



Bridging Informatics and Earth Science: a Look at Gregory Leptoukh's Contributions

Past, present (and future)



Outline

1. Early Contributions
2. Recent Work
3. Unfinished business...

Science background...

- Tbilisi State University
 - 1975: M.S. in Theoretical Physics
 - 1985: Ph. D. in Cosmic Ray Physics*
 - Monte-Carlo simulation of cosmic ray propagation
- 1994: North Carolina State
 - Molecular dynamics computer simulation
- 1997: NASA, Goddard Distributed Active Archive Center
 - Data browser for SeaWiFS



*Includes work at Moscow State University and Moscow Institute of Theoretical and Experimental Physics

In the beginning, size was the issue...

Volume Requirements

40 GB/day (!)

vs.

Distribution Capacity

Approach: Reduce volume of useless bits flowing out

The screenshot shows a NASA Image Viewing interface. At the top, there are navigation buttons: 'Previous Image', 'Next Image', 'Previous Day', 'Next Day', 'Previous Week', and 'Next Week'. Below these, the image ID is '2009032000749 C1A_10488' and the browser path is 'https://image.gsfc.nasa.gov/...'. A central image shows a vertical strip of a reef. Below the image, there is a legend: 'Legend = SARAGREIF SeaWiFS Level 1A RPTT (satellite color remote data, Aug 199, 1999)'. Metadata includes: 'Replacement Flag = ORIGINAL', 'Processing Time = 19990316000000', 'Start Time = 19990308700000', 'End Time = 19990311004740', 'Startswath Latitude = 01.89433', 'Endswath Latitude = 15.70829', 'Startswath Longitude = -155.21194', 'Endswath Longitude = -86.93894', 'Station Name = MOTHERS BAY AQUARIUM RESEARCH INSTITUTE', 'Station Latitude = NA18', and 'Station Longitude = -115.900'. An 'Additional Resources' section contains a link to 'https://image.gsfc.nasa.gov/...'. At the bottom, there are logos for NASA, SeaWiFS, and TOF, along with contact information for the Data Center and a last updated date of '2007-08-21 22:48:14'.



Greg was always out front, informatics-wise...

Email from Greg, circa 1998

One i
of a
After
issue

One issue that I'd rather men
of a Java "rubber band" _veral

CM problems:

- a. Stability of Java
- b. Compilers on all baselines (cost?)
- c. Change Makefile
- d. Dir structure for classes
- e. Classes library
- f. Prologues
- g. Proliferation of different flavors
- h. Modularization



The Other Data Problem: Usability

- Data in Hierarchical Data Format (HDF)
 - Also with an HDF-EOS layer on top
- Required use of an API (C or Fortran) to read data
- Terminology
 - Data variable names
 - Stride, offset, grid, swath, fill value, SDS, ...

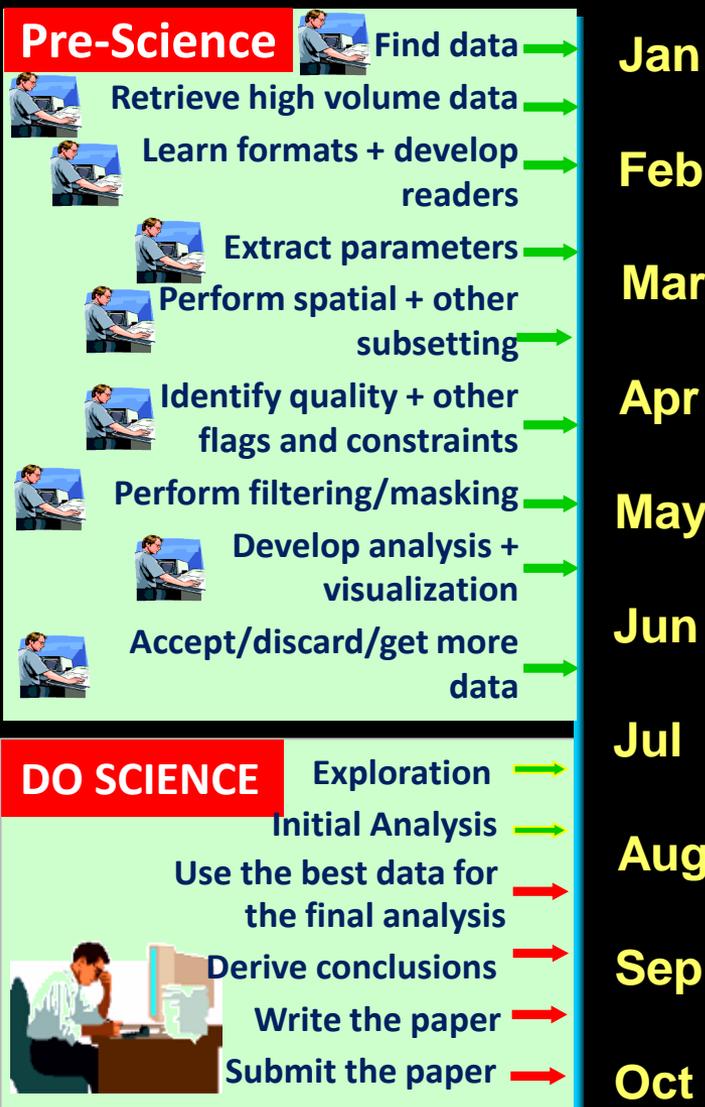
Leptoukh – Kaufman Collaboration



Yoram Kaufman
1948 - 2006

- Yoram wanted to explore MODIS Aerosol data without having to download data or write code
- Kick-started the MODIS Online Visualization and Analysis System (MOVAS)
 - Seed funding
 - Initial requirements
 - Became Giovanni

The Giovanni Way





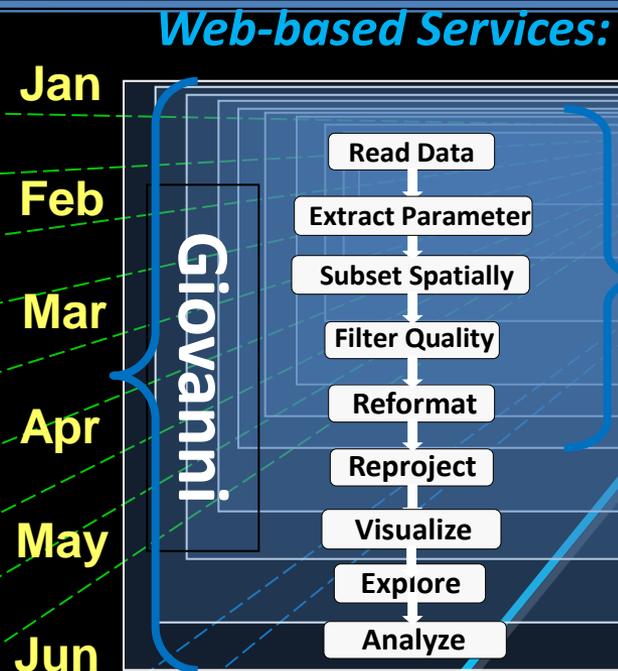
The Giovanni Way

Pre-Science

- Find data
- Retrieve high volume data
- Learn formats + develop readers
- Extract parameters
- Perform spatial + other subsetting
- Identify quality + other flags and constraints
- Perform filtering/masking
- Develop analysis + visualization
- Accept/discard/get more data

DO SCIENCE

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



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

Web-based tools like Giovanni let scientists *shrink* time needed for pre-science preliminary tasks: *data discovery, access, manipulation, visualization, and basic statistical analysis.*

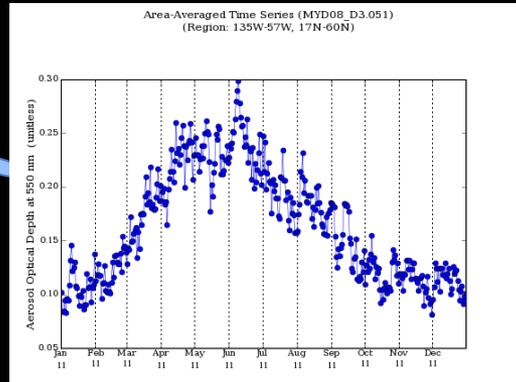
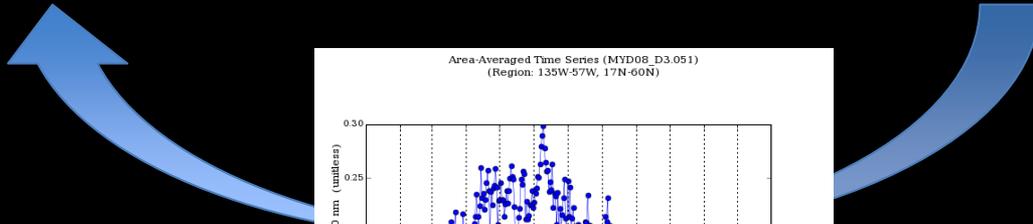
Scientists have *more time to do science!*

User-Driven Development

Add <Feature>?



User-Driven Development

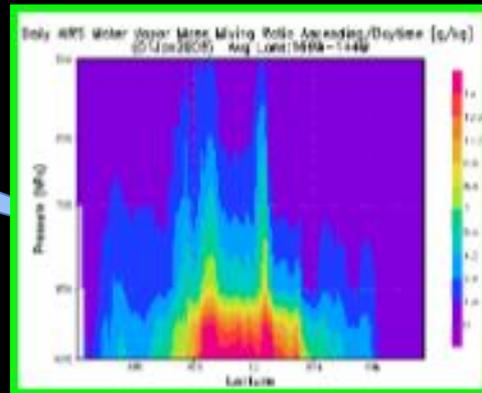


User-Driven Development

Add <Feature>?



User-Driven Development



User-Driven Development

Add <Feature>?





Some Giovanni Features

Visualization & Analysis

- Time-averaged map
- Area-averaged Time series
- Hovmoller
- Animation
- Vertical profile
- Vertical cross-section
- Climatology + Anomaly

Comparison Features

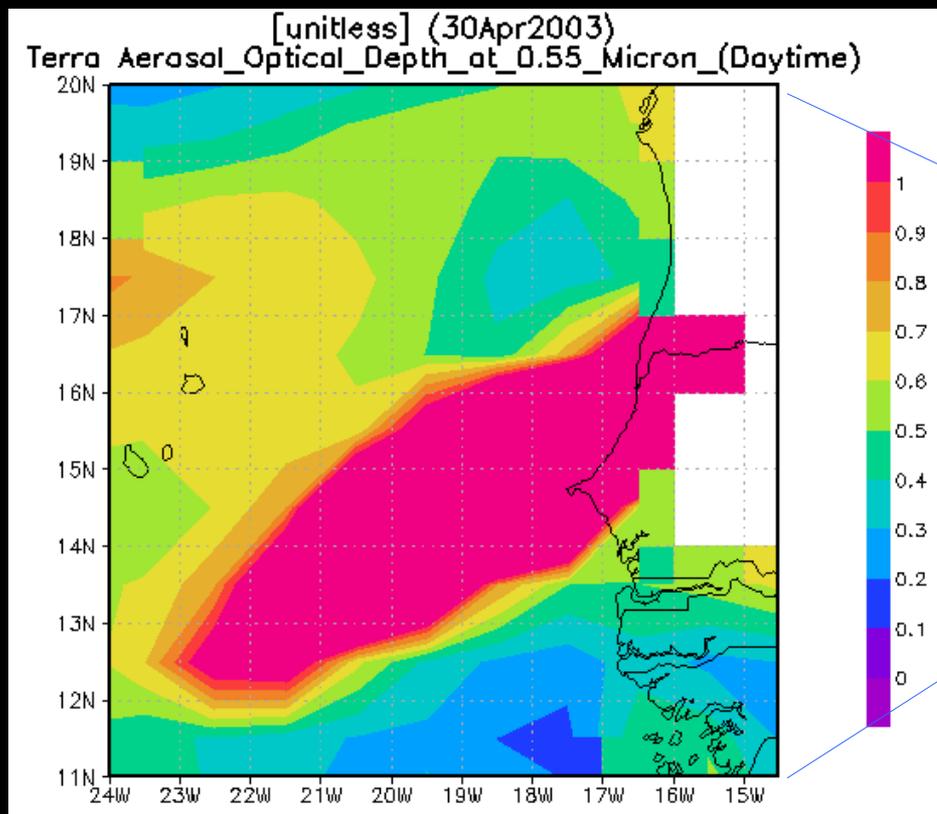
- Scatterplot + linear regression
- Difference Map
- Difference Time Series
- Correlation Map
- Map Overlays

Other Features

- Download data: netCDF, ASCII
- Create KML file
- Show lineage

Basic Visualization: Time-averaged Maps

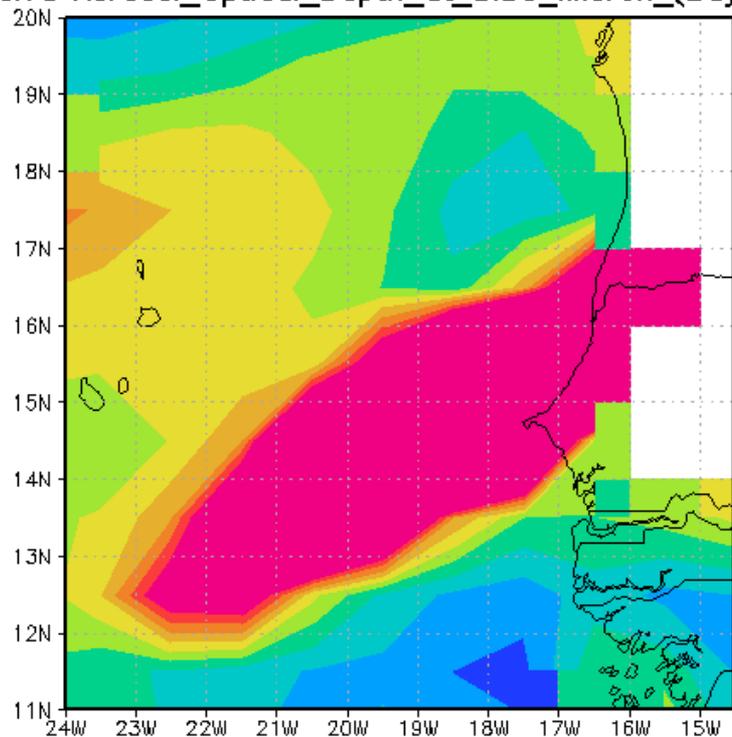
Dust Storm, 30 Apr 2003



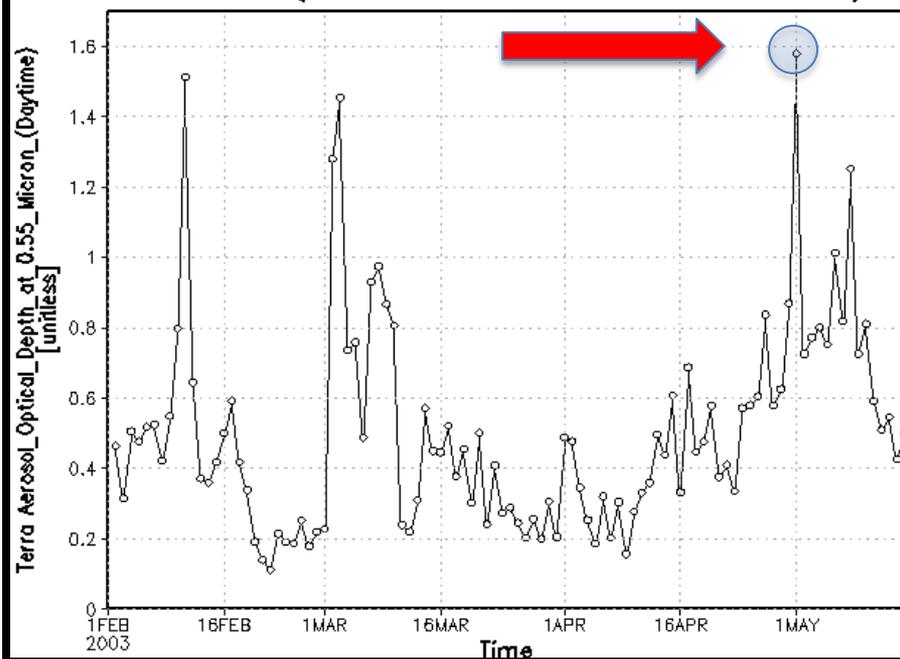
Exploring in time: Area-Averaged Time Series

