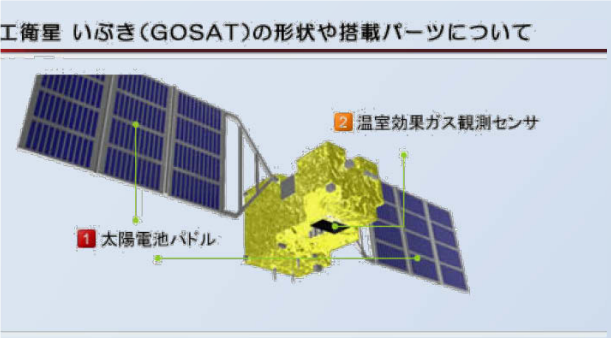
What can you do with a data filter?

V 2.10

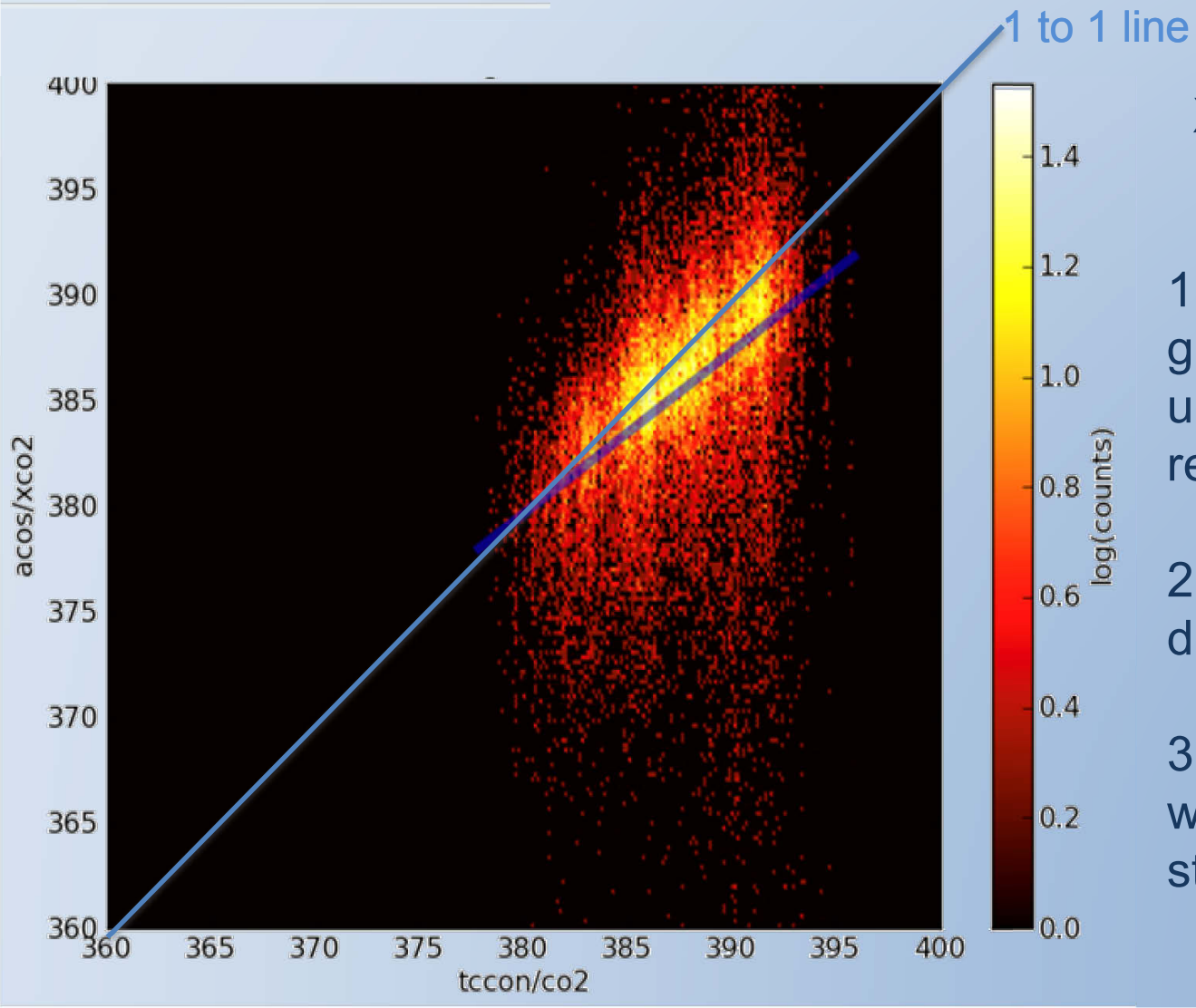
Lukas Mandrake

Jet Propulsion Laboratory, California Institute of Technology.
Copyright 2013 California Institute of Technology

ACOS vs TCCON



TCCON
total carbon column observing network



XCO2 from ACOS and TCCON differ

1)Can we make a filter to guide users to avoid undesired ACOS retrievals?

2)Can we “fit out” disagreement?

3)Can we figure out when/why xco2 differs so strongly?

Past Work

Debra Wunch's paper for 2.8 (2.9)

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	$2.4 \times$ albedo_o2_fph $- 1.13 \times$ albedo_strong_co2_fph < 1
Retain scenes unless with nonzero X_{CO_2} uncertainties	xco2_uncert $\neq 0$

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

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

Chris O'Dell's memo for 2.10

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	

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

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

Two paths

1) Gather all B2.10 L2 (Land, H Gain) & TCCON from 2008-2012

1) Add Temperature @ 700 hPa to each ACOS and TCCON point
   - NCEP is every 6 hours, 2.5 deg x 2.5 deg grid across planet

2) Find TCCON coincidences between ACOS and TCCON
   - Debra's rules: ±30° lon, ±10° lat, ±5 days, ±2K T700

3) Simultaneously solve for Filter & Regression parameters
   - Filter transparency %, # filter rules, fit error, # fit terms

4) Create variable filter that best estimates sounding agreement with TCCON/Southern Hemisphere Approx

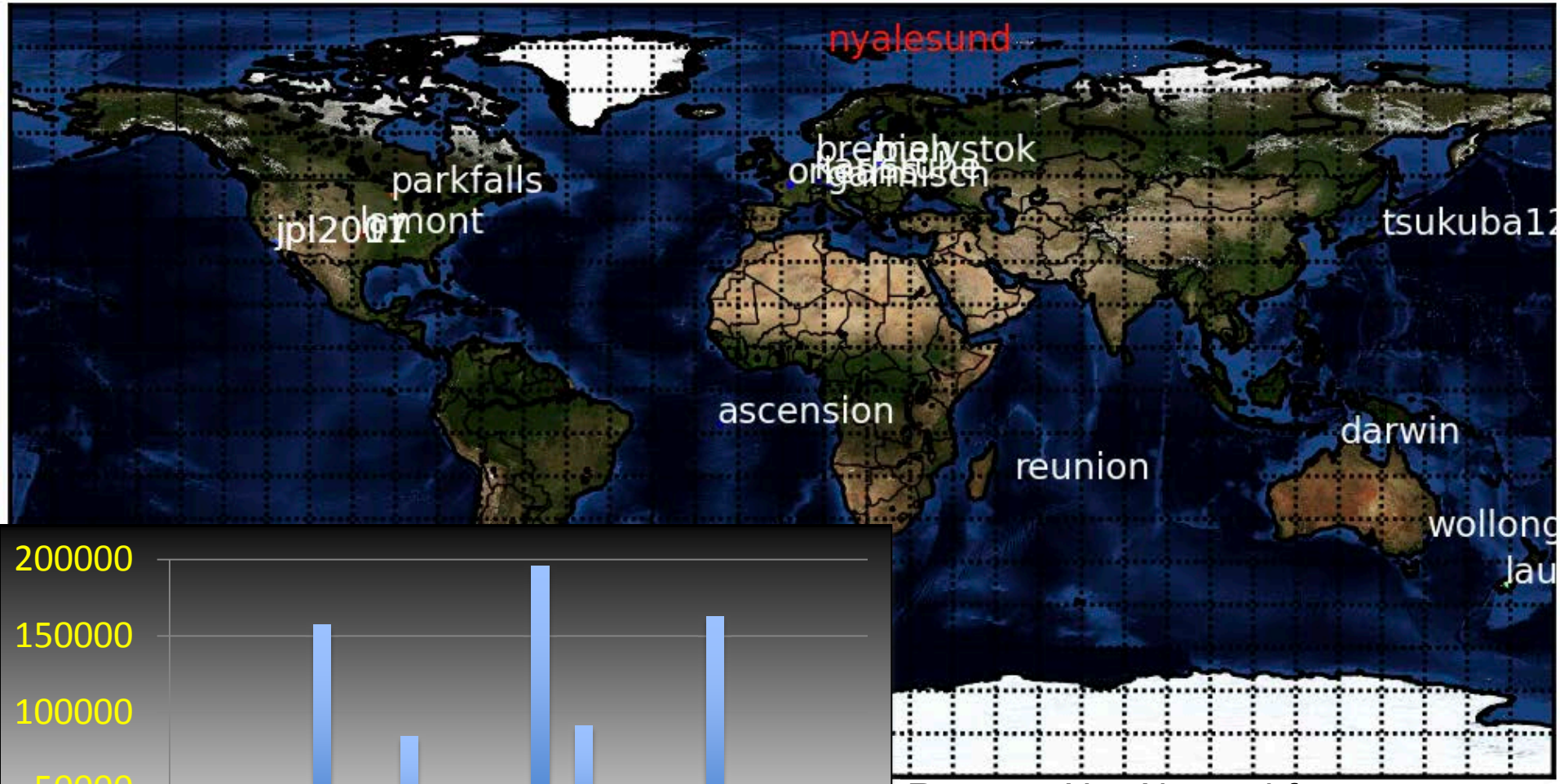
Data Prep

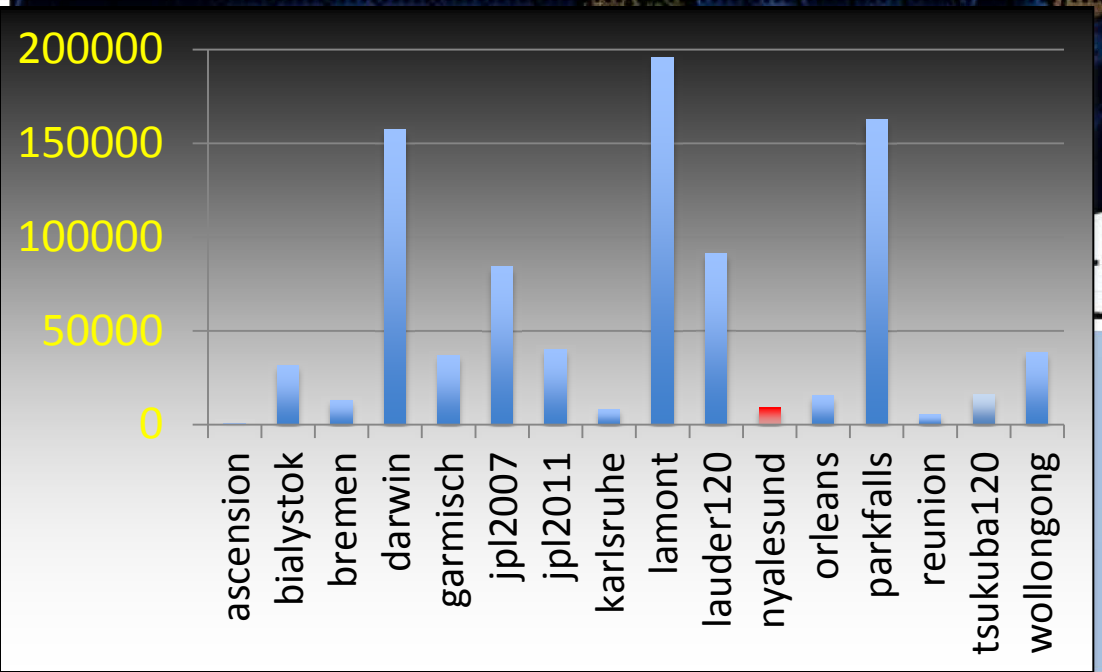
Bias Correction

Quality Estimation

Data Prep

TCCON data distribution

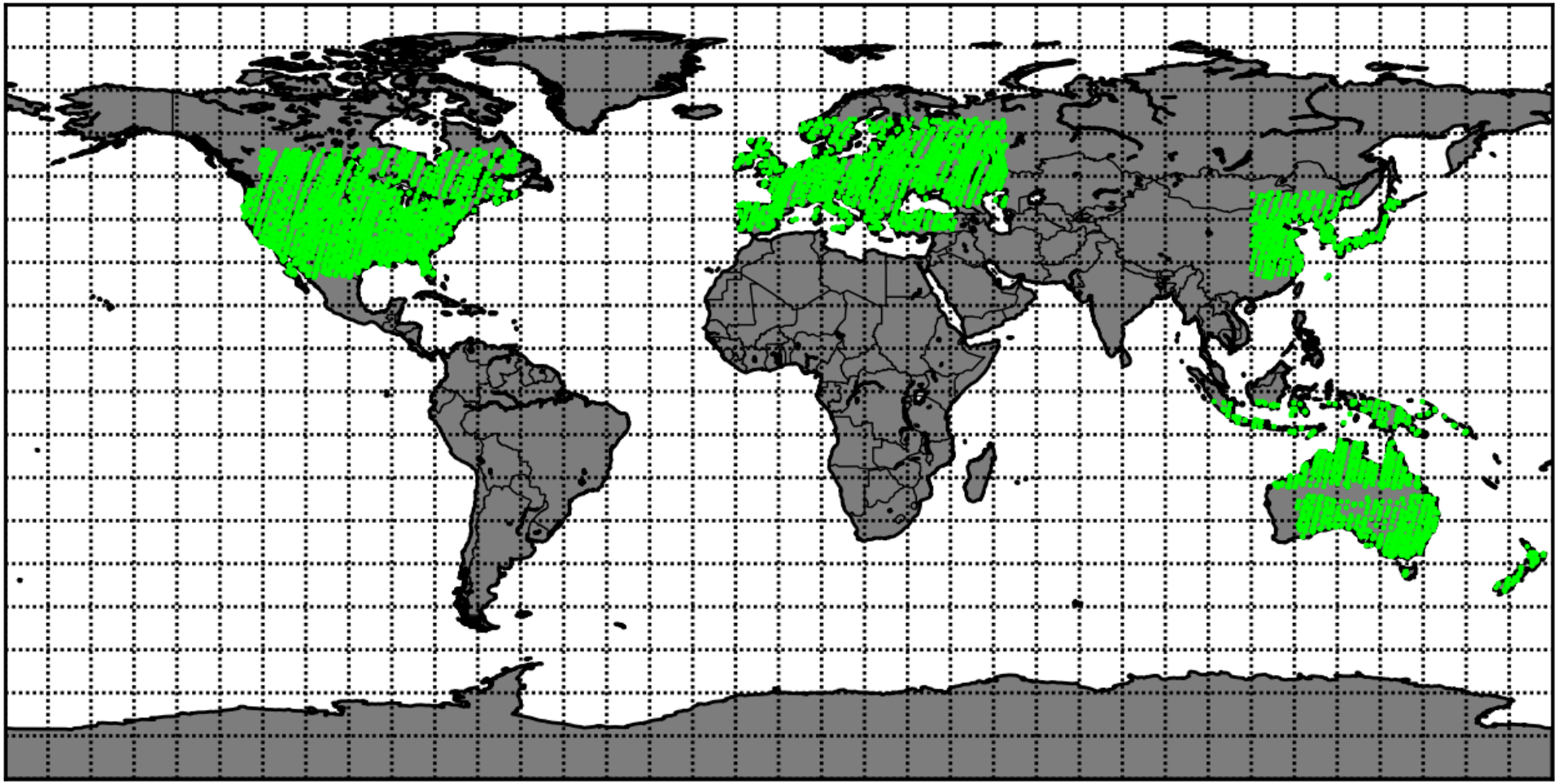




Remove Ny_Alesund for Ice issues.

Total TCCON: 645,986

ACOS TCCON Coincidences

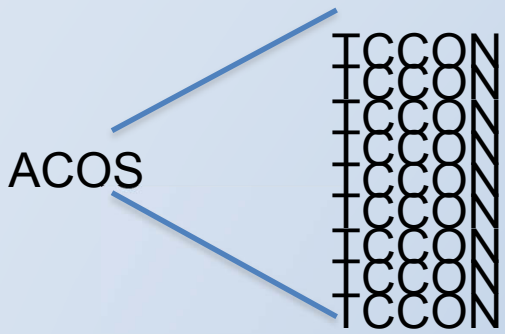


ACOS data: H gain, land, not cloudy, IMAP ok, converged, and coincidence criteria = 62,583 soundings

We will be filtering more than this later

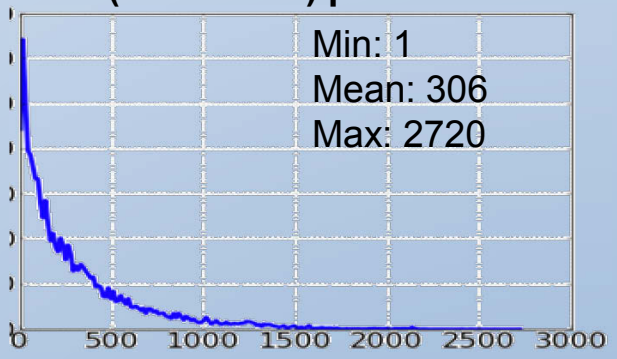
Initial Coincidences: TCCON Consistency

For each ACOS point, take mean of all TCCON within coincidence criteria

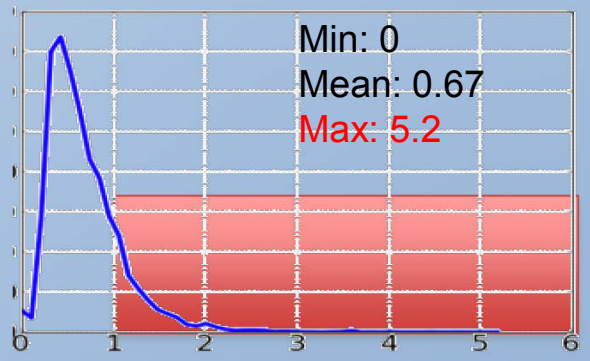


Even TCCON self-agreement has some problems

NUM(tccon CO2) per coincidence



STDEV(tccon CO2) per coincidence



Implement new coincidence filter:
STDEV(tccon) must be < 1 ppm

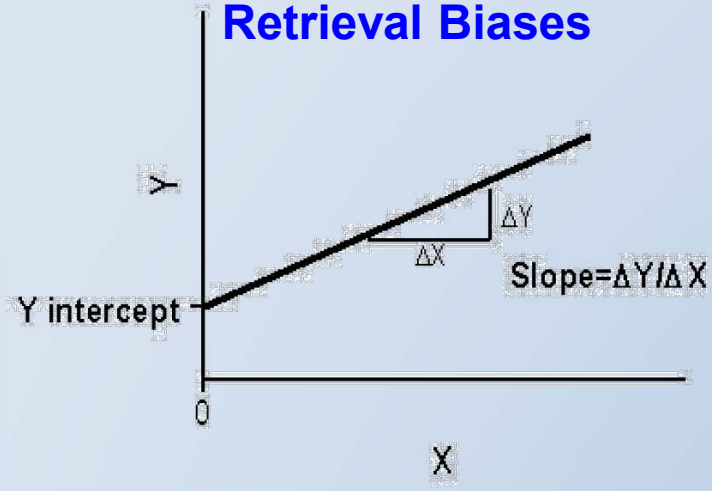
Filter Fit

Bias Correction

How do filtration and bias correction relate?

