Giovanni: Earth Science Data and Analysis Tool

Gregory Leptoukh

NASA Goddard Space Flight Center
Goddard Earth Sciences
Data and Information Services Center (GES DISC)

Goddard Interactive Online Visualization ANd aNalysis Infrastructure (Giovanni)

- With Giovanni and a few mouse clicks, easily obtain information on the atmosphere around the world.
- No need to learn data formats to retrieve and process data.
- Try various combinations of parameters measured by different instruments.
- All the statistical analysis is done via a regular web browser.

http://giovanni.gsfc.nasa.gov/

Caution: *Giovanni is a rapidly evolving data exploration tool!*

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:

Giovanni

Mirador

- Read Data
- Extract Parameter
- Subset Spatially
- Filter Quality
- Reformat
- Reproject
- Visualize
- Explore
- Analyze

The Giovanni Way:

- Minutes
- Days for exploration

DO SCIENCE

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

GES DISC tools allow scientists to ***compress*** the time needed for pre-science preliminary tasks: *data discovery, access, manipulation, visualization, and basic statistical analysis*.

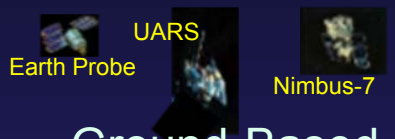
Scientists have ***more time to do science!***

Leptoukh

Comprehensive Data Systems

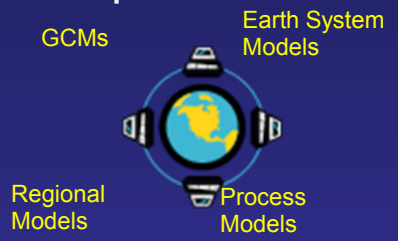
Current Satellites (NASA, NOAA, ESA, JAXA)

Legacy Satellites

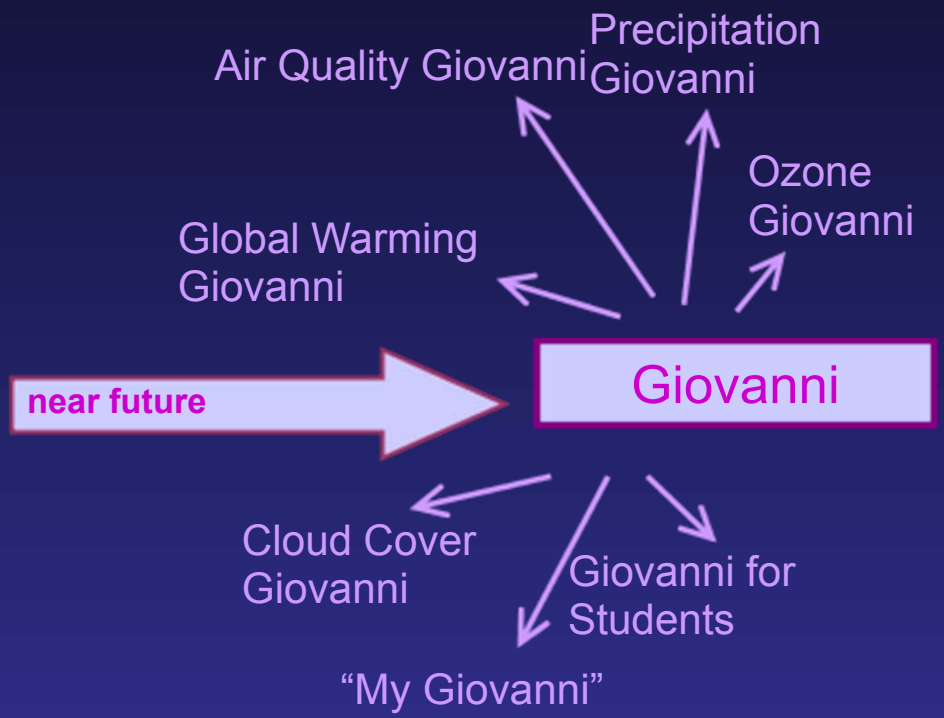


Ground-Based Observations

Computer Models



Comprehensive Data Systems provide an environment for working with *all* sources of relevant data (satellite, ground-based and models) across the full range of temporal and spatial scales.

