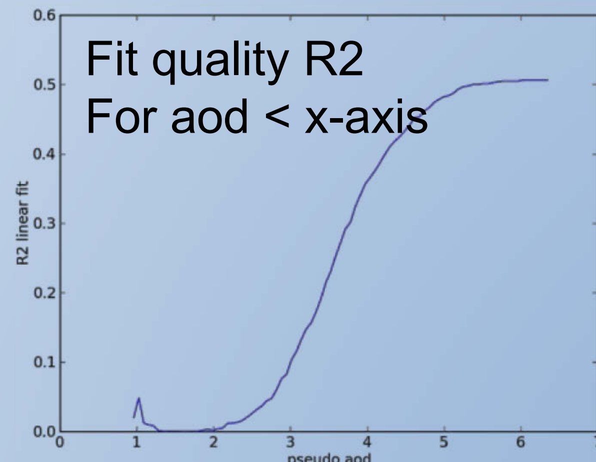
Simulation: Clouds vs. Not Clouds



Imperfectly separated populations
Red (cloud) has strong bias f(aod)
Green (clear) has no bias f(aod)

Filtration of red is appropriate
Fitting to remove effect is not





In this case, because two populations have different systematic bias, and are not fully separated, any bias correction does not improve the green distribution.

Problem with Filtration vs. Fitting

Clouds, our major contamination source,
are HIGHLY fittable (high R^2)