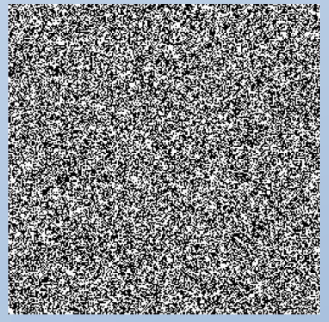
Mental Model of Data

Predictable Physics & Retrieval Biases



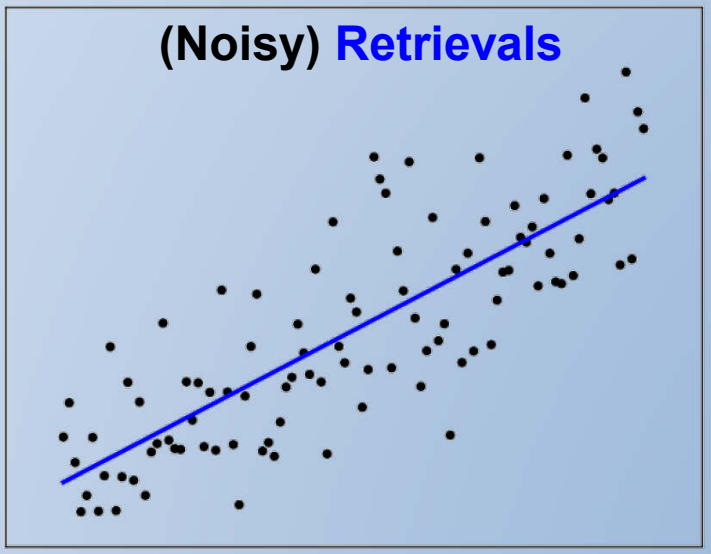
+

Noise (aerosols, Unmodelled complexities, Measurement error)



=

(Noisy) Retrievals



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

Filter, then Fit, then Judge

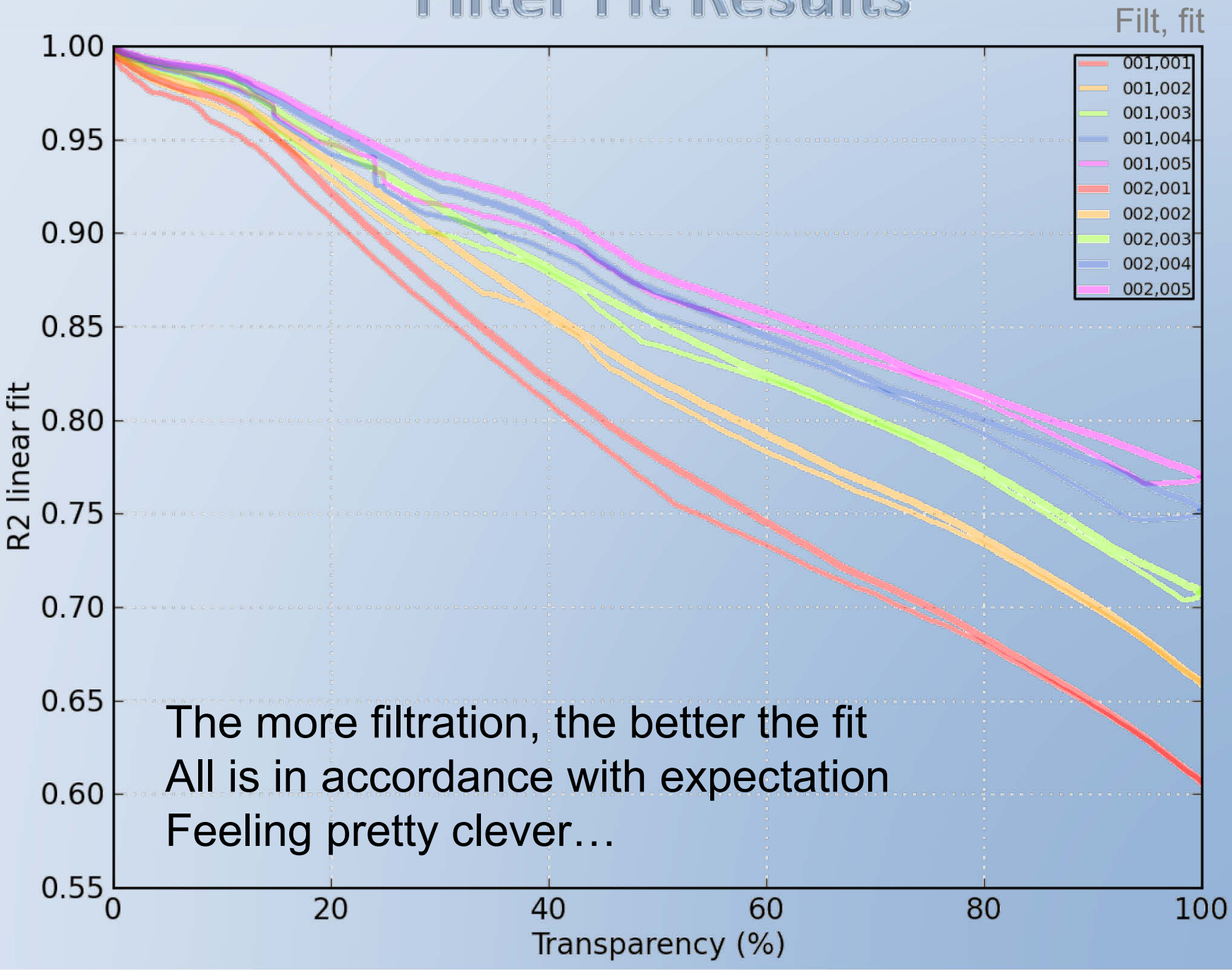
Use the (m-feature) threshold filter & (n-feature) linear fit
Randomly search all filter and fit feature spaces

Metric is R^2 of the final fit…

- better filtration (features and thresholds) = better fits
- better fit features = better fits

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

Filter Fit Results



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 HIGH dof_co2_profile first

First to go? Bright desert regions

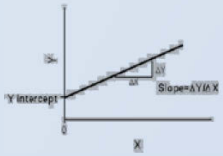
I made a cloud/aerosol SELECTOR

Now why would my code find that solution?

All I told it to do was improve the fit…

(New) Mental Model of Data

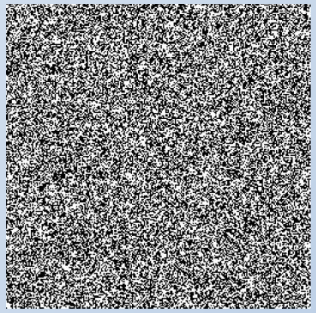
Predictable Physics & Retrieval Biases



(airmass, etc.)

+

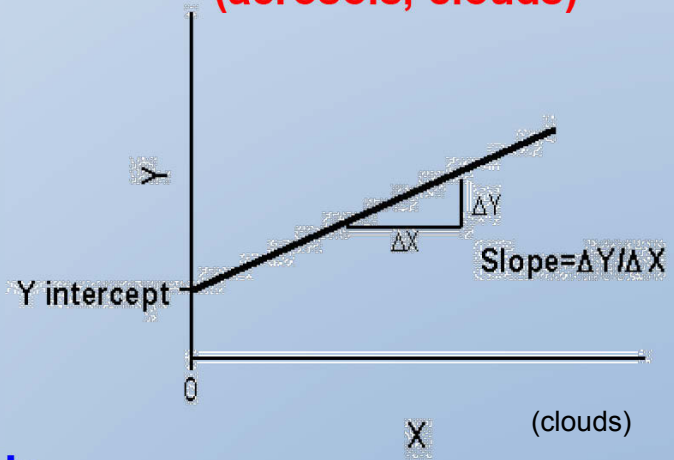
Actual sounding variability



(unmodeled physics, surface topology, etc.)

+

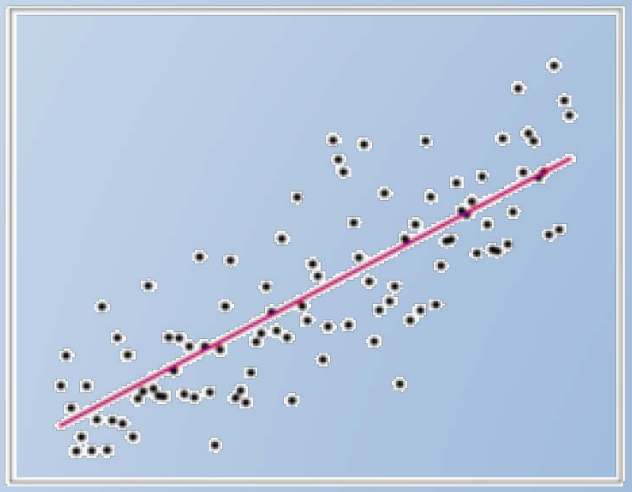
Systematic Contamination (aerosols, clouds)



(clouds)

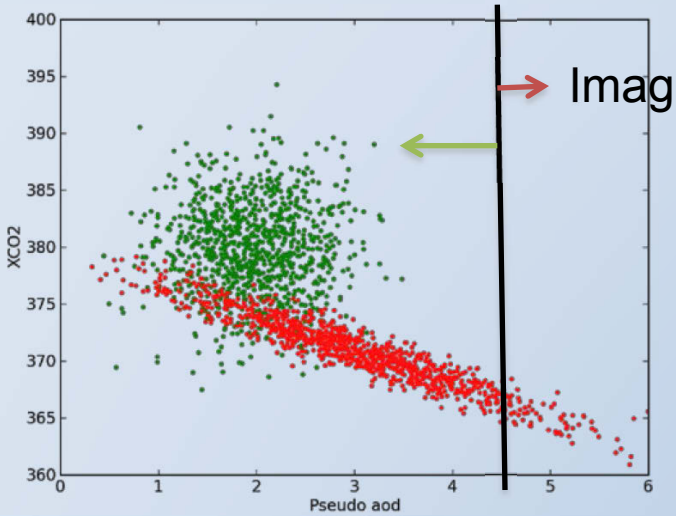
=

(Noisy) Retrievals



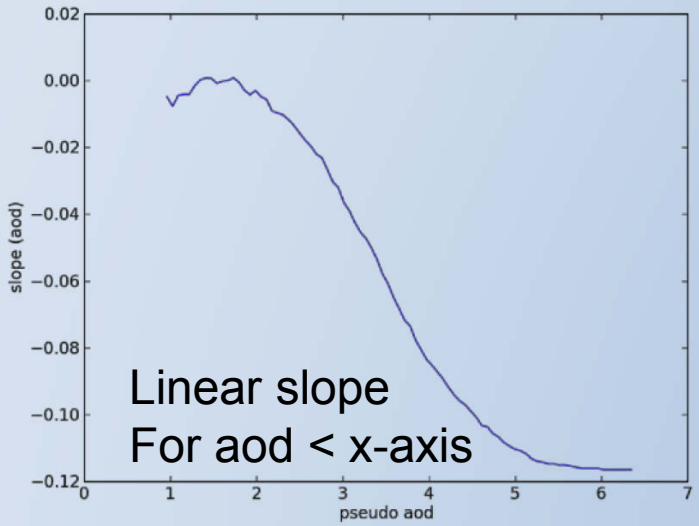
Conclusion: We have a much more difficult task to discover any underlying XCO2 bias beneath a much larger systematic bias due to aerosol/cloud contamination

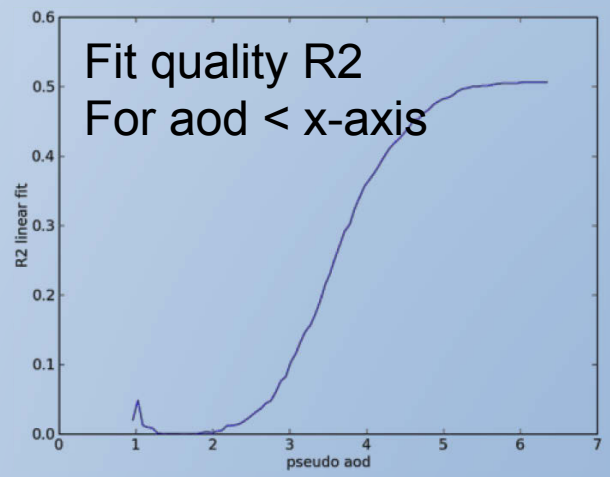
Simulation: Clouds vs. Not Clouds



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

Bright deserts and other “known good” solutions
were tossed immediately

FILTER SOLUTION: Two Pass

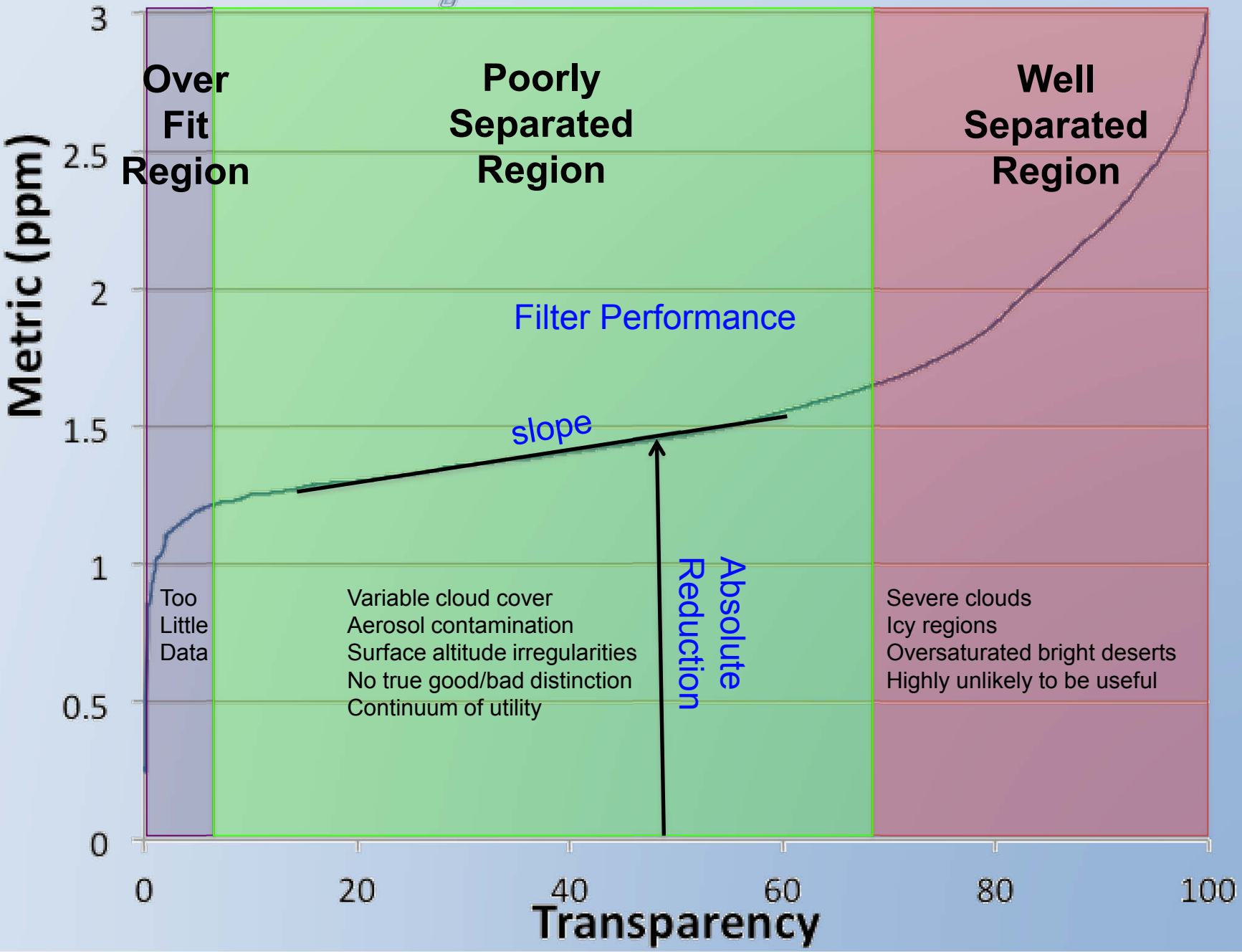
1) Create filters based on reducing the ACOS-TCCON diff alone