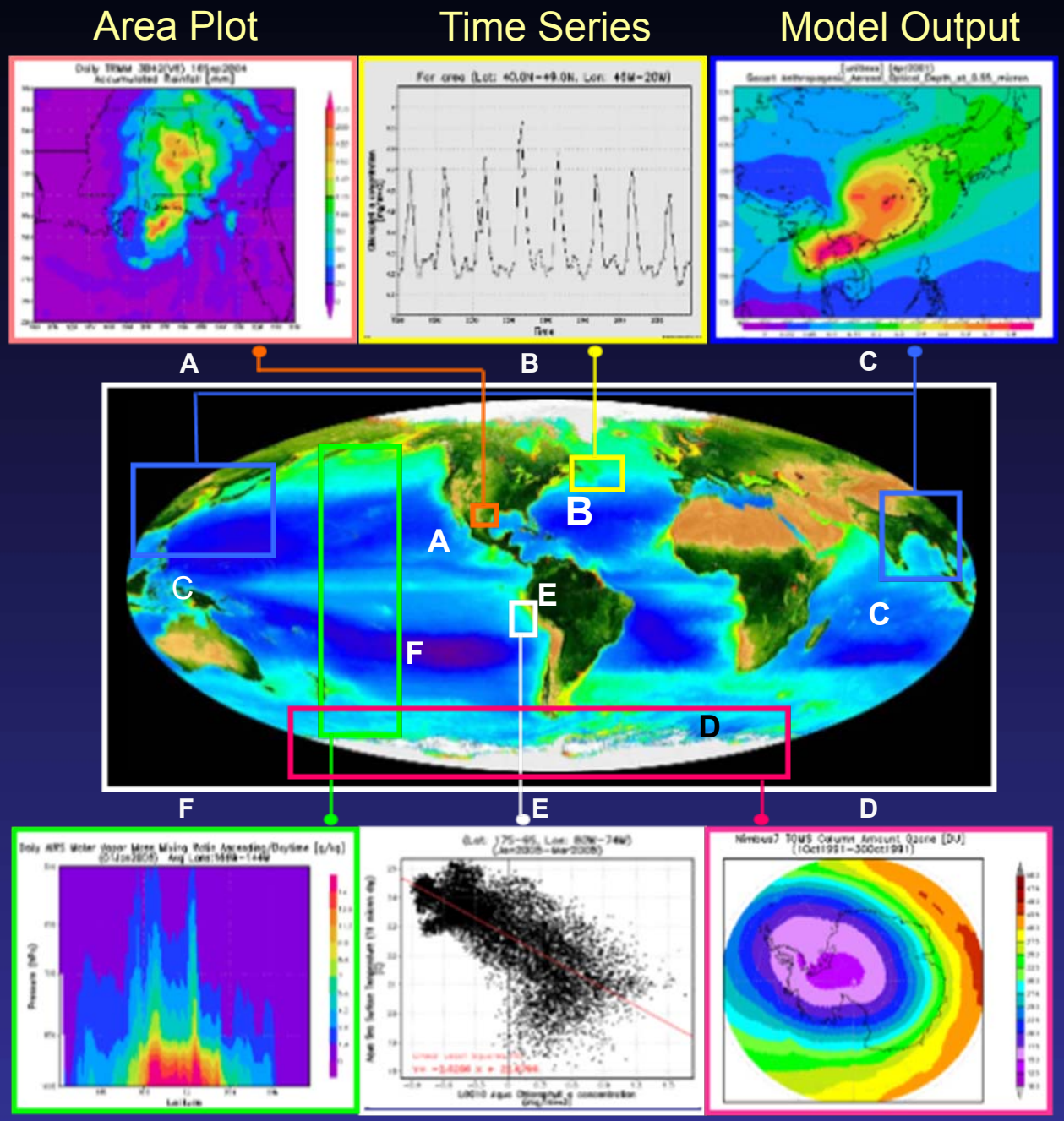


Giovanni now

- Almost 30 customized Giovanni portals
- Thousands of geophysical parameters
- Data from:
  - ~ 20 space-based instruments
  - ~ 50 models
  - EPA and Aeronet stations
- Multiple visualization and statistical analysis functionalities including data intercomparison
- Data lineage
- Subsetted data downloads in multiple formats

Science Analysis

Giovanni provides a suite of statistical analysis and visualization tools for the comparison of regional and global datasets.