Dust Storm, 30 Apr 2003

[unitless] (30Apr2003)
Terra Aerosol_Optical_Depth_at_0.55_Micron_(Daytime)



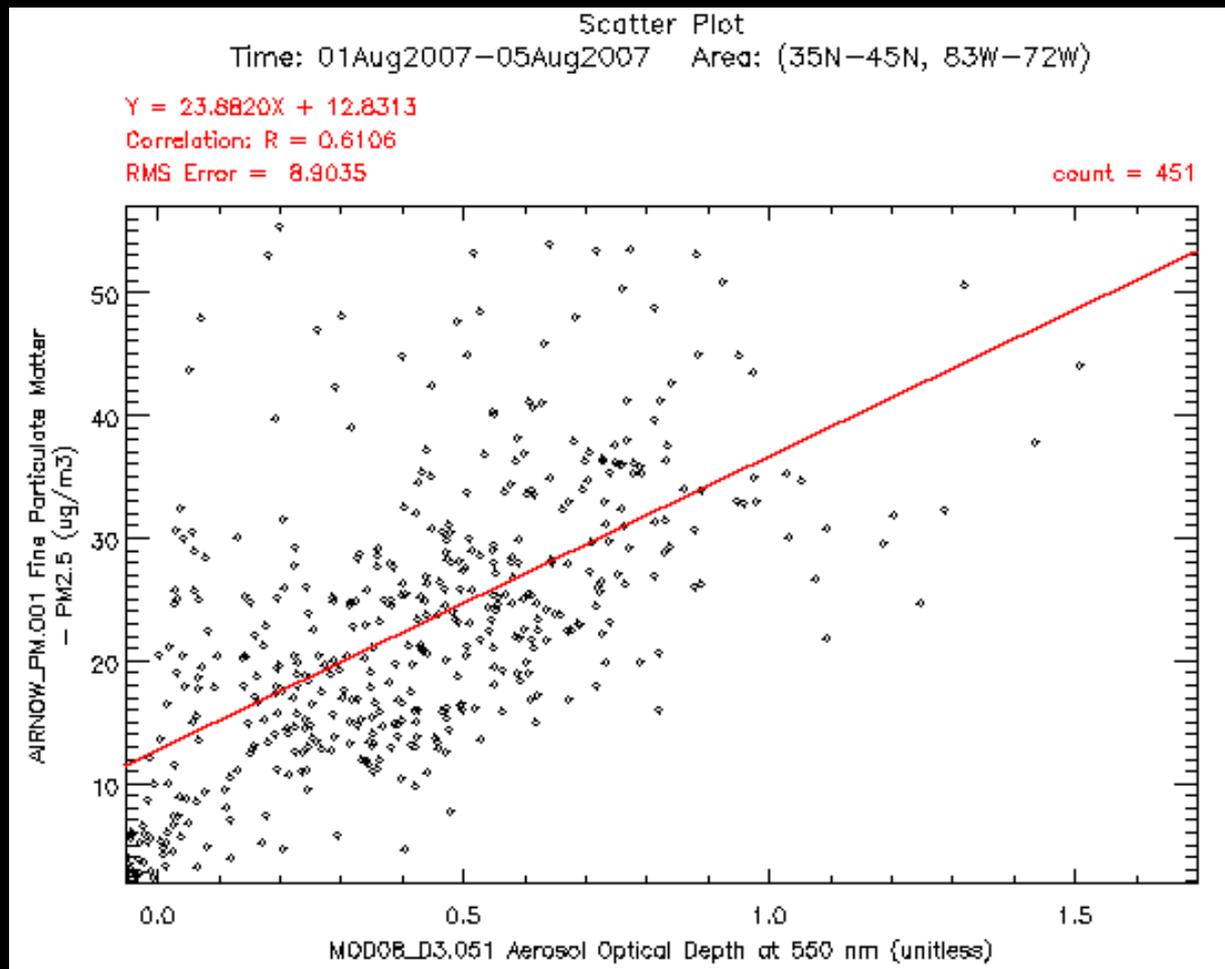
For area (Lat: 11.0N–20.0N, Lon: 24W–14.5W)



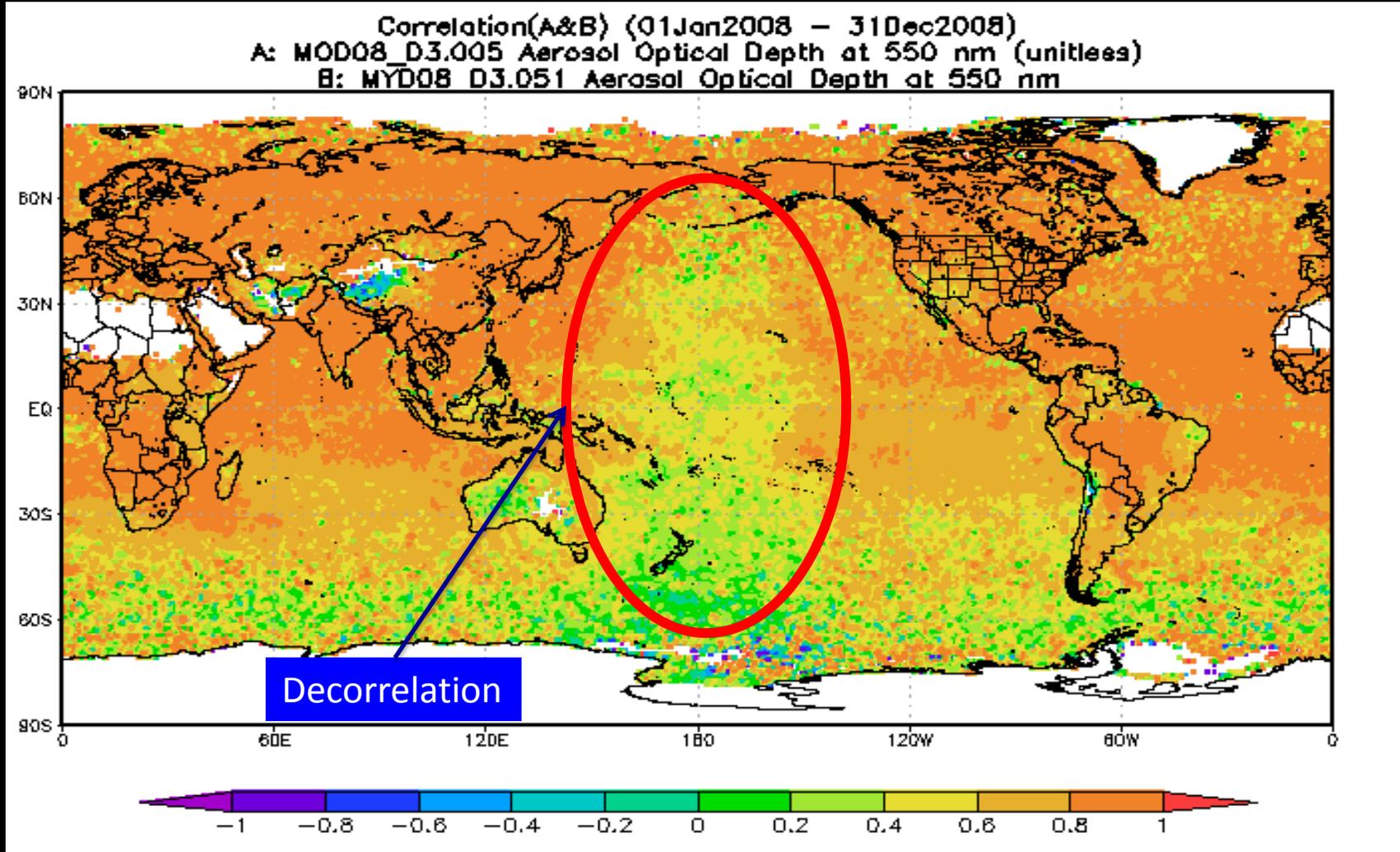
Multi-sensor Analysis: X-Y Scatterplot

AirNow PM 2.5 vs. MODIS Terra Aerosol Optical Depth
01 Aug 2007 to 05 Aug 2007

How well do they correlate?



Exploring Correlation with Correlation Maps: MODIS Aerosols, Aqua v. Terra



Feature: Download Data

- Giovanni as workflow tool
- DIY publication-ready figures

[Visualization Results](#) | **Download Data** | [Product Lineage](#) | [Acknowledgment Policy](#)

Download source data products and data products derived from Giovanni processing stages. For simplicity purposes, only the initial retrieval and final rendering phases are currently accessible for downloading. Supported download formats are HDF, NetCDF(NCD), ASCII, and KMZ (ASCII is available only when the array size is within about half-million points). To **download multiple files** at once, select the desired files (from any section) by clicking on their associated checkboxes, and then click '**Download in Batch**'. **Note:** that 'n/a' means that a file size or other column value is not available; 'saa' means that a file is exactly the same as the previous one in the list. Also, not all services and data products support all download file formats.

Initial Data Retrieval Download in Batch

Data Product	Start Time	File Size (b)	Download Files		
			<input type="checkbox"/> HDF	<input type="checkbox"/> NCD	<input type="checkbox"/> ASC
MOD08_D3.051 (Optical_Depth_Land_And_Ocean_Mean)	2008-11-14T00:00:00Z	100561	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MOD08_D3.051 (Optical_Depth_Land_And_Ocean_Mean)	2008-11-15T00:00:00Z	100561	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MOD08_D3.051 (Optical_Depth_Land_And_Ocean_Mean)	2008-11-16T00:00:00Z	100561	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MOD08_D3.051 (Optical_Depth_Land_And_Ocean_Mean)	2008-11-17T00:00:00Z	100561	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MOD08_D3.051 (Optical_Depth_Land_And_Ocean_Mean)	2008-11-18T00:00:00Z	100561	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Download in Batch

Download Files

Two Dimensional Map Plot

Input Files	Start Time	File Size (b)	Download Files
G3Corr0 (g3_correlation_0)	2008-01-01T00:00:00Z	n/a	<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>
G3Corr0 (g3_correlation_0_count)	2008-01-01T00:00:00Z	525252	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Output Files KMZ

Correlation.MOD08_D3.051_A.MYD08_D3.051_B.2008-01-01.0001.gif	41590	<input checked="" type="checkbox"/>
---	-------	-------------------------------------



Feature: DIY OMI Gridded Products

UV Aerosol Index (OMTO3G.003)	<input type="checkbox"/> Solar Zenith Angle (Deg) <input type="text" value="0.0"/> <input type="text" value="70.0"/> <input type="checkbox"/> Viewing Zenith Angle (Deg) <input type="text" value="0.0"/> <input type="text" value="70.0"/> <input type="checkbox"/> Radiative Cloud Fraction <input type="text" value="0.0"/> <input type="text" value="1.0"/> <input type="checkbox"/> Ocean Glint Angle (Deg) <input type="text" value="20.0"/>	Set filter values for the named OMI parameter
Aerosol Absorption Optical Depth (OMAERUVG.003)	<input type="checkbox"/> Solar Zenith Angle (Deg) <input type="text" value="0.0"/> <input type="text" value="70.0"/> <input type="checkbox"/> Viewing Zenith Angle (Deg) <input type="text" value="0.0"/> <input type="text" value="70.0"/> <input type="checkbox"/> Lambert Equivalent Reflectivity <input type="text" value="0.0"/> <input type="text" value="1.0"/>	Set filter values for the named OMI parameter
Aerosol Absorption Optical Depth (OMAERUVG.003) Return to plot	Parameter Min <input type="text" value="0.0"/> Parameter Max <input type="text" value="0.2"/>	Set parameter preference values
UV Aerosol Index (OMTO3G.003) Return to plot	Parameter Min <input type="text" value="0.0"/> Parameter Max <input type="text" value="5.0"/>	Set parameter preference values



Feature: Lineage

Visualization Results

Download Data

Product Lineage

A

Browse the processing details of the *latlonplot_corr.xml* visualization service.

Data Fetching

Fetches data file(s) using spatial constraints of South: -90 North: 90 East: 180 West: -180 and to Aerosol Optical Depth at 550 nm from MYD08_D3.051
Aerosol Optical Depth at 550 nm from MOD08_D3.051

Preprocessor

The original data files are reformatted to HDF-4 format. Scaling factors are applied, and, in some

Grids Regridding

The following are currently the defaults in the regridding, in order of priority: If all datasets have the

Grid Subsetting

Extracted spatial subset of each parameter in previous step using spatial constraint of South: -90

Anomaly

Anomalies are computed as the differences between a parameter and a selected climatology over

Dimension Collapse

Averaged parameter(s) over the selected spatial area of South: -90 North: 90 East: 180 West: -1

Time Stitch

This step constructs a timeseries for a parameter over the entire temporal range of the selection

Correlation Map

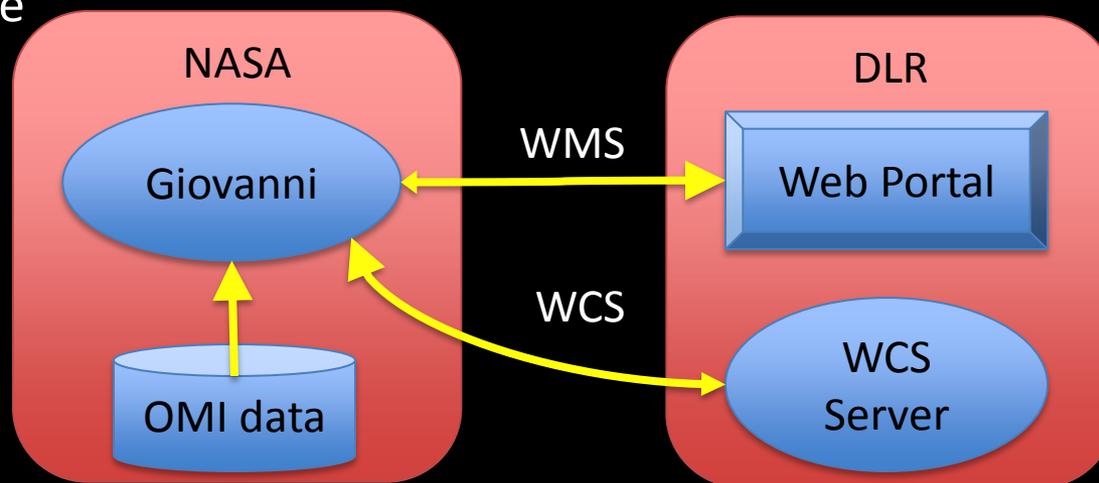
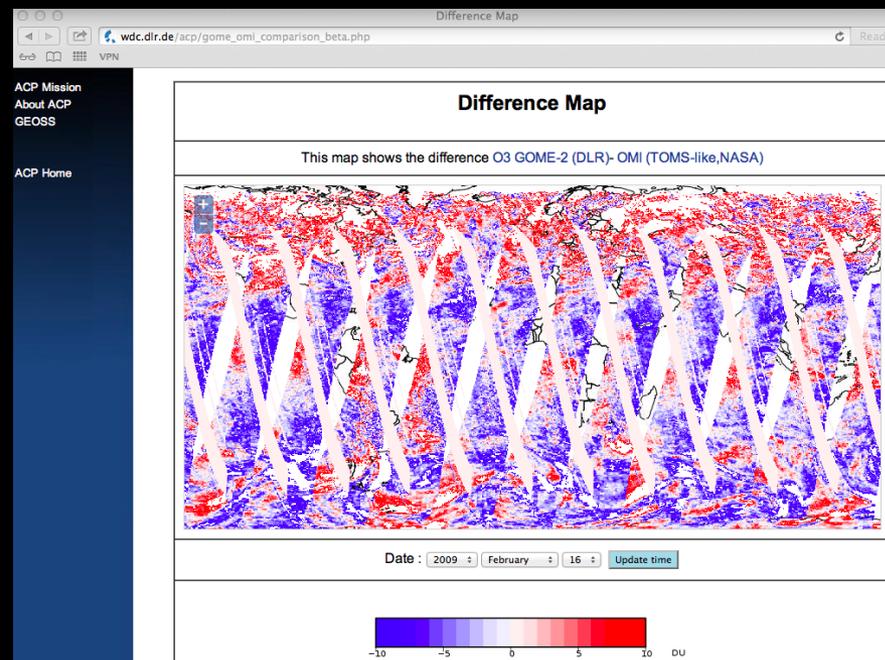
Calculated correlation coefficients at each grid point.

