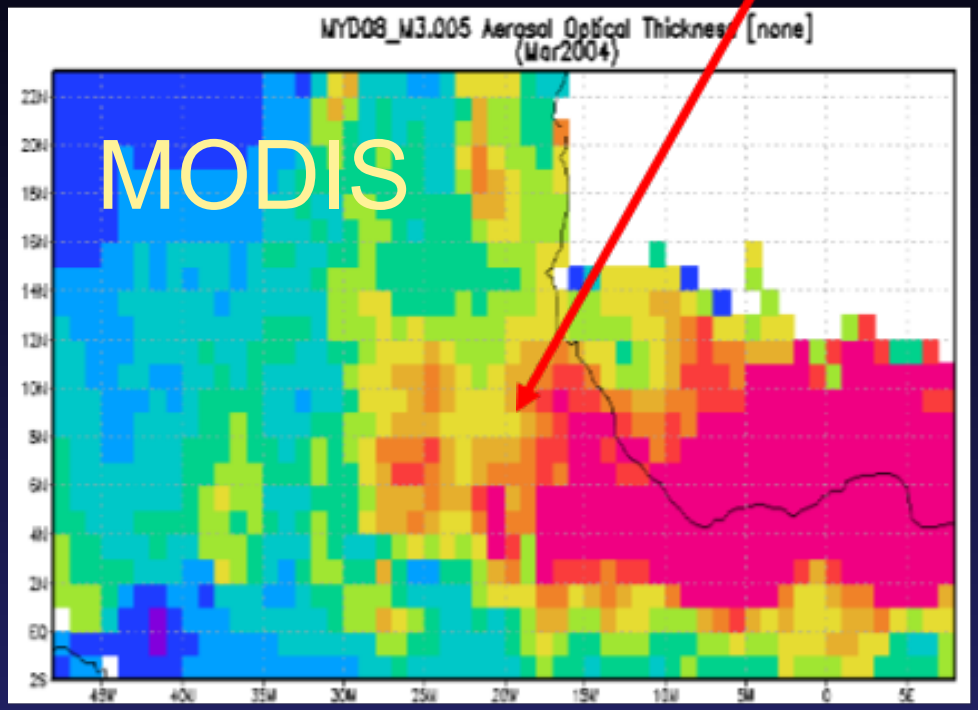
Data Lineage: Why it is important, plus stories from the ESTO MDSA project

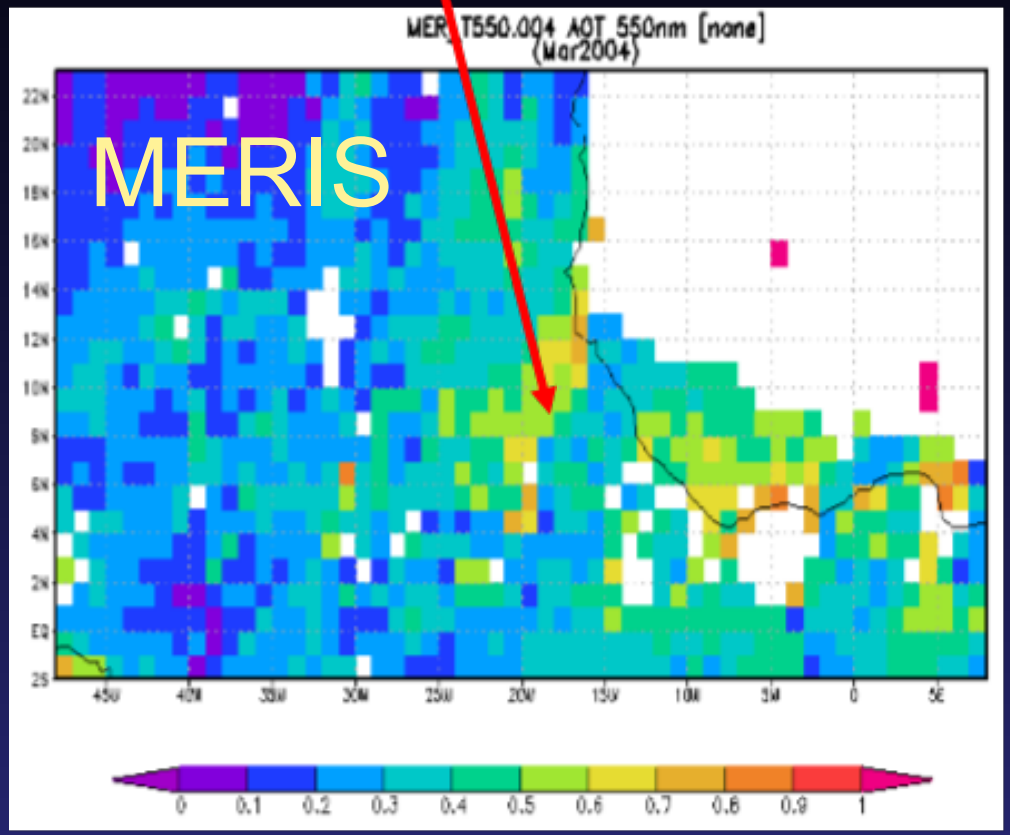
Gregory Leptoukh, Christopher Lynnes
NASA GSFC
Peter Fox
Rensselaer Polytechnic Institute

Case 1: MODIS vs. MERIS

Same parameter

Same space & time





Different results – why?

Provenance aspect: A threshold used in MERIS processing effectively excludes high aerosol values.

Outline

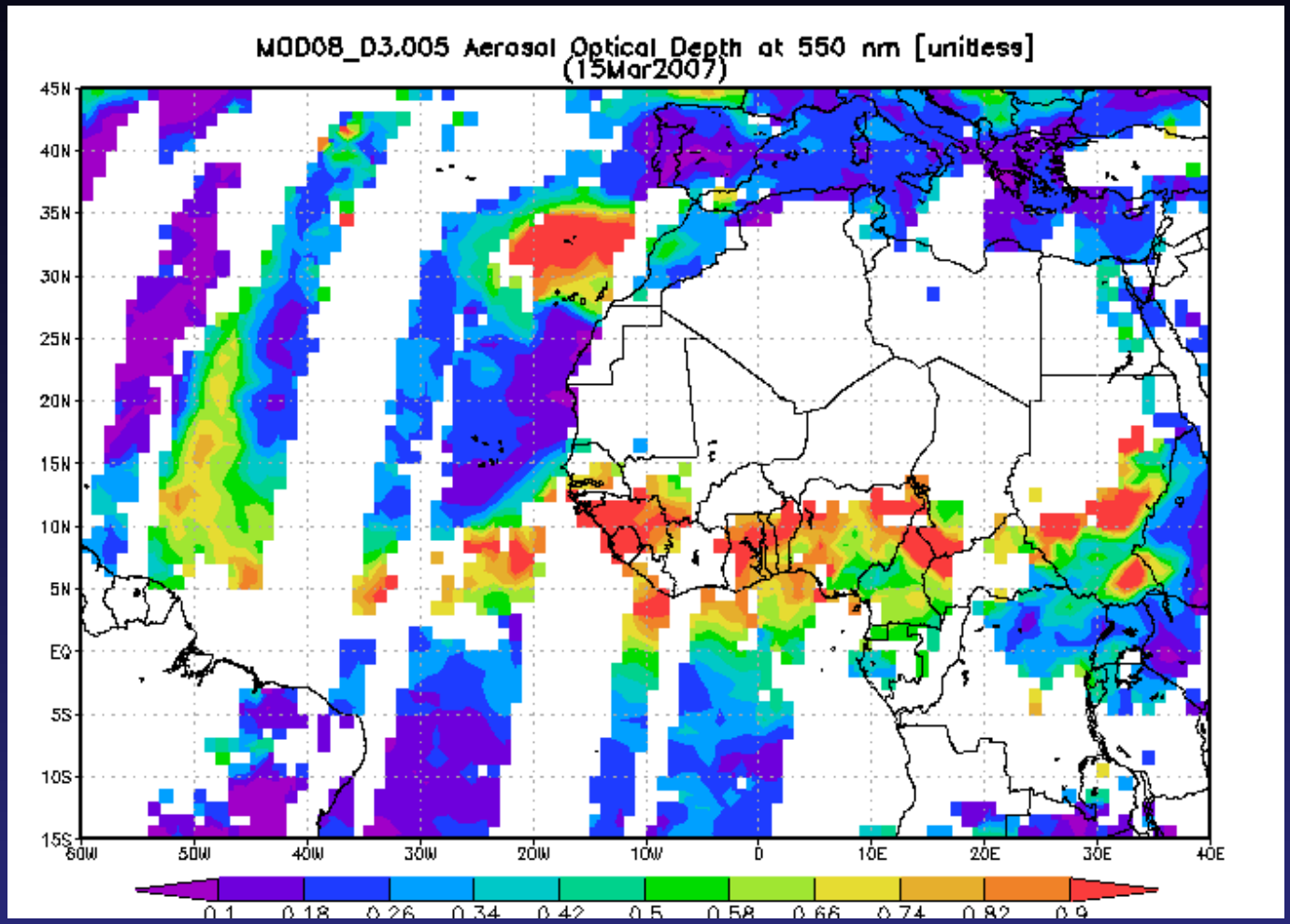
- Case 1: MODIS vs. MERIS
- Decadal Survey missions: data from multiple sensors
- Why Provenance is needed:
  - knowledge for using the data
- Case 2: Temporal aggregation
- Collecting and delivering provenance
- Harmonizing Multi-sensor provenances:
  - Joint provenance ≠ prov1 + prov2
- Case 3: Orbital characteristics and Dataday
- Conclusions