Multi-Sensor Data Systems

Several Giovanni instances represent our effort to move into *comprehensive data systems*:

- A-Train Data Depot
- Aerosol Giovanni
- Air Pollution Giovanni
- TOVAS (TRMM, GPCP)
- NEESPI
- Aerosol Data Fusion
- Hurricane Portal (prototype)

Giovanni A-Train Data Depot

http://gdata1.gsfc.nasa.gov/daac-bin/G3/gui.cgi?instance_id=atrain

CloudSat-collocated previews of data from 8 instruments and **ECMWF**:

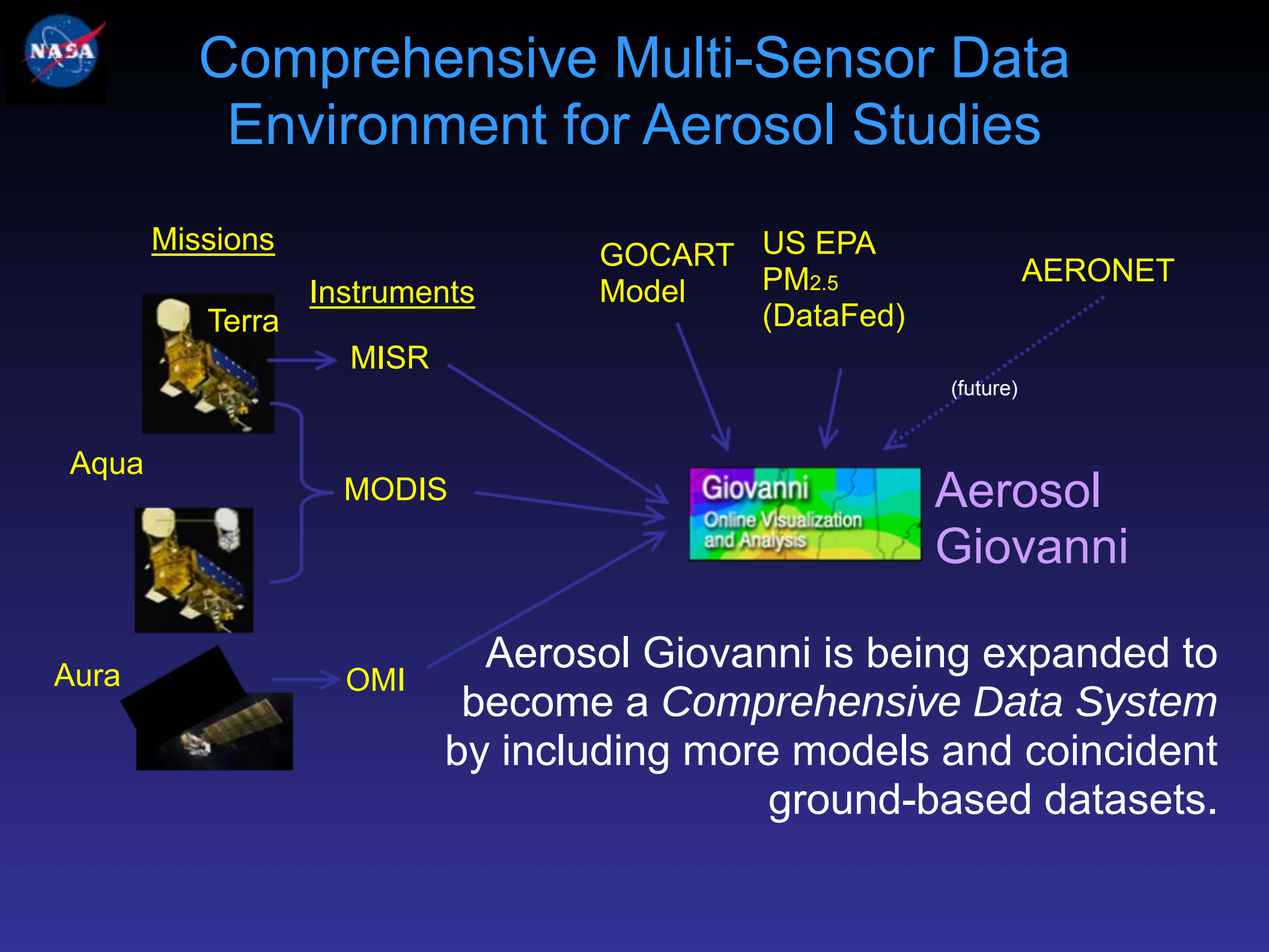
- MODIS/Aqua*
- AIRS
- AMSR-E*
- CloudSat
- CALIPSO
- POLDER/PARASOL*
- MLS
- OMI*