Filter + fit solution graded on the quality of fit makes a CLOUD SELECTOR

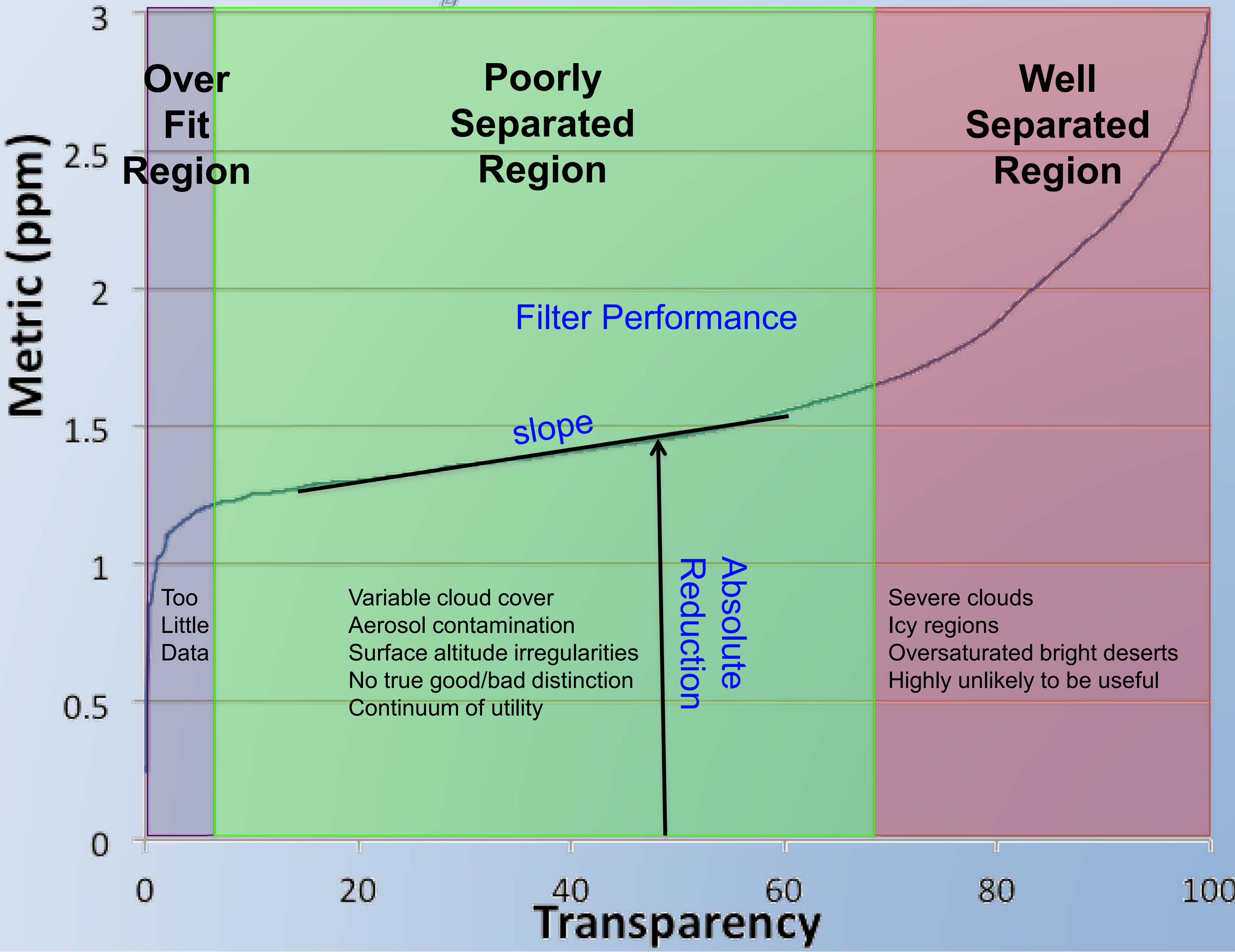
FILTER SOLUTION: Two Pass

1) Create filters based on reducing the RMS truth difference alone

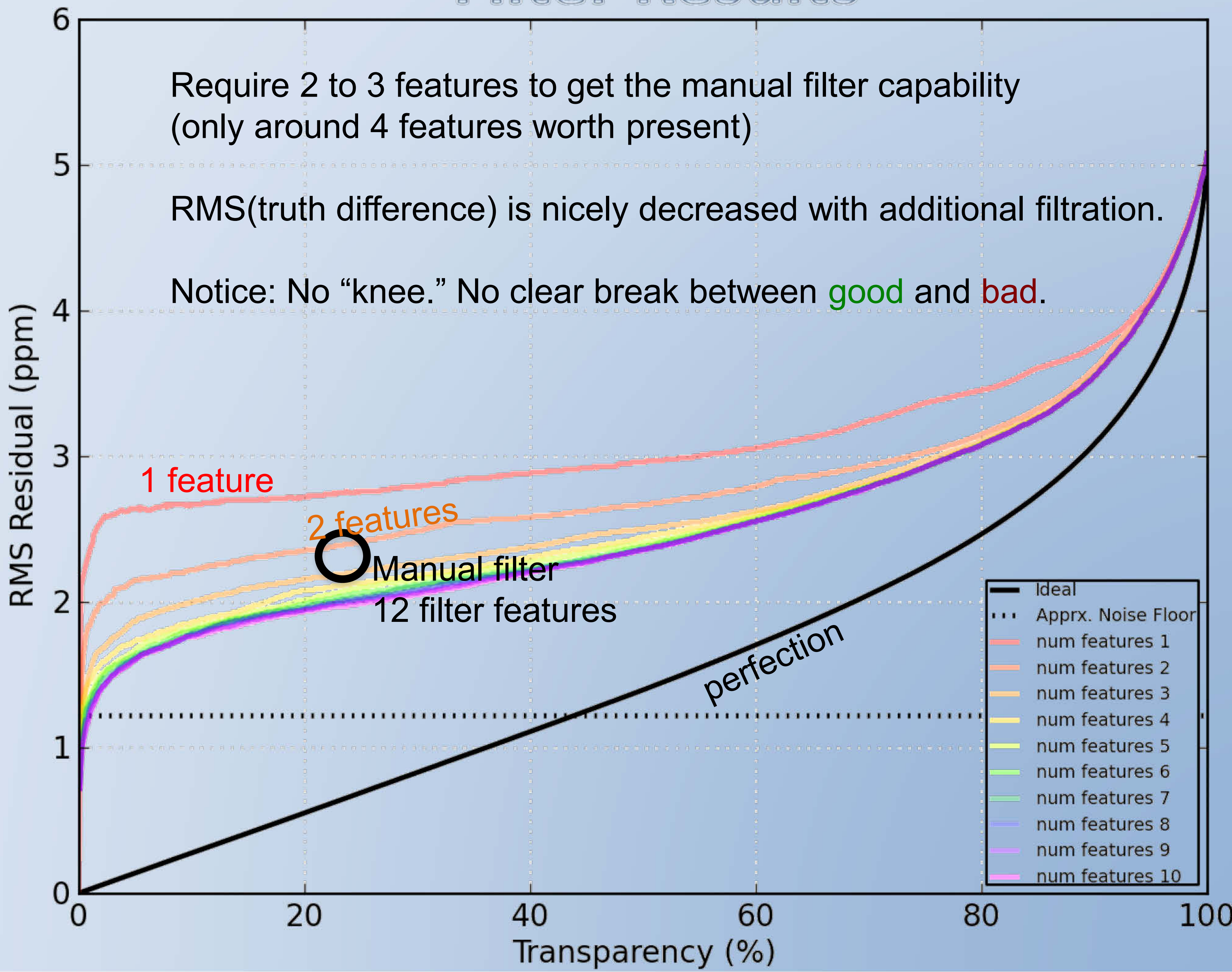
2) Select the best fit for each filter solution afterwards

