# Warn Levels: Ordering Data for Custom Filtration

Dr. Lukas Mandrake, Dr. Gary Doran – Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109



# What You Need To Know

## What's a Warn Level?

A number for each delivered OCO-2 sounding
0 to 19

Estimates sounding contamination
(lower is better)

WL = 0-5 is best quality data

WL 6-15 can be useful but has rising issues

WL 16-19 is likely and increasingly useless

## Why can't I use a Quality Flag?

**You can, but they aren't made for you:**
Data provider guesses which data "good enough"

- Passes inflexible, small % of total data -

What if you need 30% instead of 20%?

What if you want even more filtration to 10%?

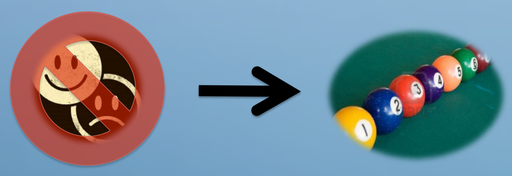
## How do Warn Levels help?

**Alternative to Quality Flag: Data Ordering**

Just provide users a data ordering

No good/bad decisions, no cutoff, no lost data

User decides how far into the ordering to use



## How to Use Warn Levels?

1 Decide requirements beforehand: how much data volume / coverage or scatter / error is needed or tolerable?

2 Begin admitting WL = 0, 1, 2, … into project. Monitor above statistics.

3 Stop when data volume / coverage are acceptable, or when scatter / error become intolerable.

## You're done!

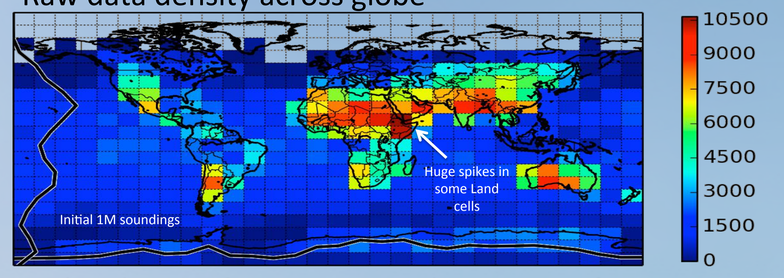
You now have a custom filter for your analysis