Two Dimensional Map Plot

Generated image(s) with options:
Map Projection = latlon

Standards Support

- Giovanni Output
 - Download Formats
 - Keyhole Markup Language
 - netCDF / CF1
 - OGC Service Formats
 - Web Map Service
 - Web Coverage Service
- Giovanni Input
 - Web Coverage Service
 - OPeNDAP



Giovanni Impact

- Science Research

- Applications:

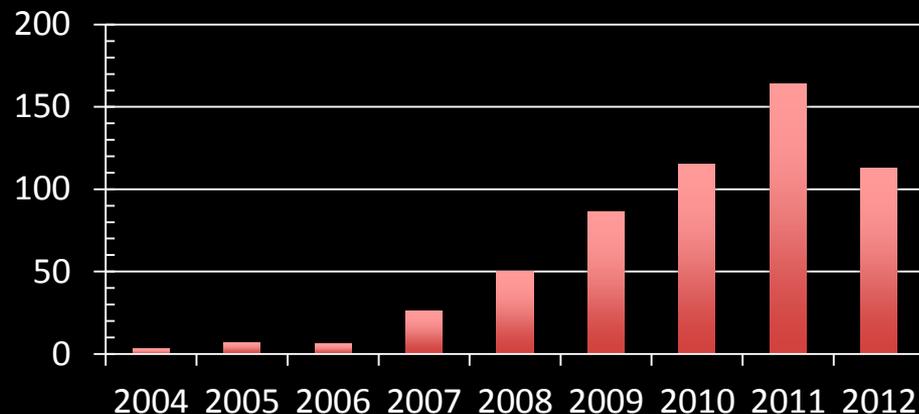
- Air Quality
- Agriculture
- Water Resources

- Education:

Data-enhanced Investigations for Climate Change Education (DICCE)

- Informatics: ...coming up next

Publications





Giovanni as Informatics Laboratory

- Does it make it “too easy” to get results?
 - Spurious comparisons in multi-sensor analysis?
- Motivates attention to documentation
 - Provenance and citation of dynamic results
 - Documentation for variables



Recent Work

- Multi-Sensor Analysis
- Community Work
- Collaboration Enablement



Multi-Sensor Analysis

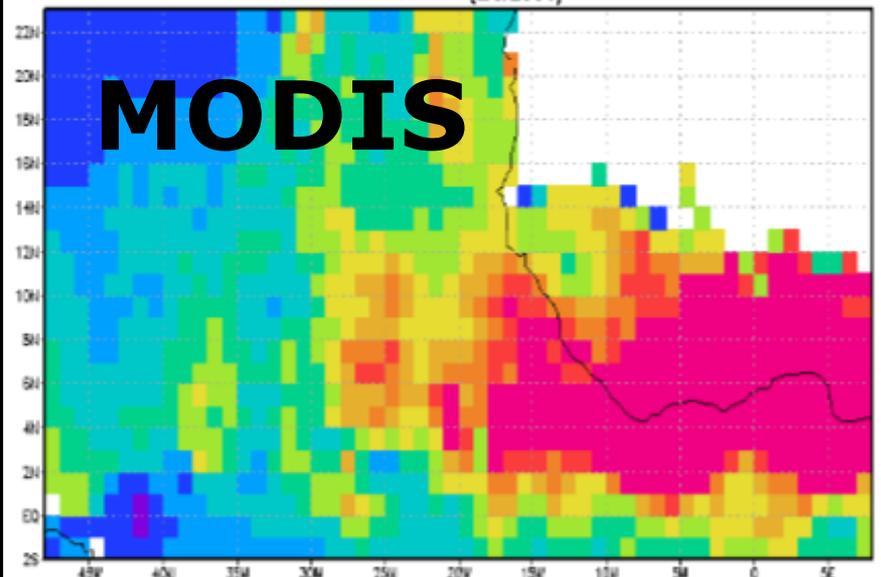
- Multi-sensor Data Synergy Advisor (MDSA)
- AeroStat

MDSA Problem Statement

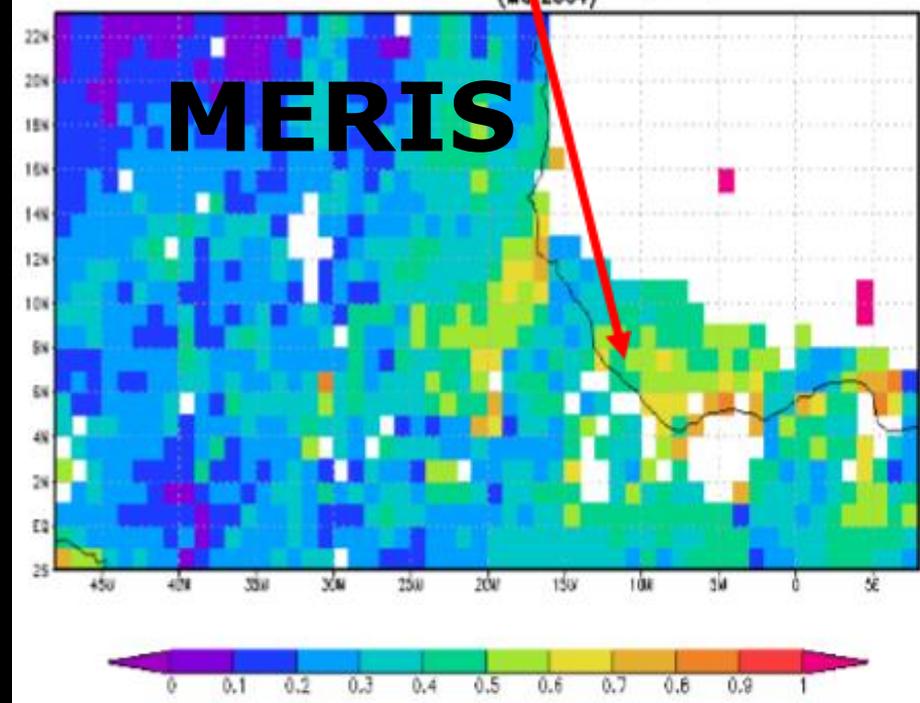
Same measurement
(Aerosol Optical Depth)

Same space & time

MYD08_M3.005 Aerosol Optical Thickness [none]
(Mar2004)



MER_T550.004 AOT 550nm [none]
(Mar2004)



Different provenance



Different results – why?



MDSA provides users with key info about similarities and known issues when comparing or merging datasets

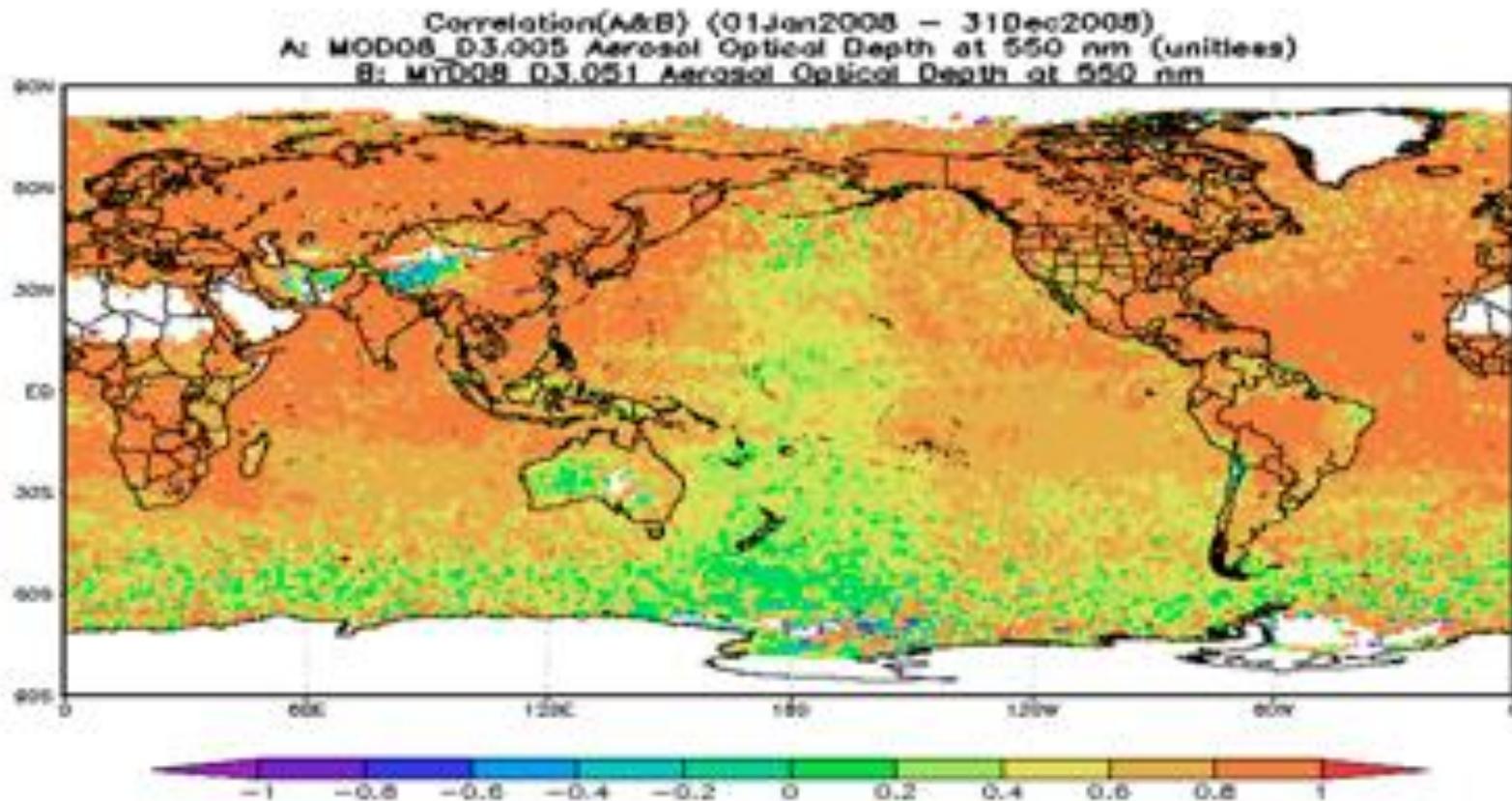
MDSA Similarity Report

About your selected parameters:

	Parameter A	Parameter B	Match
Platform	MODIS	MODIS	
Orbit	Aqua	Terra	
Time	13:30:00	10:30:00	
Swath	ascending	descending	
Data Fetching	Data Fetching	Data Fetching	
Giovanni operations	End Subset <ul style="list-style-type: none">Time AveragingTwo-Dimensional Map Plot	End Subset <ul style="list-style-type: none">Time AveragingTwo-Dimensional Map Plot	

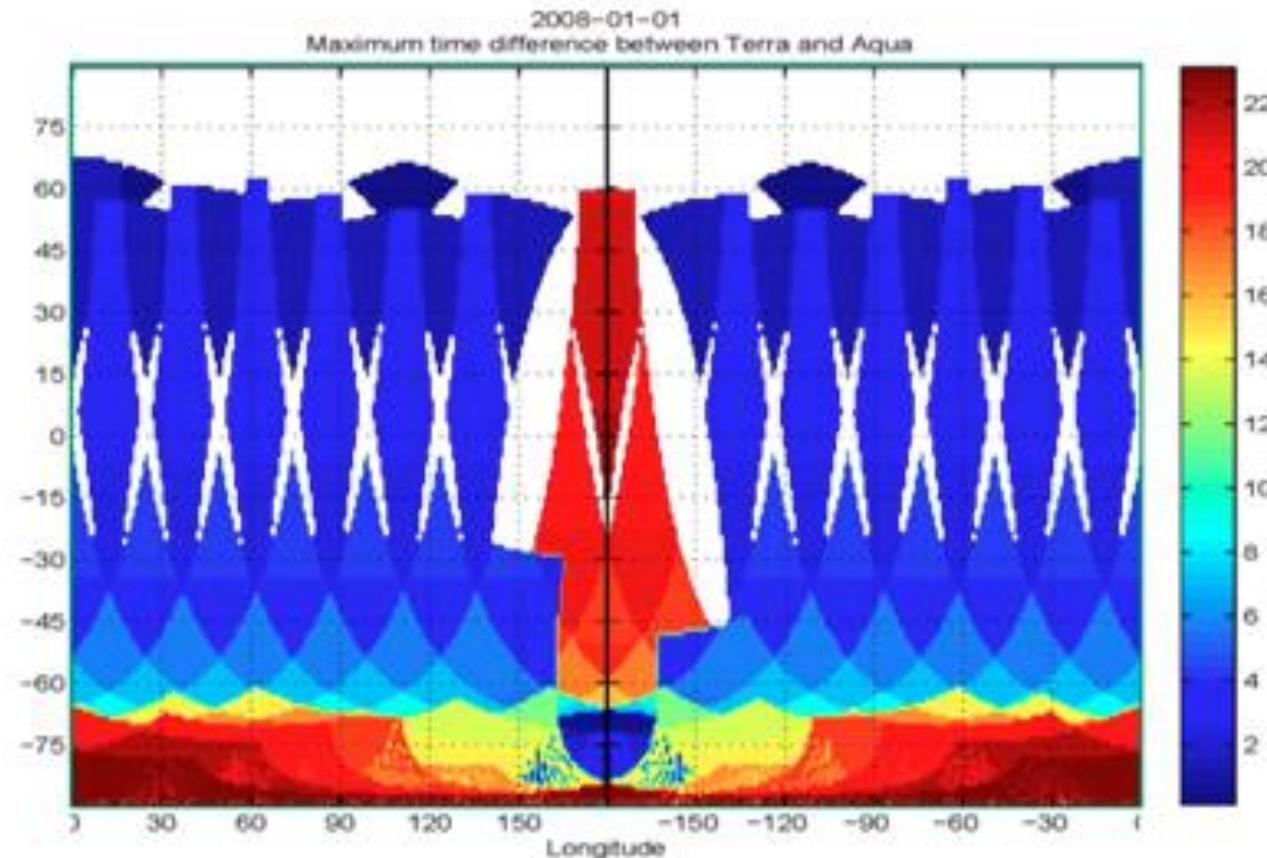


MDSA provides users with key info about similarities and known issues when comparing or merging datasets





MDSA provides users with key info about similarities and known issues when comparing or merging datasets



Maximum time difference between Aqua and Terra observations



MDSA provides users with key info about similarities and known issues when comparing or merging datasets

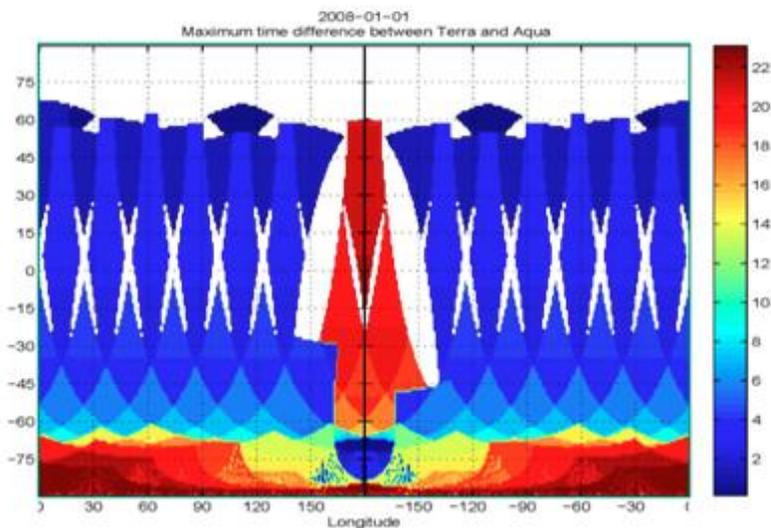
MDSA Similarity Report

Abc
Par
Dat
D
T
S
In
P
G

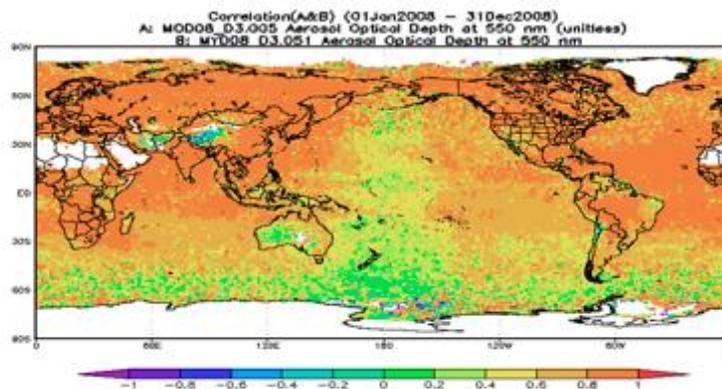
MDSA Known Issues Report

Known issues:

The difference of EQCT and Day Time Node, modulated by data-day definition, caused the included overpass time difference, which makes the artifact difference. See sample images:



Maximum time difference between Aqua and Terra observations



Correlation of MODIS Terra and MODIS Aqua Aerosol Optical Depth at 550 nm



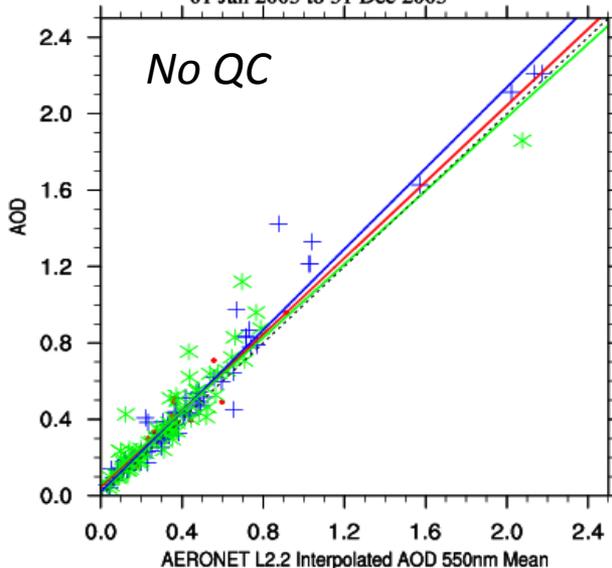
AeroStat Goals

- Compare and combine aerosol data
- Compare aerosol satellite with ground truth
- Merge aerosol measurements from multiple instruments
- Adjust for inter-instrument biases when merging

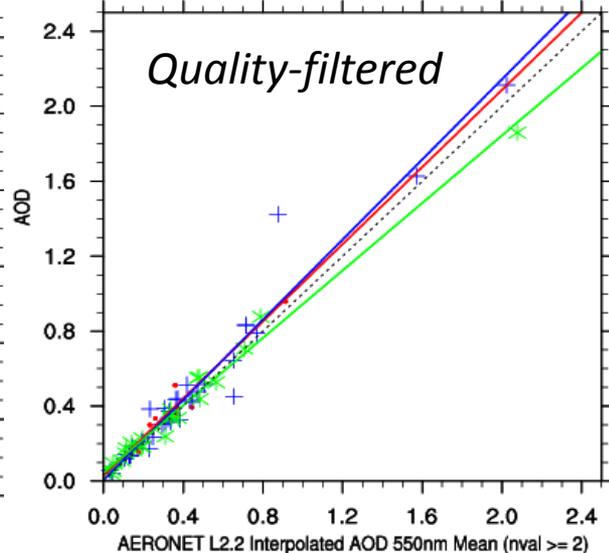


AeroStat System

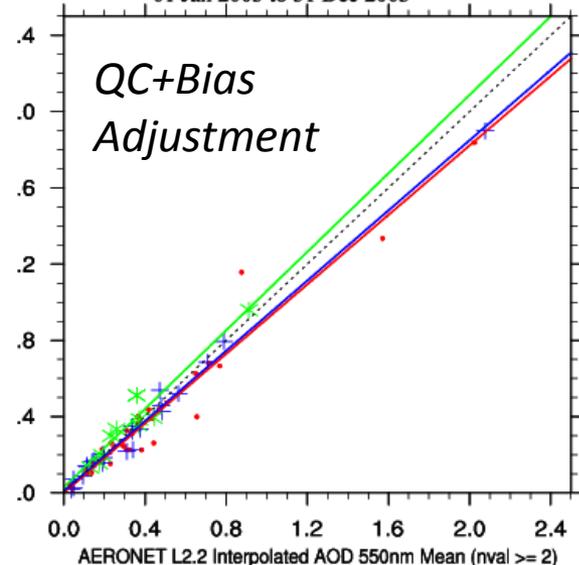
AOD at Capo_Verde (Lat=16.733 Lon=-22.935 Alt=60m)
01 Jan 2003 to 31 Dec 2003



AOD at Capo_Verde (Lat=16.733 Lon=-22.935 Alt=60m)
01 Jan 2003 to 31 Dec 2003



AOD at Capo_Verde (Lat=16.733 Lon=-22.935 Alt=60m)
01 Jan 2003 to 31 Dec 2003



- × MYD04.051 AOD Dark Target Ocean 550nm Mean
- $y=0.961x+0.059, \text{RMS}=0.091, \text{R}^2=0.897, \text{N}=73$
- + MOD04.051 AOD Dark Target Ocean 550nm Mean
- $y=1.059x+0.02, \text{RMS}=0.09, \text{R}^2=0.965, \text{N}=73$
- MISR.022 AOD 558nm Mean
- $y=0.999x+0.045, \text{RMS}=0.076, \text{R}^2=0.863, \text{N}=15$
- 1:1

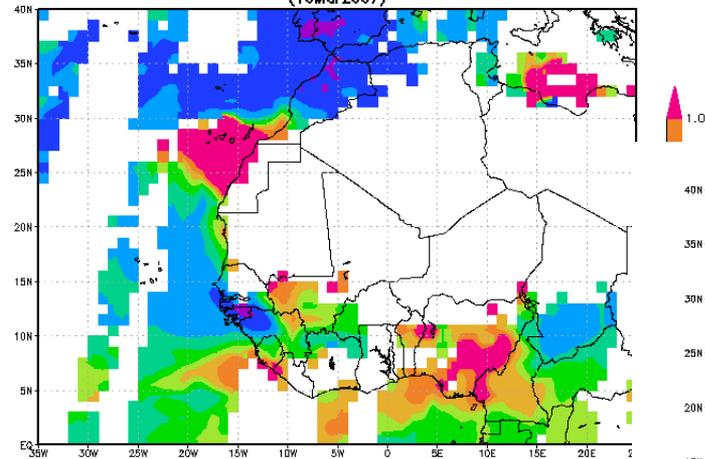
- × MYD04.051 AOD Dark Target Ocean 550nm Mean (nval >= 5, QAavg-o >= 1)
- $y=0.901x+0.043, \text{RMS}=0.047, \text{R}^2=0.984, \text{N}=23$
- + MOD04.051 AOD Dark Target Ocean 550nm Mean (nval >= 5, QAavg-o >= 1)
- $y=1.069x+0.005, \text{RMS}=0.103, \text{R}^2=0.945, \text{N}=35$
- MISR.022 AOD 558nm Mean (nval >= 2, QAb<=1)
- $y=1.029x+0.029, \text{RMS}=0.063, \text{R}^2=0.915, \text{N}=11$
- 1:1

- MISR.022 AOD 558nm Mean (nval >= 2, QAb<=1)
- $y=1.029x+0.029, \text{RMS}=0.063, \text{R}^2=0.915, \text{N}=11$
- MYD04.051 AOD Dark Target Ocean 550nm Mean (nval >= 5, QAavg-o >= 1, BIAS_NN)
- $y=0.92x+0.009, \text{RMS}=0.04, \text{R}^2=0.989, \text{N}=23$
- MOD04.051 AOD Dark Target Ocean 550nm Mean (nval >= 5, QAavg-o >= 1, BIAS_NN)
- $y=0.911x+0, \text{RMS}=0.084, \text{R}^2=0.949, \text{N}=35$
- 1:1

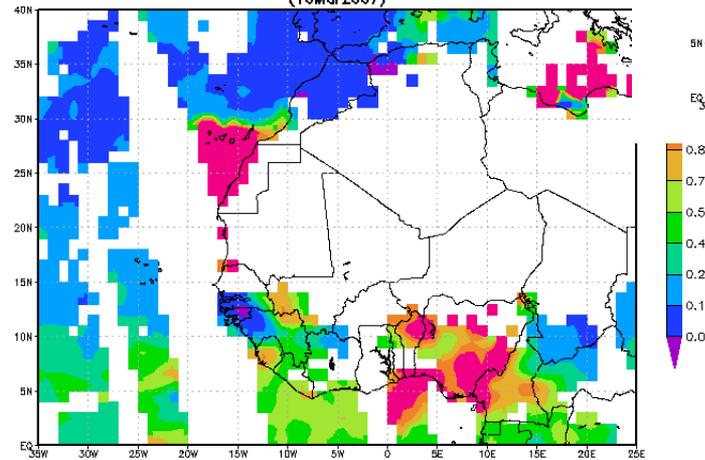
Aerostat Problem Statement

MODIS Dark Target

MOD08_D3.051 Aerosol Optical Depth at 550 nm [unitless]
(10Mar2007)

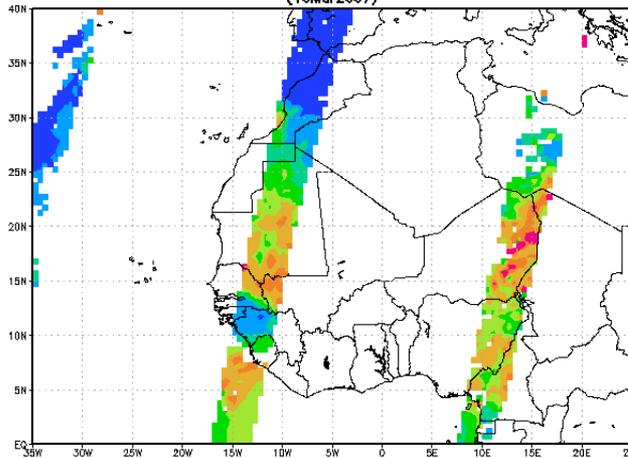


MYD08_D3.051 Aerosol Optical Depth at 550 nm [unitless]
(10Mar2007)



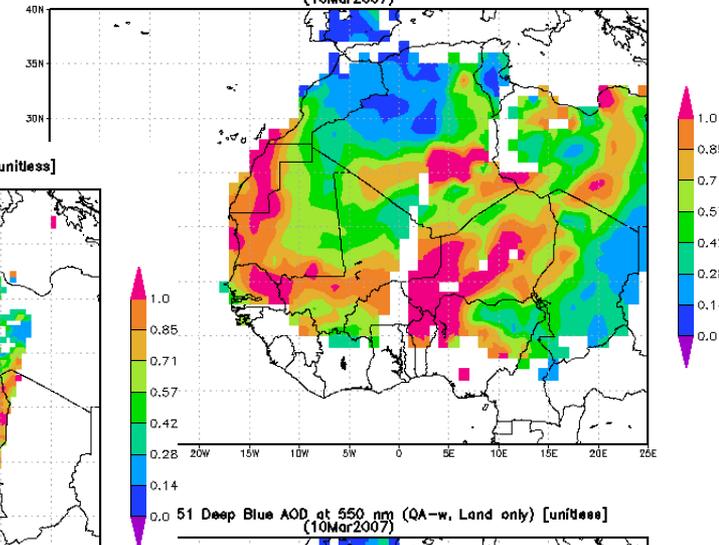
Terra
MISR

MIL3DAE.004 Aerosol Optical Depth at 555 nm (Green Band) [unitless]
(10Mar2007)

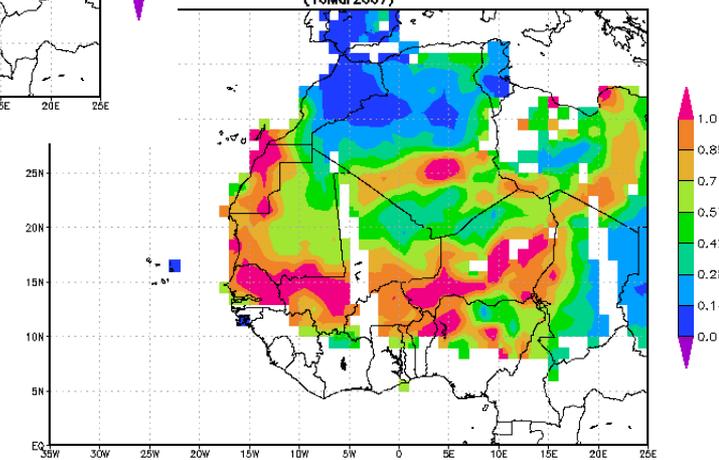


Aqua

MOD08_D3.051 Deep Blue AOD at 550 nm (QA-w, Land only) [unitless]
(10Mar2007)



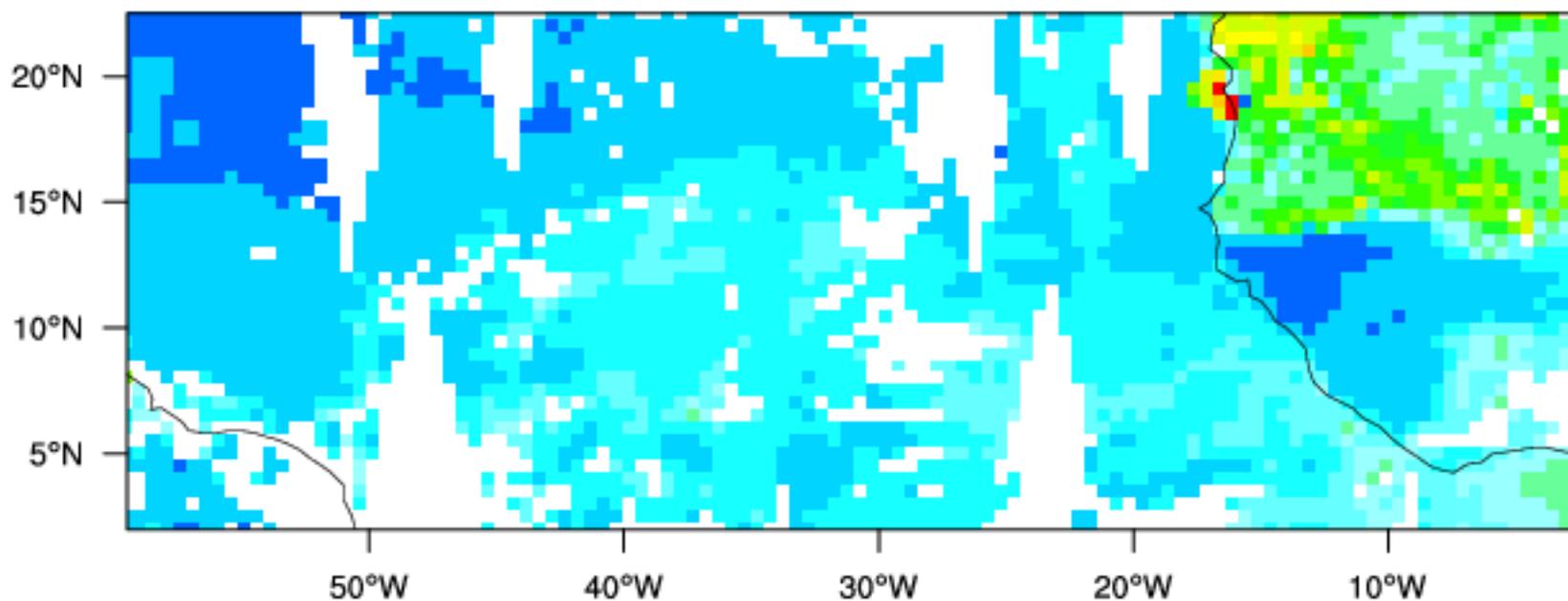
51 Deep Blue AOD at 550 nm (QA-w, Land only) [unitless]
(10Mar2007)