Two Pass Filter Fit

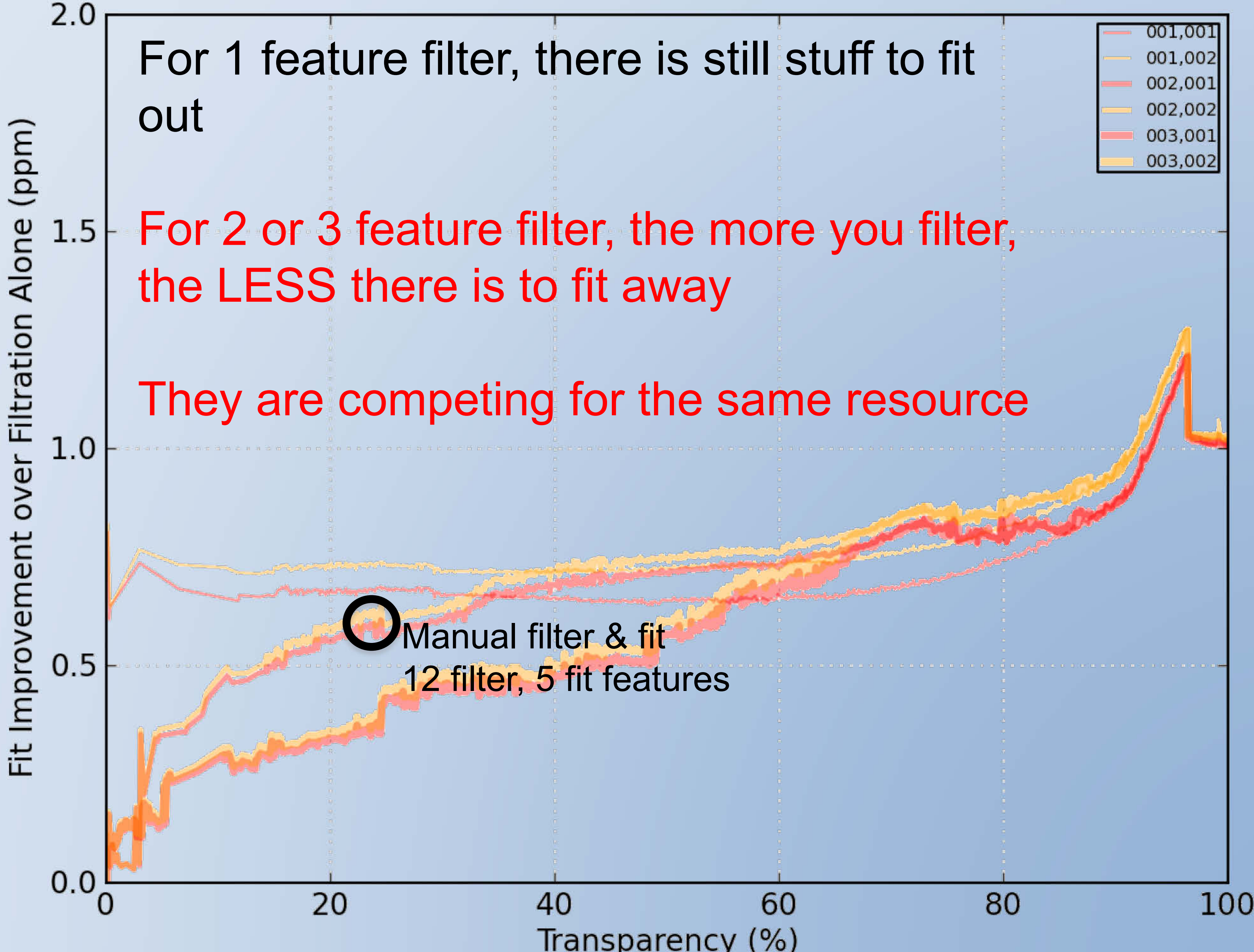
Anatomy of a Filter Curve



Filter Results



Fit - Filter Results



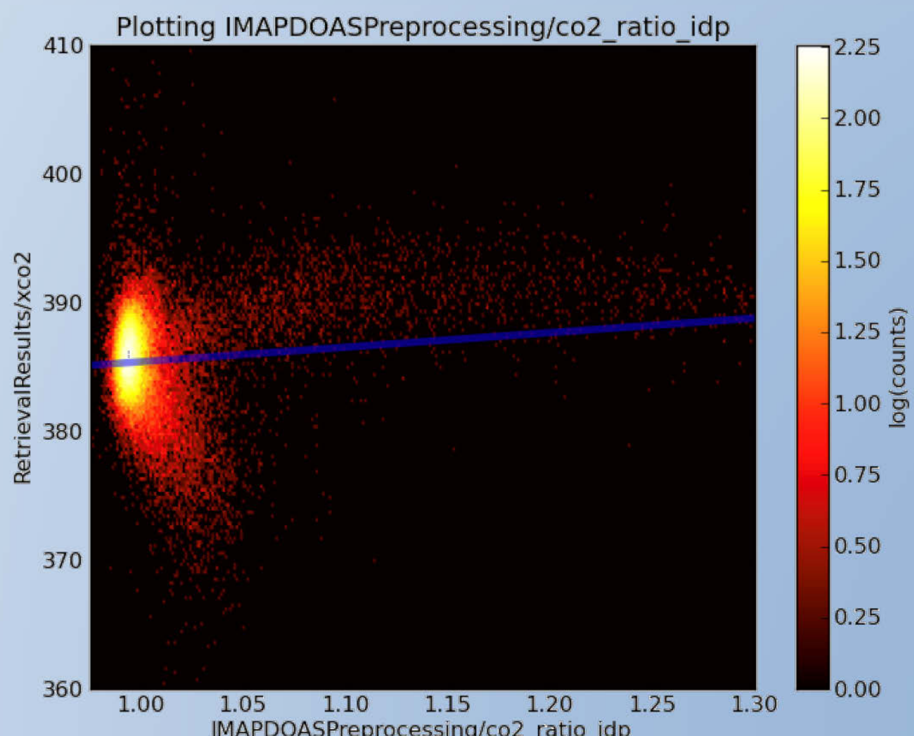
Don't bias correct

We have complex data with multiple overlapping populations

Our features are (not yet) able to separate them fully

Some have high bias and some do not

A single overall bias correction is not correct for any one population



We've learned so far

Filtration is powerful BUT:

Binary good/bad makes little sense

Bias correction (fitting) is ill-advised and strongly filter-dependent

You have around 12 predictive features, but only ~3 are needed for over 90% of the prediction

So… can you help us make a filter or not?