When collaborating:
"I used Warn Levels < N" to define filtration strategy

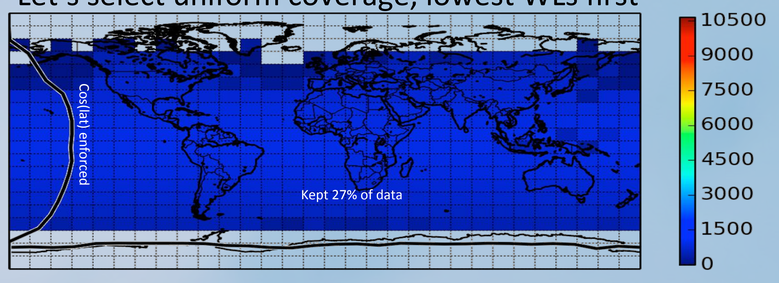
# Watching Them Work

## Selecting your dataset (OCO-2 v5 data)

Raw data density across globe

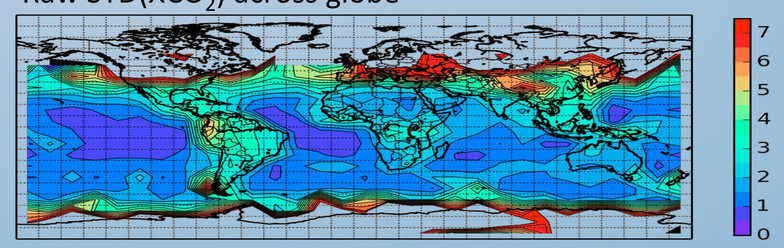



Let's select uniform coverage, lowest WLs first

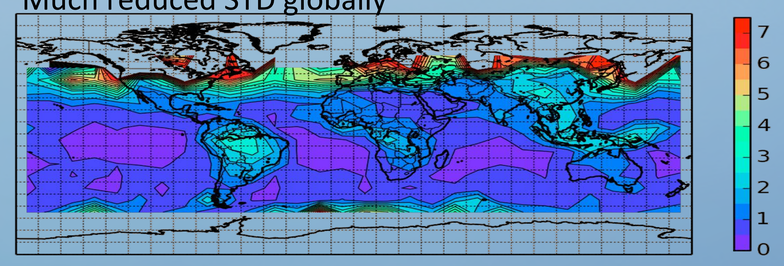



## What did using lowest WL first buy you?

Raw STD($XCO_2$) across globe

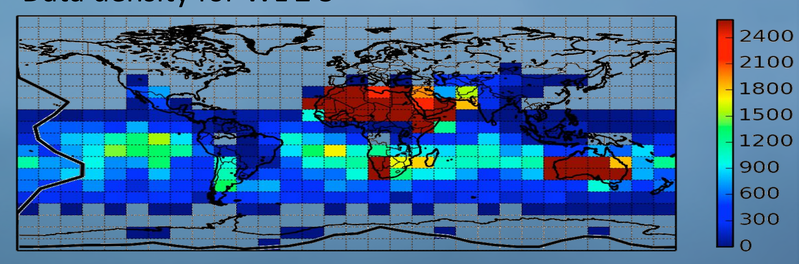



Much reduced STD globally




## Want high-quality data, don't care where?

Data density for WL ≤ 5




## Warn Level Summary:

- Show where the highest quality data resides
- Intelligently order data for bin fill / user selection
- Reduce error metrics (such as STDEV)
- Easily communicate to colleagues which data used

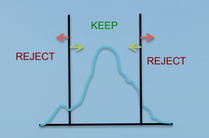
# How Are They Made?

DOGO — Data Ordering Genetic Optimization

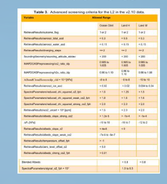
★ Winner JPL Software of the Year

## What do human experts do?

Define threshold rules




A lot of rules



Final Product: Quality Flag


Much expert time & effort yields single filter based on many rules. Binary output: yes or no for each sounding.

## DOGO -1- The Metrics

**A DOGO Metric measures something unpleasant about the data**
What DOGO minimizes when ordering the data
Explored three:

- Minimize STDEV(XCO2) in the southern hemisphere
- Minimize STDEV(XCO2) in individual Small Areas (see Wennberg)
- Minimize RMS difference with TCCON truth (where available)

Final Warn Levels merge the knowledge from all 3 metrics

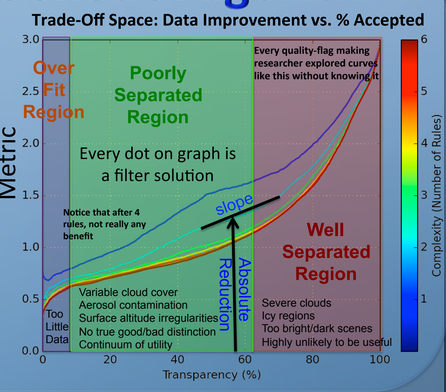
## DOGO -2- The Genetic Algorithm

- Creates x10,000's of optimal filters
- Like 1000's of experts cooperating

- Each filter accepts X% data = transparency
- Each filter reduces metric

- Graph to produce Trade-off Curve
- Fundamental shape of data quality
- Each dot on graph is a valid filter

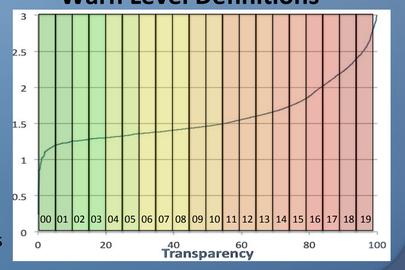
**Trade-Off Space: Data Improvement vs. % Accepted**




## DOGO -3- Defining Warn Levels

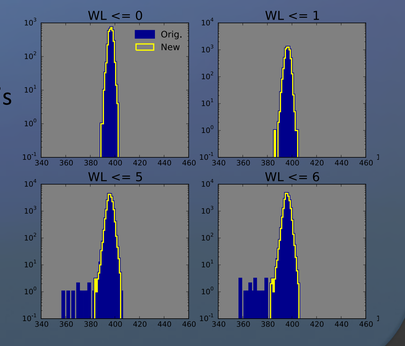
- Use trade-off curves to create WL's
- WL's chop this graph every 5% transparency
- Must combine curves from all Metrics
- Discovers the features that predict problems
- Discovers how many there are (complexity)
- Discovers feature thresholds that define WL's

**Warn Level Definitions**




## DOGO -4- Removing Outliers

- WL's are defined by improving "bulk statistics"
- Lone outliers that don't influence bulk metric may linger
- Identify outliers using model
- Re-categorize outliers to higher WL's
- Produces cleaner WL histograms
- Removes "surprise" outliers, especially at low WL's




New in OCO-2 v7.1 data!