2) Select the best fit for each filter solution afterwards

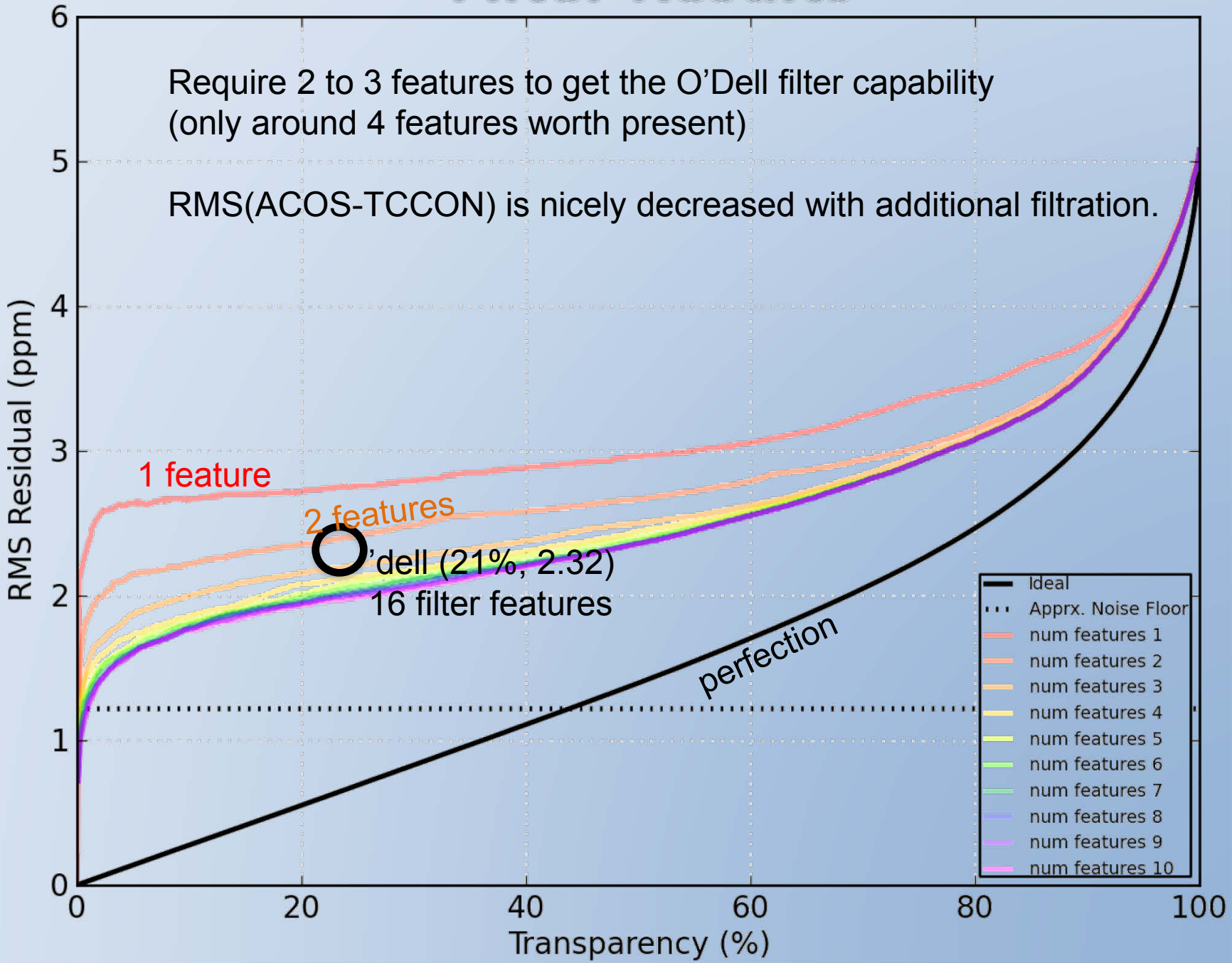
Two Pass Filter Fit

Bias Correction

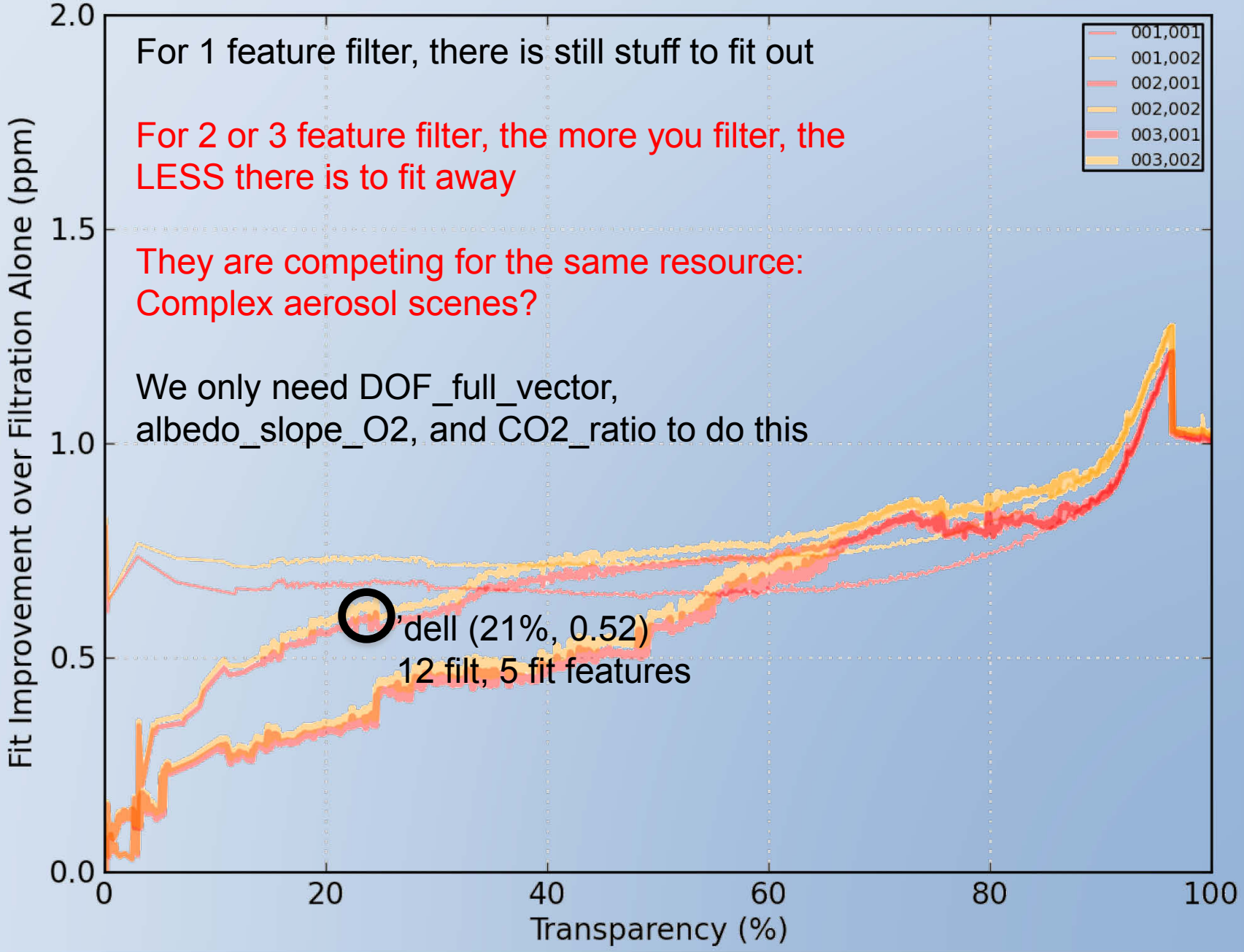
Anatomy of a Filter Curve



Filter Results



Fit - Filter Results



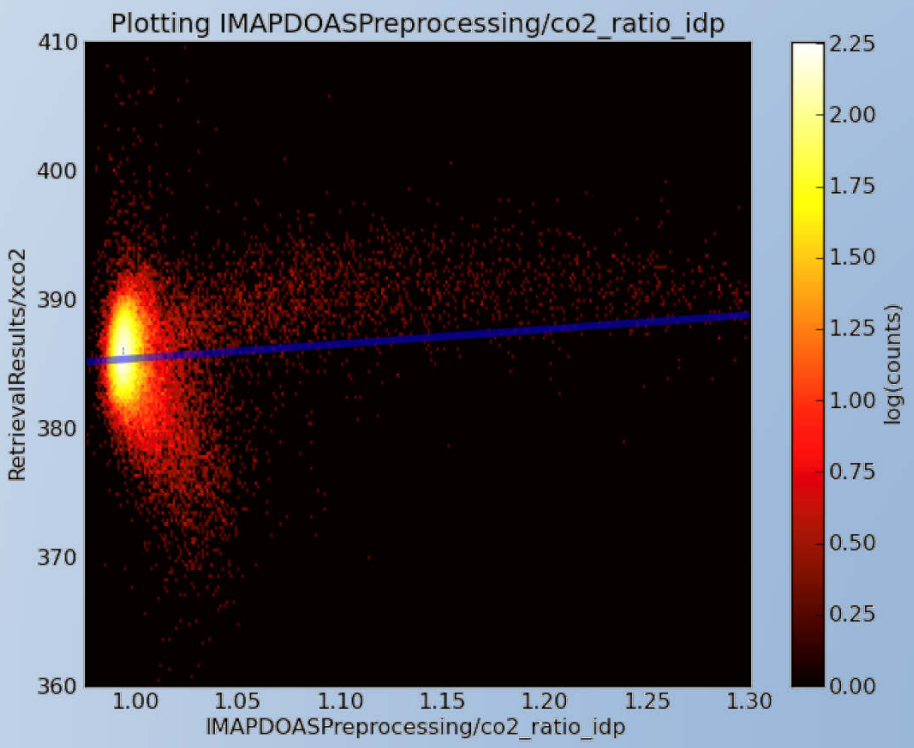
Don’t bias correct

We have a 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

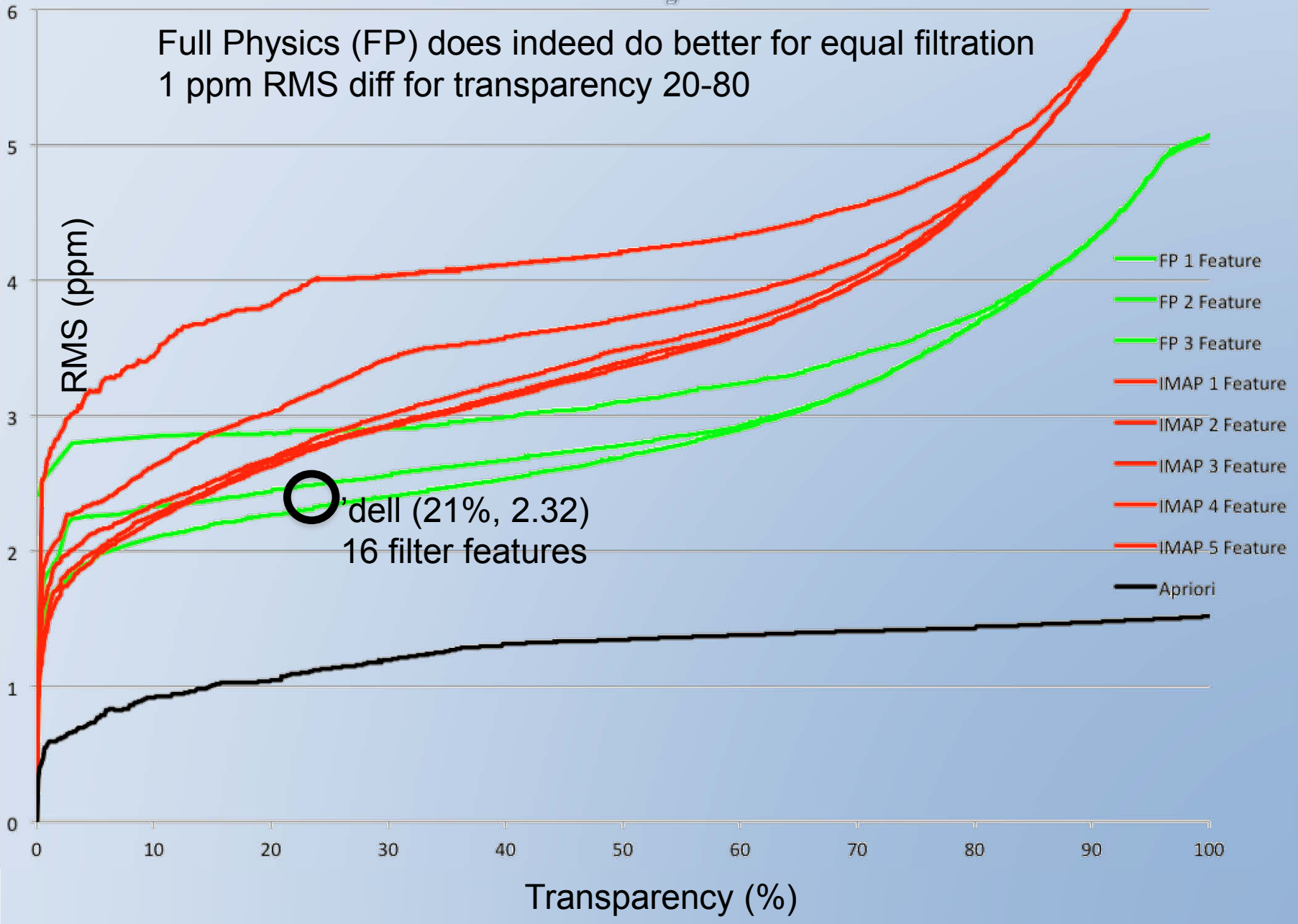


Bonus Issue:

Does IMAP/DOAS do just as well?

Don’t use any Full Physics outputs…

IMAP vs. Full Physics Filter Results

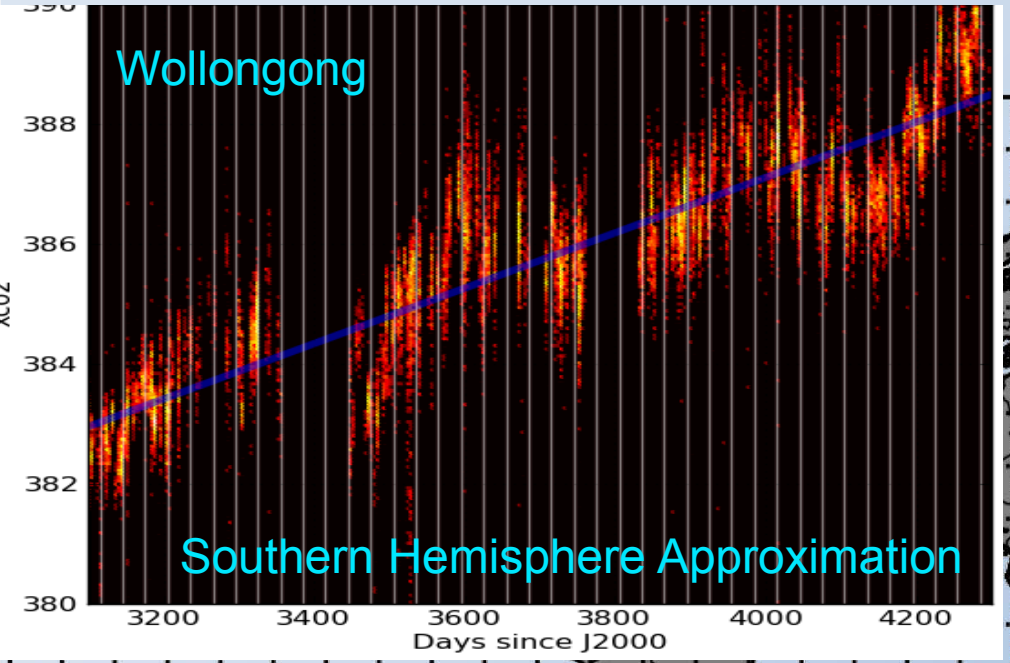


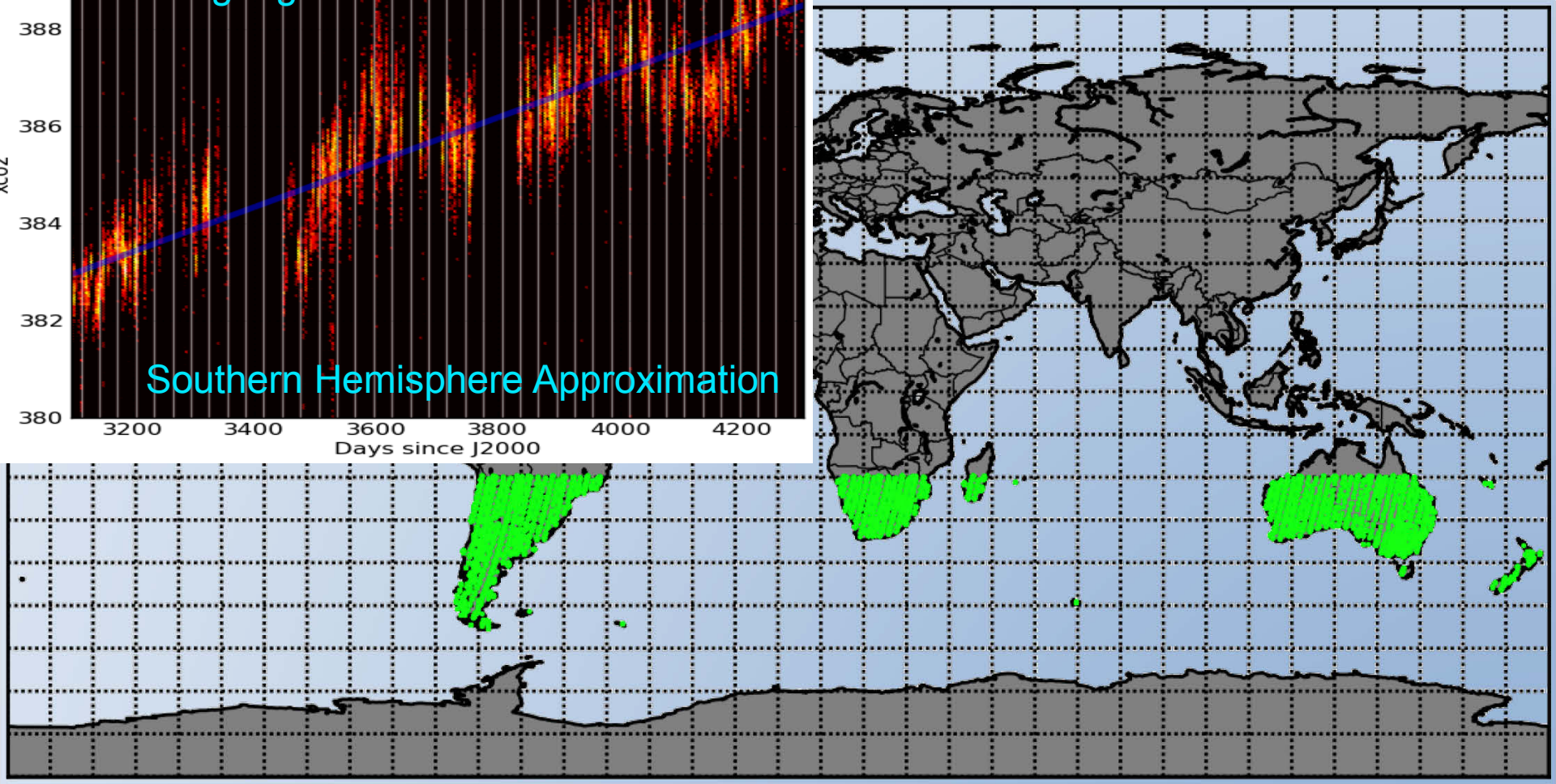
Reqtrieval Quality Estimation

What filter to use?

Quality Estimation

Two Metrics: SHA & TCCON





ACOS data: H gain, land, not cloudy, IMAP ok, converged, and southern latitudes (-20 to -60) = 72,341 soundings

Metric: Reduce mean(monthly STD(XCO2)

Method

Genetic algorithm + threshold filters to minimize:

SHA: the Mean Monthly STDEV (MMS)
of surviving data

ACOS-TCCON: the RMS of surviving data

Then examine features that dominate for each

Useful Features

Southern Hemisphere Approximation	
ABandCloudScreen	surface_pressure_offset_cld
IMAPDOASPreprocessing	co2_ratio_idp
IMAPDOASPreprocessing	h2o_ratio_idp
RetrievalResults	aerosol_ice_aod
RetrievalResults	aerosol_water_aod
RetrievalResults	albedo_slope_strong_co2
RetrievalResults	albedo_strong_co2_fph
SoundingGeometry	sounding_altitude_stddev

TCCON Comparison	
ABandCloudScreen	dP_cloud
IMAPDOASPreprocessing	h2o_ratio_idp
RetrievalResults	aerosol_ice_aod
RetrievalResults	aerosol_total_aod
RetrievalResults	aerosol_water_aod
RetrievalResults	albedo_slope_o2
RetrievalResults	dof_full_vector

Land only

Roughly 6 groups of information types being targeted

Desire ~ 5 features that minimize both metrics well

Keep in mind potential for Glint quality flag

Selected Features

TCCON & SHA	
IMAPDOASPreprocessing	co2_ratio_idp
RetrievalResults	aerosol_ice_aod
RetrievalResults	albedo_slope_o2

These three features work together to obtain most of the filtration power

Combining the Metrics

Have two different filter classes based on different metrics (SHA/TCCON)

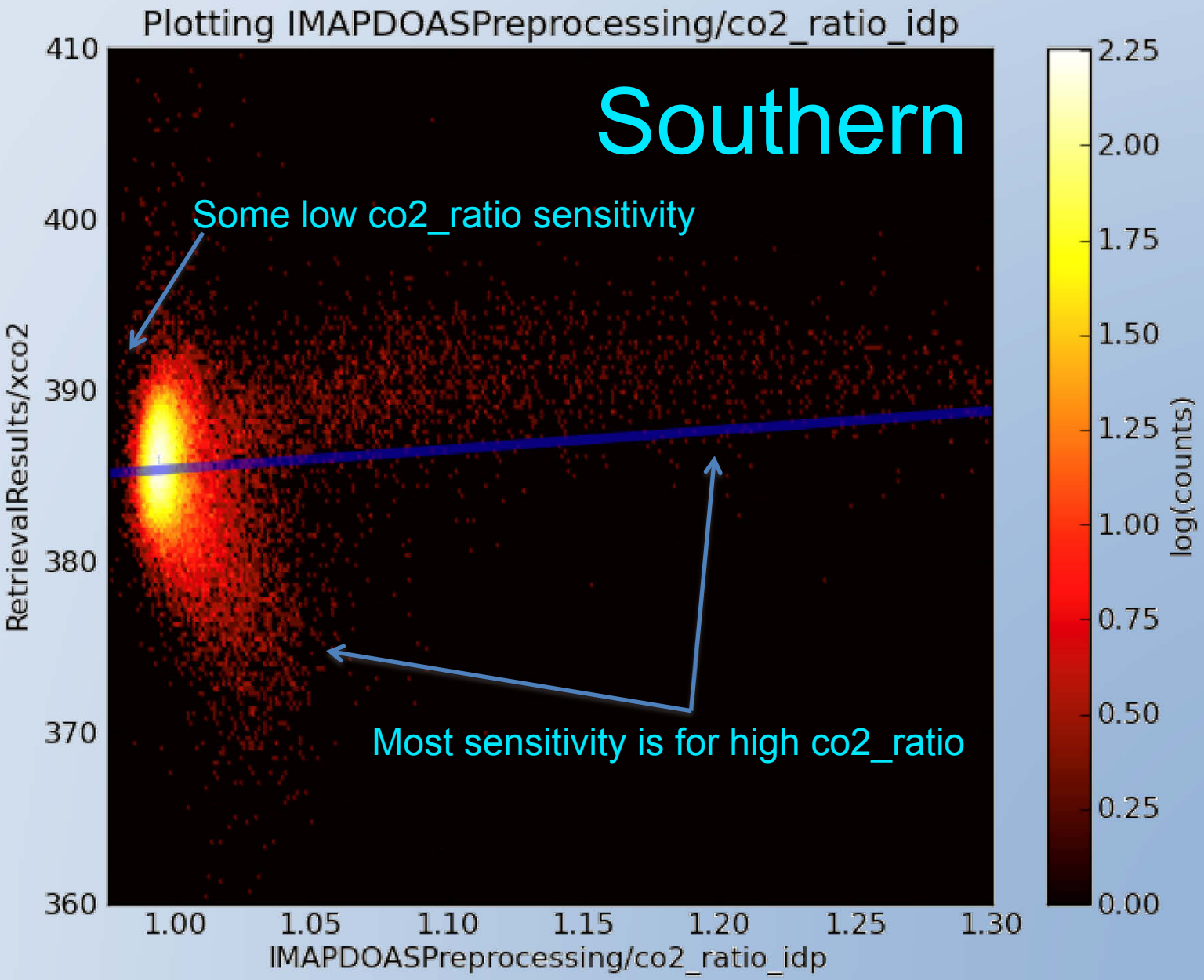
We must:

1) Unify their thresholds

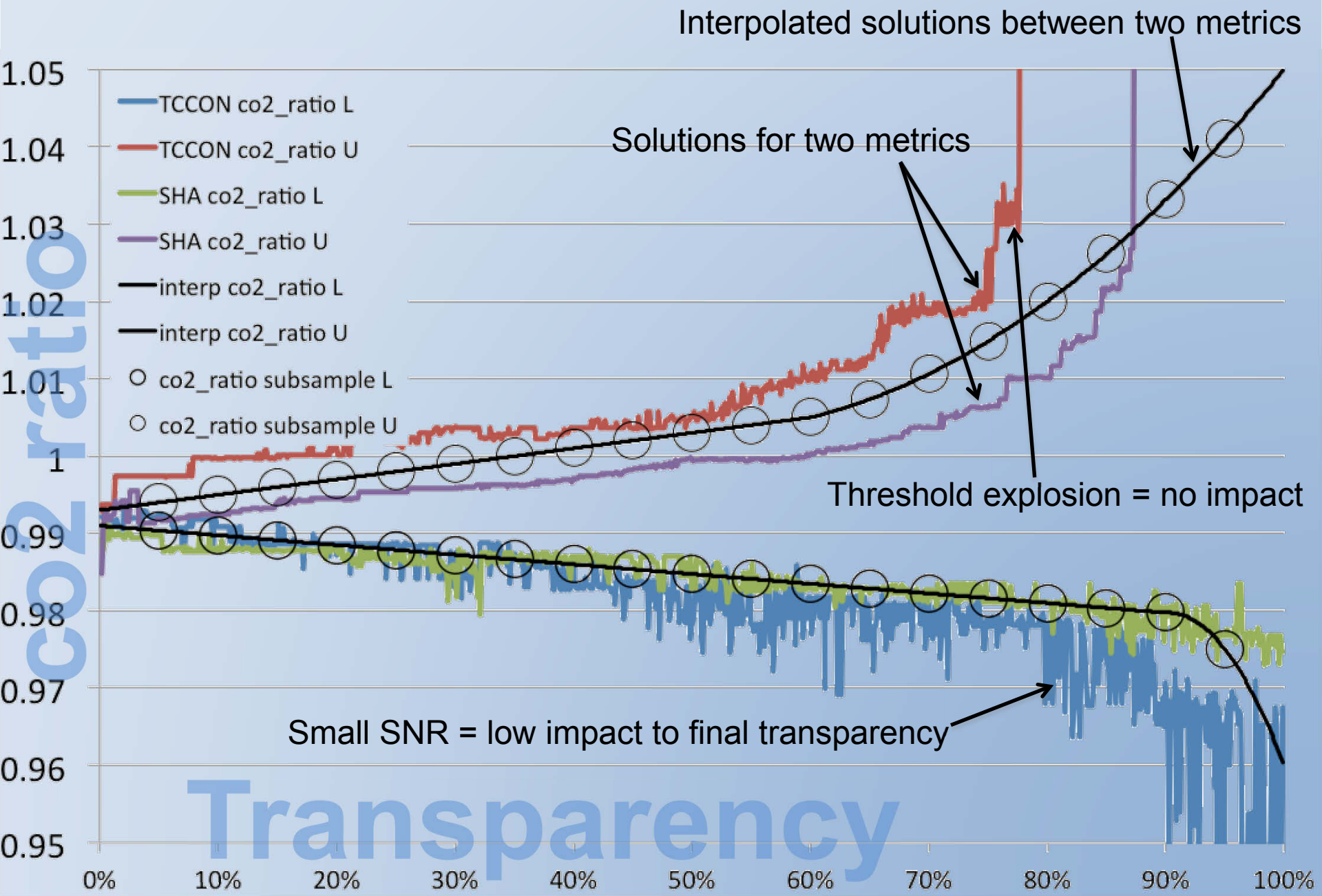
2) Subsample the thresholds into warning levels