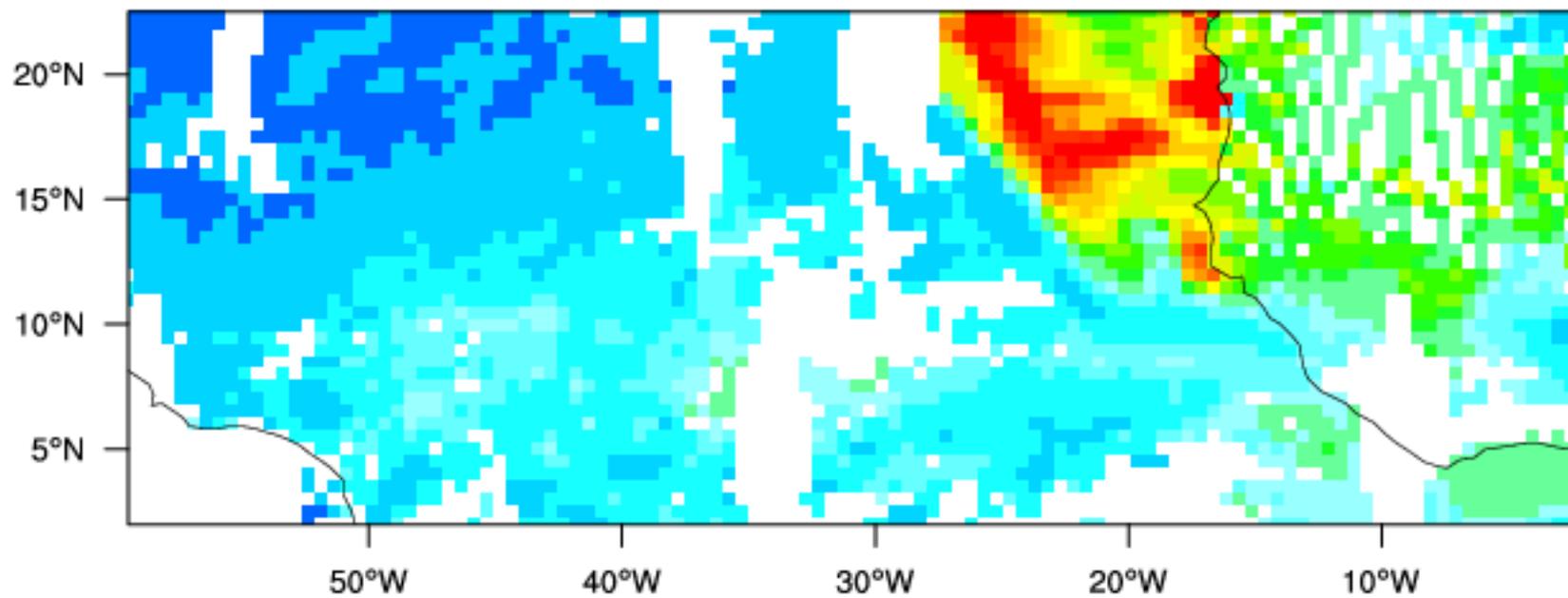
Merged Maps: 1 Mar 2003

AOD Giovanni 0.50x0.50 deg for 01 Mar 2003

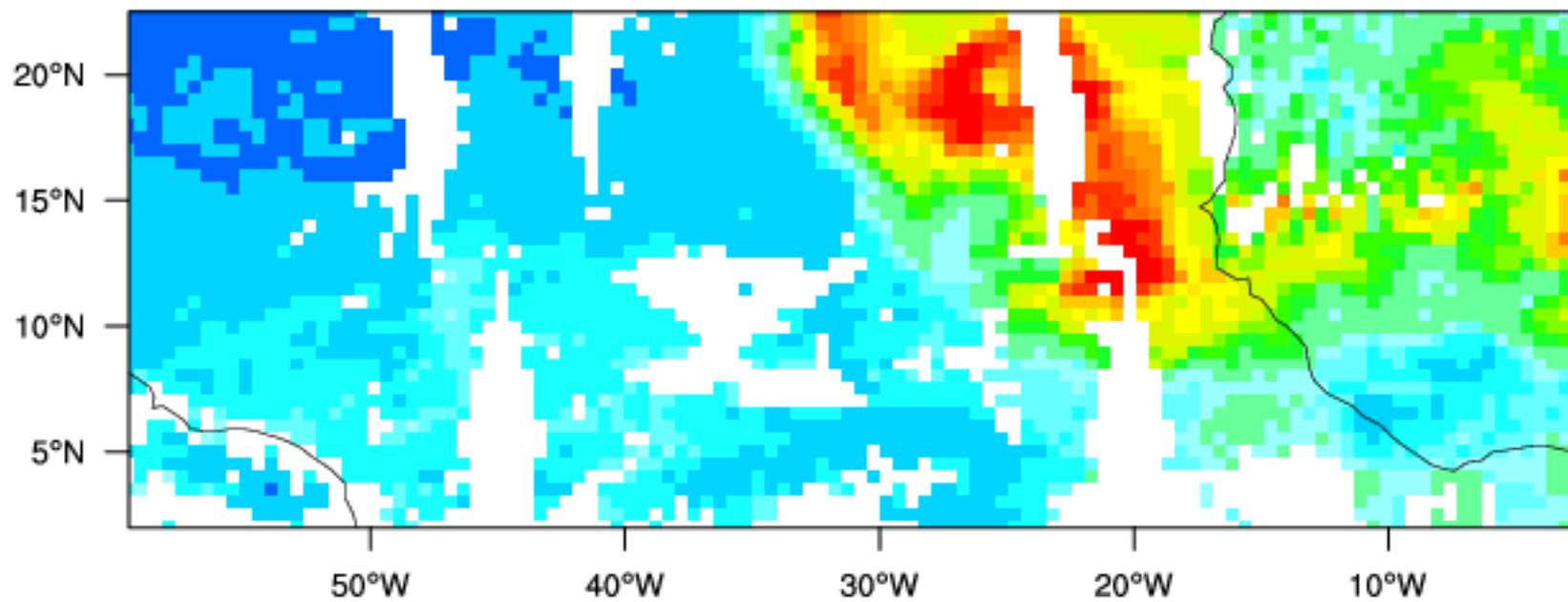
Merged AOD from MOD04 Dark Target Land@550nm*, MOD04 Dark Target Ocean@550nm, MOD04 Deep Blue@550nm*, MISR_AM1_AS@558nm*, MYD04 Deep Blue@550nm*, MYD04 Dark Target Land@550nm*, MYD04 Dark Target Ocean@550nm (*=filtered)



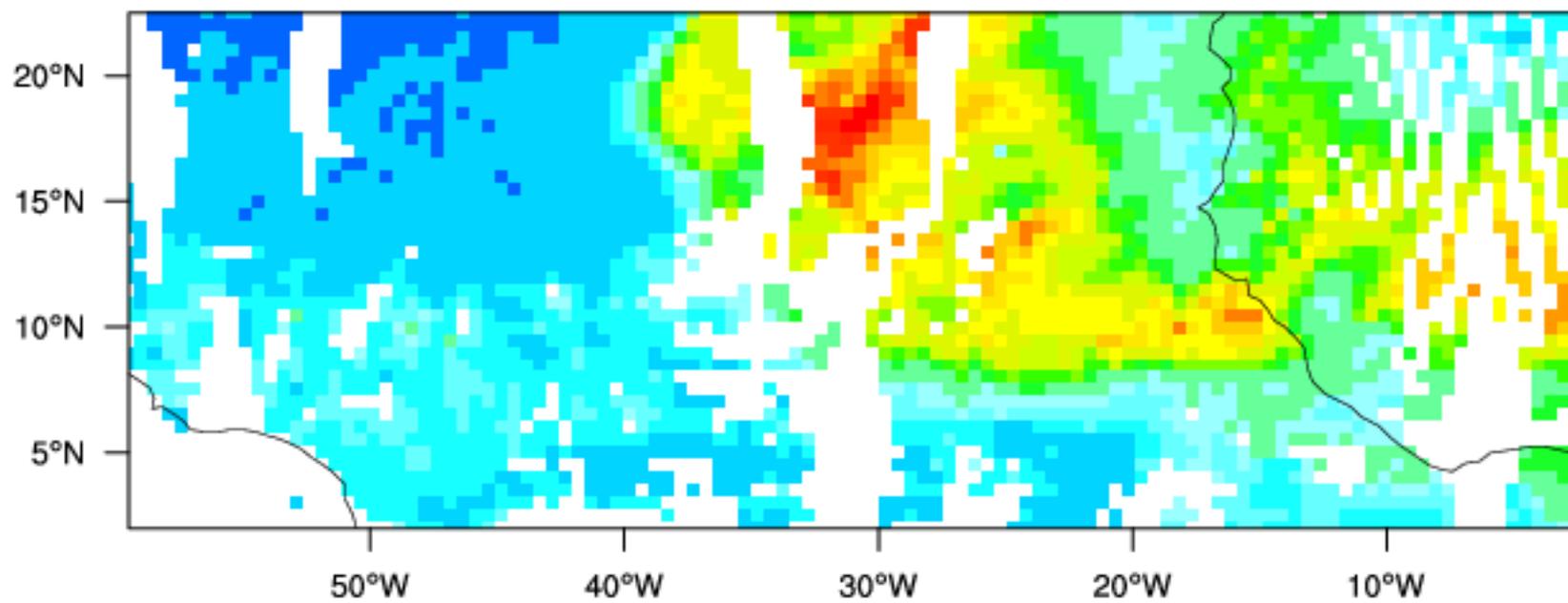
2 Mar 2003



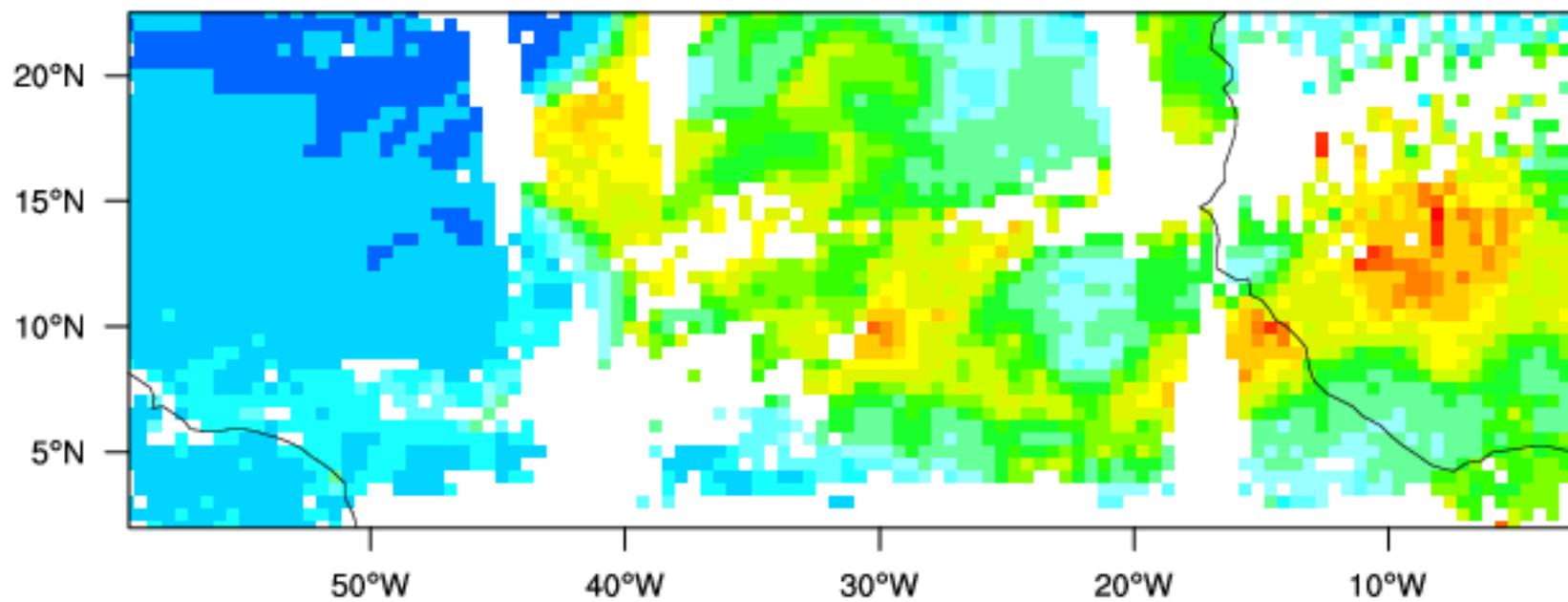
3 Mar 2003



4 Mar 2003



5 Mar 2003





Collaborations in Informatics

- Community Giovanni Portals
 - Northern Eurasian Earth Science Partnership Initiative (NEESPI)
- Community Participation
 - Earth and Space Science Informatics (ESSI)
 - Earth Science Information Partners (ESIP)
 - NASA Earth Science Data Systems Working Group (ESDSWG)
 - GEWEX Aerosol Panel
- Mentor for interns and fellows
- Collaboration Features
 - gSocial