Decadal Survey missions: data from multiple sensors

Data from multiple sources need to be used together:

- ACE
- NPP and NPOESS
- Geo-Cape
- European and other countries' satellites
- Models

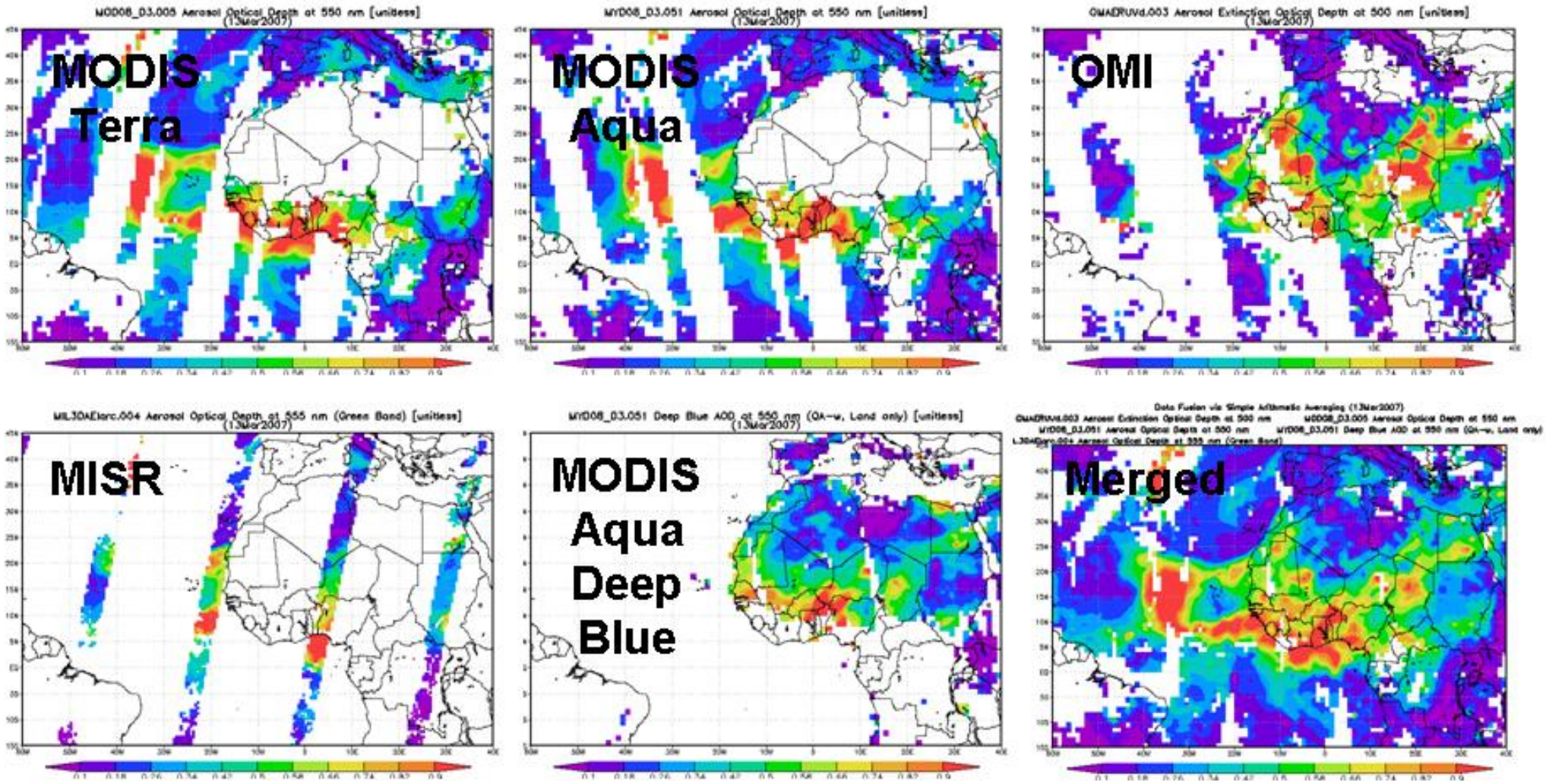
MODIS Terra Aerosol Animation



Data from a single sensor don't provide sufficient spatial coverage

Merged multi-sensor aerosol data

March 13, 2007



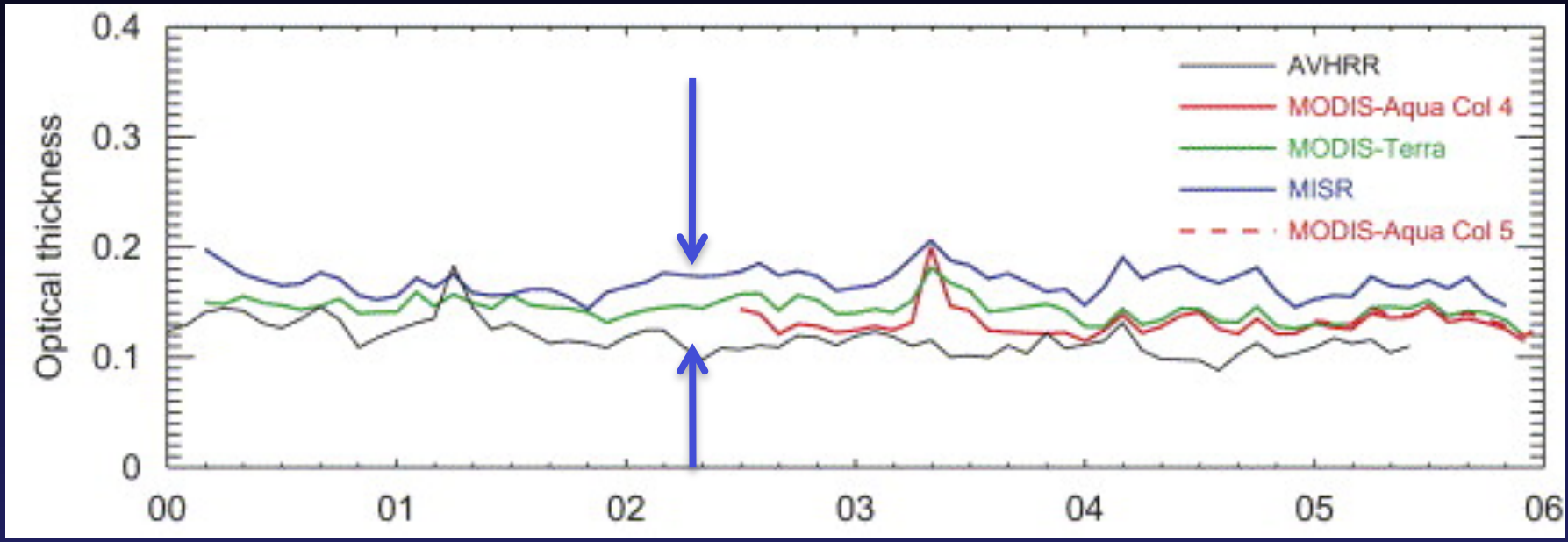
Merged AOD data from 5 retrieval algorithms (4 sensors: MODIS-Terra, MODIS-Aqua, MISR, and OMI) provide almost complete coverage.