Here come some busy graphs. Take a deep breath…

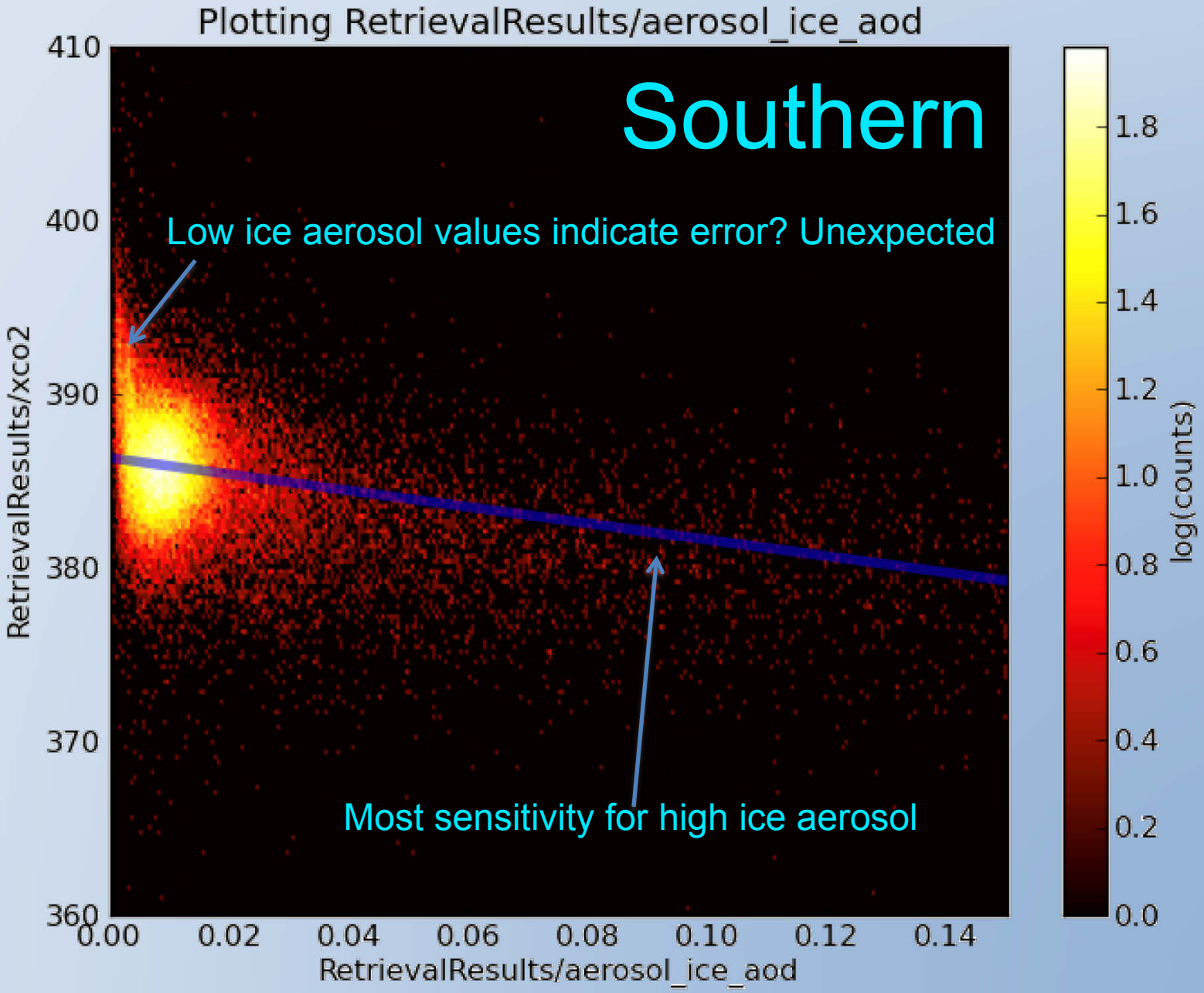
Feature 1: CO2_ratio



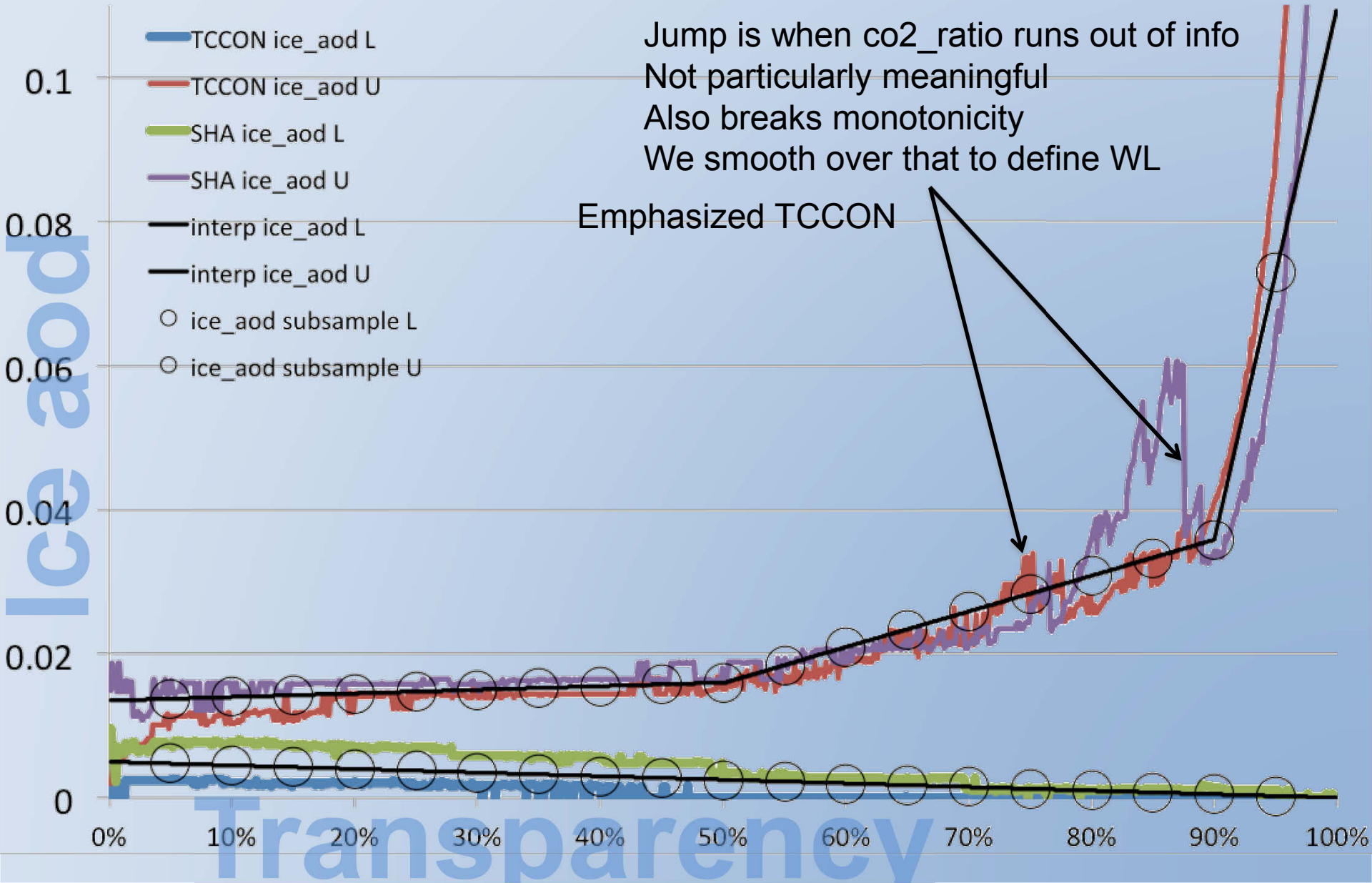
CO_2 Ratio: Thresholds and Warn Levels



Feature 2: ice_aerosol_AOD

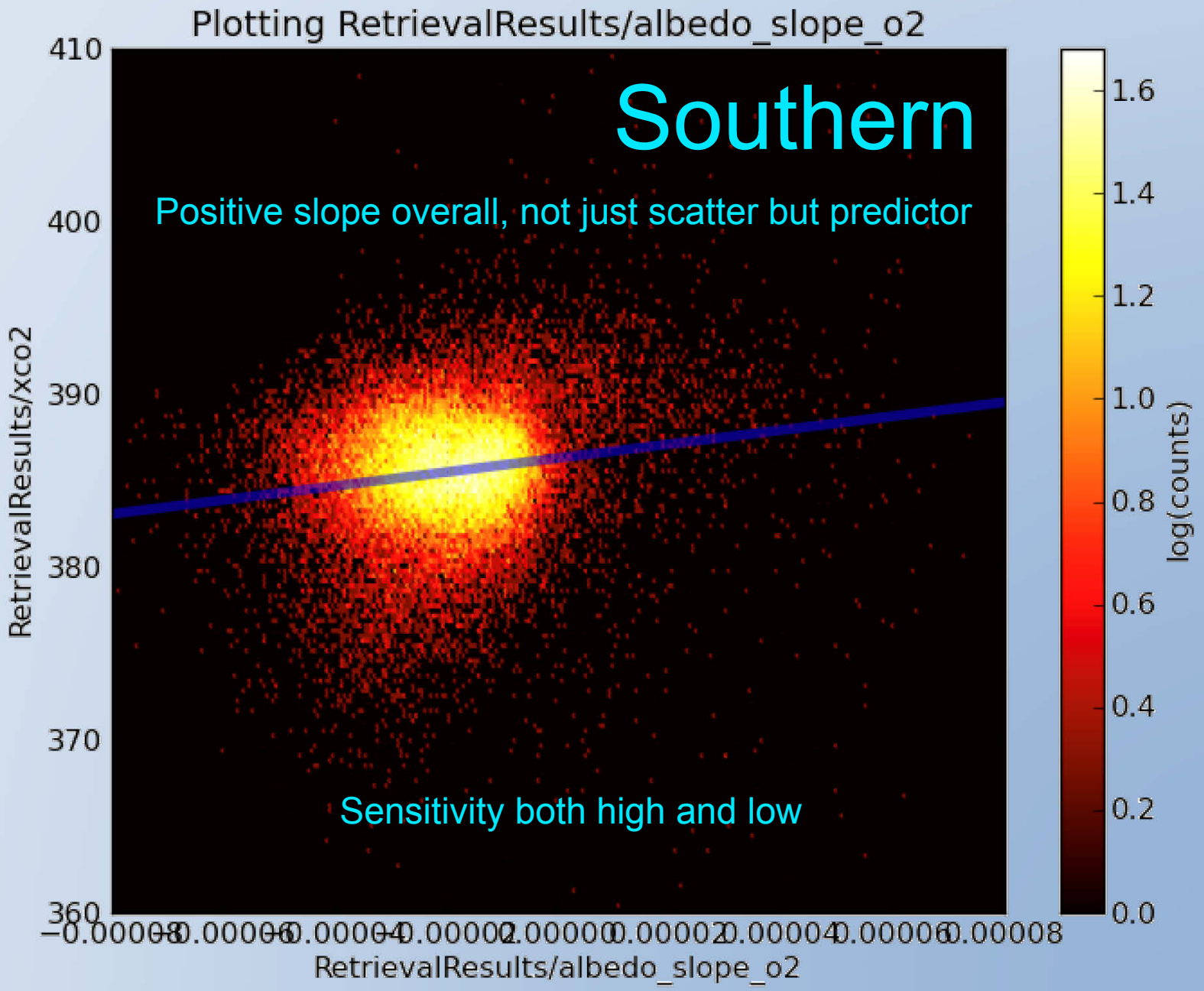


Ice aod: Thresholds and Warn Levels

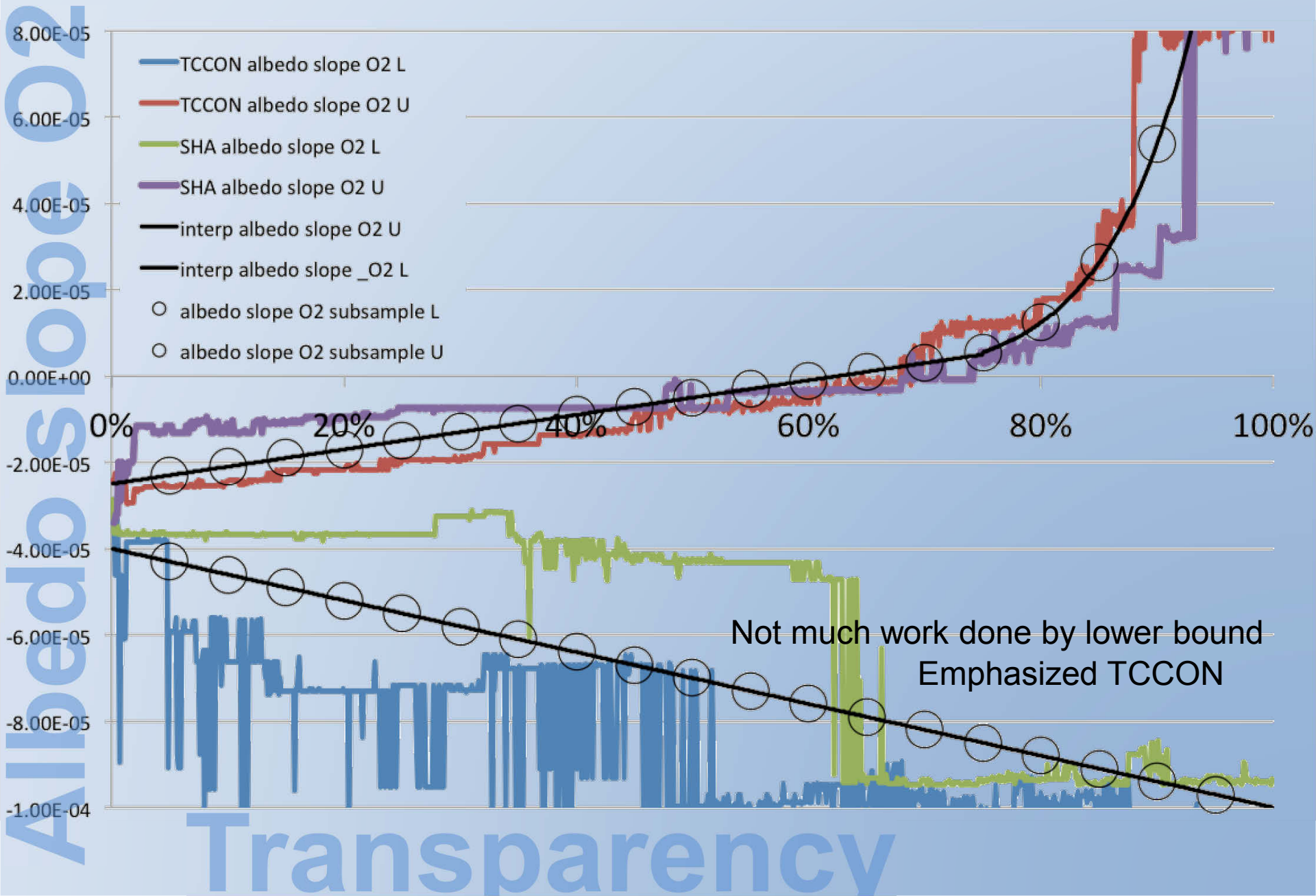


Transparency

Feature 3: albedo_slope_O2



Albedo slope O2: Thresholds and WLs



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
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

Least Conservative

Most Conservative

IMAPDOASPreprocessing co2_ratio_idp

RetrievalResults aerosol_ice_aod

RetrievalResults albedo_slope_o2

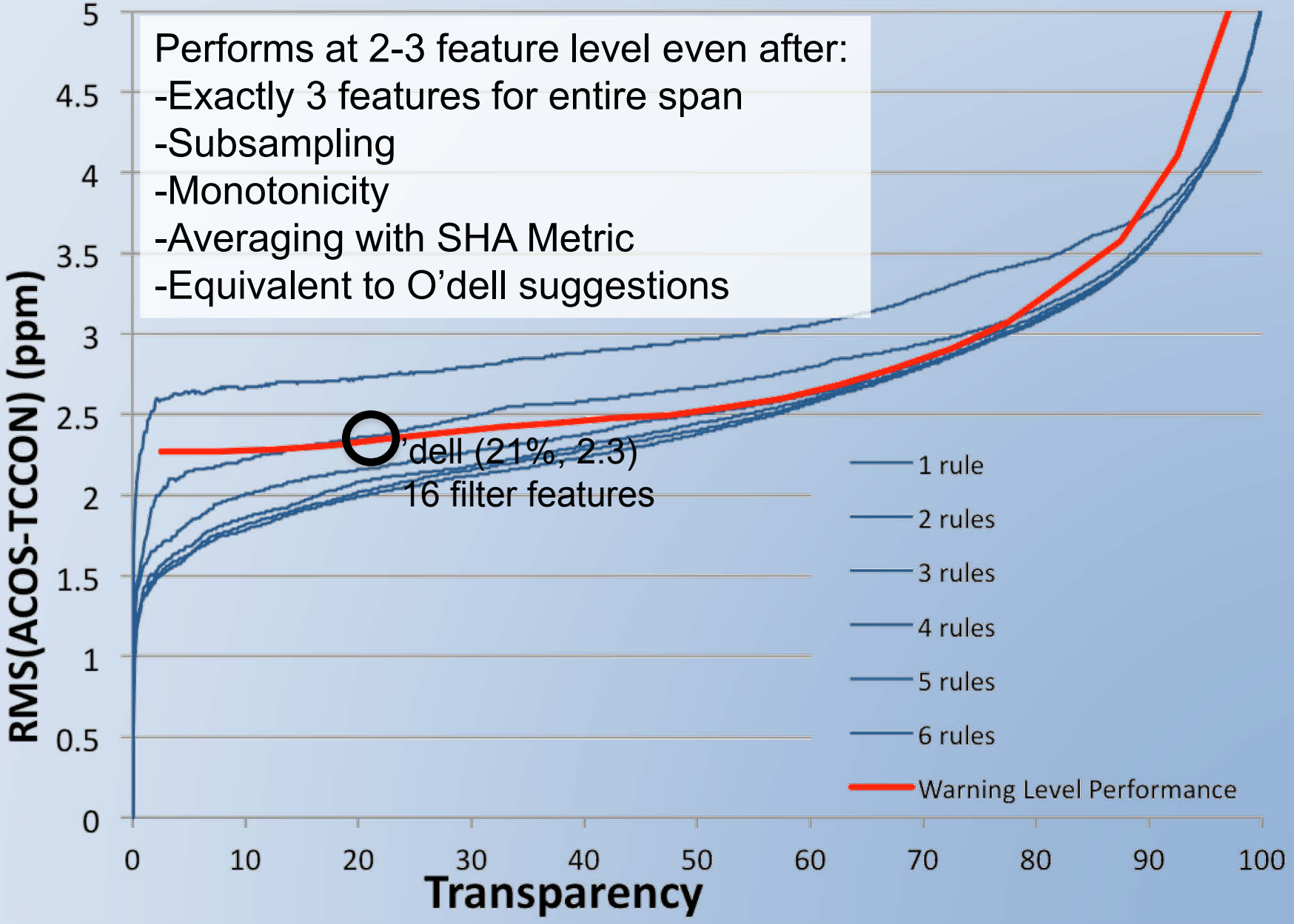
Evaluate Resulting Filter

The Warn Levels / Quality Estimation System are now defined.