Something better

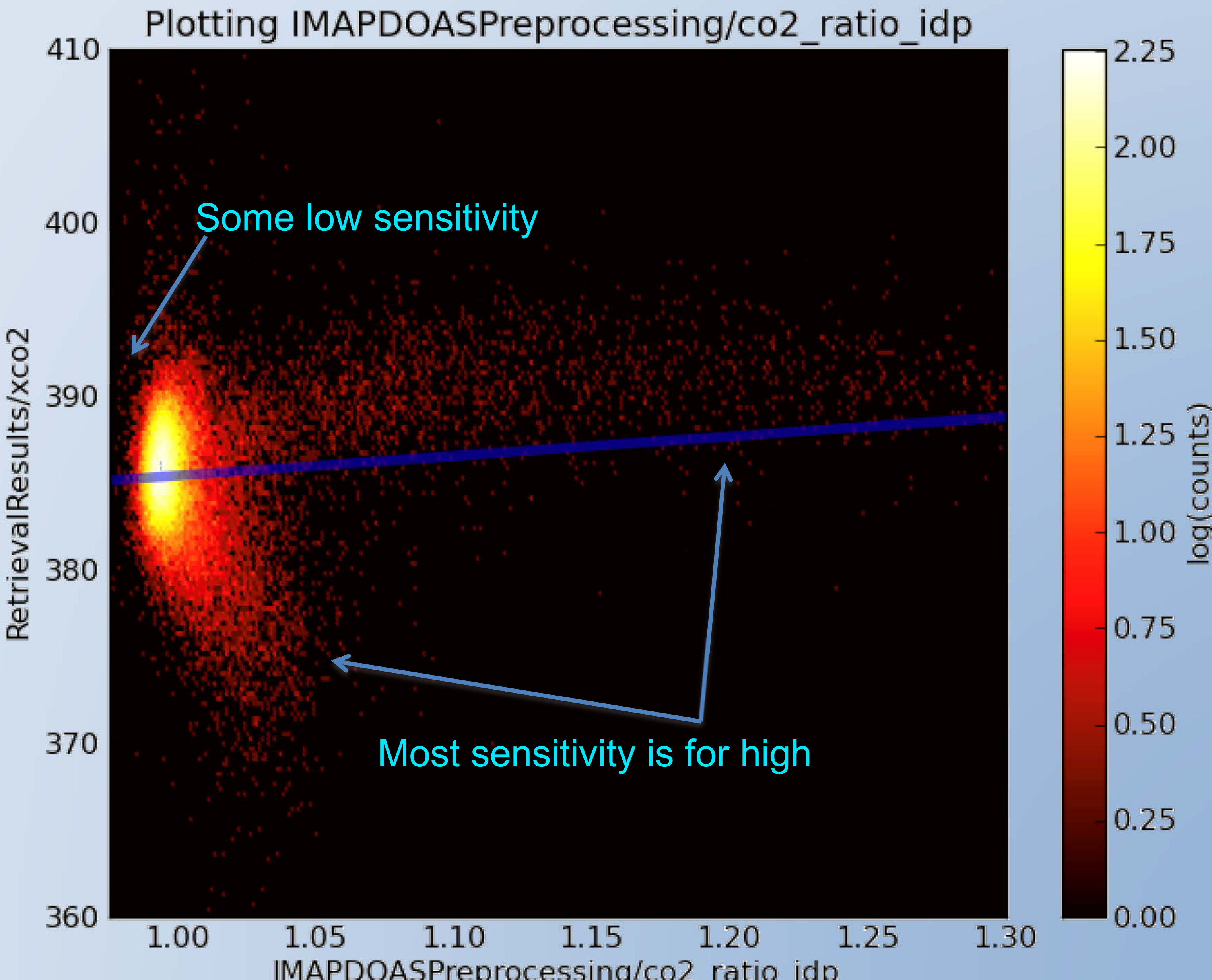
Just order the data instead

No good / bad assignment

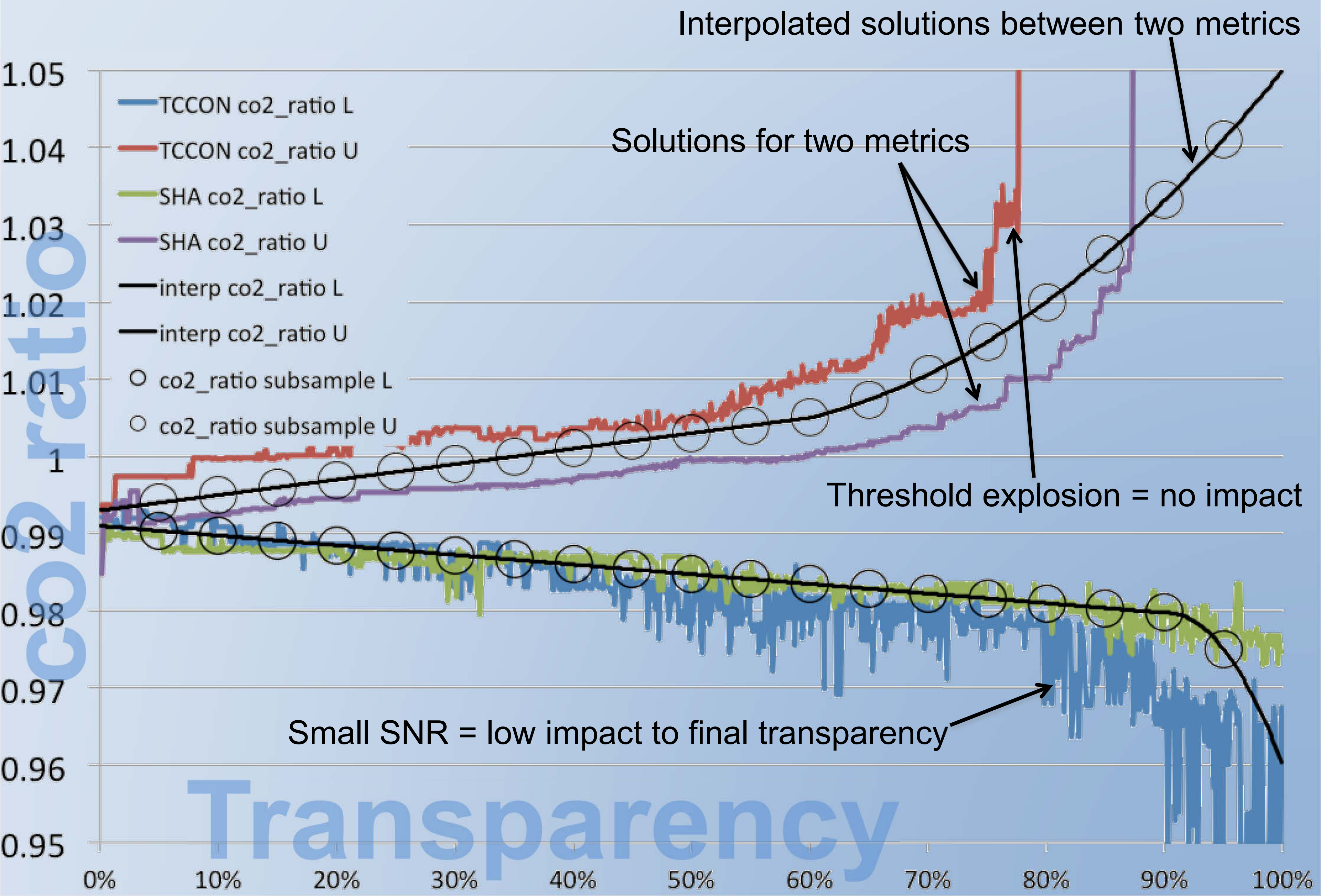
List of most to least fit data according to metric

User decides how far into the ordering to use

Feature 1 example



Thresholds and Warn Levels



Filter Definition:

“If sounding passes this filter but none below, it’s this warn level”

~ Transparency	Warn Level	co2_ratio L	co2_ratio U	ice_aod L	ice_aod U	albedo L	albedo U	
93%	18	0.9750	1.0411	0.0003	0.0729	-0.000097	0.000105476	Least Conservative
88%	17	0.9798	1.0331	0.0005	0.0358	-0.000094	5.37342E-05	
83%	16	0.9804	1.0261	0.0008	0.0333	-0.000091	2.63391E-05	
78%	15	0.9810	1.0200	0.0010	0.0309	-0.000088	1.23644E-05	
73%	14	0.9816	1.0148	0.0013	0.0284	-0.000085	5.5277E-06	
68%	13	0.9823	1.0106	0.0015	0.0259	-0.000082	0.000003	
63%	12	0.9829	1.0073	0.0018	0.0234	-0.000079	0.000001	
58%	11	0.9835	1.0050	0.0020	0.0209	-0.000076	-0.000001	
53%	10	0.9841	1.0040	0.0023	0.0184	-0.000073	-0.000003	
48%	9	0.9848	1.0030	0.0025	0.0160	-0.00007	-0.000005	
43%	8	0.9854	1.0020	0.0028	0.0158	-0.000067	-0.000007	
38%	7	0.9860	1.0010	0.0030	0.0155	-0.000064	-0.000009	
33%	6	0.9866	1.0000	0.0033	0.0153	-0.000061	-0.000011	
28%	5	0.9873	0.9990	0.0035	0.0150	-0.000058	-0.000013	
23%	4	0.9879	0.9980	0.0038	0.0148	-0.000055	-0.000015	
18%	3	0.9885	0.9970	0.0040	0.0145	-0.000052	-0.000017	
13%	2	0.9891	0.9960	0.0043	0.0143	-0.000049	-0.000019	
8%	1	0.9898	0.9950	0.0045	0.0140	-0.000046	-0.000021	
3%	0	0.9904	0.9940	0.0048	0.0138	-0.000043	-0.000023	Most Conservative