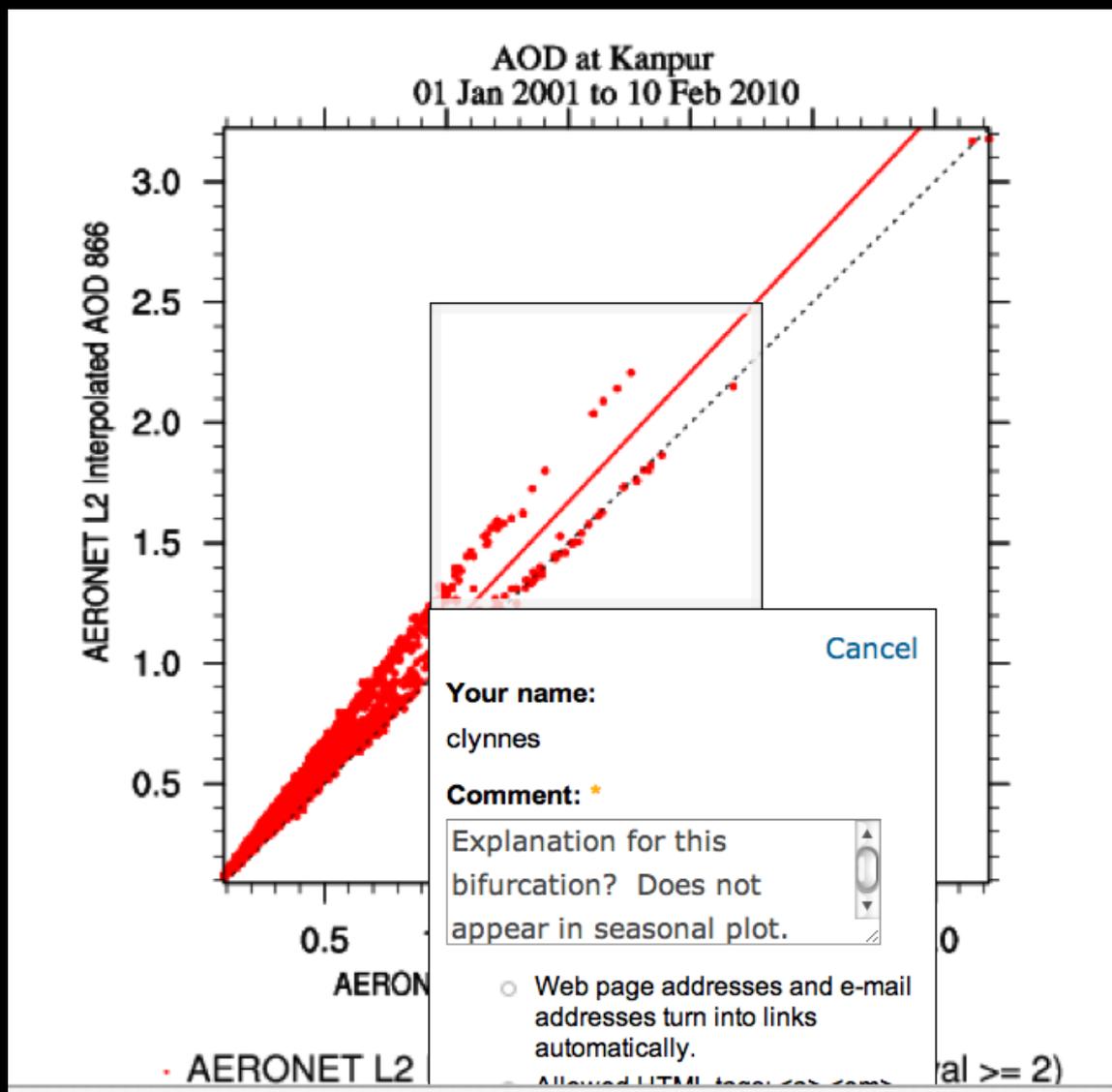


Collaborative Annotation with gSocial*

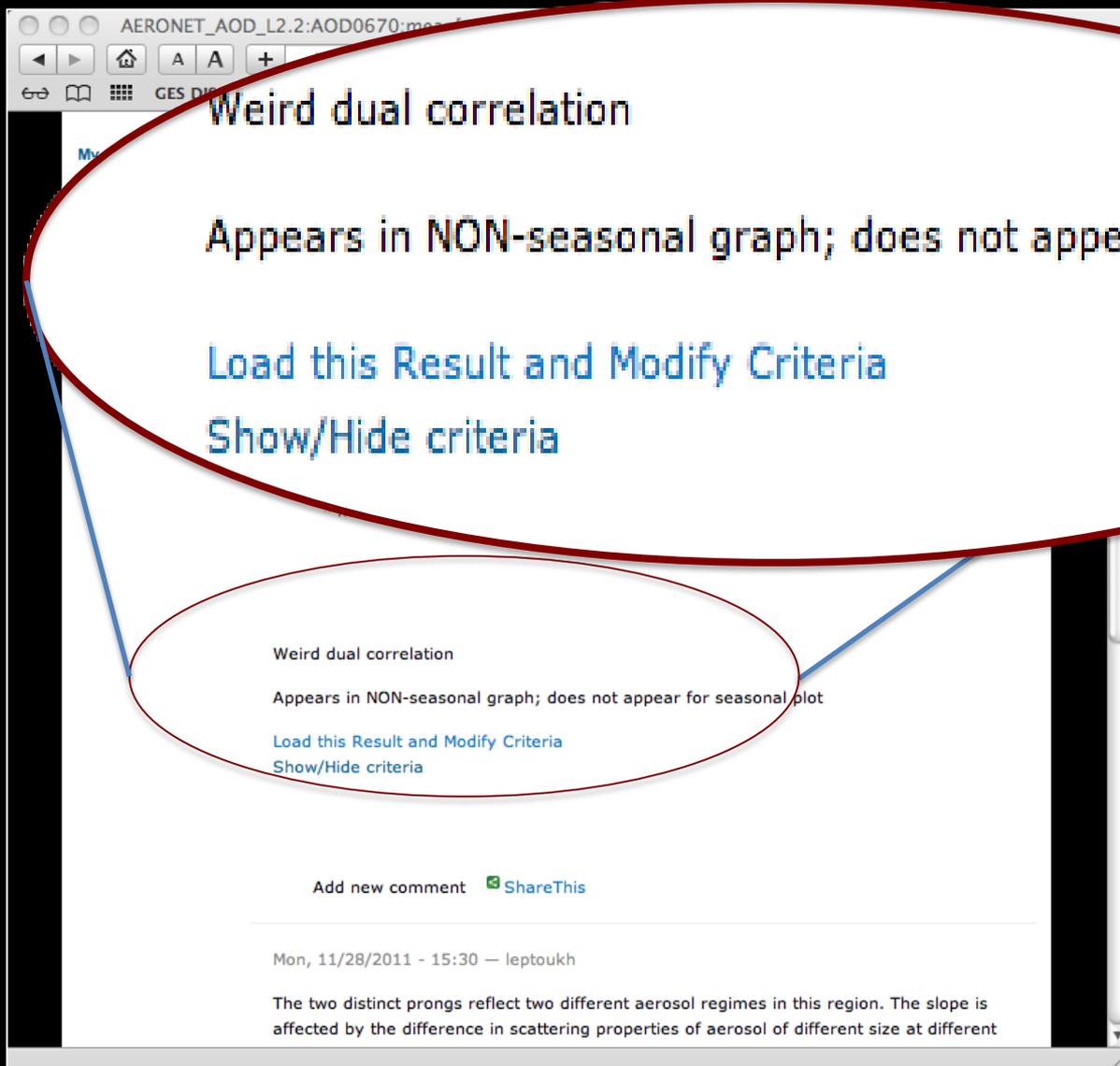
- Results
 - Annotation of results graphics
 - Discussion of results
 - Reproducing results
- Designed for AeroStat, but...
- ...Can be integrated with other REST based services

*Implemented by Daniel DaSilva, NASA/GSFC

Annotating the result's graphic



Discussing the Result



AERONET_AOD_L2.2:AOD0670:mon...

GES DI...

Weird dual correlation

Appears in NON-seasonal graph; does not appear for seasonal plot

[Load this Result and Modify Criteria](#)

[Show/Hide criteria](#)

Weird dual correlation

Appears in NON-seasonal graph; does not appear for seasonal plot

[Load this Result and Modify Criteria](#)

[Show/Hide criteria](#)

Add new comment [ShareThis](#)

Mon, 11/28/2011 - 15:30 — leptoukh

The two distinct prongs reflect two different aerosol regimes in this region. The slope is affected by the difference in scattering properties of aerosol of different size at different

Discussing the Result

AERONET_AOD_L2.2:AOD0670:mean[nval%20%3E (200...01 to 2010-02-01T23:59:59Z) | Social Giovanni

https://gsocial.ecs.nasa.gov/?q=node/176

GES DISC Yammer Yammer : GESDISC GESDISC Feed NASA News Su...ry (1,322) EOSDIS NASA

01 Jan 2001 to 01 Feb 2010

My groups

Not a member of any groups.

Mon, 11/28/2011 - 15:30 — leptoukh

The two distinct prongs reflect two different aerosol regimes in this region. The slope is affected by the difference in scattering properties of aerosol of different size at different

Appears in NON-seasonal graph; does not appear in seasonal graph

[Load this Result and Modify Criteria](#)

[Show/Hide criteria](#)

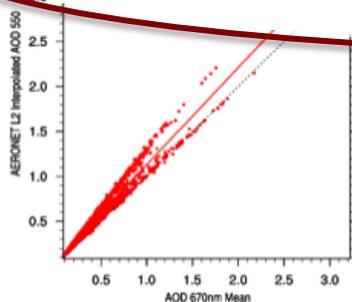
[Add new comment](#) [ShareThis](#)

Mon, 11/28/2011 - 15:30 — leptoukh

The two distinct prongs reflect two different aerosol regimes in this region. The slope is affected by the difference in scattering properties of aerosol of different size at different

Reproducing the Result

Load this Result and Modify Criteria



- AERONET L2 Interpolated AOD 550 Mean (nval >= 2)
- $y=1.078x+0.054, \text{RMS}=0.071, \text{R}^2=0.948, \text{N}=2209$
- 1:1

Weird dual correlation

Appears in NON-seasonal graph; does not appear for seasonal plot

[Load this Result and Modify Criteria](#)
[Show/Hide criteria](#)

Add new comment [ShareThis](#)

Mon, 11/28/2011 - 15:30 — leptoukh

The two distinct prongs reflect two different aerosol regimes in this region. The slope is affected by the difference in scattering properties of aerosol of different size at different



Unfinished Business...

- Underway
 - Data Quality
 - Giovanni-4: Faster, better, cheaper
- Still nascent...
 - myGiovanni



Data Quality

- Characterizing Quality
 - Variations in bias
 - Quality needs for different user types
 - Additional quality dimensions
- Collaborations on Quality
 - QA4EO
 - ESIP Information Quality Cluster
 - Now chaired by Brent Maddux (U. Wisconsin)



Characterizing Quality with Quality Labels

Quality Facts Report

localhost:8080/mdsa/quality/report/01ffc5bfba33e7139ffbd4b7185f9

Quality Facts

Product: MODIS Terra Daily Collection 5.1, Aerosol Optical Depth at 550 nm

Accuracy (vs Aeronet)	
45% within expected error (Africa above Equator) ¹	bad compliance
66% within expected error (East Asia mid-latitudes) ¹	good compliance
76% within expected error (Europe - Mediterranean) ¹	very good compliance
slope of linear regression fit (MODIS vs Aeronet) = 0.93 (Global) ¹	low bias underestimation
67% within expected error (Global, land-only) ¹	good compliance
64% within expected error (Global, ocean-only) ²	good compliance
81% within expected error (Indian Subcontinent) ¹	very good compliance
52% within expected error (Peninsular Southeast Asia) ¹	marginal compliance

Measurement Characteristics

Platform: [Terra](#)

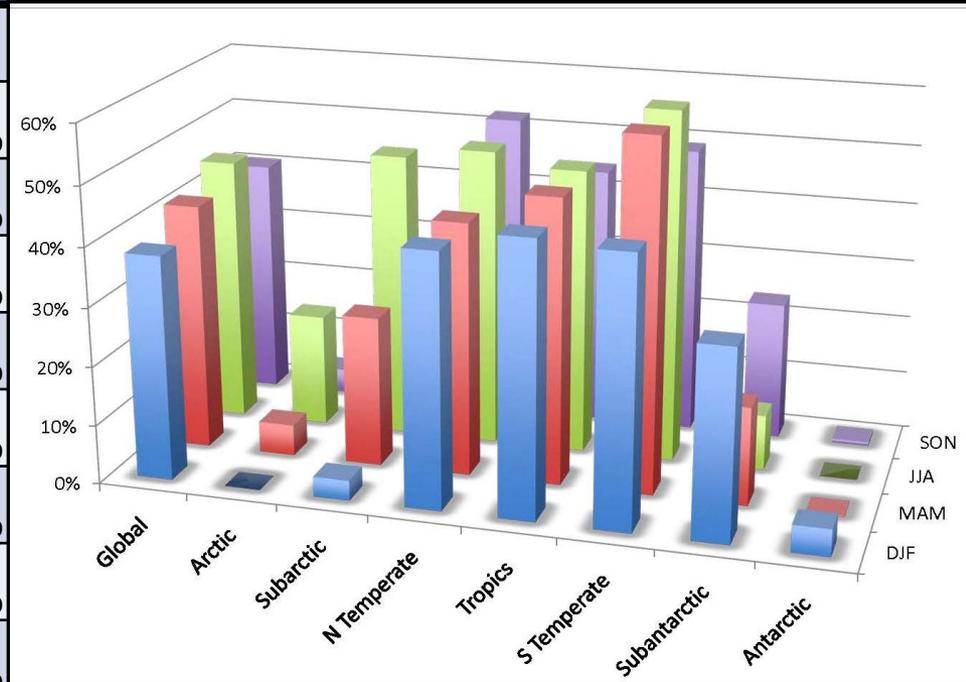
Generated for a request for 20-90 deg N, 0-180 deg E



Spatial Completeness

MODIS Aqua AOD Average Daily Spatial Coverage by Region + Season

	MODIS Aqua AOD			
	DJF	MAM	JJA	SON
Global	38%	42%	45%	41%
Arctic	0%	5%	19%	4%
Subarctic	3%	26%	49%	25%
N Temperate	43%	43%	51%	52%
Tropics	46%	48%	49%	44%
S Temperate	45%	59%	60%	49%
Subantarctic	32%	17%	10%	24%
Antarctic	5%	0%	0%	1%



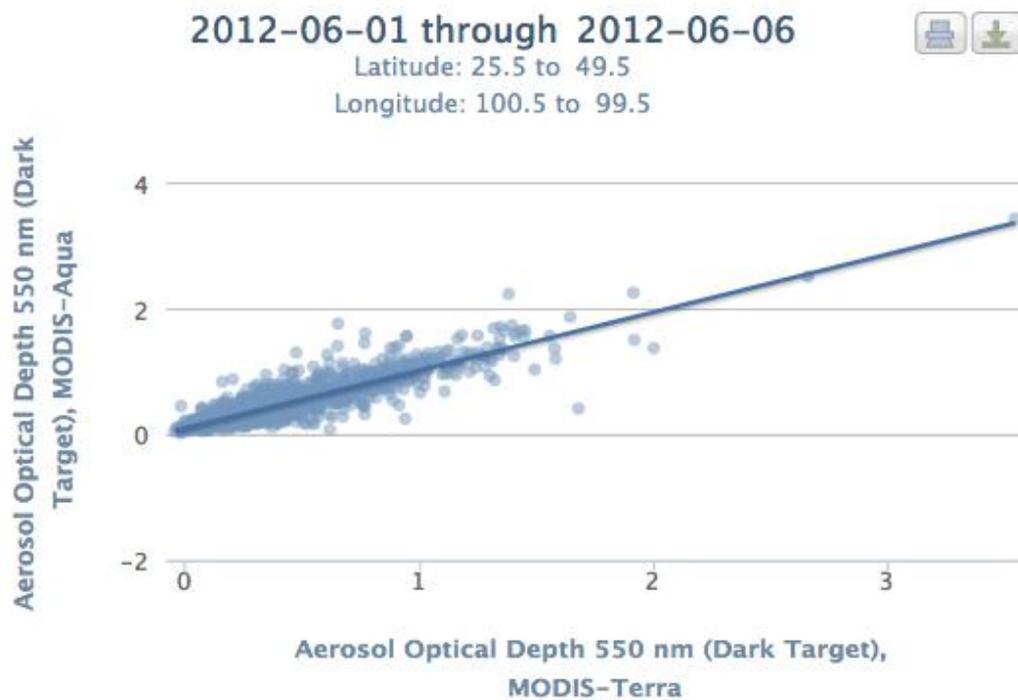
This table and chart is Quality Evidence for the Spatial Completeness (Quality Property) of MODIS Aqua Dataset



Giovanni-4: Next Generation

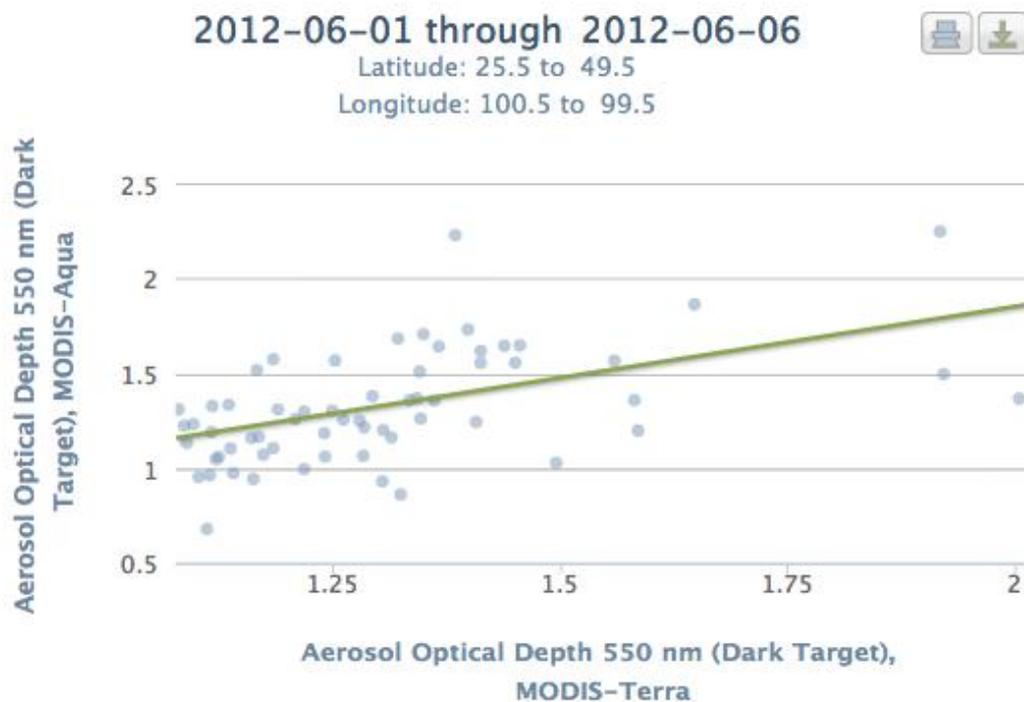
- Emphasis on data exploration
- Collaboration with end users...