*On-line archive of pre-processed collocated subsets available:

- http://mirador.gsfc.nasa.gov/cgi-bin/mirador/collectionlist.pl?&keyword=atrain
- ftp://atrain.gsfc.nasa.gov/data/s4pa/

Aerosols

Comprehensive Multi-Sensor Data Environment for Aerosol Studies



Aerosol Giovanni is being expanded to become a *Comprehensive Data System* by including more models and coincident ground-based datasets.

Aerosol data in Giovanni

Giovanni is already a powerful tool for the analysis of satellite aerosol datasets

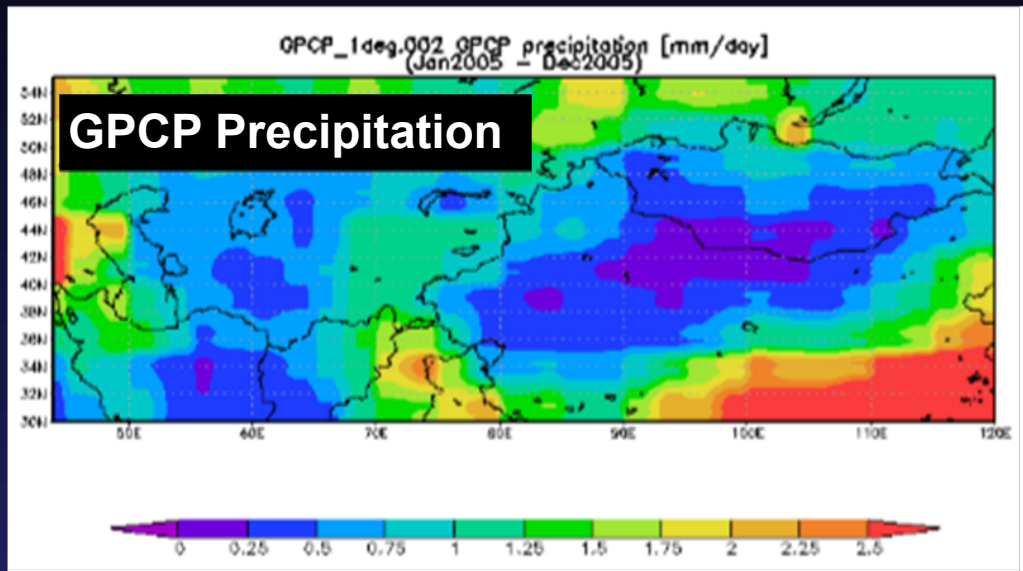
It evolves to include additional ground-based, validation campaign and model aerosol data