What is Data Lineage and why it's needed?

Data Lineage or Provenance is *the source of data,* including *the execution history of the processes that produced them*

- Data by themselves without provenance are not sufficient to make accurate scientific conclusions
- Without this provenance, data users will not trust the data and/or may use data incorrectly
- Documenting steps leading to the final product is paramount

Case 2: Temporal aggregation

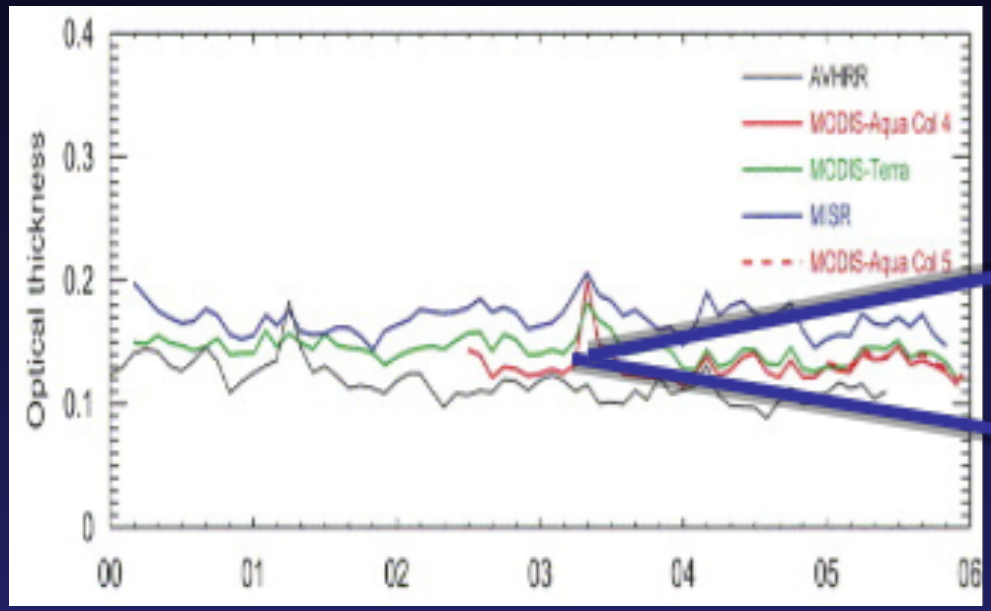


Time series of the global mean values of the AOD over the oceans from Mishchenko et al., 2007

Differences in Aerosol Optical Depth (AOD) between various sensors seemingly exceed reported accuracies of each sensor

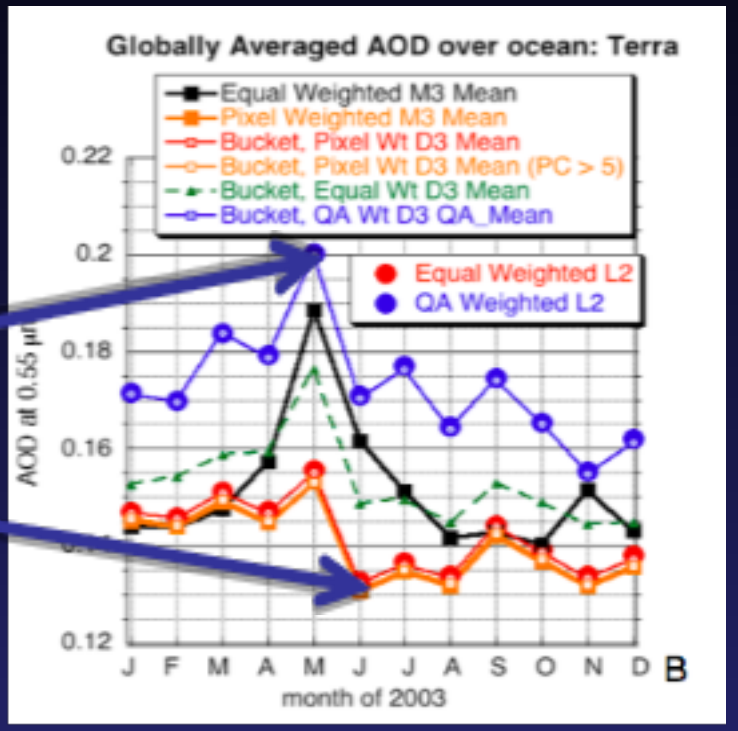
How sensitive are MODIS aerosols to different time aggregations?

AOD difference between sensors



Mishchenko et al., 2007

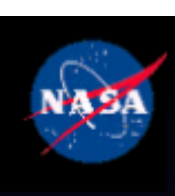
MODIS Terra only AOD: difference between diff. aggregations



Levy, Leptoukh, et al., 2009

Result: Very sensitive. The AOD difference can be up to 40%.

Provenance aspect: Must record even apparently minor differences in aggregation



Collecting and Delivering Data Provenance

Where to find the knowledge about data?

- It is scattered in scientific papers, the actual code, unwritten assumptions, folklore, etc.
- Assess sensitivity of the results to variations in processing algorithms/steps…
- Work closely with scientists to guarantee science quality

How to deliver provenance?

- Deliver to users together with the data
- Present to users in a convenient, easy-to-read fashion
- Provide recommendations for different data usage (applications vs. climate studies)

Data from multiple sensors: harmonization

- It is not sufficient just to have the data and their provenance from different sensors in one place
- Before data can be compared and fused, many items need to be harmonized:
  - Data: format, grid, spatial and temporal resolution
  - Metadata: standard fields, units, scales, quality?
  - Provenance: what to do with it?

Product A	Product B
Good	3
Bad	2
Ugly	1
	0

Are these quality flags compatible?