IMAPDOASPreprocessing	co2_ratio_idp
RetrievalResults	aerosol_ice_aod
RetrievalResults	albedo_slope_o2

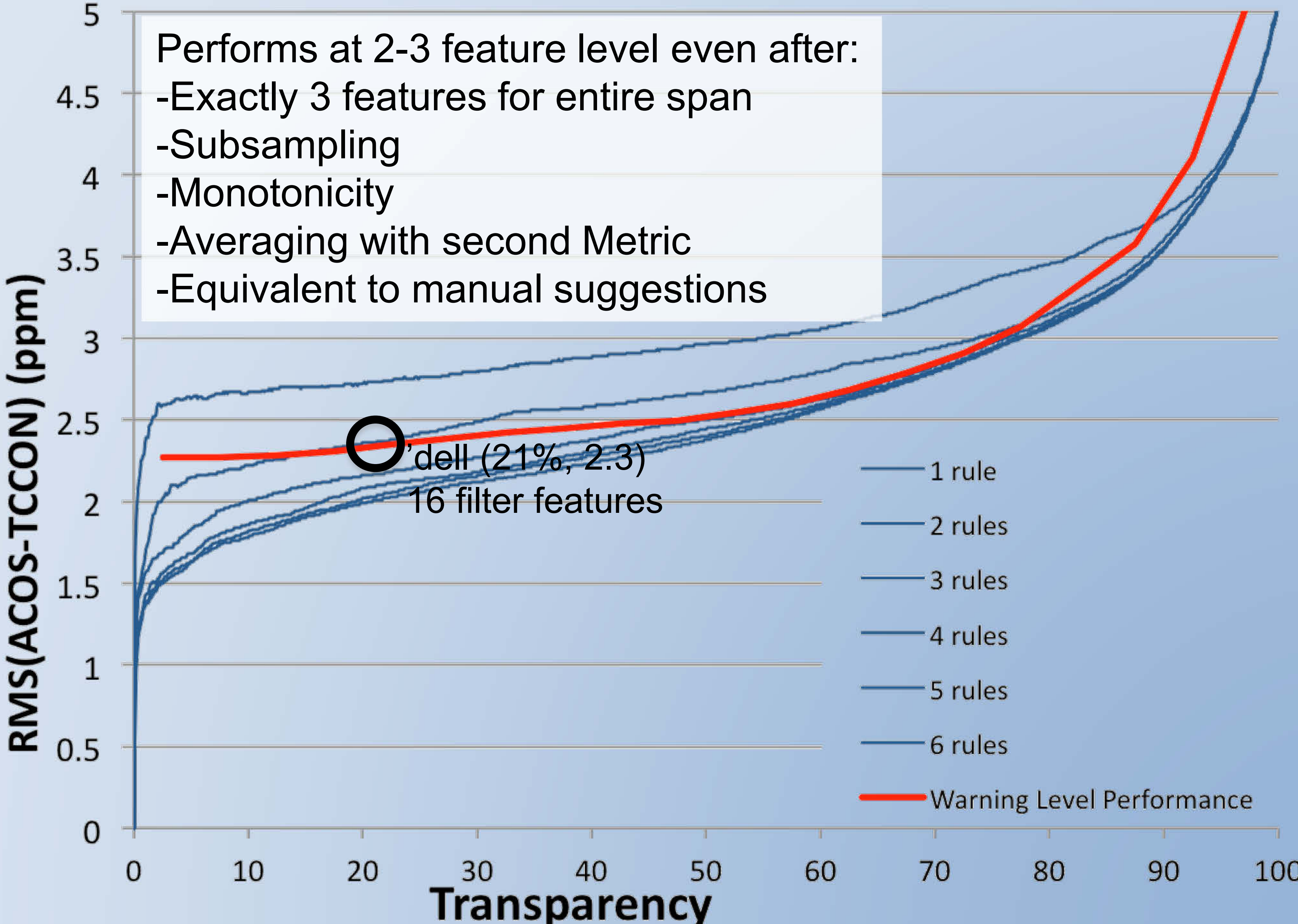
Evaluate Resulting Filter

The Warn Levels / Quality Estimation System are now defined.