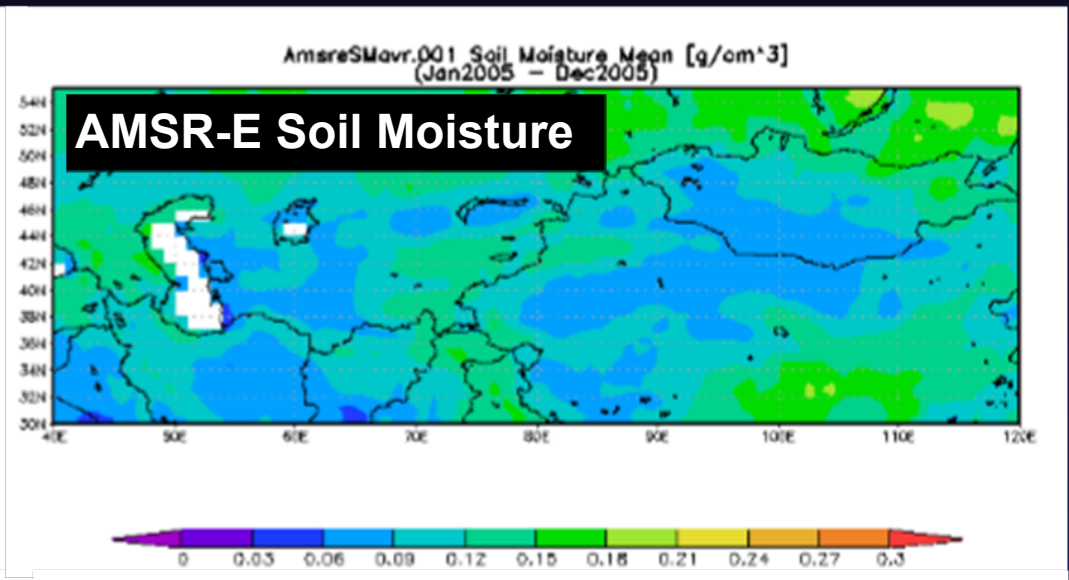
- already included in Giovanni 3.08
- in preparation, prototype or testing
- to be included

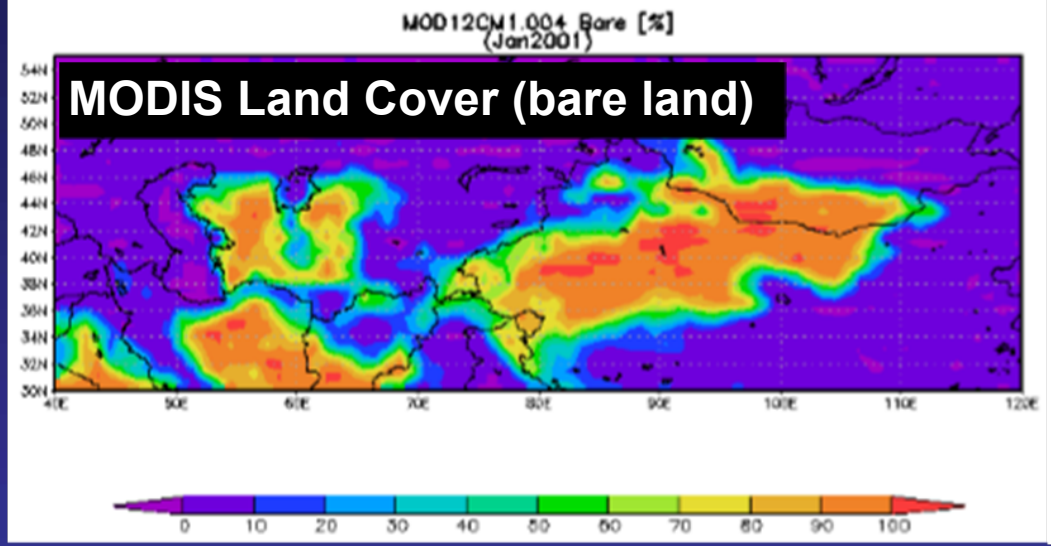
Aerosol Dataset		Status		
MODIS	Terra	■		
MODIS	Aqua	■		
OMI	Aura	■		
TOMS	N7/EP	■		
CALIOP	Calipso	■		
SeaWIFS	Orbview2	■		
MISR	Terra	■		
POLDER	Parasol	■	■	
MERIS	Envisat	■	■	
AVHRR	NOAA			■
APS	Glory			■
VIIRS	NPP			■
HIRDLS	Aura	■		
MODIS-Deep Blue		■		
MODIS-MAIAC				■
MODIS-NAAPS				■
AERONET			■	■
MAPSS			■	
GACP				■
EPA AirNOW $PM_{2.5}$		■	■	
Validation Campaigns				■
ISCCP				■
AEROCOM			■	■
GOCART		■		
HTAP			■	
GlobAEROSOL (ESA)				■