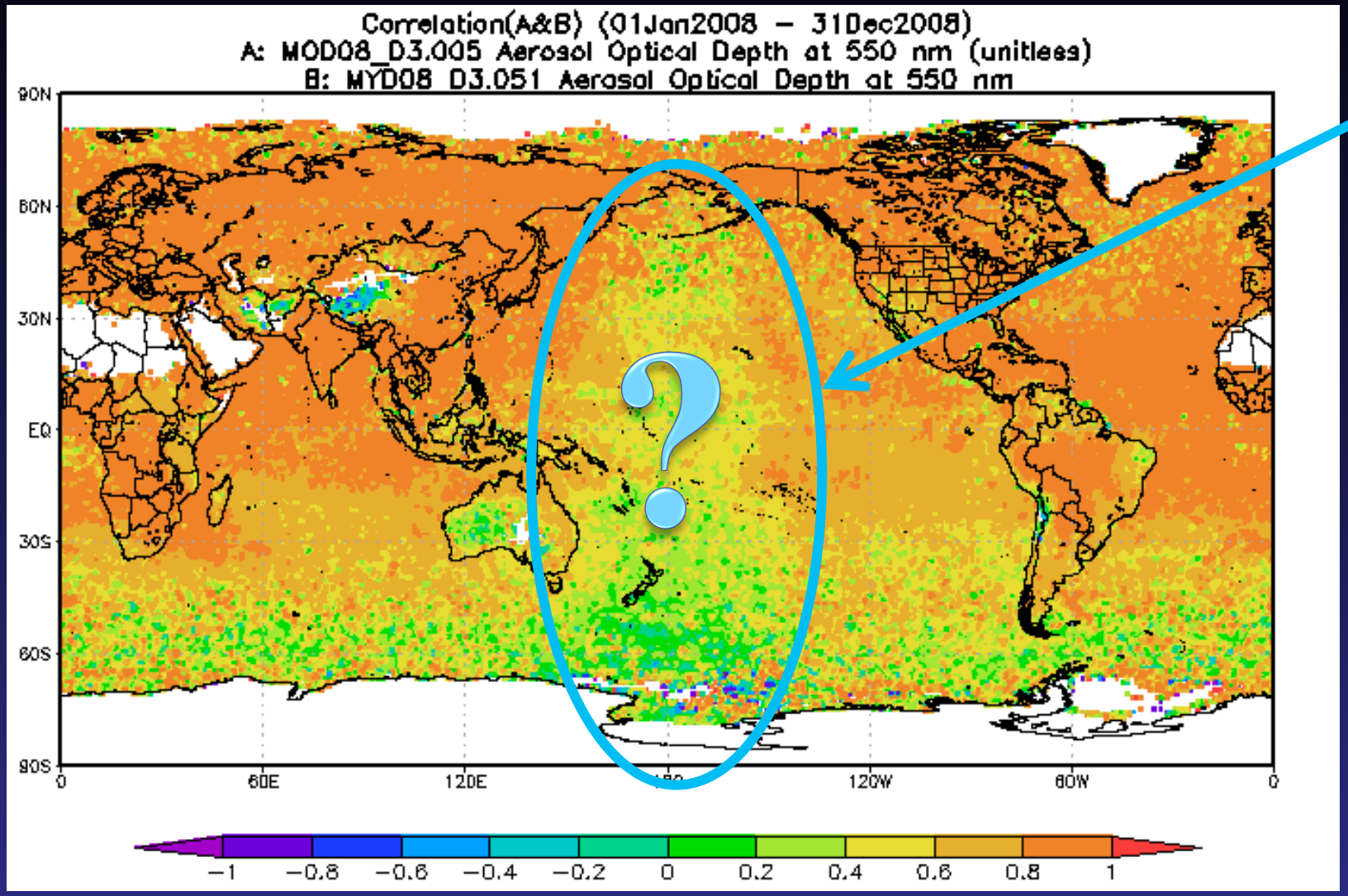
How to work with multi-sensor data?

- Capture and classify the details of measurement technique, data collection and processing
- Identify and spell out similarities and differences
- Assess importance of these differences
- Deliver all this information in such a way that a user can easily see and understand the details
- Present recommendations to guide the data usage and avoid apples-to-oranges comparison and fusion

Multi-Sensor Data Synergy Advisor (MDSA)

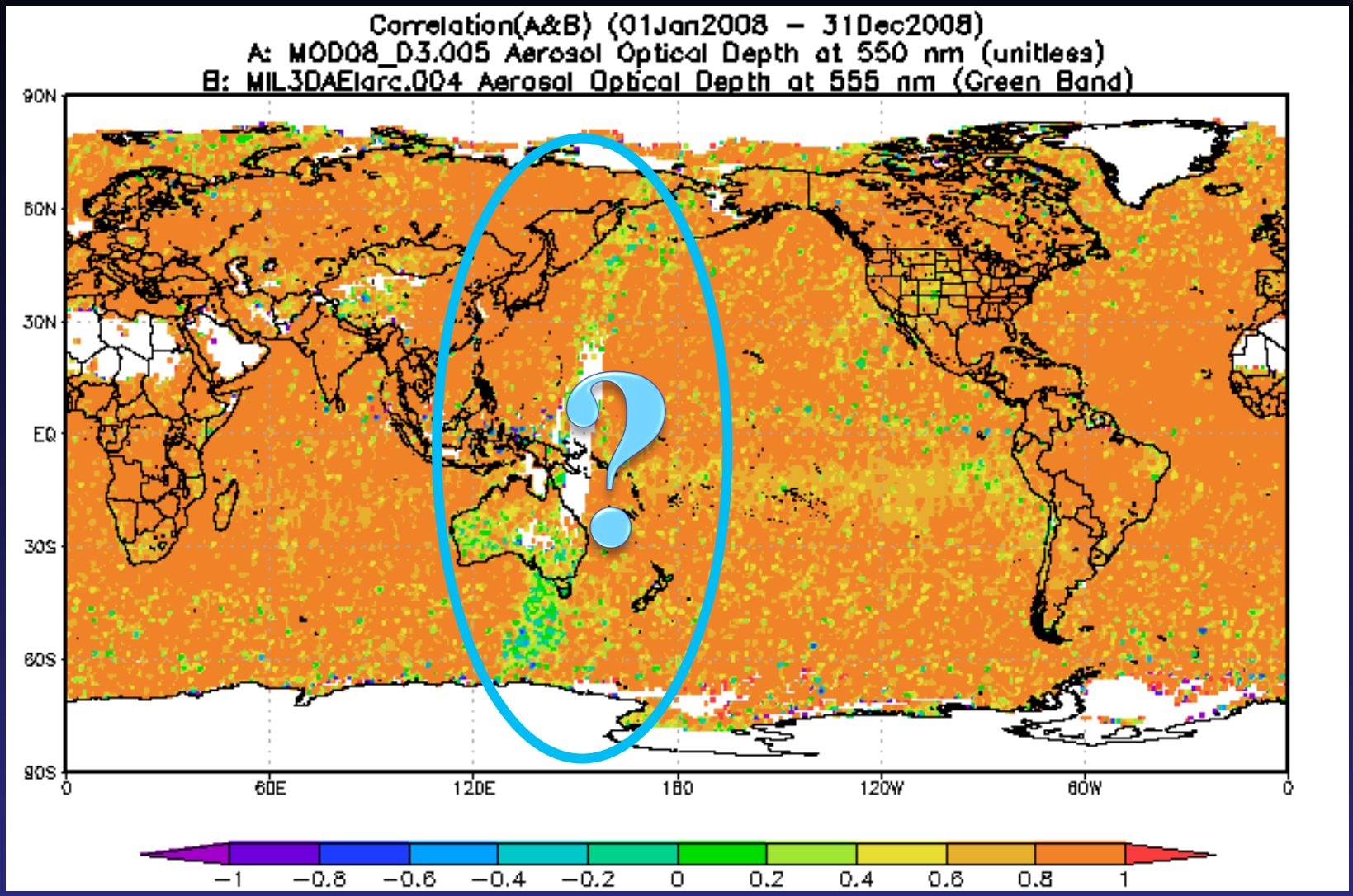
ESTO AIST Project

Case 3: Why don't MODIS Terra and Aqua Aerosols agree?