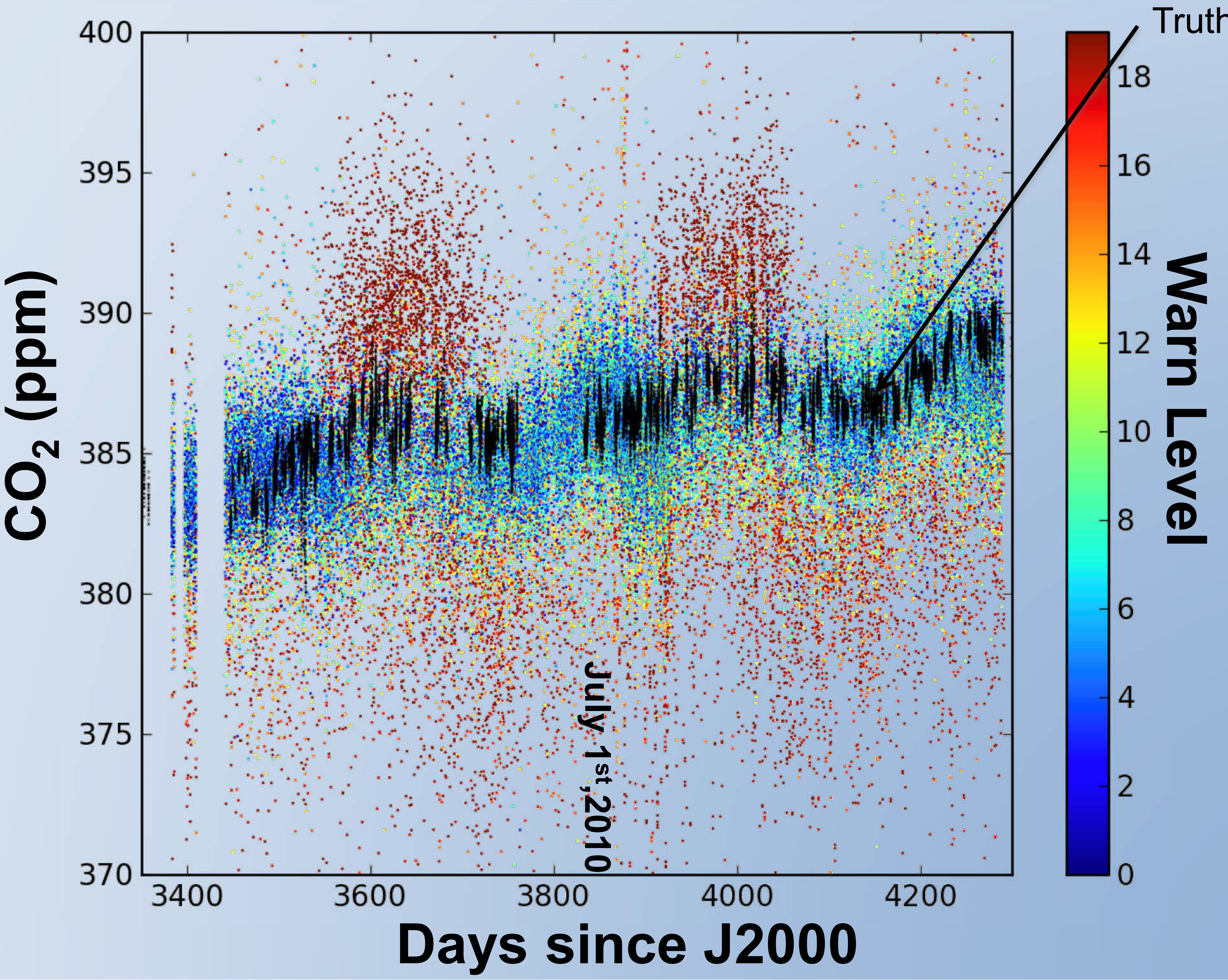
Evaluate resulting system for:

1) Metric Reduction (RMS and MMS)
2) Global/regional WL dependence
3) Temporal dependence
4) Strange behavior