Exploring Data: Interactive Scatterplot



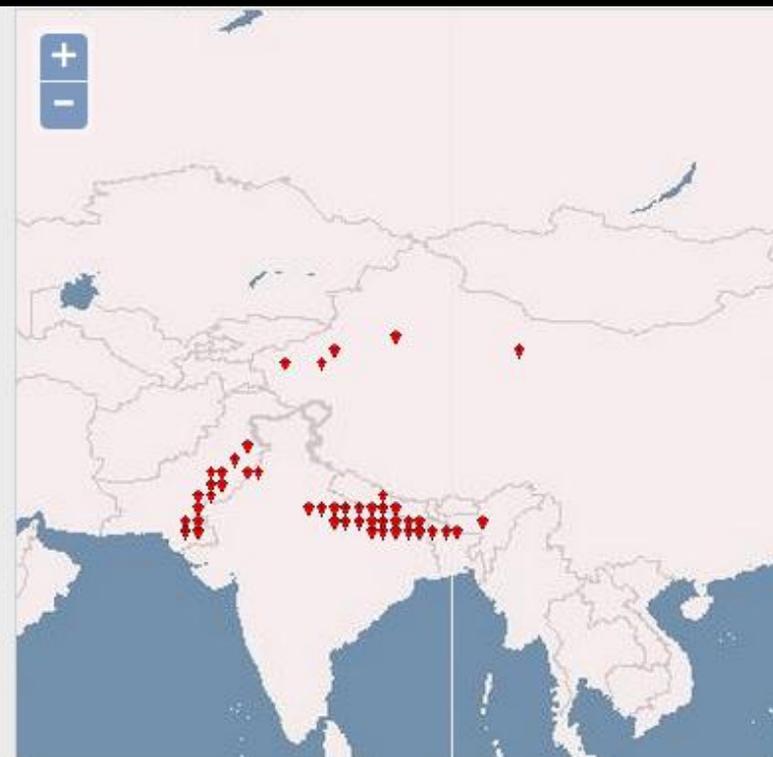
Reset Map and Chart

Zoom in on scatterplot



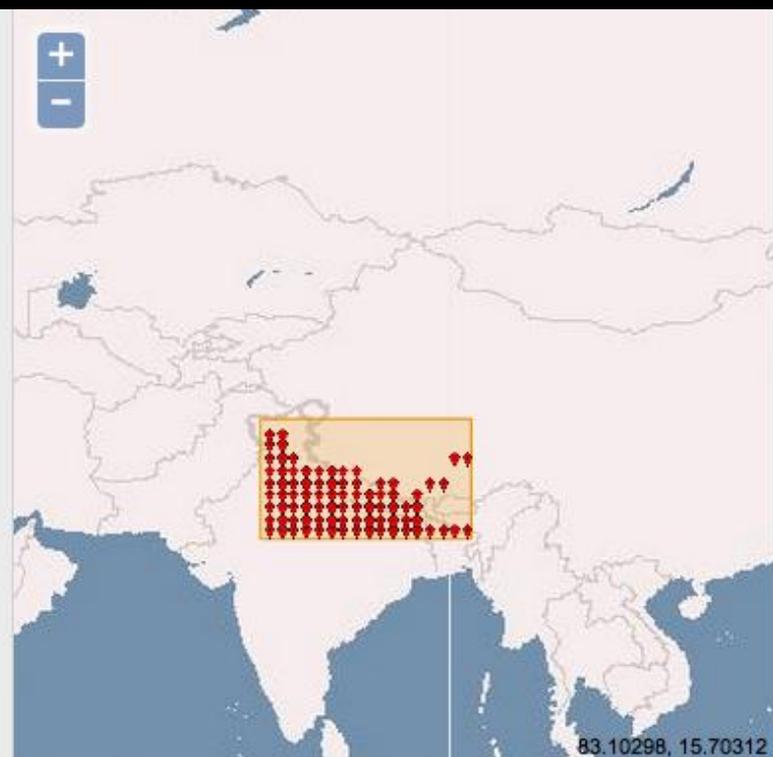
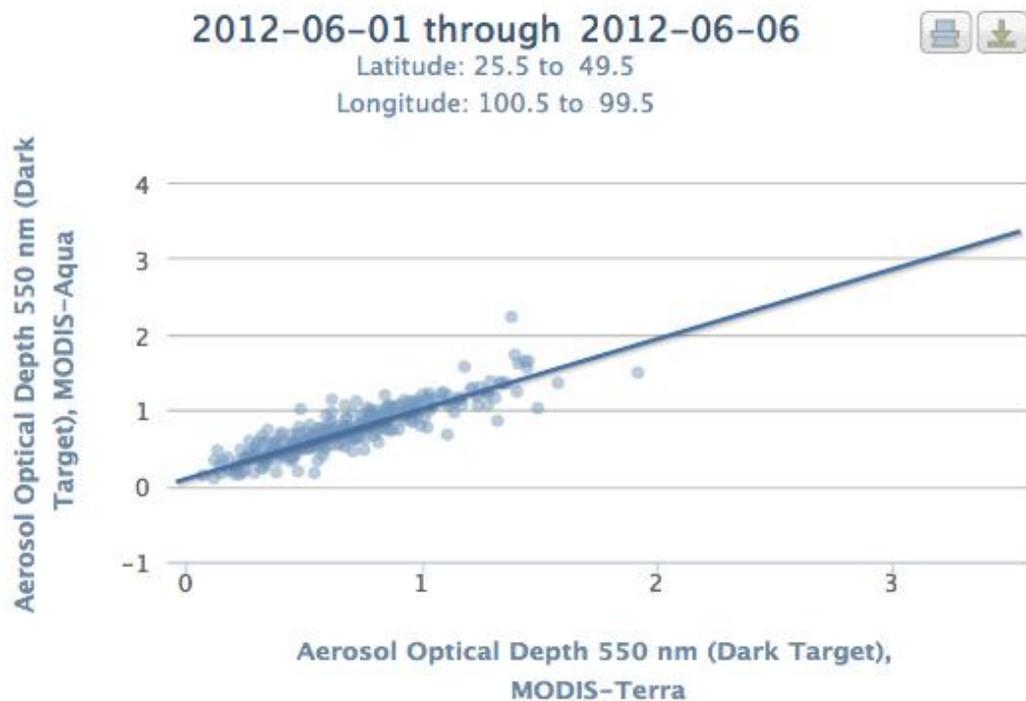
Series 1
Regression Line Equation: $y = 0.7507112587684994x + 0.3484881241472614$

Highcharts.com



Reset Map and Chart

Examine a subregion



Reset Map and Chart

Current G3 Performance...



<http://gdata1.sci.gsfc.nasa.gov>

The date range you entered covers more than one year, but at least one parameter in your query has a small time interval (i.e., it is repeated daily, hourly, minutely, etc.). This could result in a great many data being returned in the query (and could take more than a few minutes). Are you sure you want to submit this large query?

← for > 1 year of
MODIS Daily
AOD



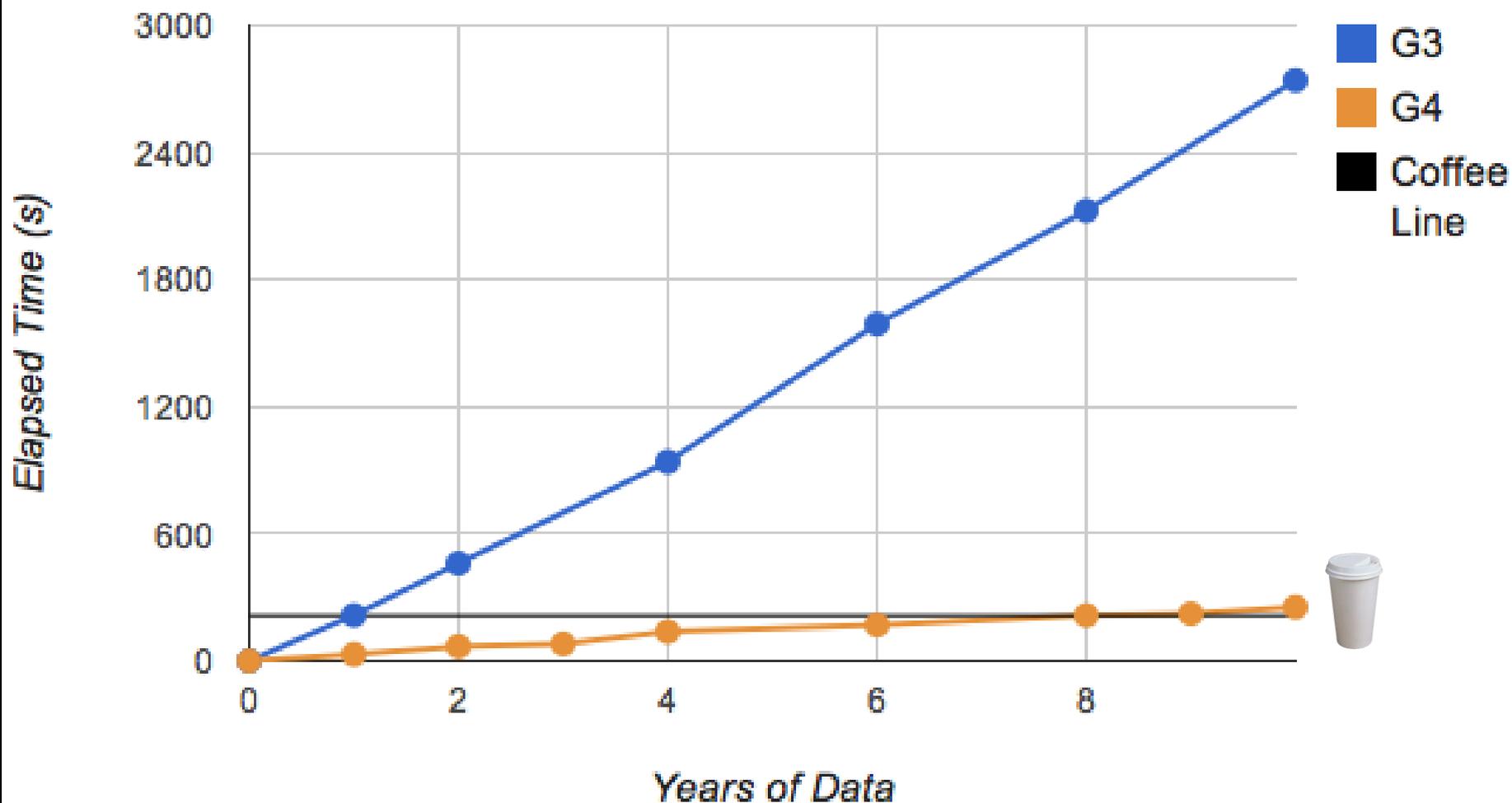
<http://gdata1.sci.gsfc.nasa.gov>

Ok. Go grab some coffee. ←

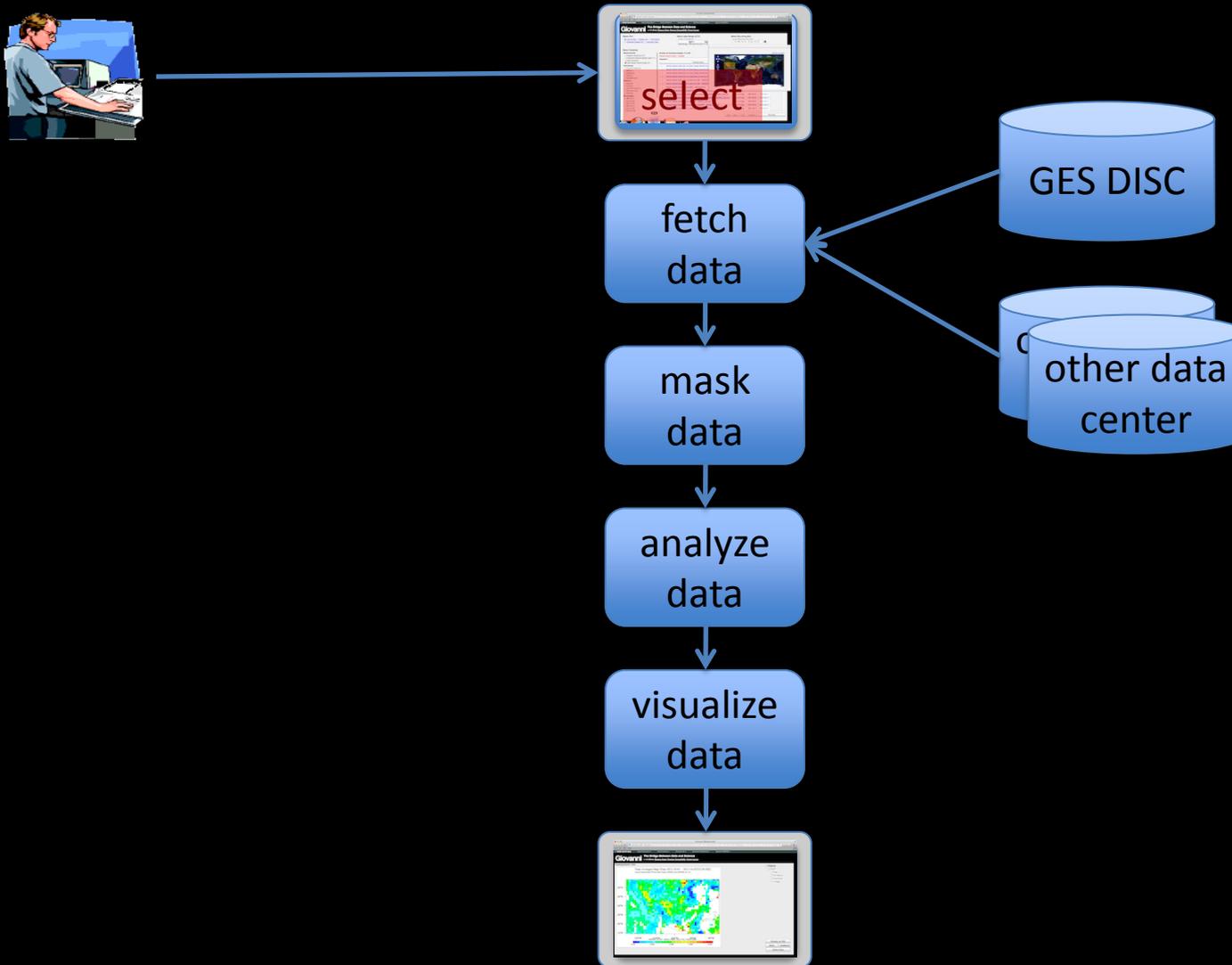
OK

Initial G3-G4 Performance Tests

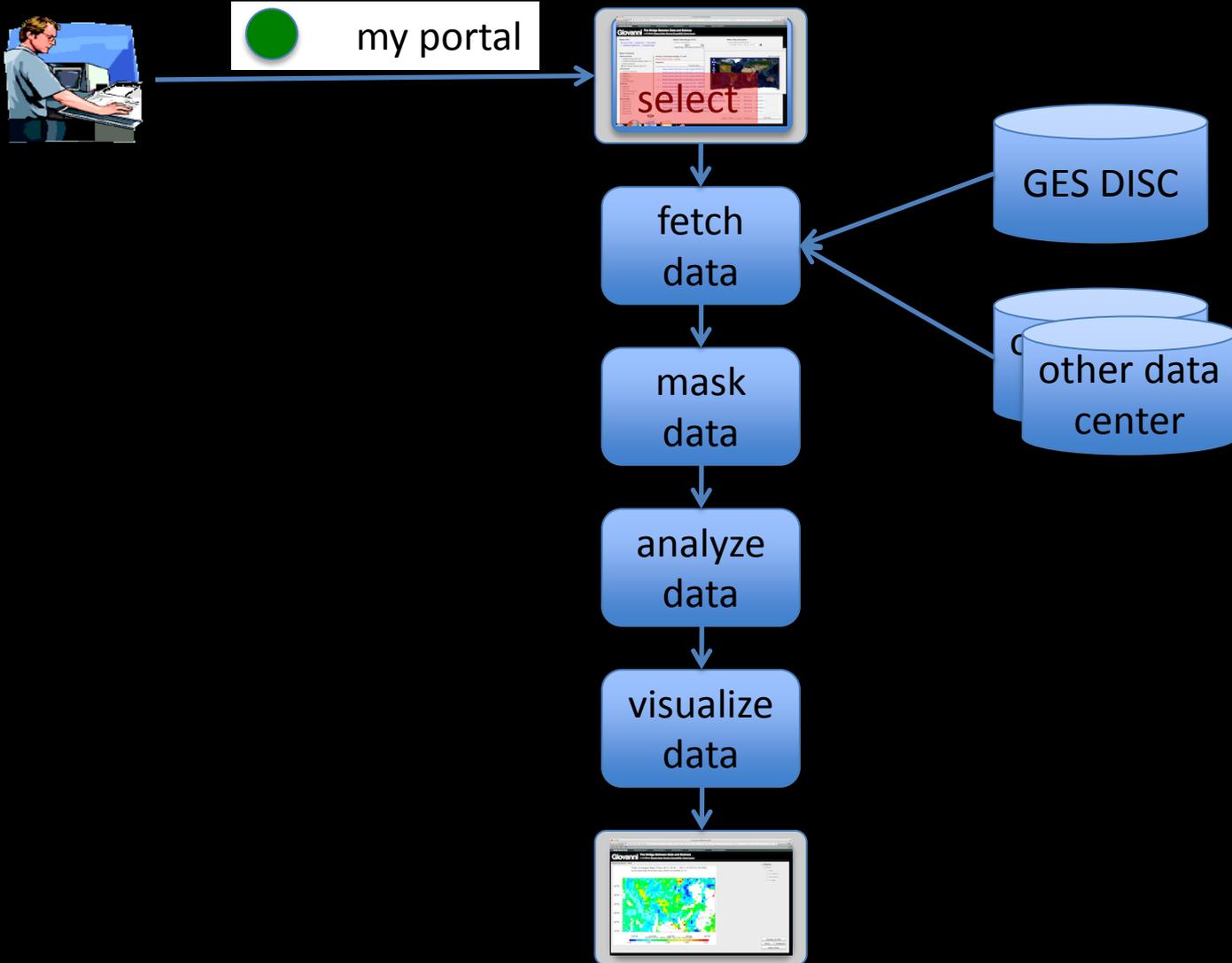
G3-G4 Comparison: Area-averaged Time Series