MODIS-Terra vs. MODIS-Aqua: Map of AOD temporal correlation, 2008

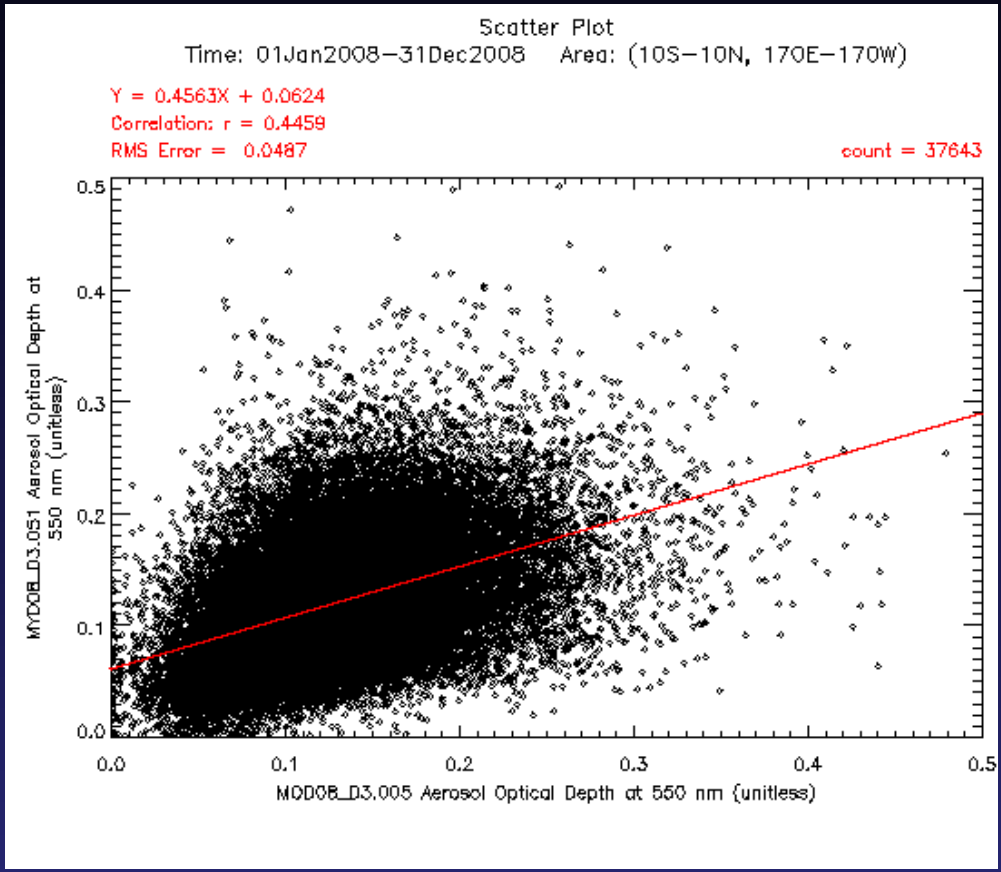
Why does MODIS Terra agree better with MISR than MODIS Aqua?



MODIS-Terra vs. MISR-Terra: Map of temporal correlation

AOD MODIS Terra vs. Aqua in Pacific

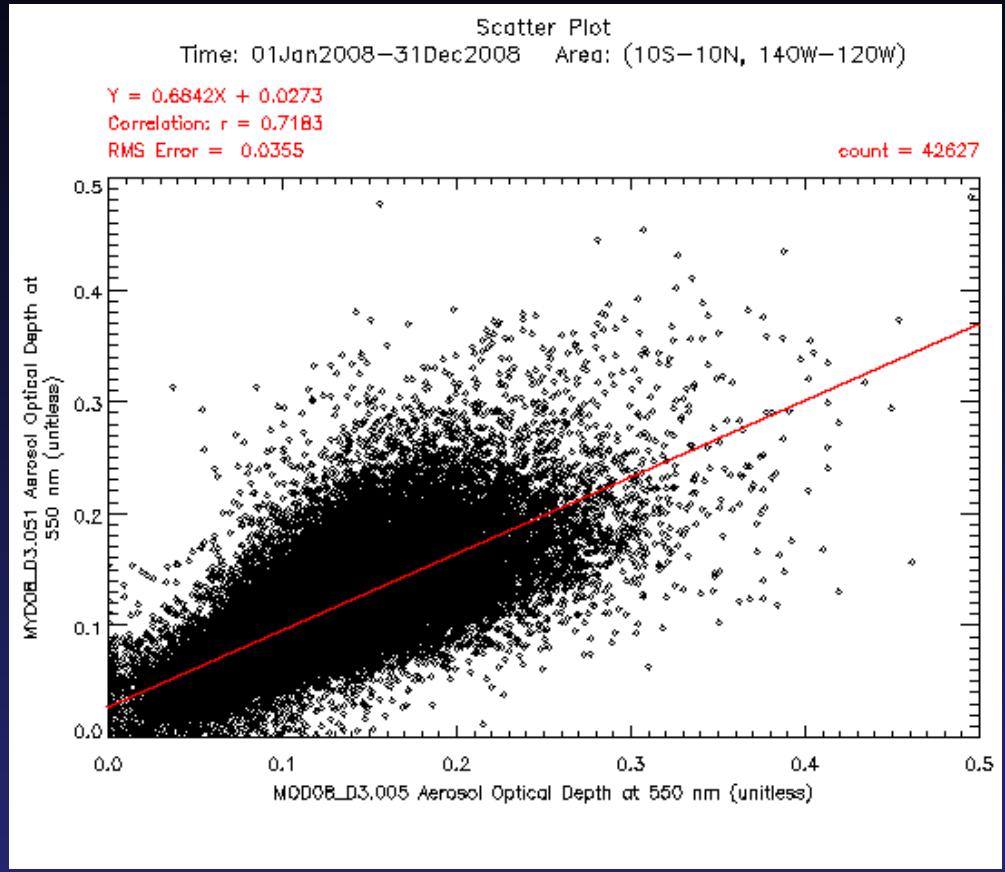
Over the dateline



$R^2 = 0.45$

RMS = 0.05

Away from the dateline



$R^2 = 0.72$

RMS = 0.036

Regressing AOD in two adjacent regions lead to different results

Level 3 Data day definitions

Level 3 gridded data are easy to use by modelers, application users, climate scientists… but also easy to get wrong conclusions….

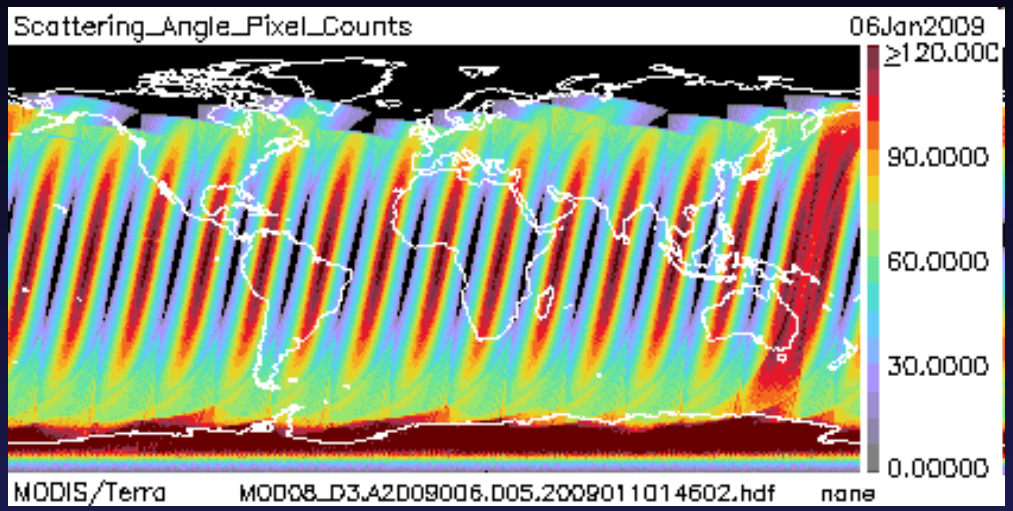
Level 3 daily products are generated by binning Level 2 data belonging to one day onto a certain spatial grid according to a dataday definition:

1. MODIS Atmospheric: all granules between 00:00 – 24:00 UTC
2. Spatial (pixel-based): uses local date/time and ensures spatial continuity. TOMS, AVHRR, AIRS, OMI, MODIS Ocean, SeaWiFS