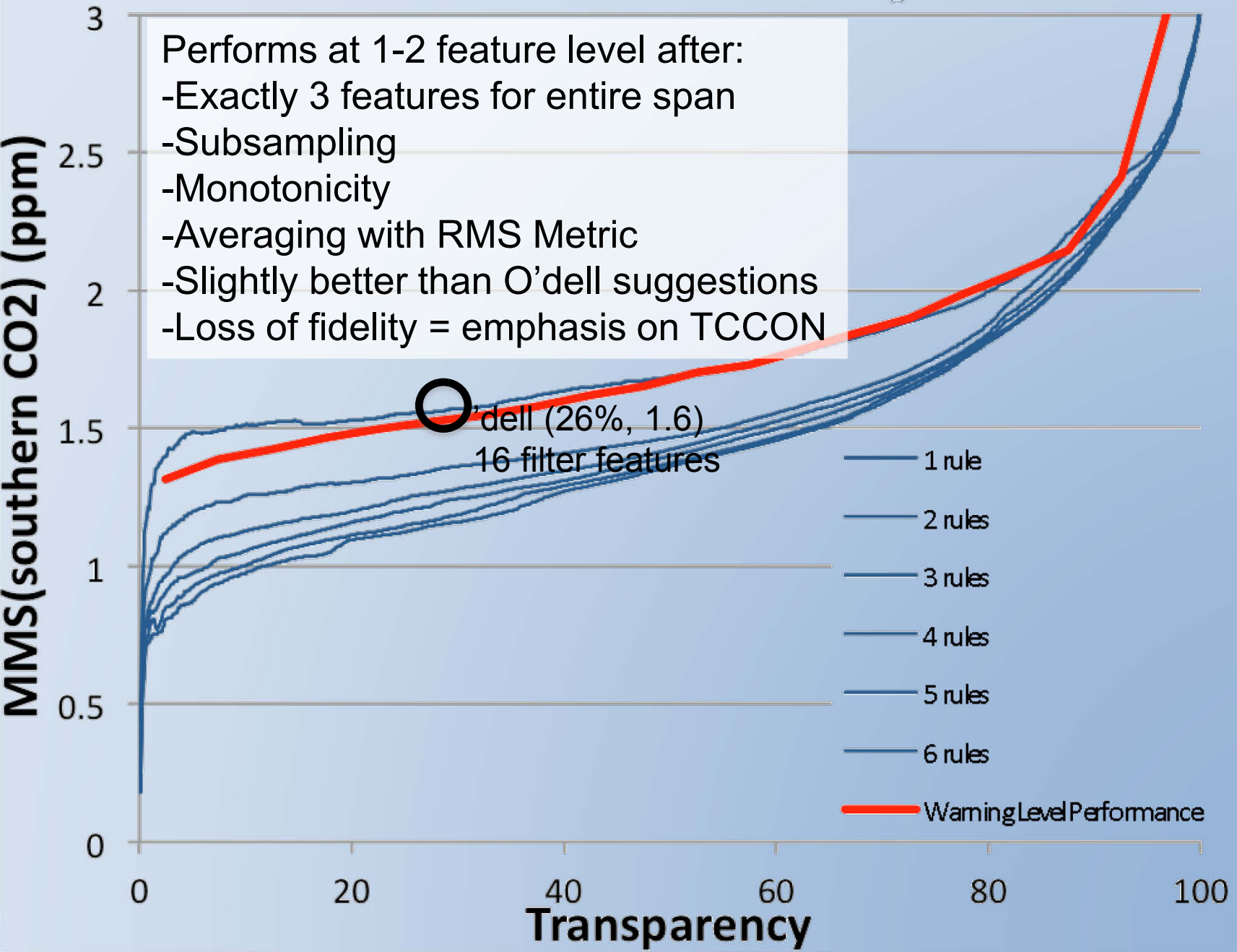
Evaluate resulting system for:

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

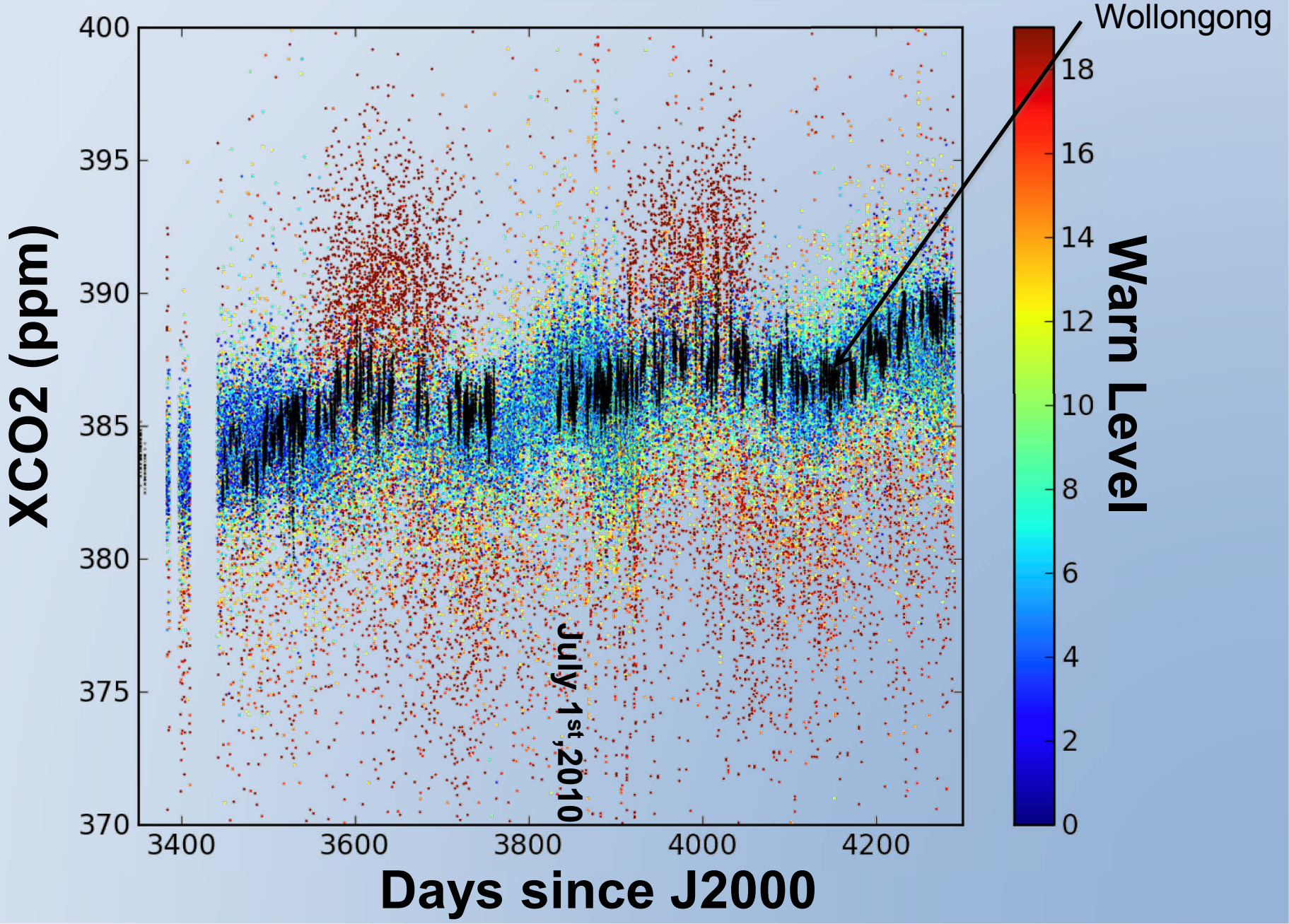
TCCON Metric Comparison



SHA Metric Comparison

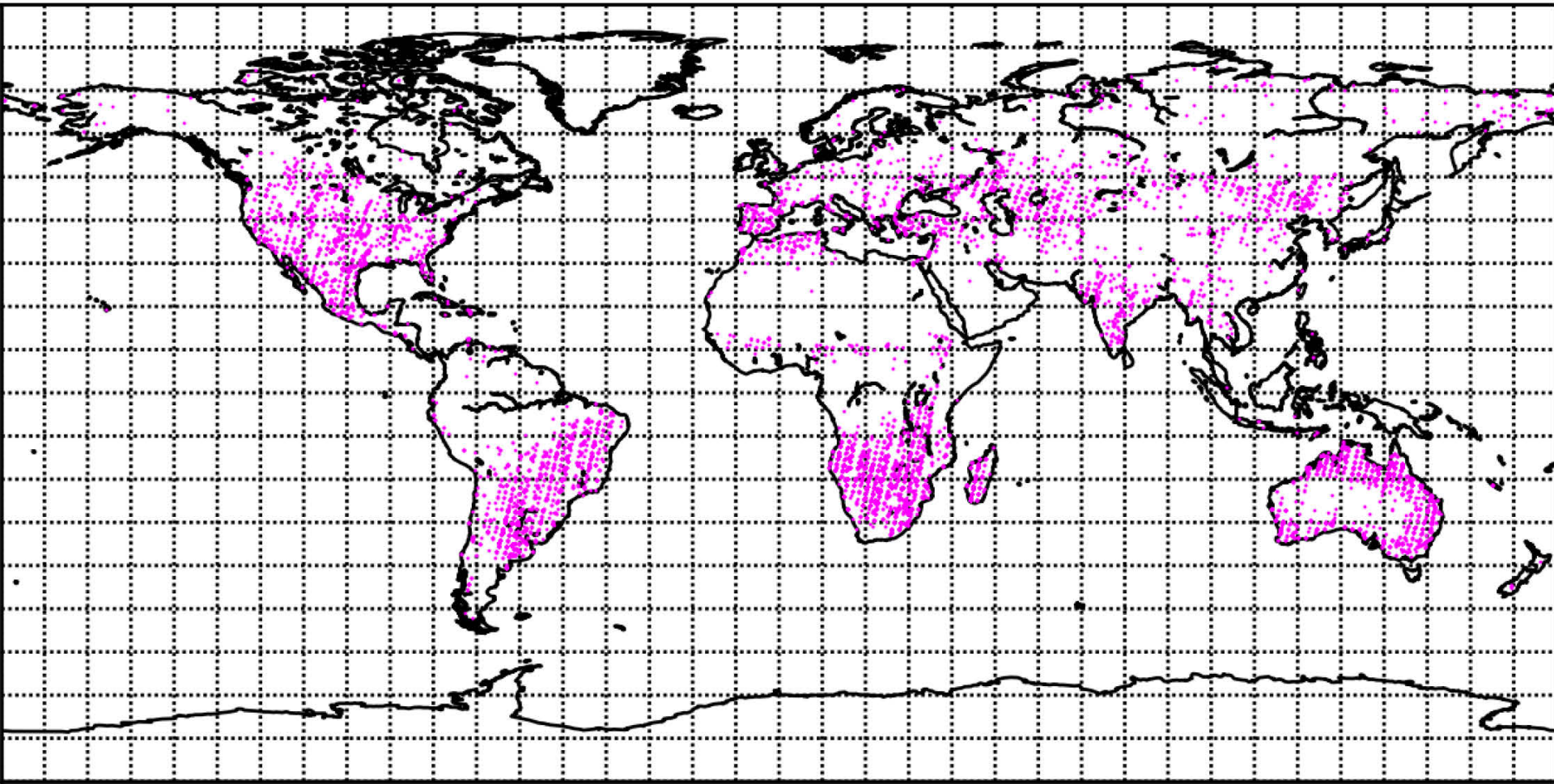


Temporal Analysis SHA



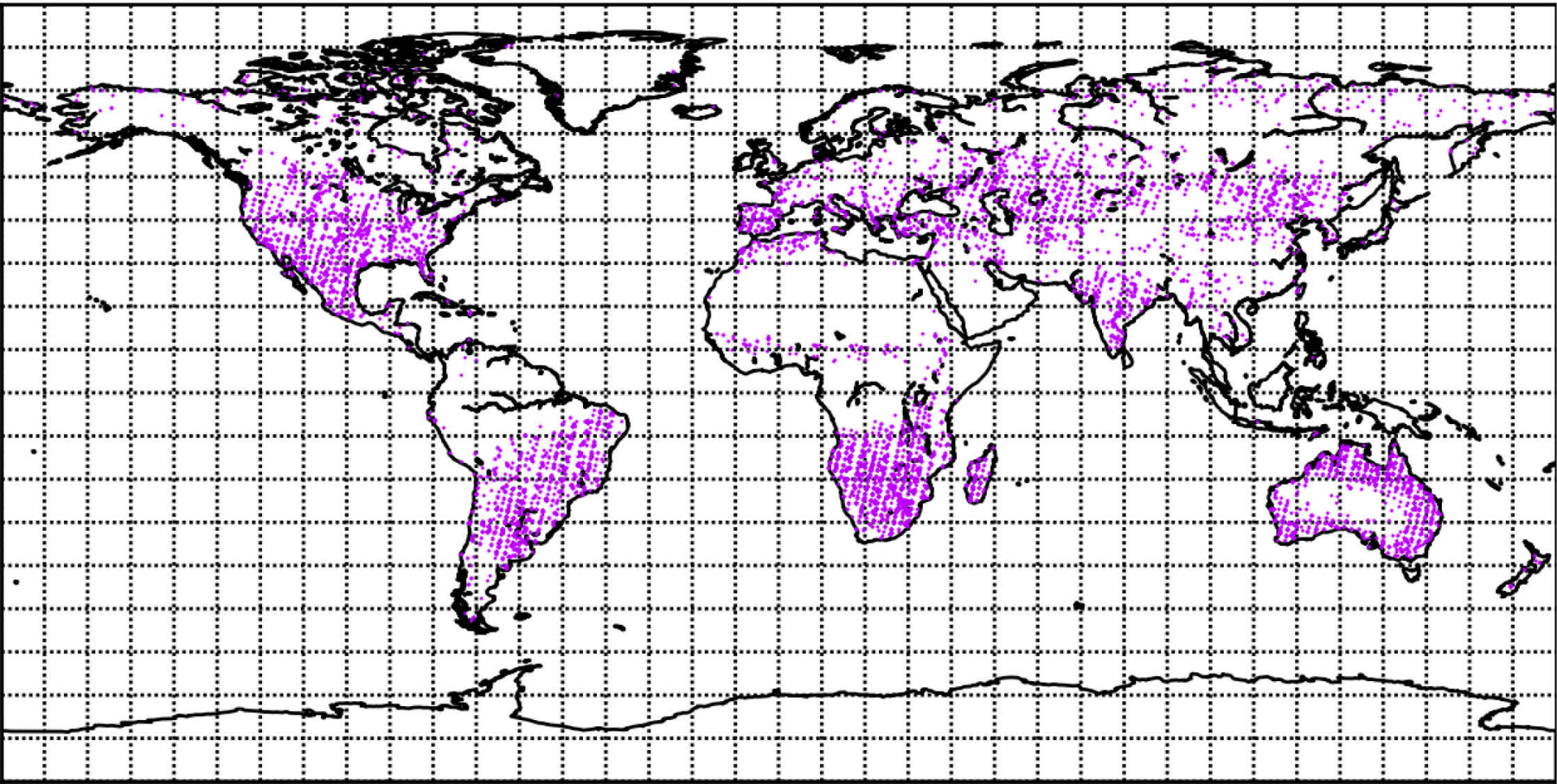
Global Spatial Analysis

Warn Level 0/19



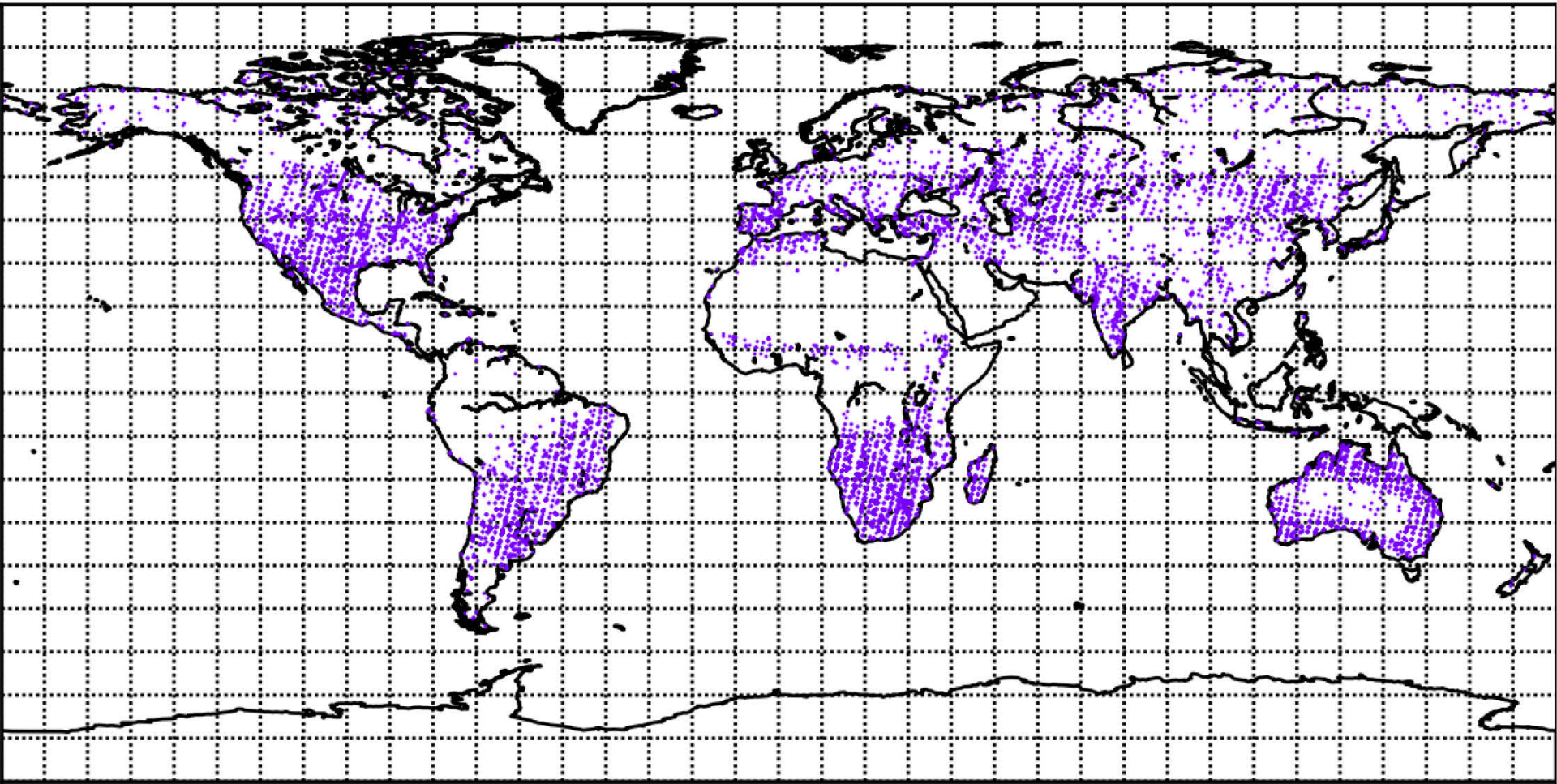
Global Spatial Analysis

Warn Level 1/19



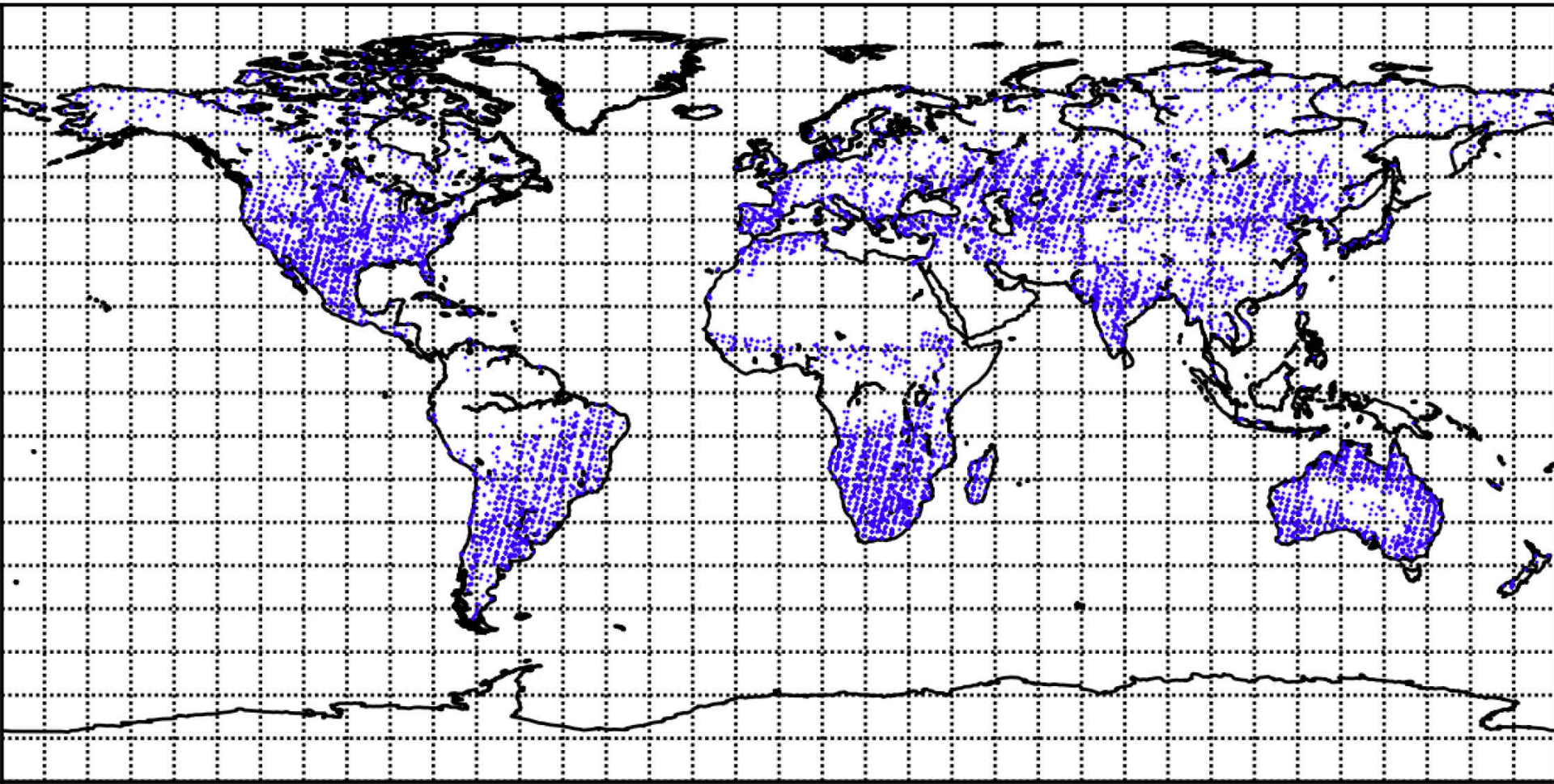
Global Spatial Analysis

Warn Level 2/19



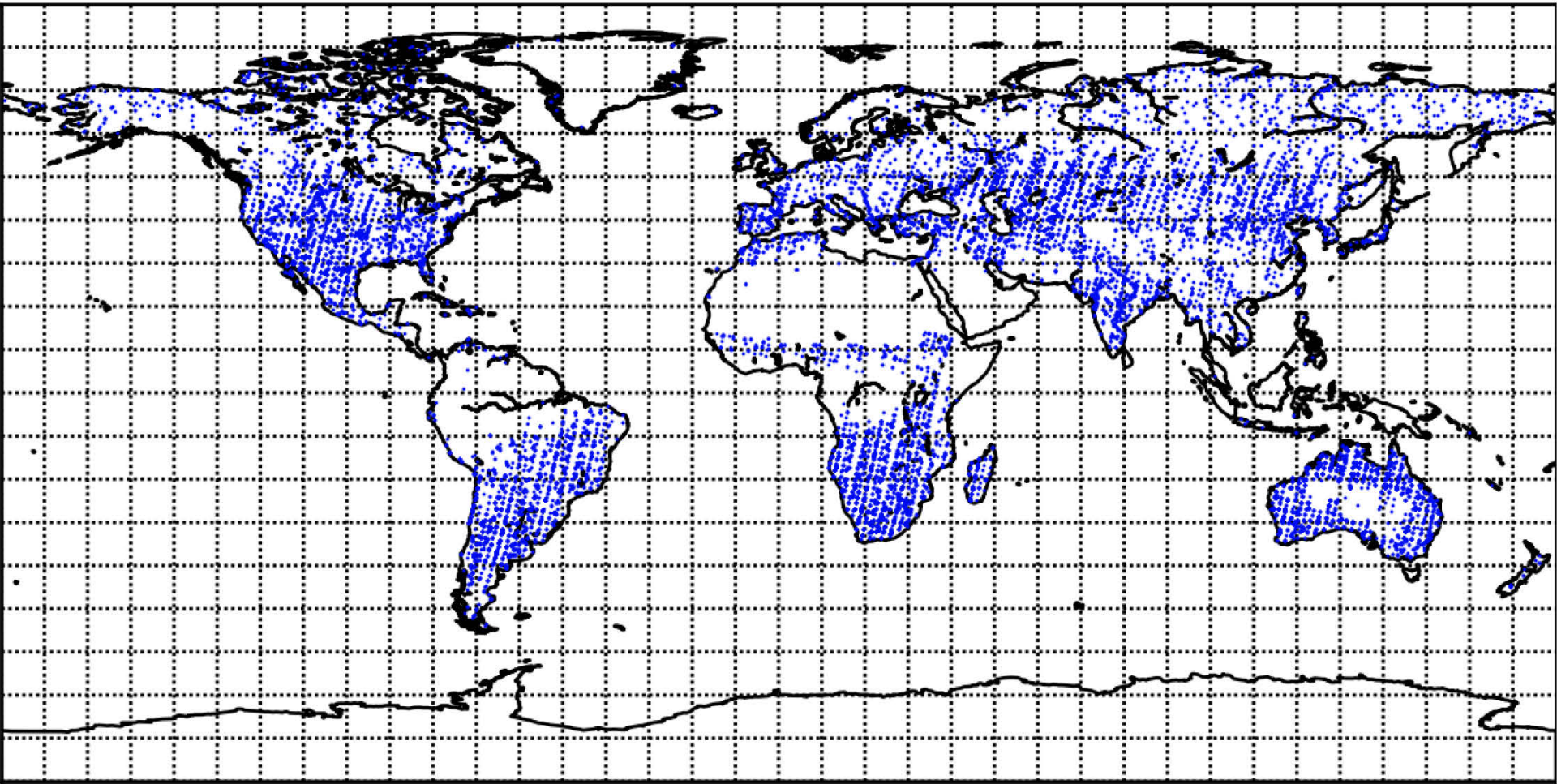
Global Spatial Analysis

Warn Level 3/19



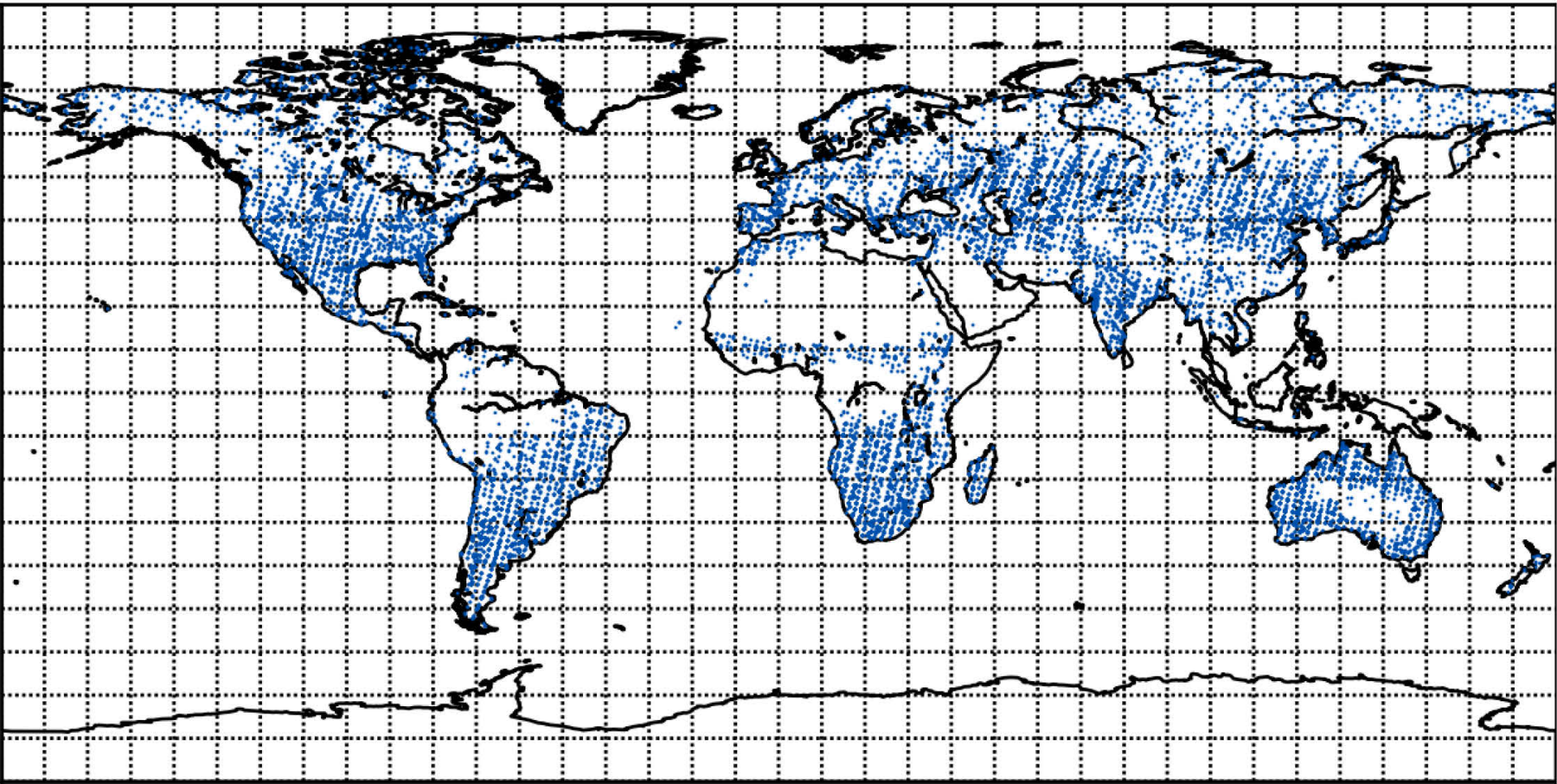
Global Spatial Analysis

Warn Level 4/19



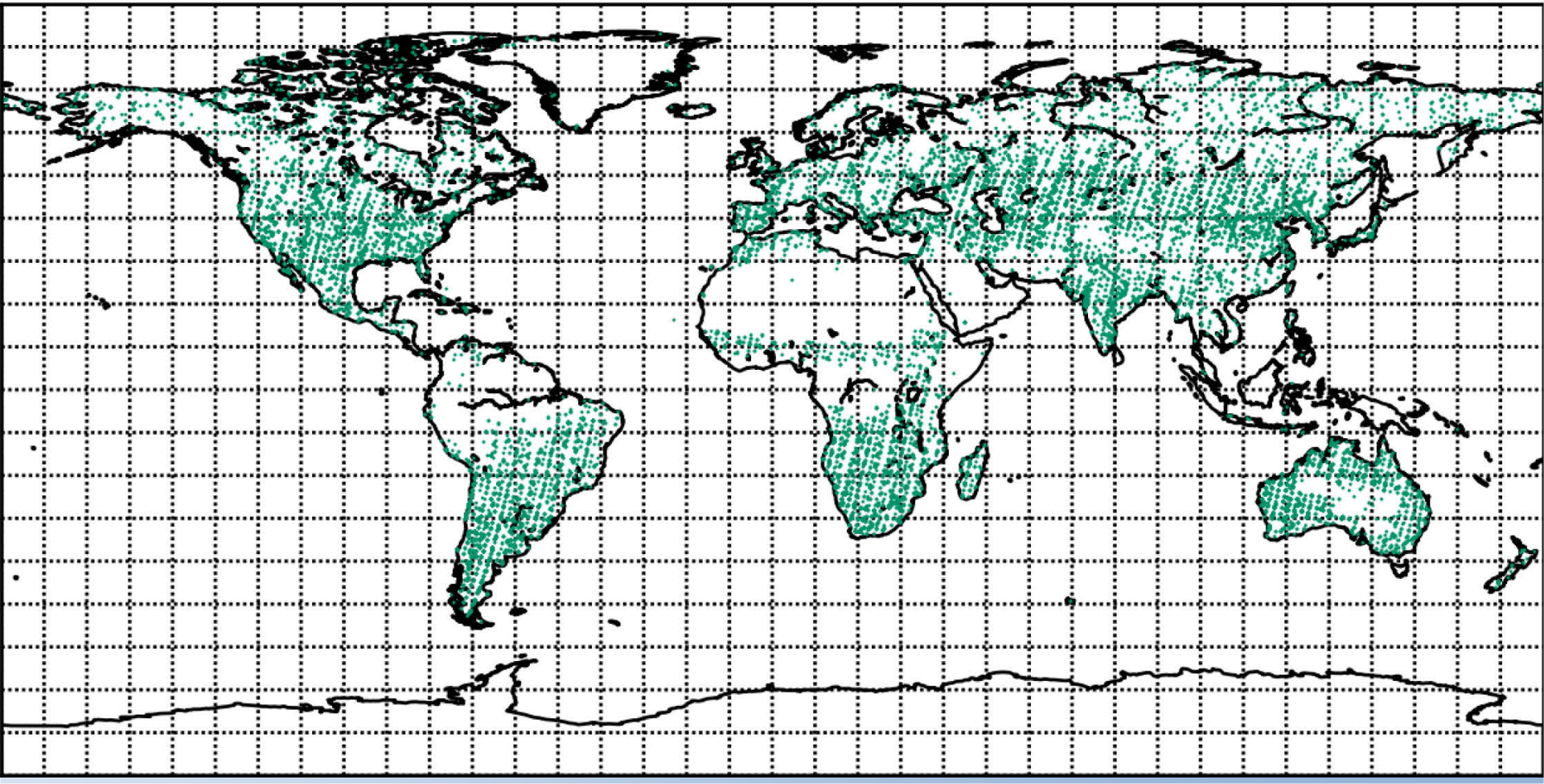
Global Spatial Analysis

Warn Level 5/19



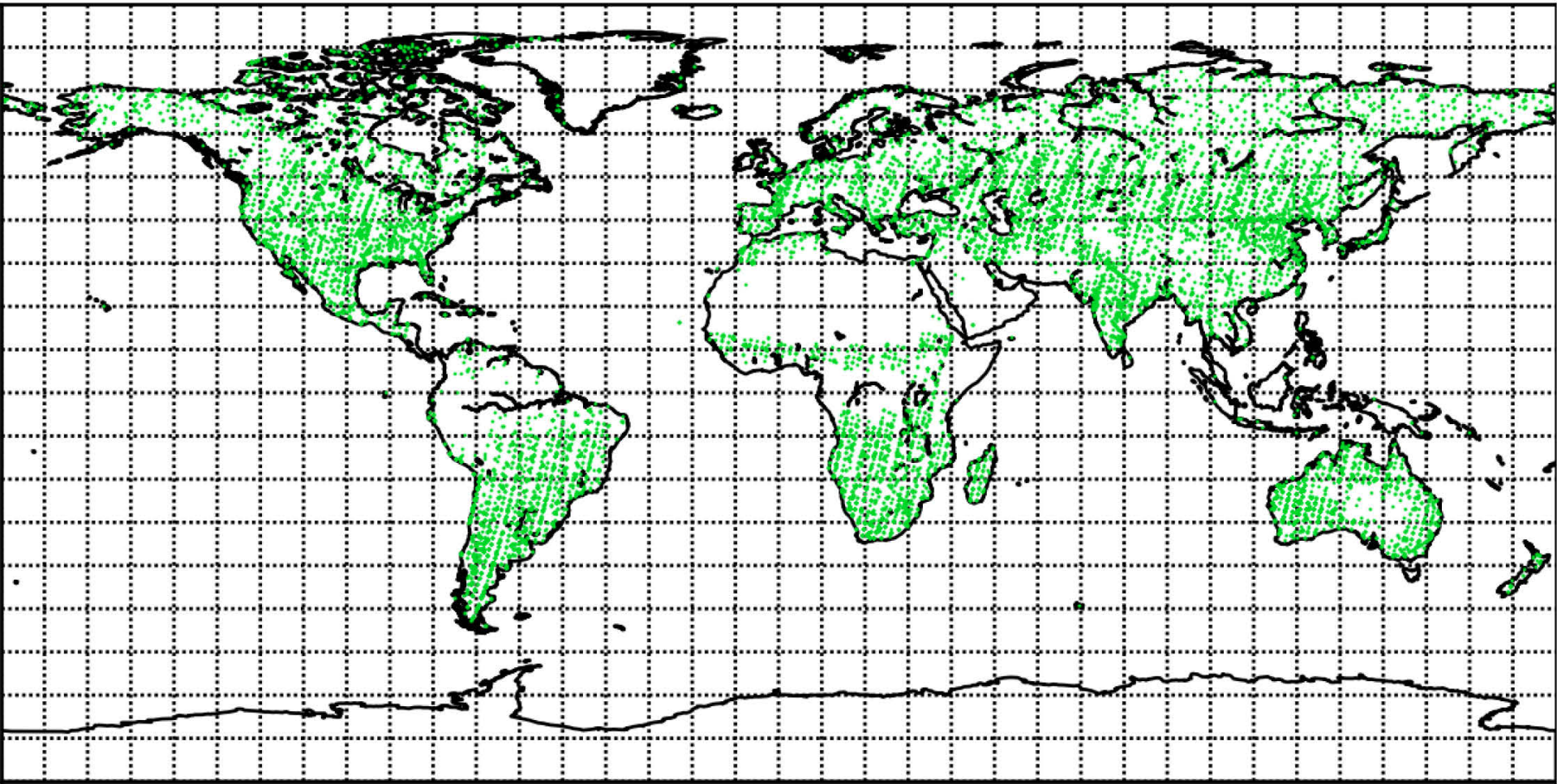
Global Spatial Analysis

Warn Level 6/19



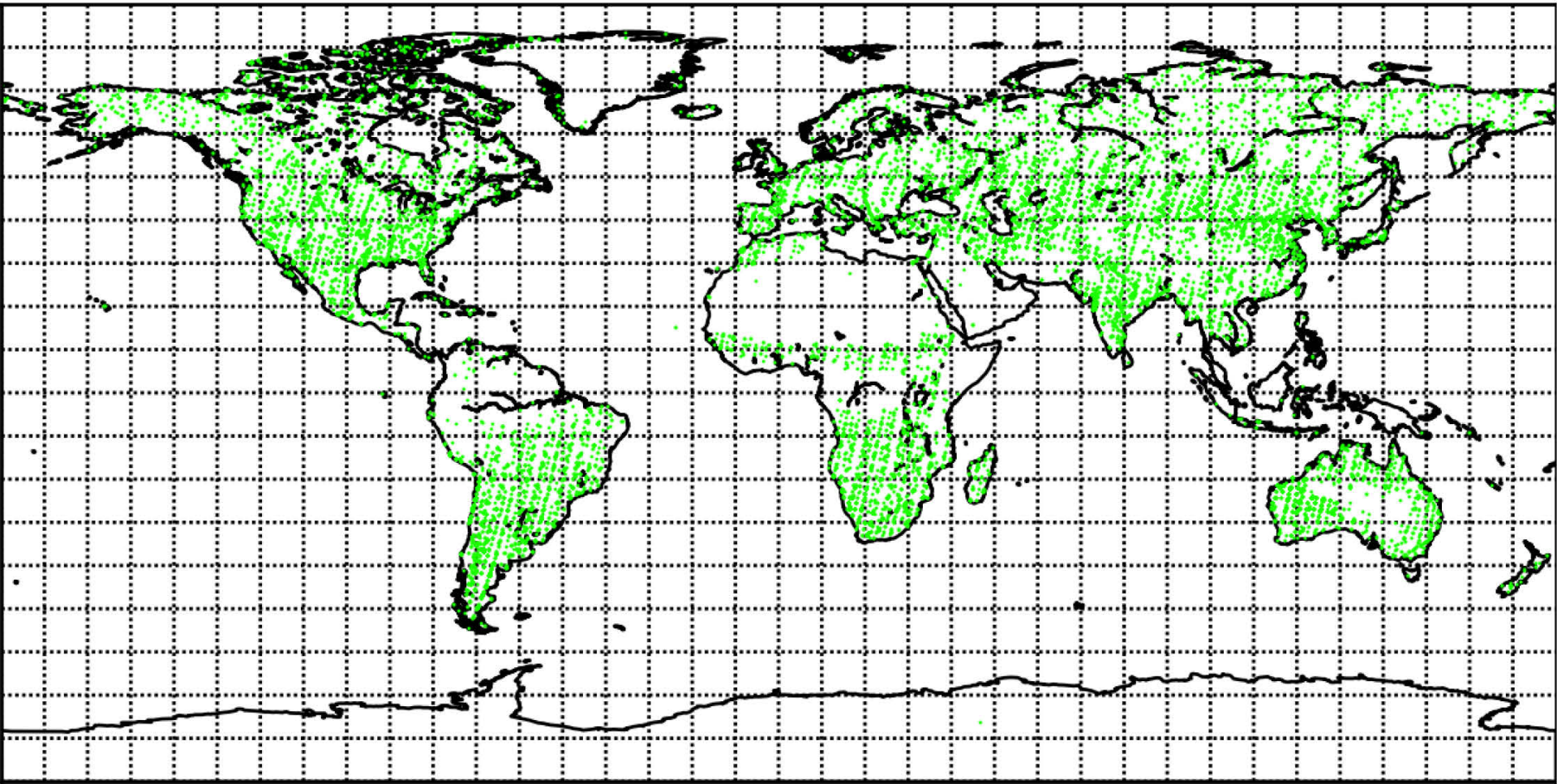