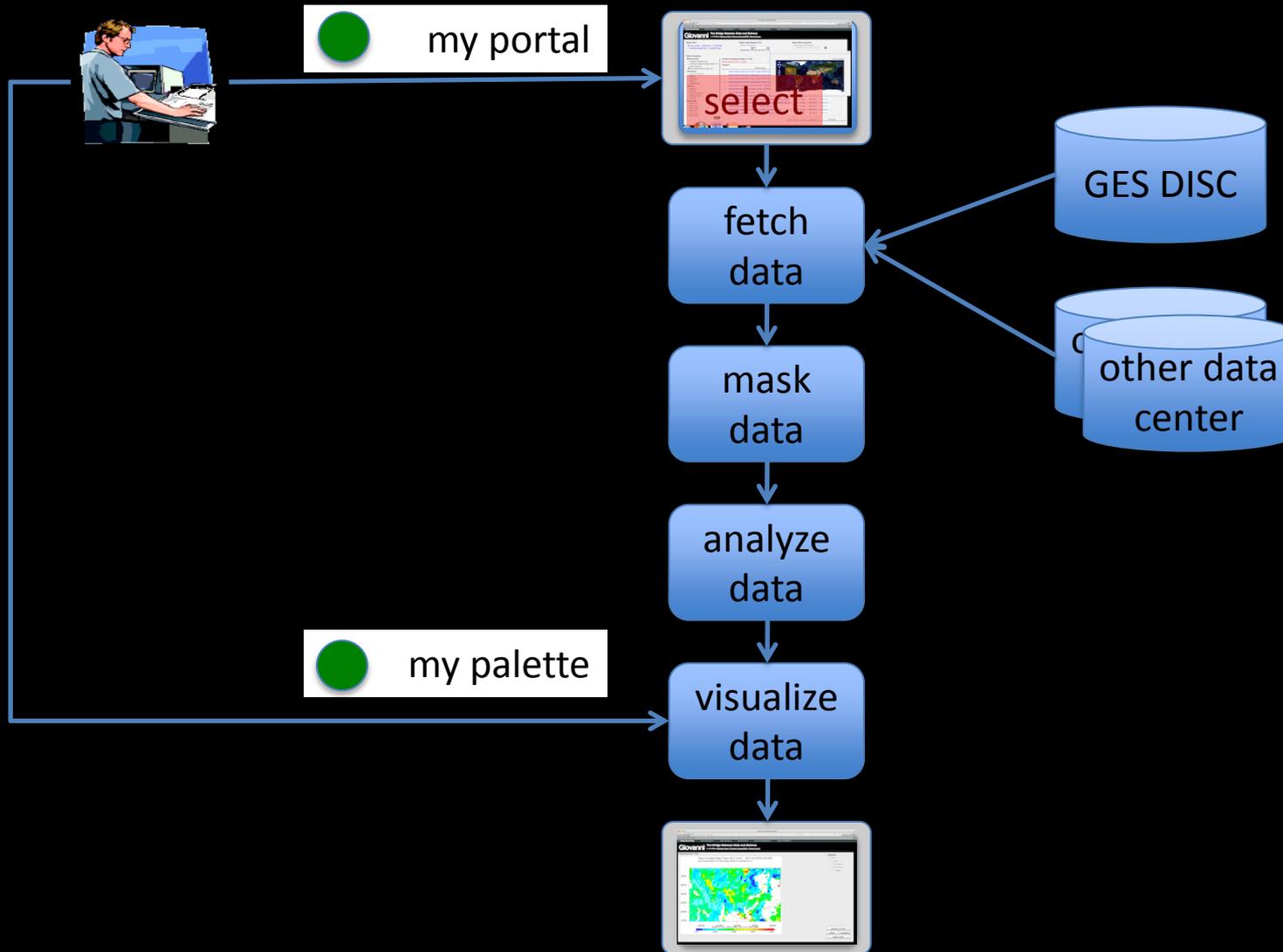
Participatory Design: myGiovanni



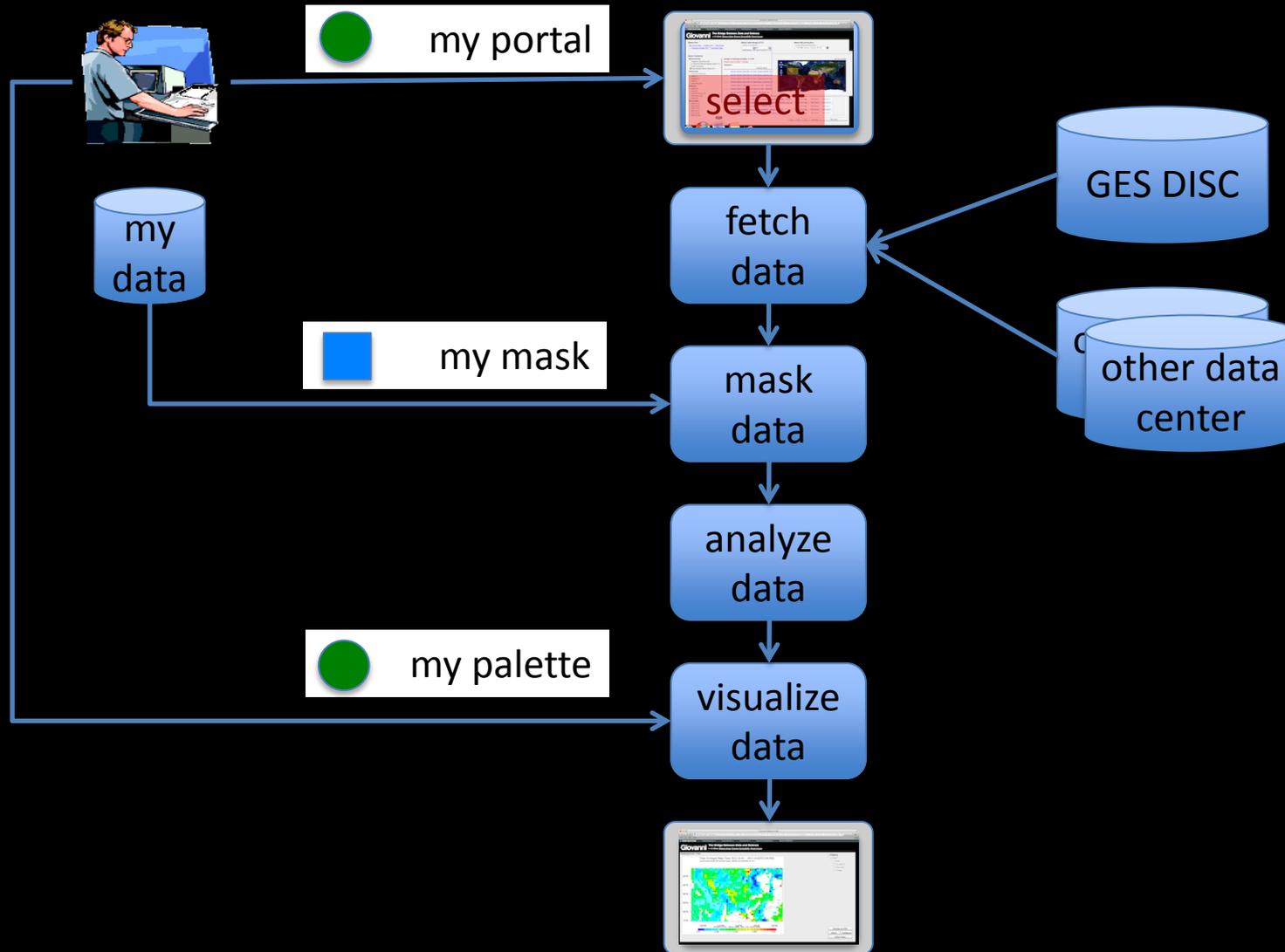
Participatory Design: myGiovanni



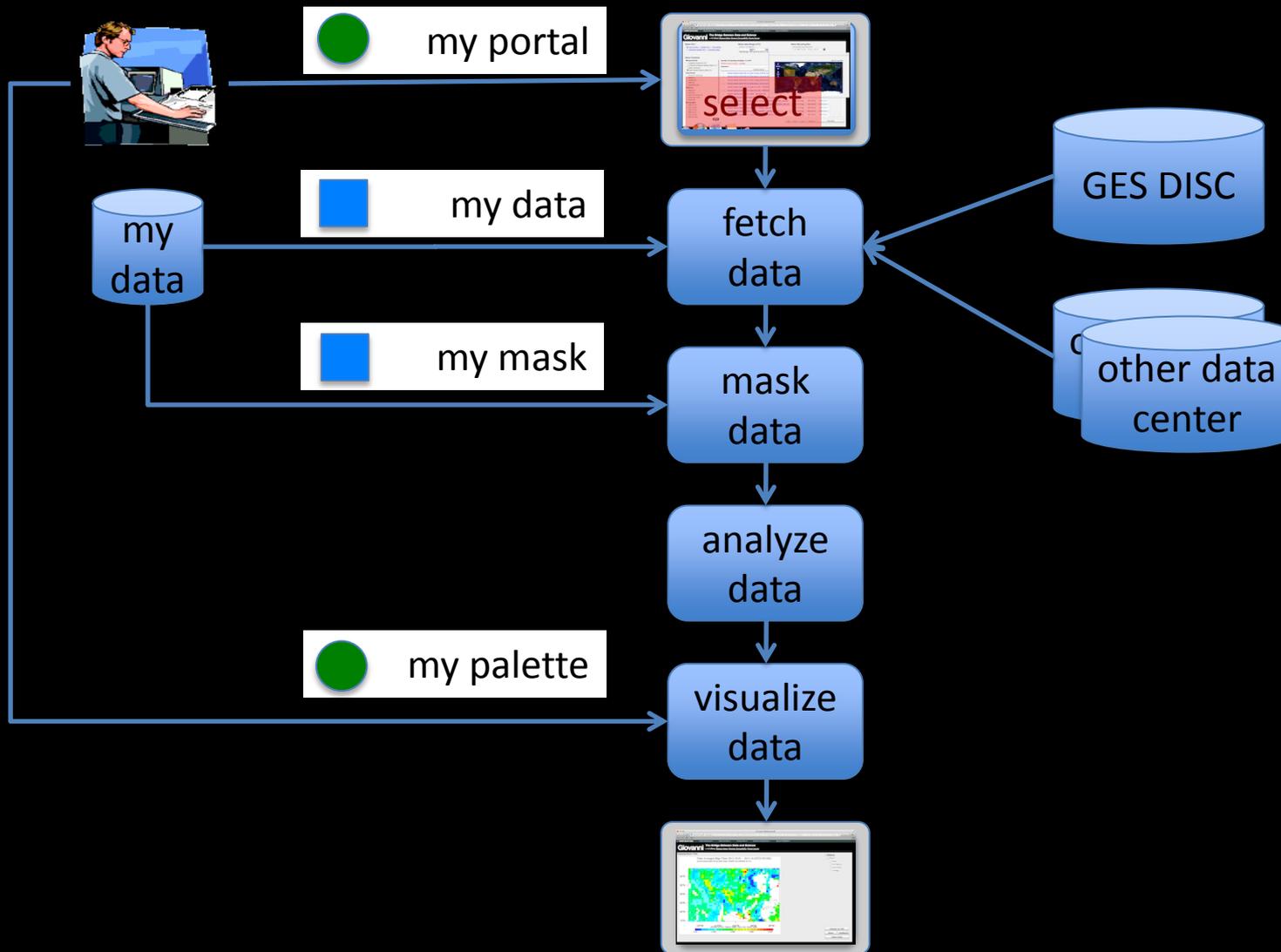
Participatory Design: myGiovanni



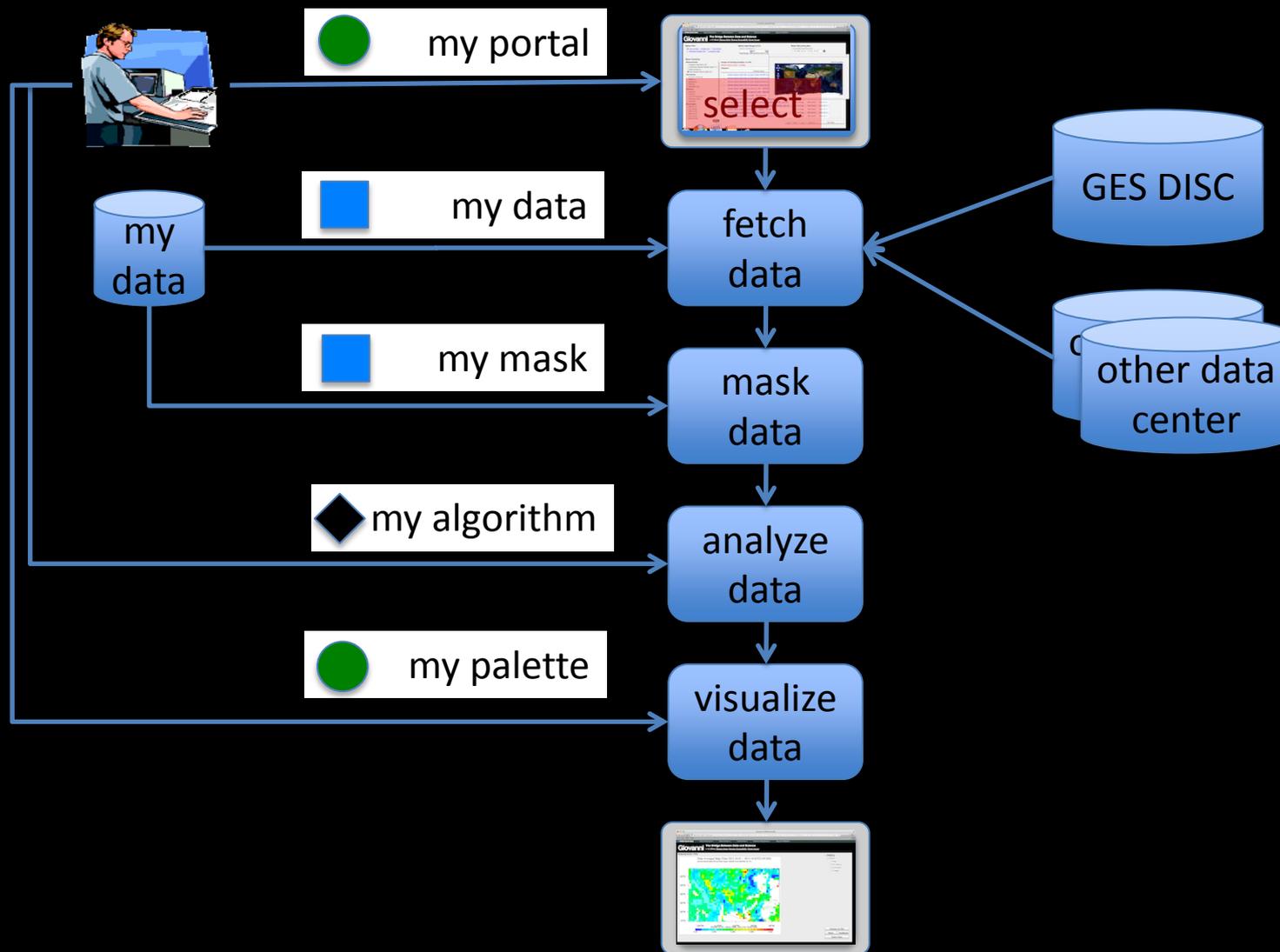
Participatory Design: myGiovanni



Participatory Design: myGiovanni



Participatory Design: myGiovanni





Charge to the Informatics Community

- Carry on:
 - Multi-sensor Data Analysis
 - Data Quality
 - Participatory design

Charge to the Informatics Community

- Carry on:
 - Multi-sensor Data Analysis
 - Data Quality
 - Participatory design
- Build Bridges
 - Data and Science
 - Informatics and Earth Science



Thank you