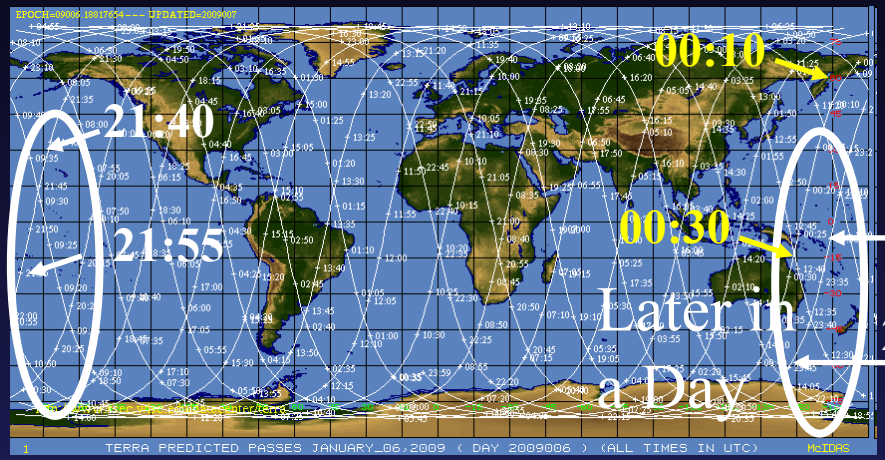
There are other flavors too…

Orbit Time Difference for Terra and Aqua 2009-01-06

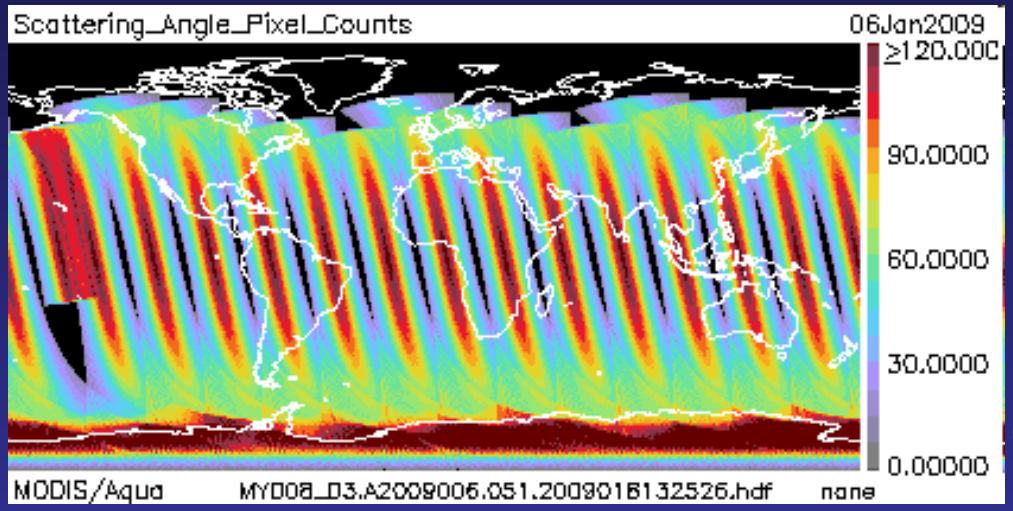
Terra



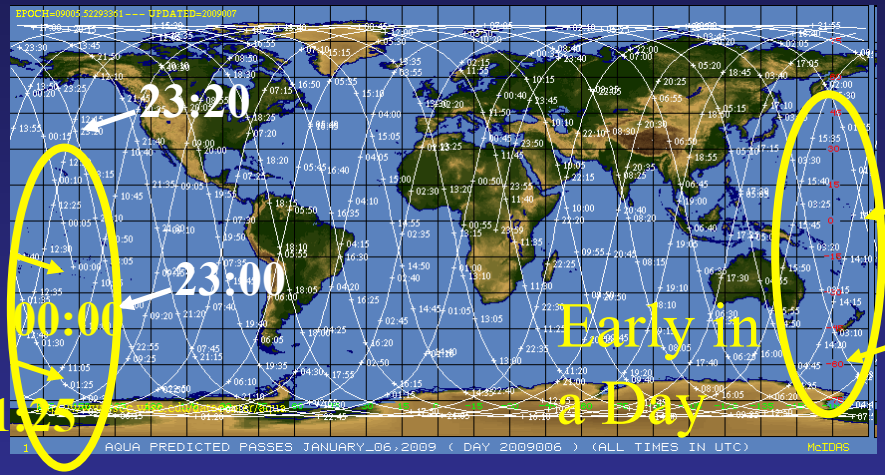
Terra



Aqua

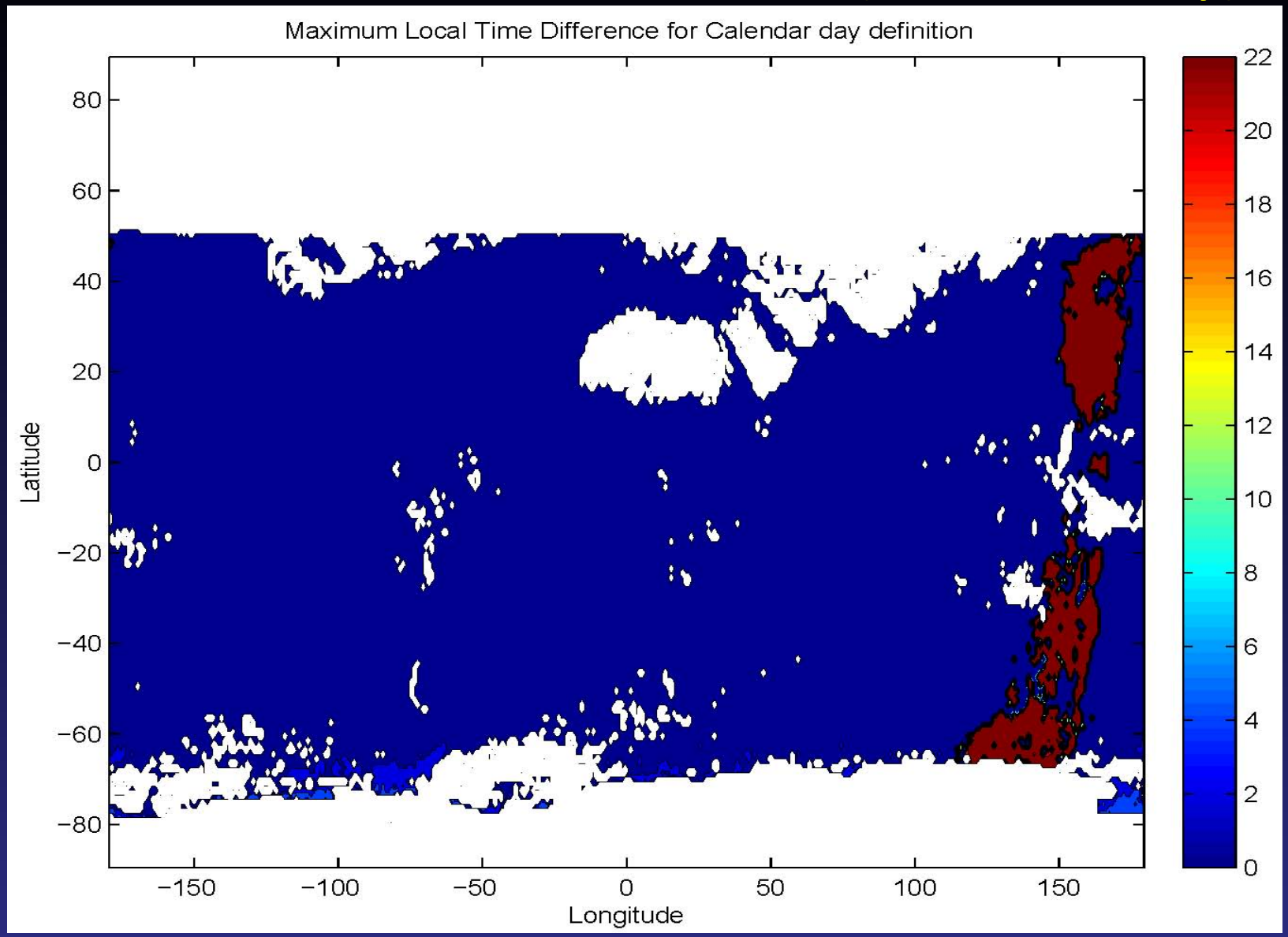