Global Spatial Analysis

Warn Level 7/19



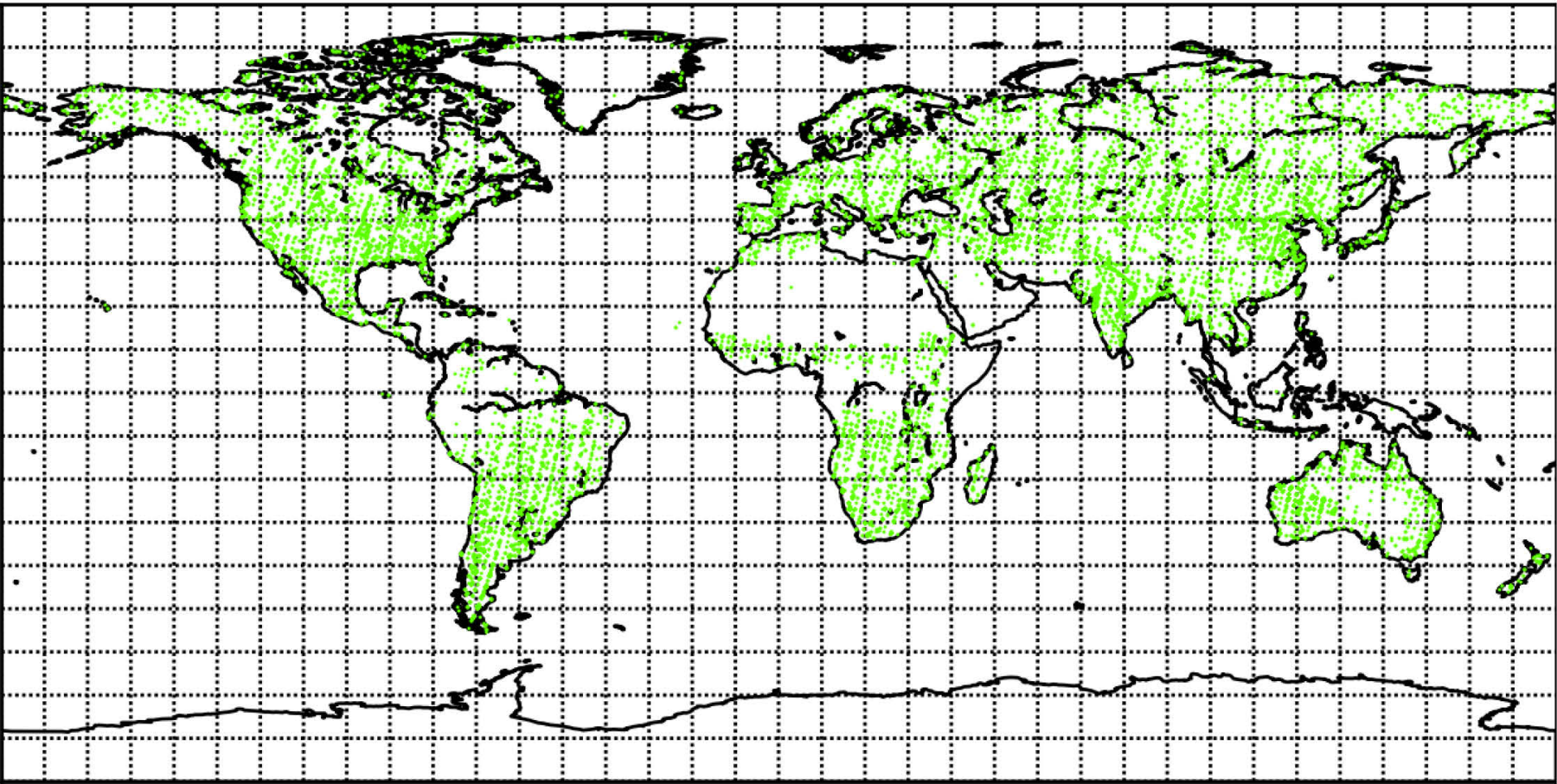
Global Spatial Analysis

Warn Level 8/19



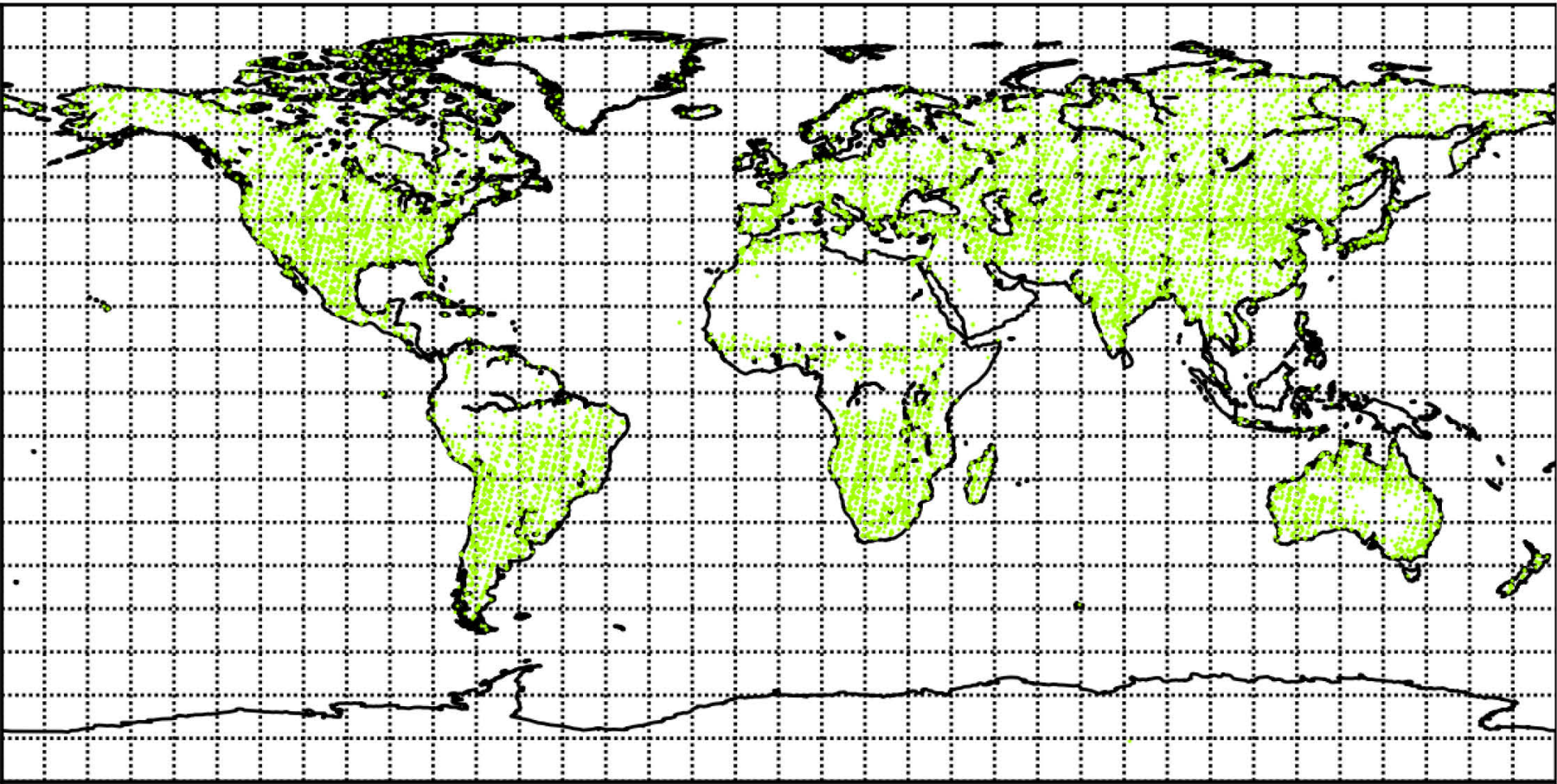
Global Spatial Analysis

Warn Level 9/19



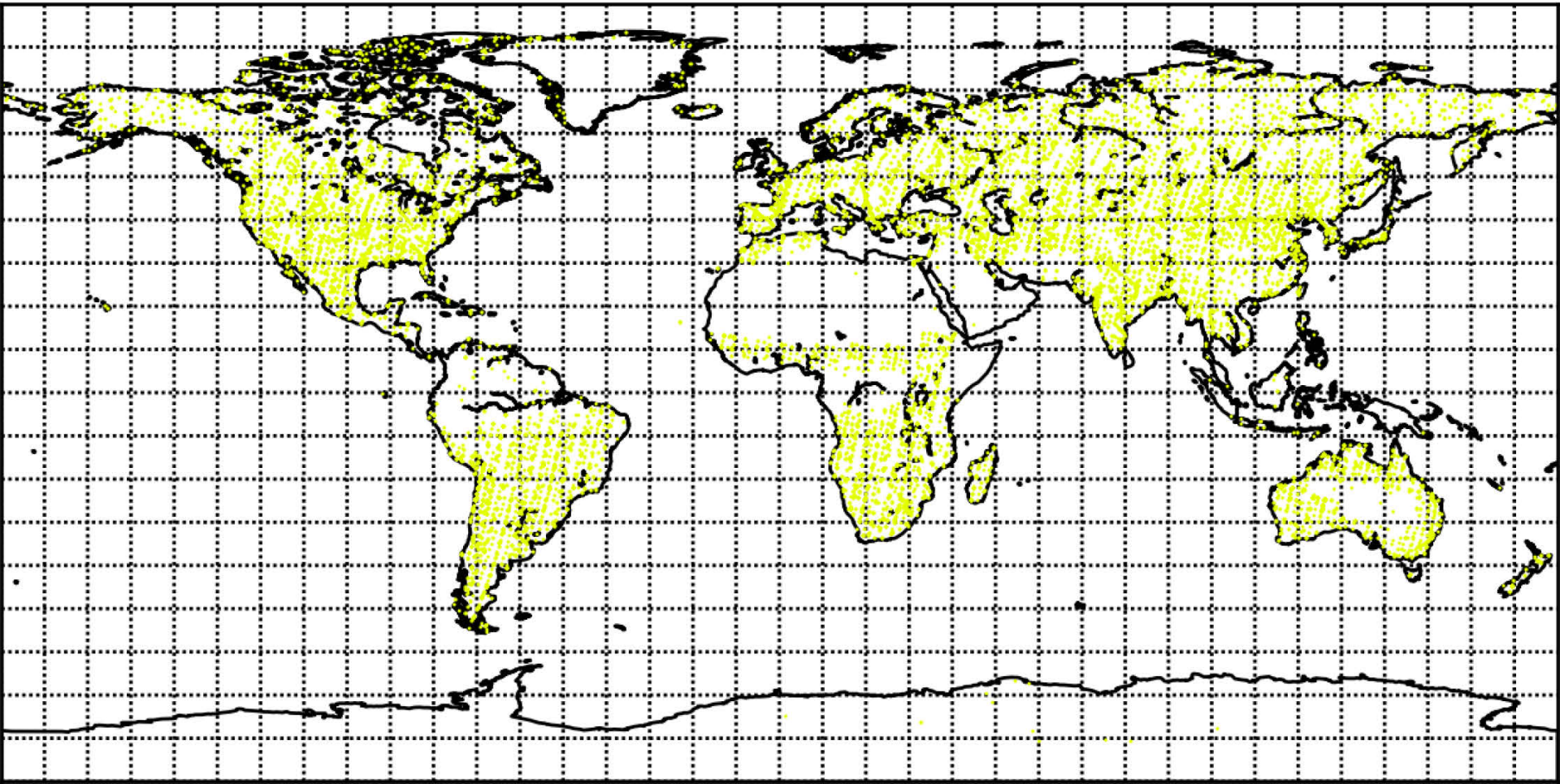
Global Spatial Analysis

Warn Level 10/19



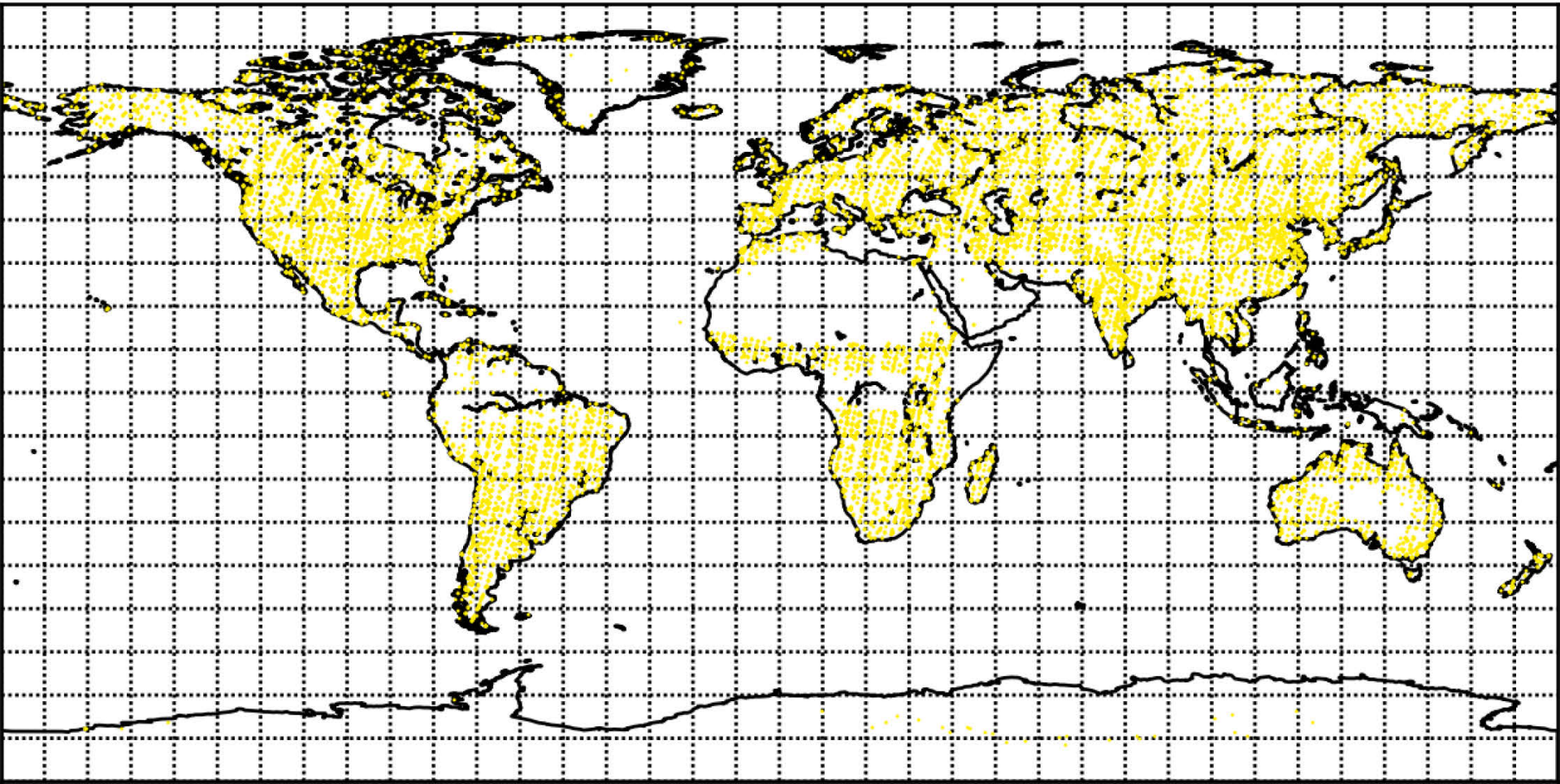
Global Spatial Analysis

Warn Level 11/19



Global Spatial Analysis

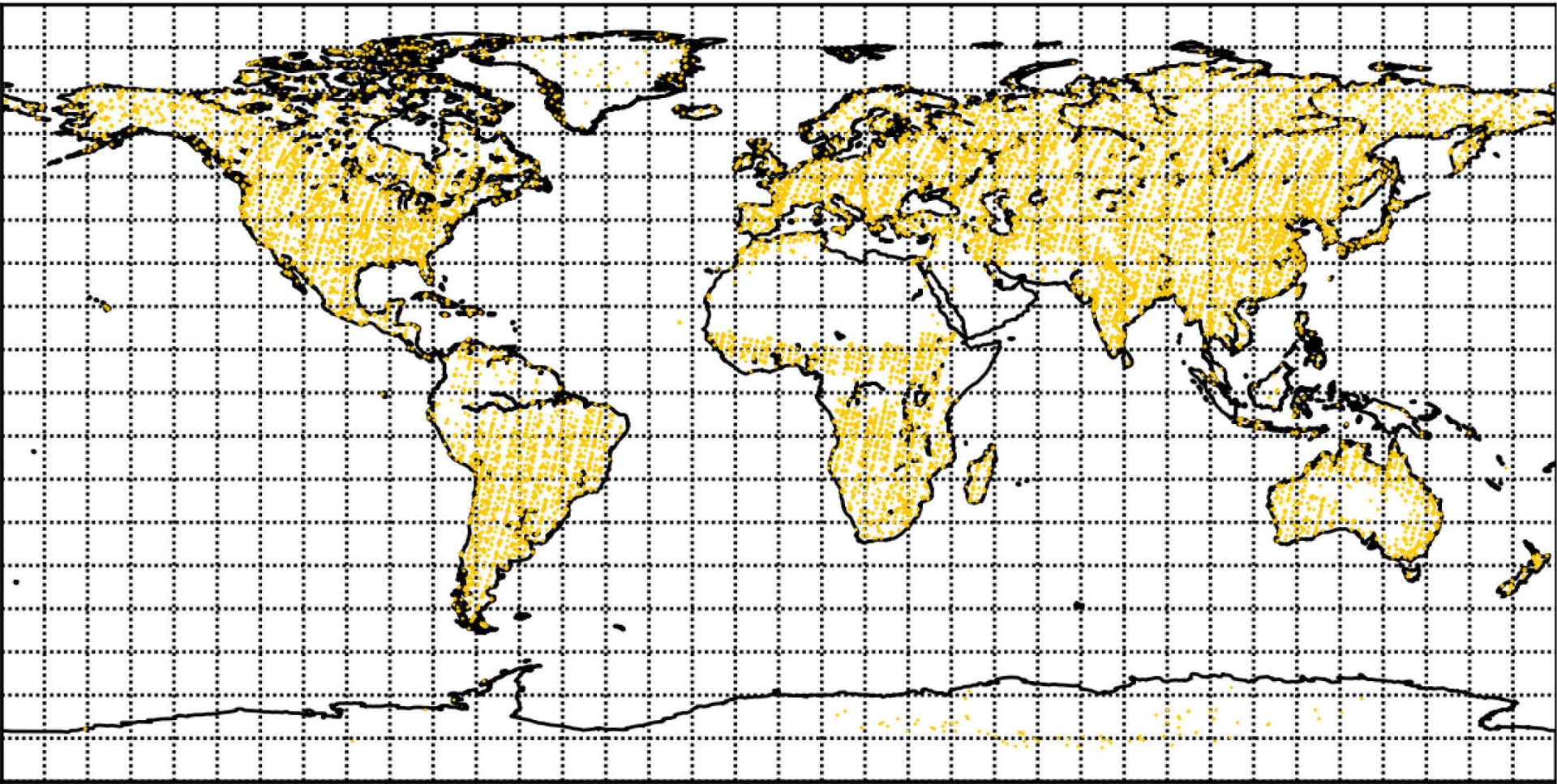
Warn Level 12/19



Global Spatial Analysis

Icy regions

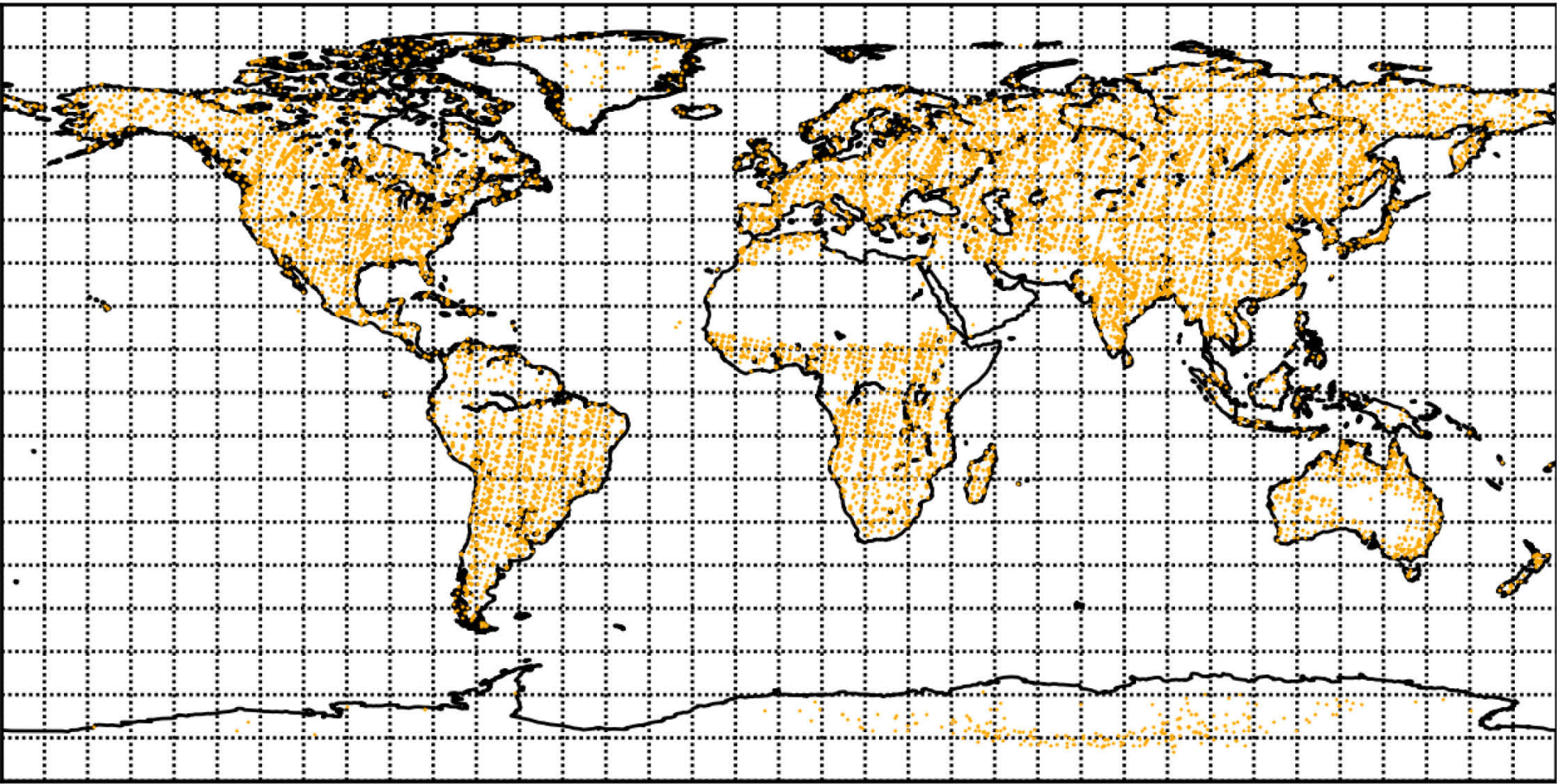
Warn Level 13/19



Global Spatial Analysis

Icy regions

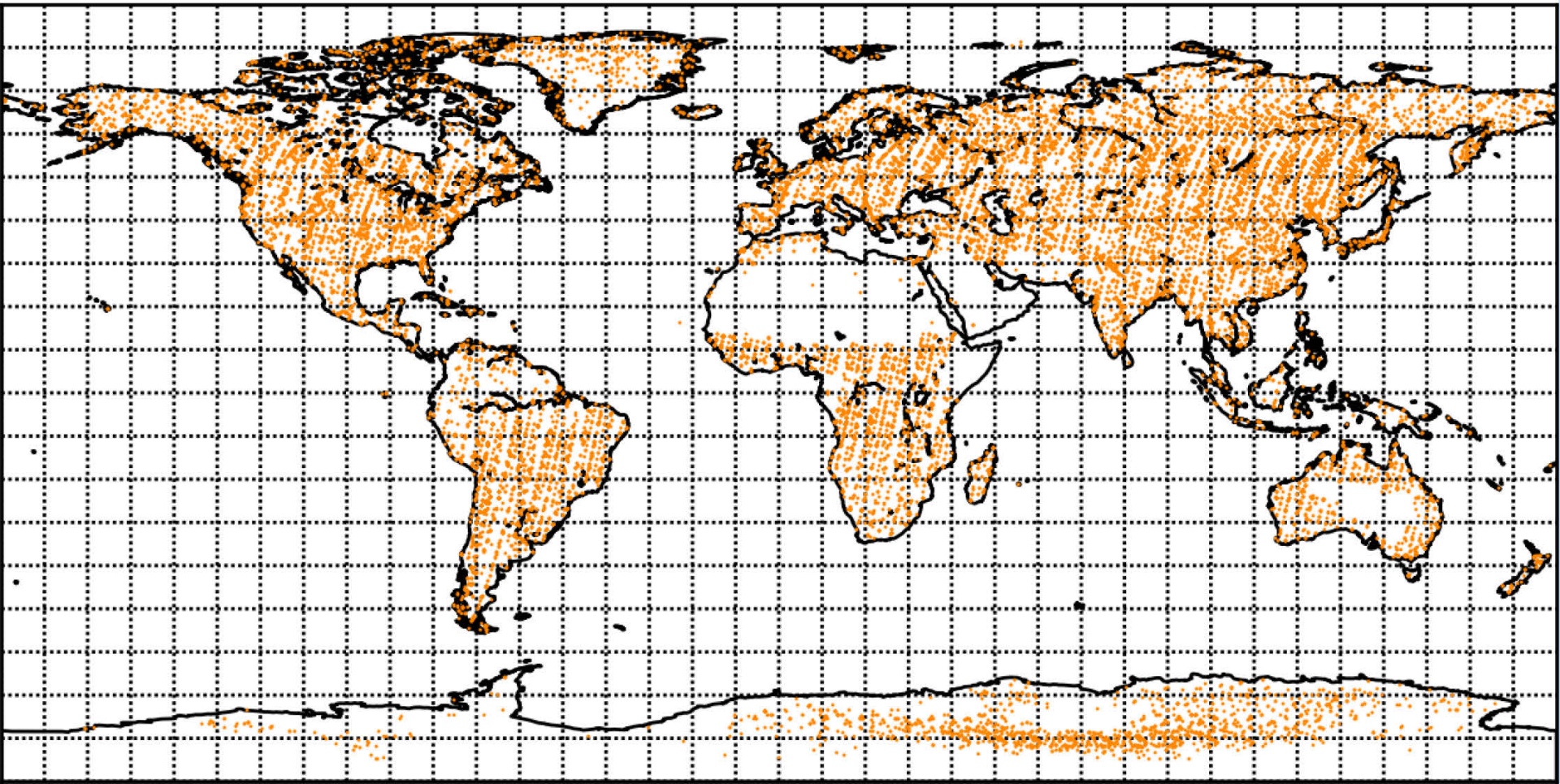
Warn Level 14/19



Global Spatial Analysis

Icy regions

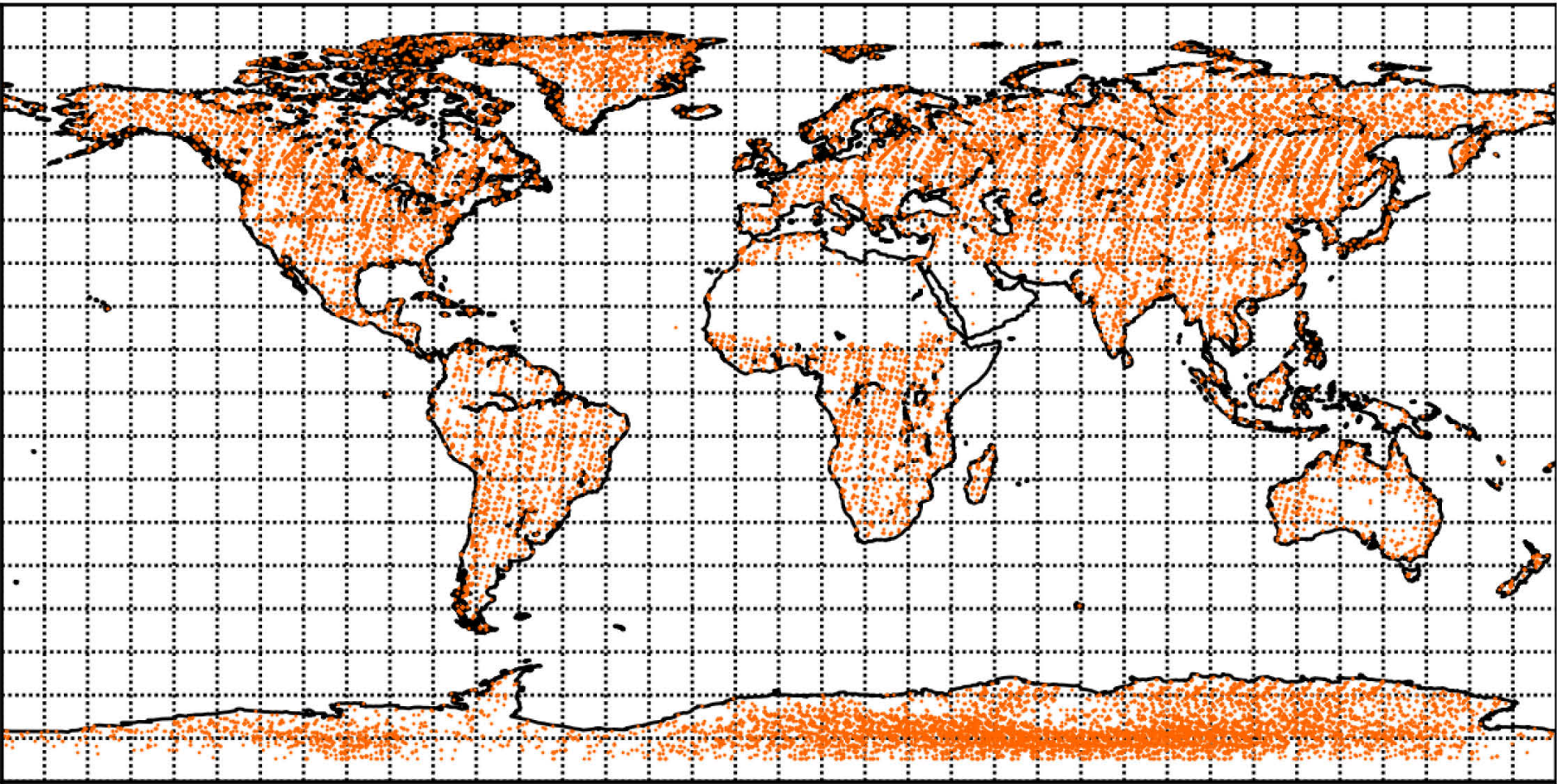
Warn Level 15/19



Global Spatial Analysis

Icy regions

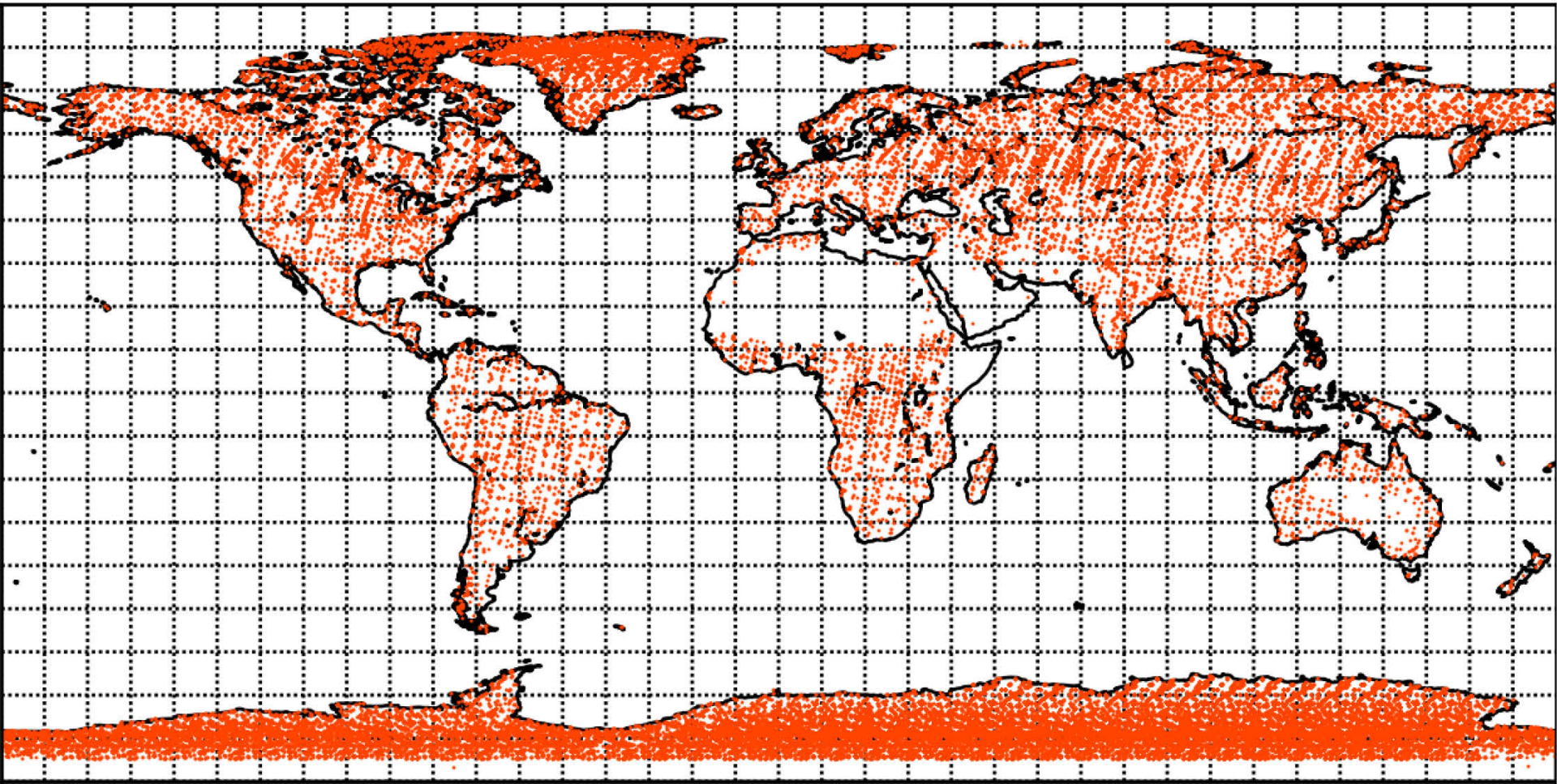