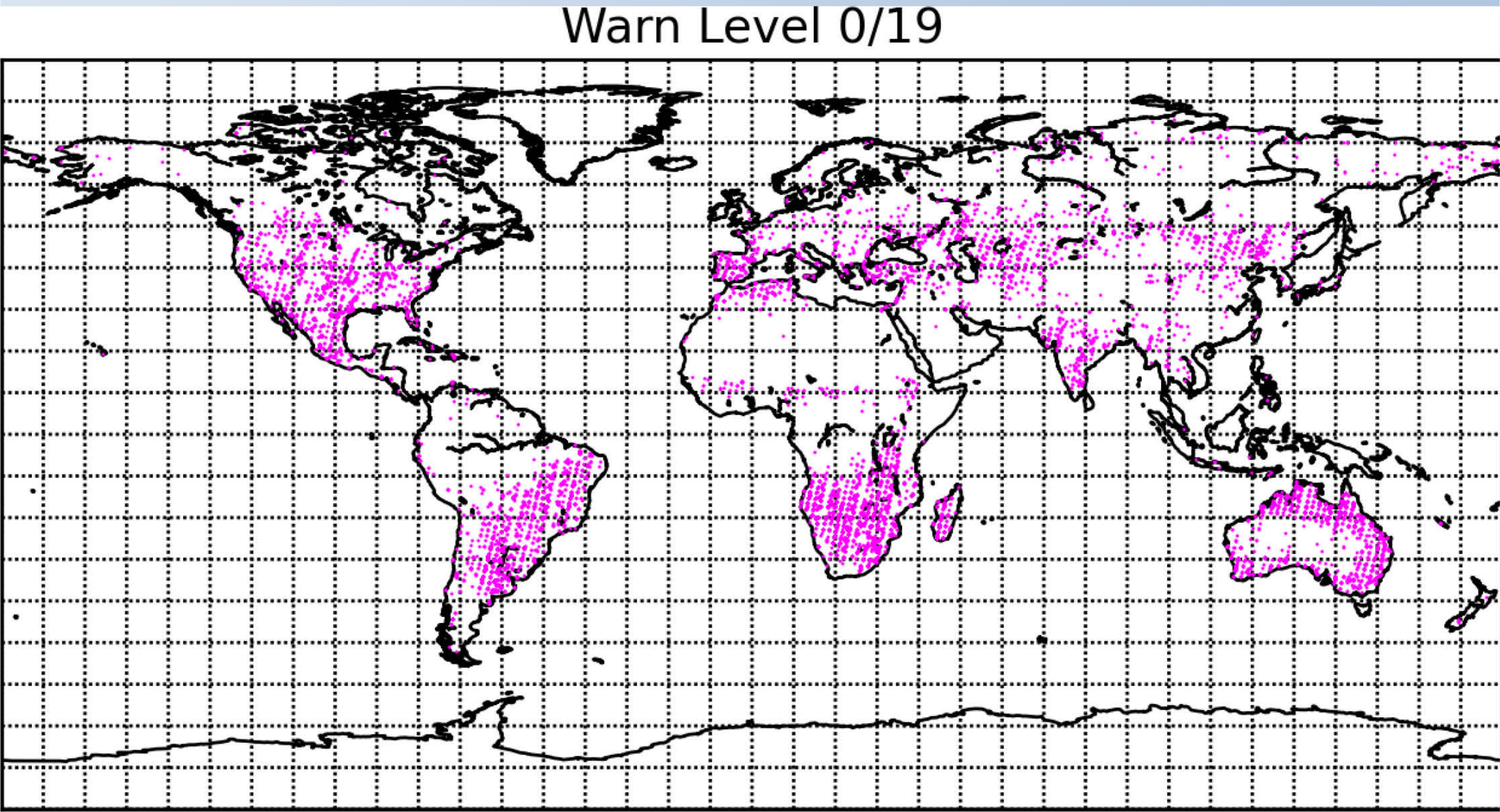
Warn Level Performance



Temporal Analysis Warn Levels

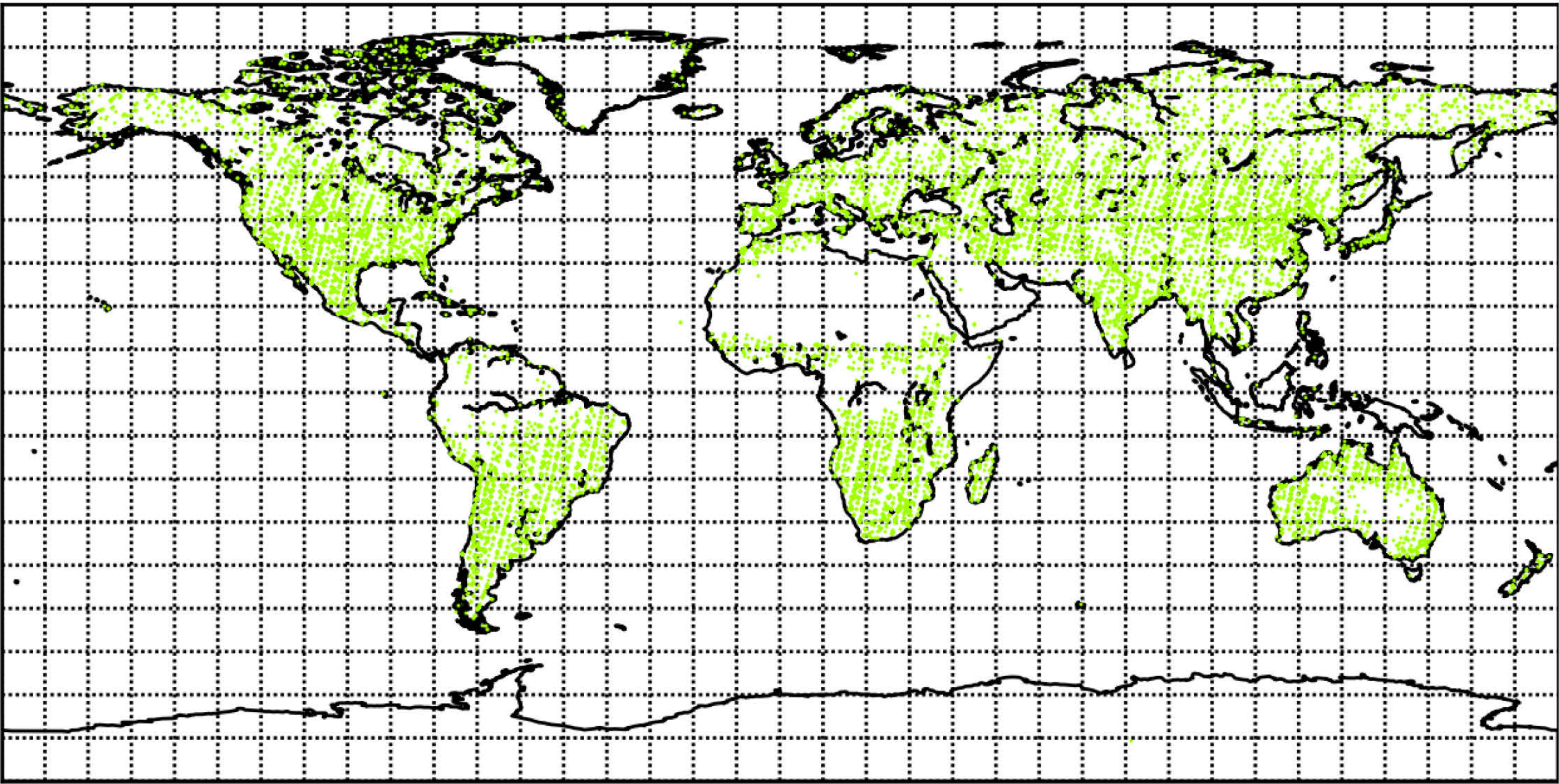


Global Spatial Analysis



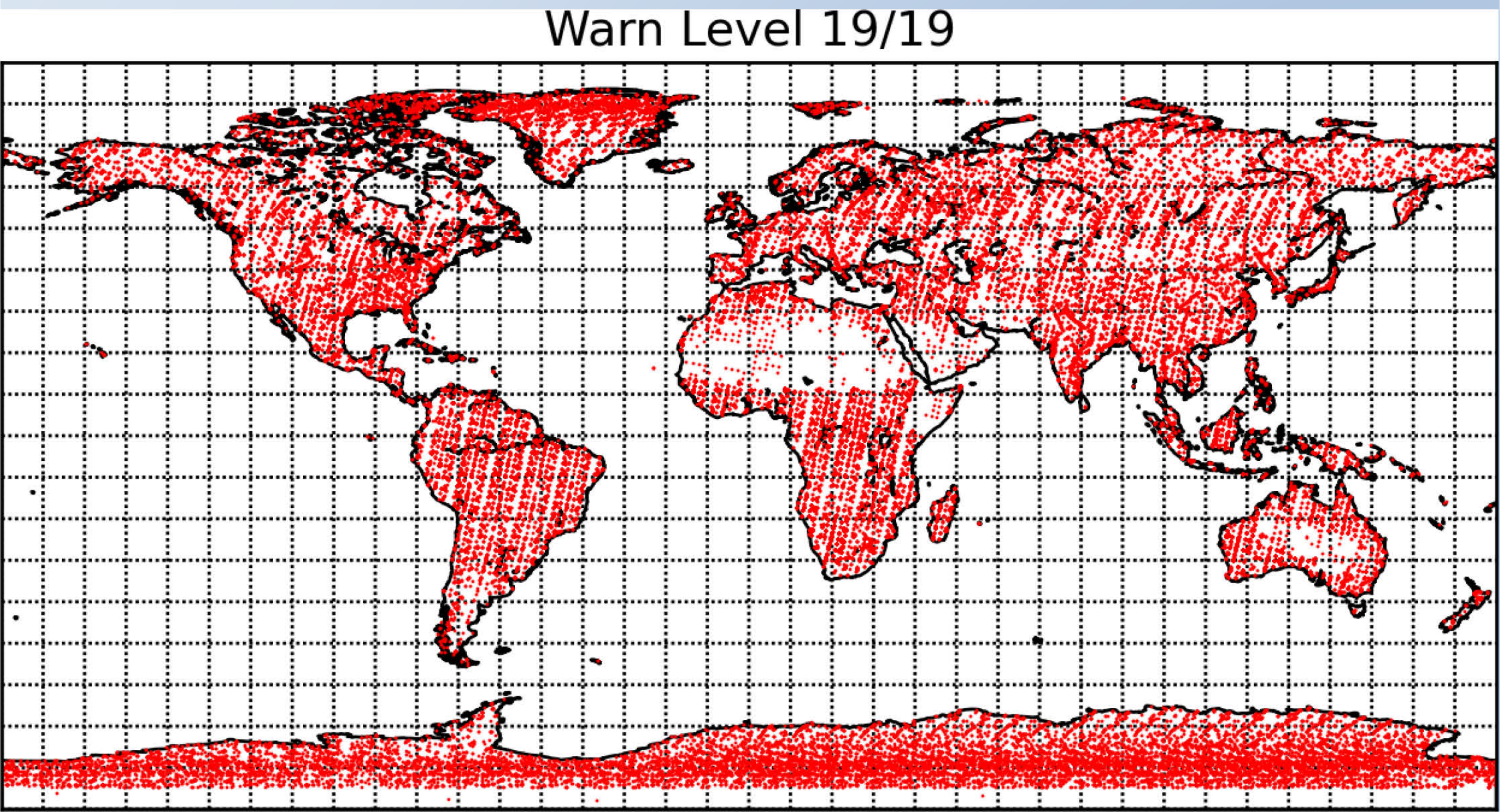
Global Spatial Analysis

Warn Level 10/19

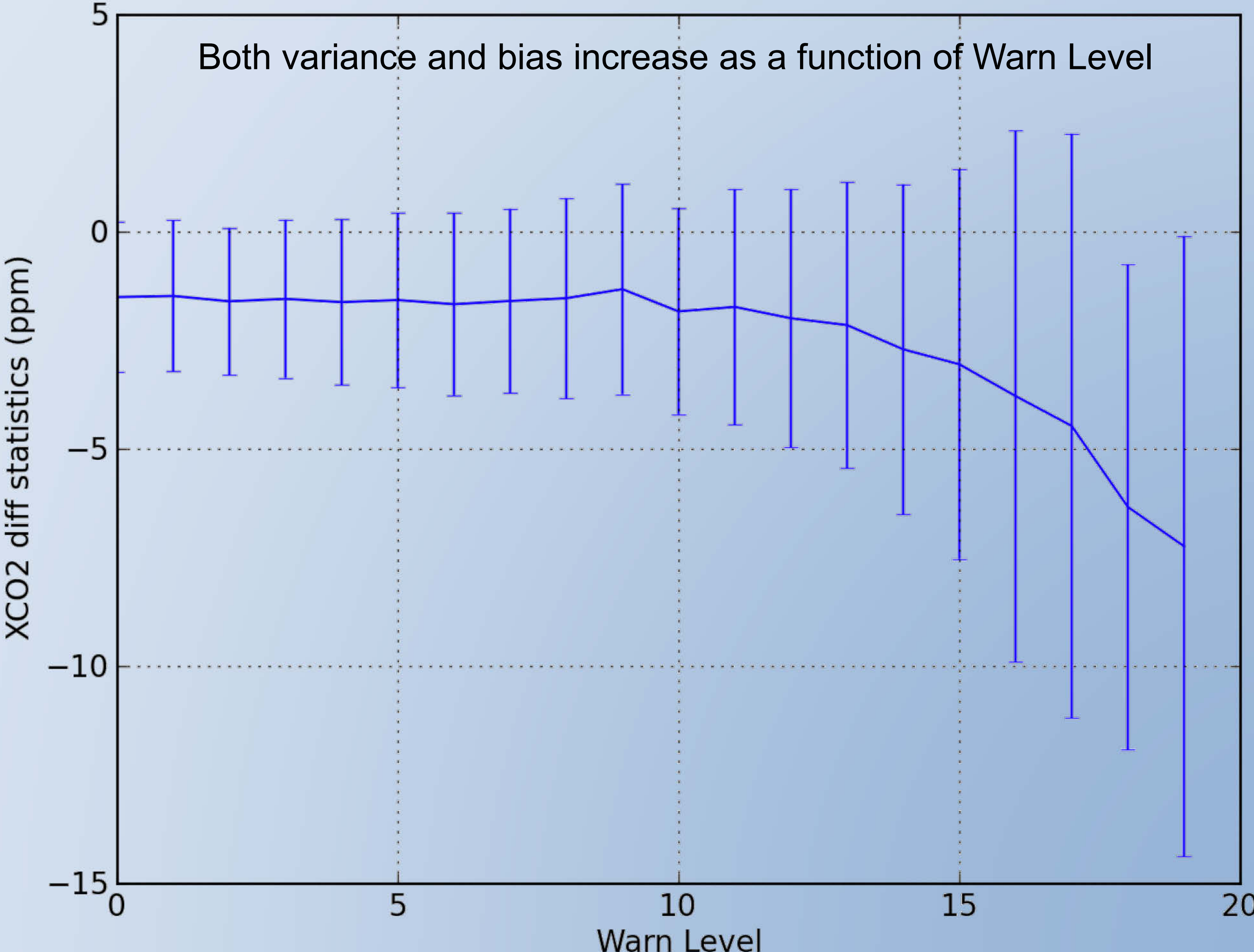


Global Spatial Analysis

Icy regions



Truth difference (Warn Level)



Solution Found

Have “Warn Levels” that:

-Perform similarly to a manually crafted expert system

-Permit dialable transparency for less/more data than above

-Identify several key features that correlate to quality of retrieval

-Create a new product that sorts soundings by likely utility

-Do not favor particular geographic regions or timespans

-Incorporate two truth metrics

-Can be used for Sounding Selection (pre-algorithm) or Quality Estimation (post-algorithm)

Future Work

What if we could classify those sub-populations and fit each of them individually?

Might resurrect bias correction

“Retrieval Identifier” to let you know what went wrong and attempt to correct…