Aqua



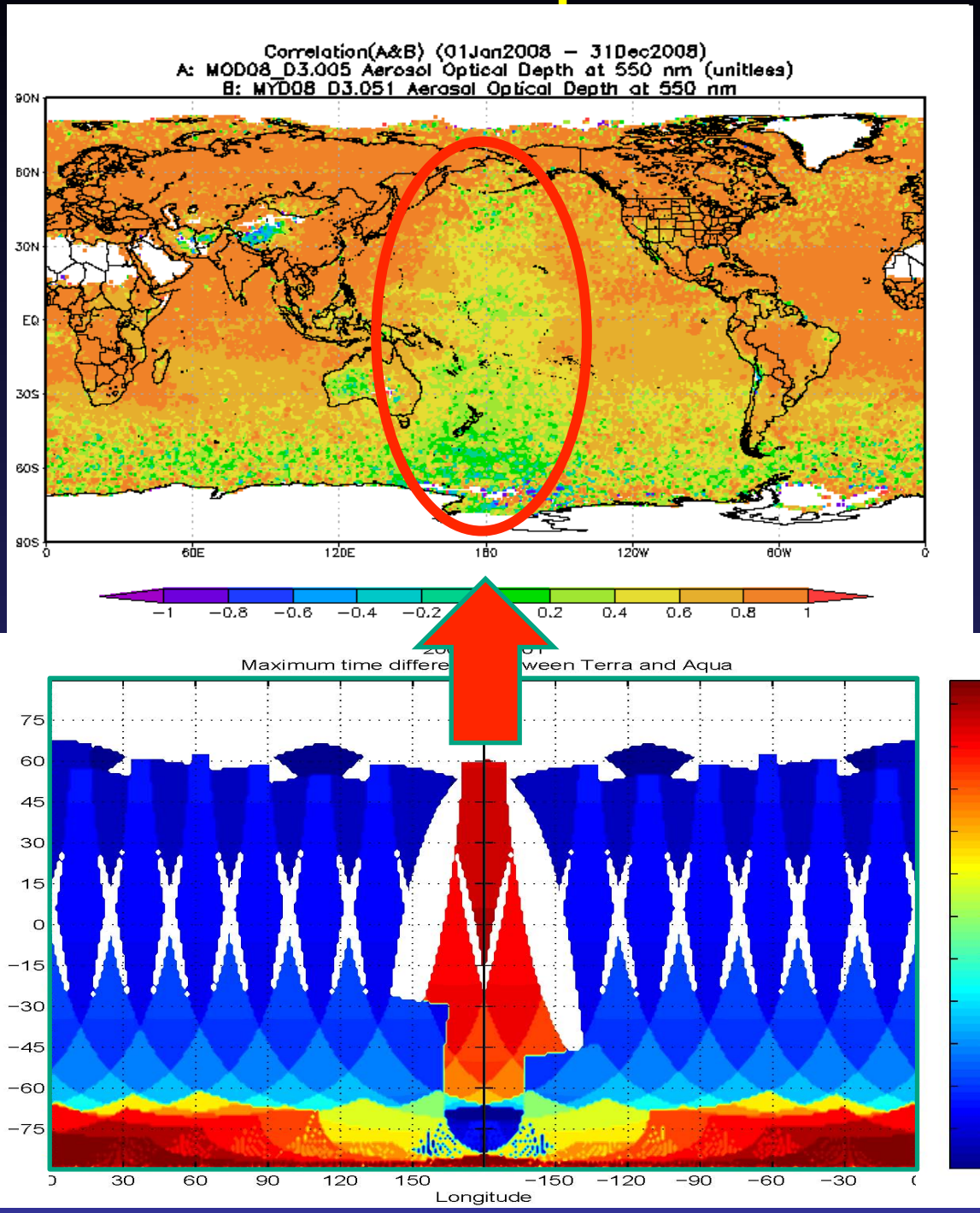
Orbit track from: http://www.ssec.wisc.edu/datacenter

Max Time diff. for Terra (calendar day)



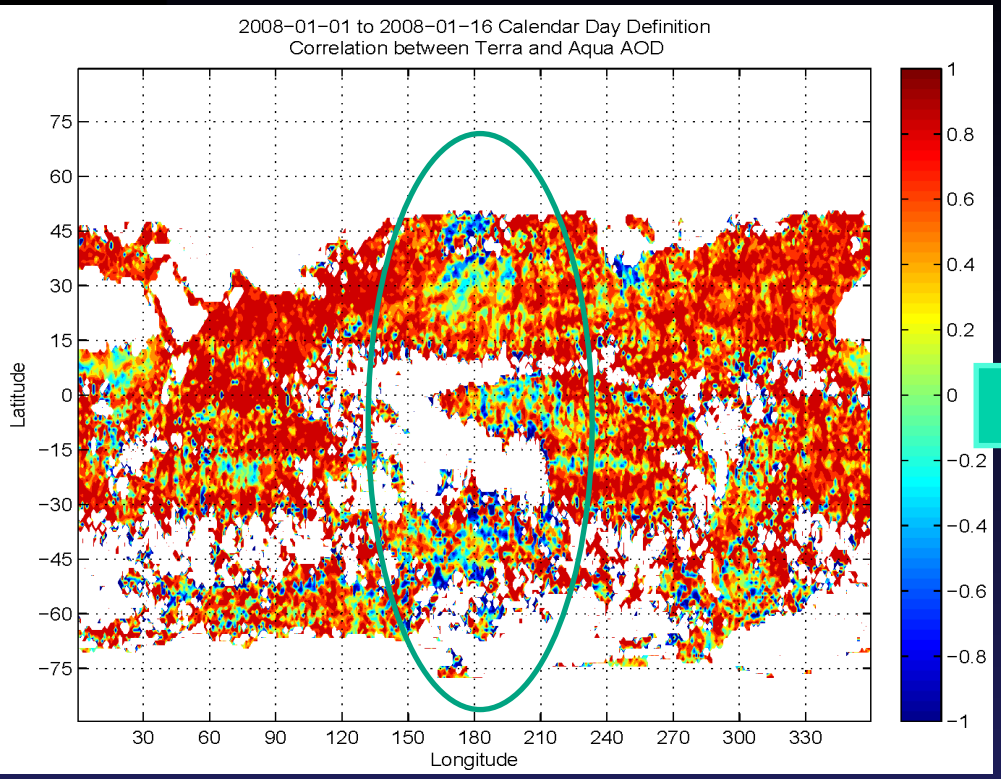
In some areas time difference can go up to more than 22 hours