Warn Level 16/19



Global Spatial Analysis

Icy regions

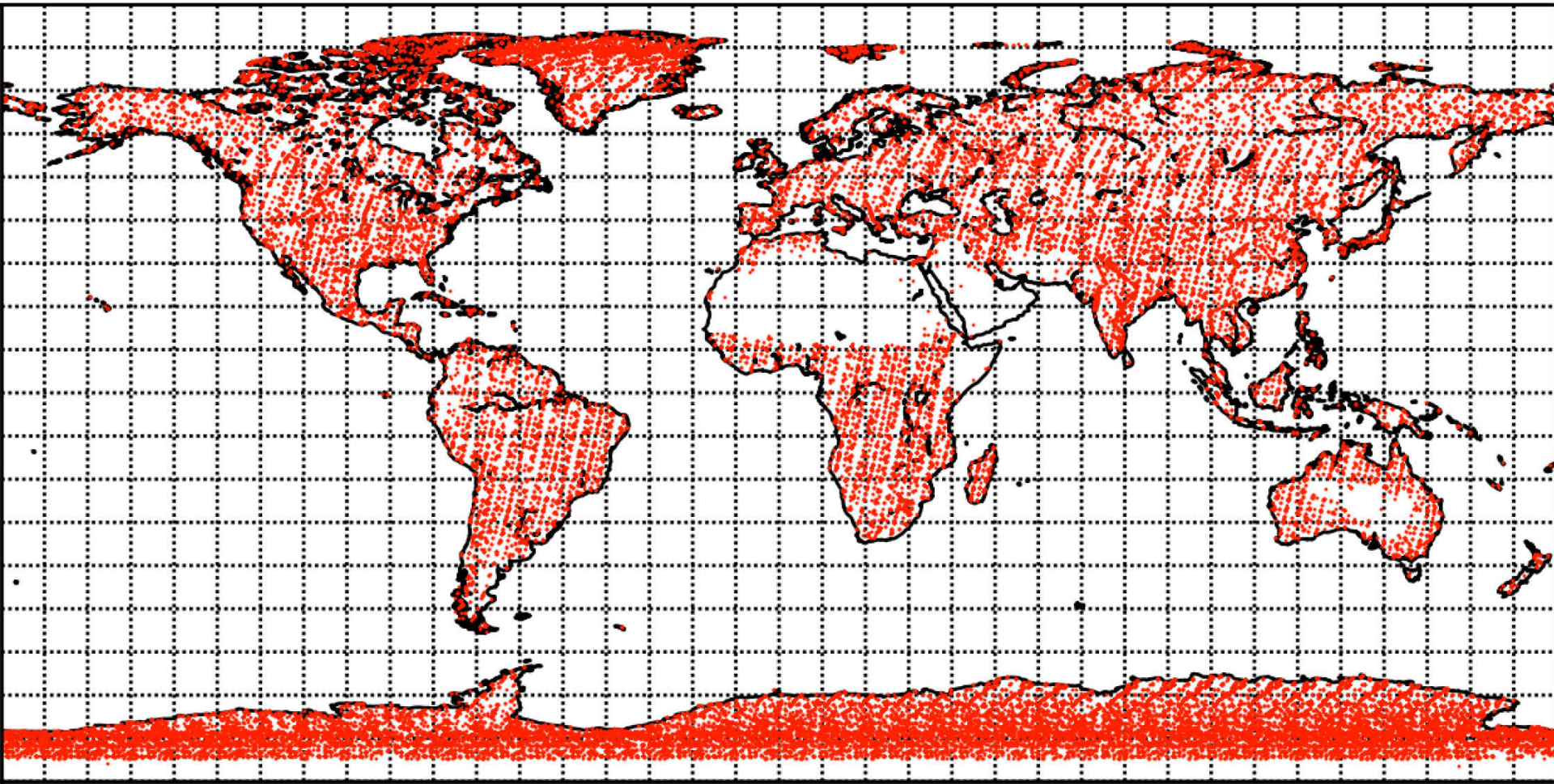
Warn Level 17/19



Global Spatial Analysis

Icy regions

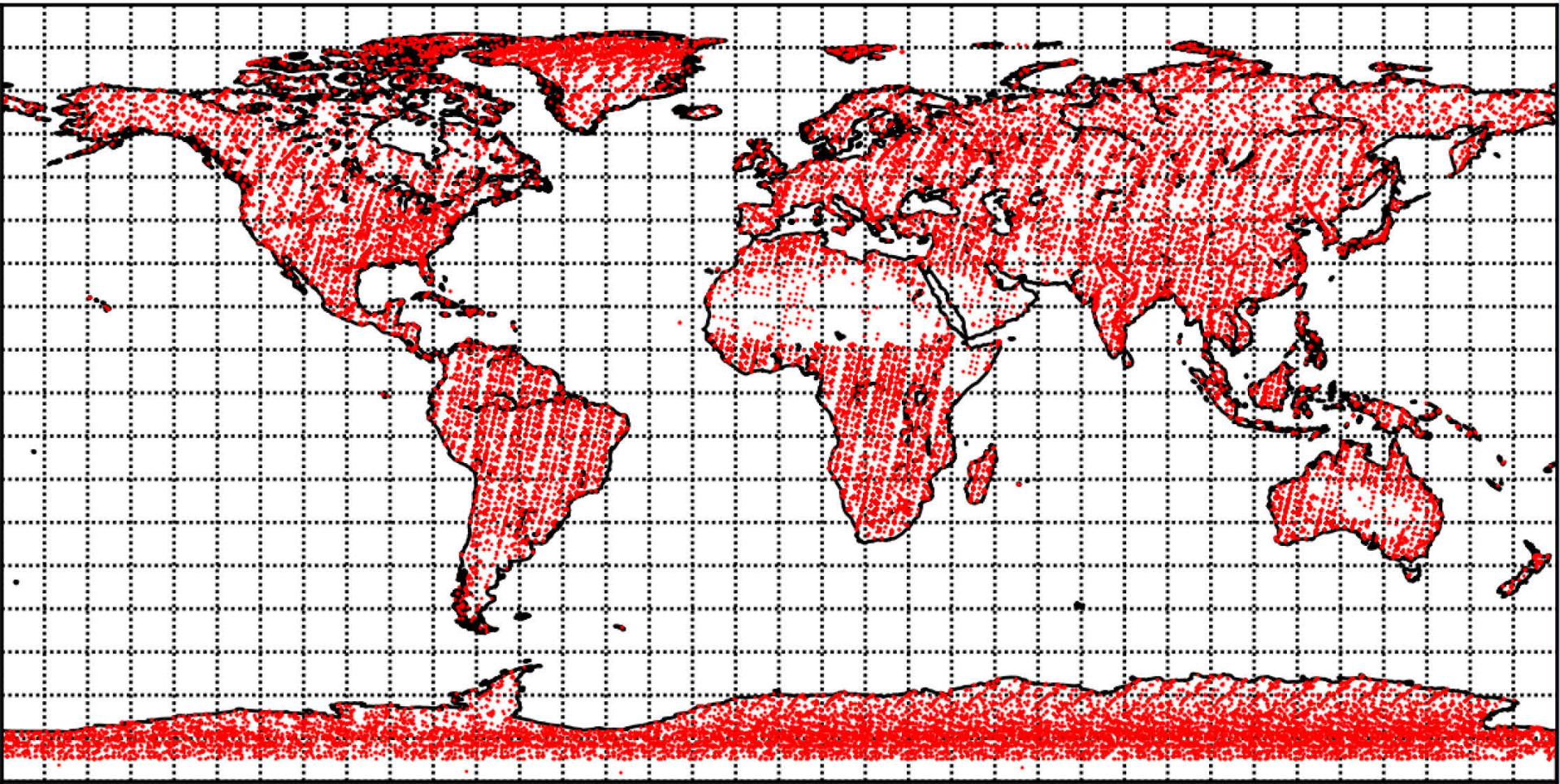
Warn Level 18/19



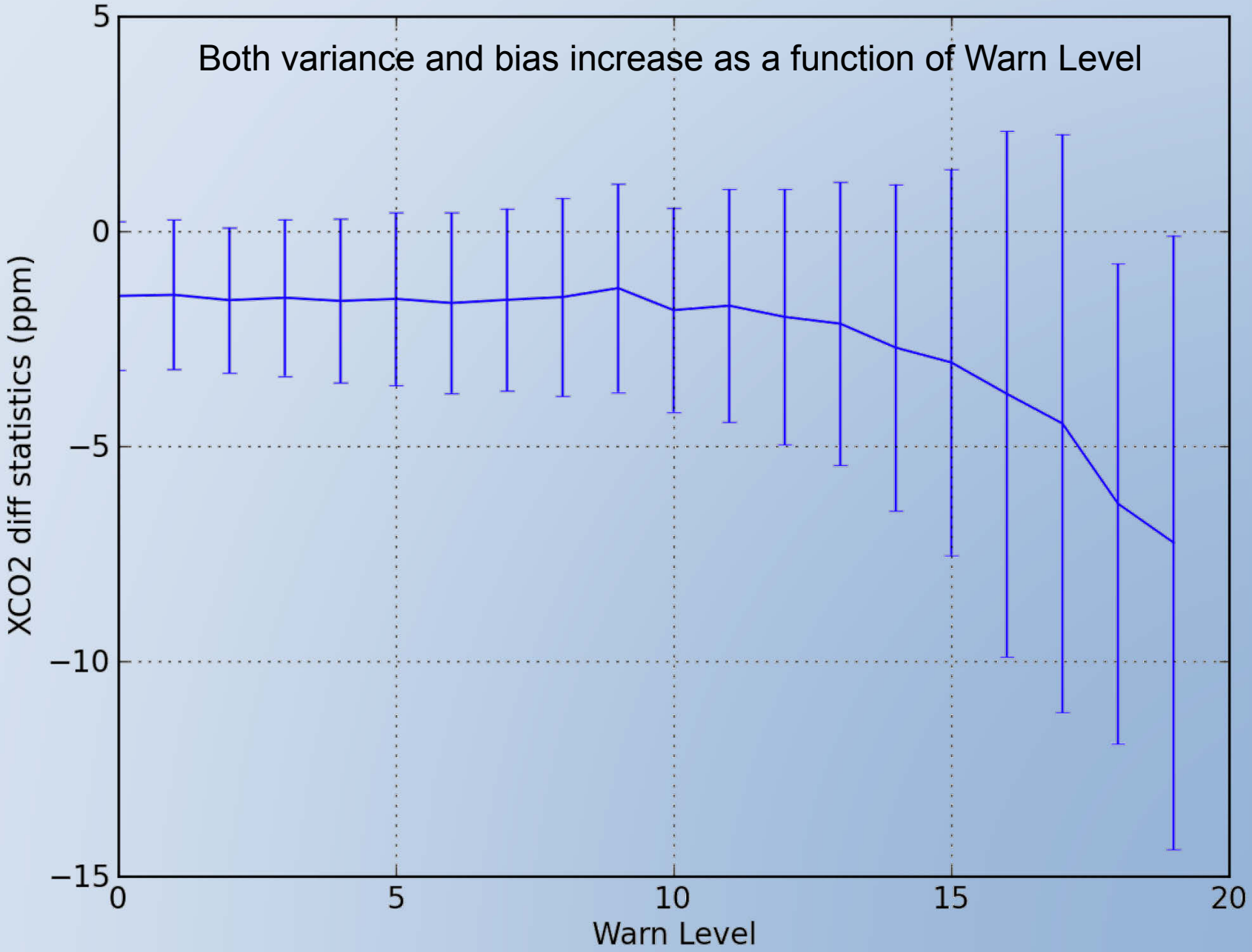
Global Spatial Analysis

Icy regions

Warn Level 19/19



ACOS-TCCON (Warn Level)



Solution Found

Have a Retrieval Quality Estimation (RQE) that:

-Performs similarly to Chris O'dell's hand-crafted expert system

-Permits dialable transparency for less/more data than above

-Identifies several key parameters that correlate to quality of retrieval

-Creates new product that sorts soundings by likely utility

-Does not favor particular geographic regions or timespans

-Incorporates both TCCON and SHA as truth metrics

-Completed for Land H-gain and Glint, but not M-gain (too little data)