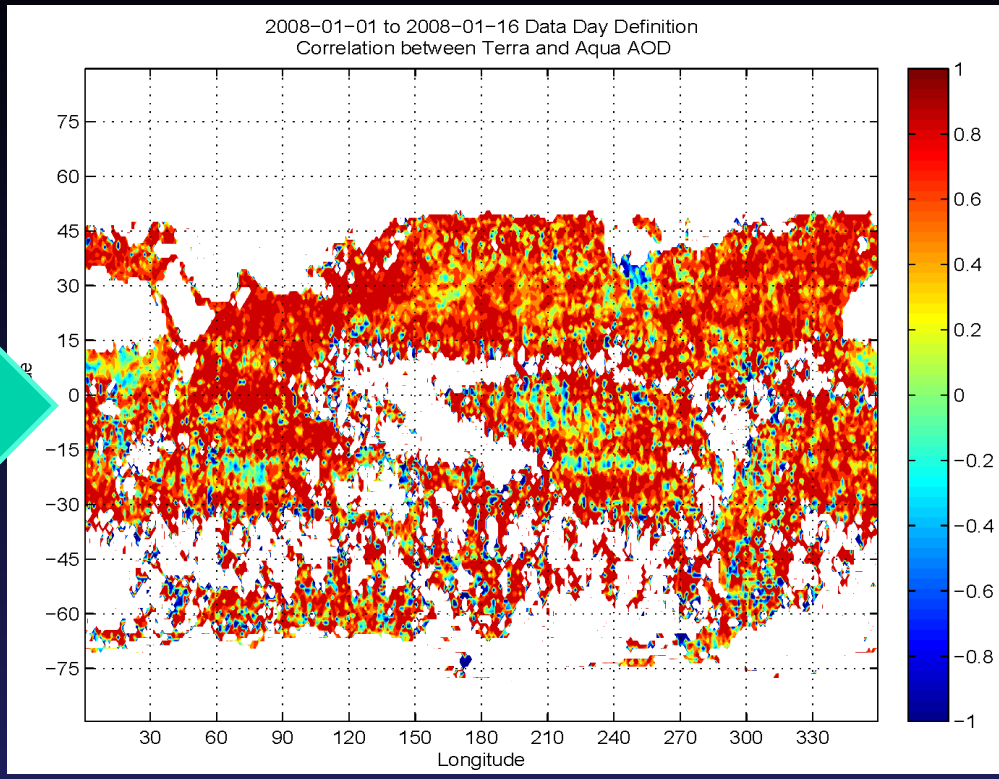
Artifact explained!



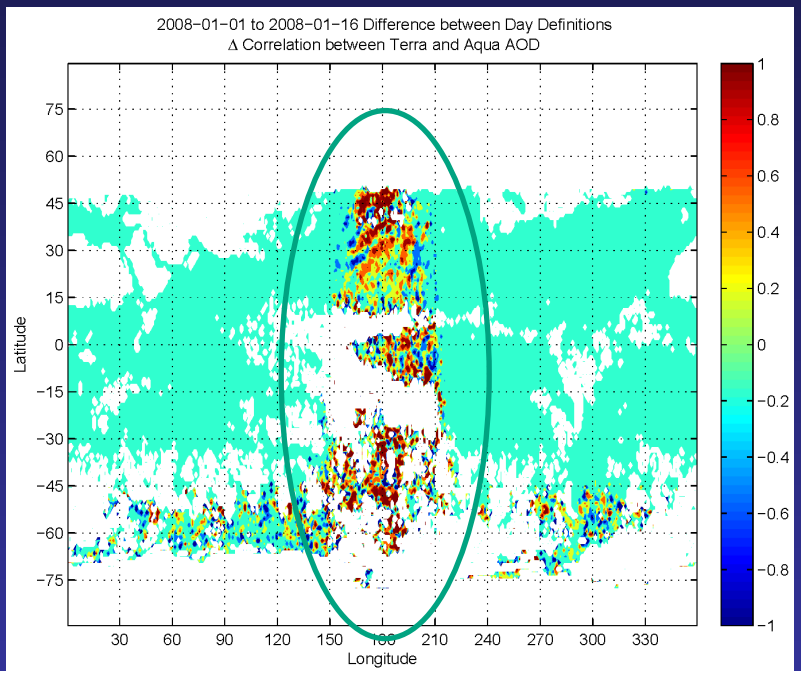
Max Δt between Terra and Aqua

Removing the artifact in 16-day AOD correlation





Calendar dataday



Spatial dataday

Artifact exposed: difference between calendar and spatial dataday defs.

Provenance aspect of difference

- All processing steps are the SAME for both MODIS
- Dataday definition is the same for both MODIS
- Processing provenance alone doesn't provide any explanation for the difference
- Difference in the Equatorial Crossing time alone is not crucial (a diff. dataday def. handles it correctly)
- Only a combination of few factors lead to the artifact: MODIS dataday def. together with Equatorial Crossing time
- It is Knowledge Provenance and the Processing one

Current impact: MODIS Science Team considers generating alternative daily and monthly products for better comparison with other sensor data

MDSA: Presenting data provenances

	Parameter A	Parameter B	Difference alert
Parameter Name :	Aerosol Optical Depth at 550 nm	Aerosol Optical Depth at 550 nm	
Dataset:	MYD08_D3.005	MOD08_D3.005	← Diff
Data-Day definition	UTC (00:00-24:00Z)	UTC(00:00-24:00Z)	The same but….
Temporal resolution	Daily	Daily	
Spatial resolution	1x1 degree	1x1 degree	
Sensor:	MODIS	MODIS	
Platform:	Aqua	Terra	← Diff
EQCT	13:30	10:30	← Diff
Day Time Node	Ascending	Descending	← Diff
Pre-Giovanni Processes :	ATBD-MOD-30	ATBD-MOD-30	
Giovanni Processes:	Spatial subset Time average	Spatial subset Time average	

Conclusions

- Data from multiple sensors provides a more complete coverage of physical phenomena
- Data provenance is needed to insure science quality
- Developing processing provenance is laborious
- Joint provenance is even a bigger challenge
- Proper capture and delivery of joint provenance improves quality of multi-sensor data utilization
- Combination of knowledge provenance and steps in processing provenance is needed to explain artifacts

Giovanni Allows Scientists to Concentrate on the *Science*

The Old Way:

Pre-Science

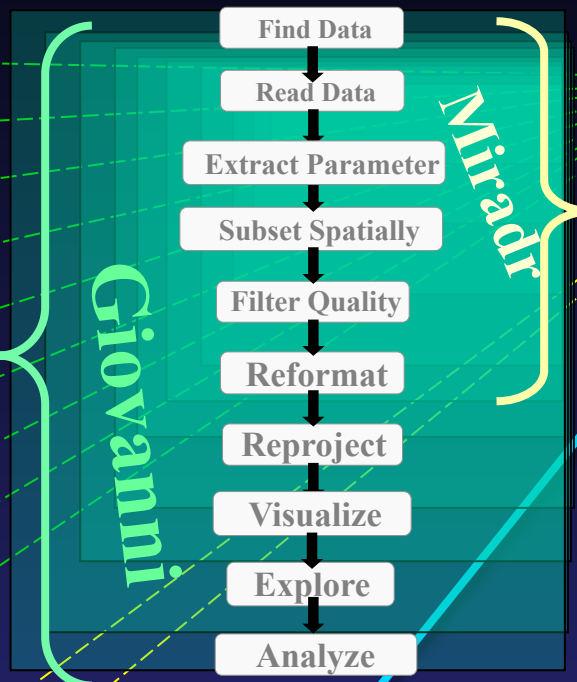
- Find data
- Retrieve high volume data
- Learn formats and develop readers
- Extract parameters
- Perform spatial and other subsetting
- Identify quality and other flags and constraints
- Perform filtering/masking
- Develop analysis and visualization
- Accept/discard/get more data (sat, model, ground-based)

DO SCIENCE

- Exploration
- Initial Analysis
- Use the best data for the final analysis
- Derive conclusions
- Write the paper
- Submit the paper

Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct

Web-based Services:



The Giovanni Way:

- Minutes
- Days for exploration

- Use the best data for the final analysis
- Derive conclusions
- Write the paper
- Submit the paper

DO SCIENCE

Giovanni takes care of technical tasks:
data discovery, access, manipulation, harmonization visualization, and basic statistical analysis.

Scientists have ***more time to do science***