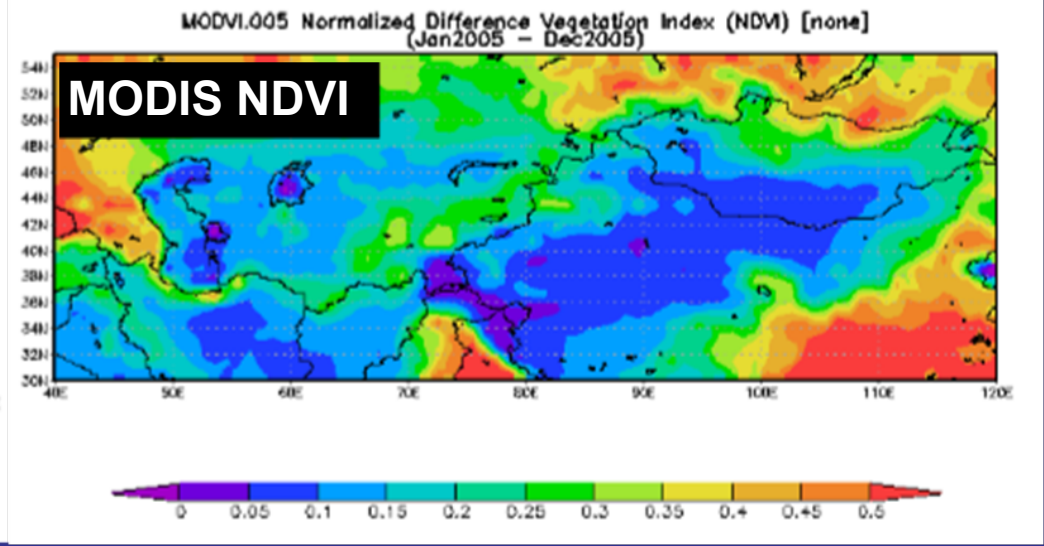
Interdisciplinary

Multi-sensor view of dry land in mid-Asia, northwestern China, and Mongolia

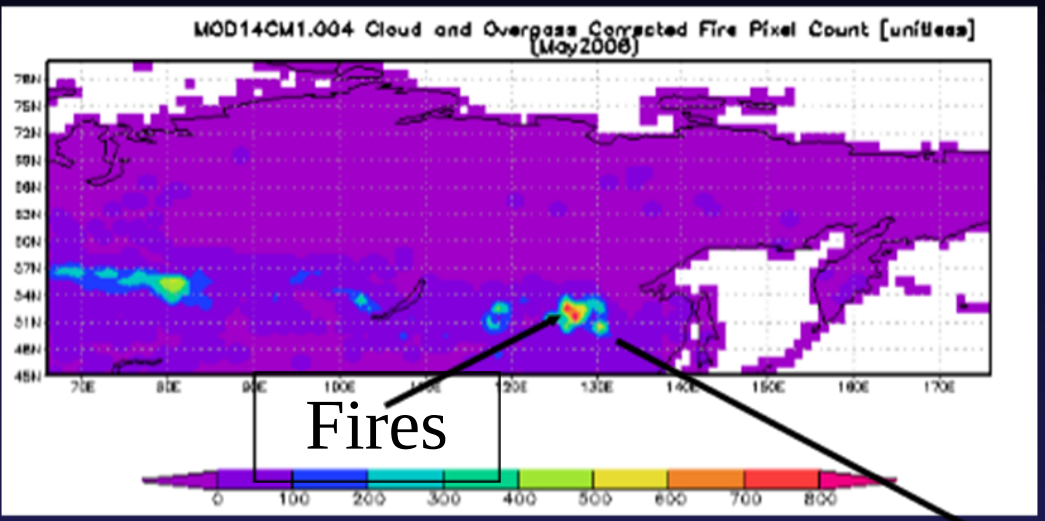




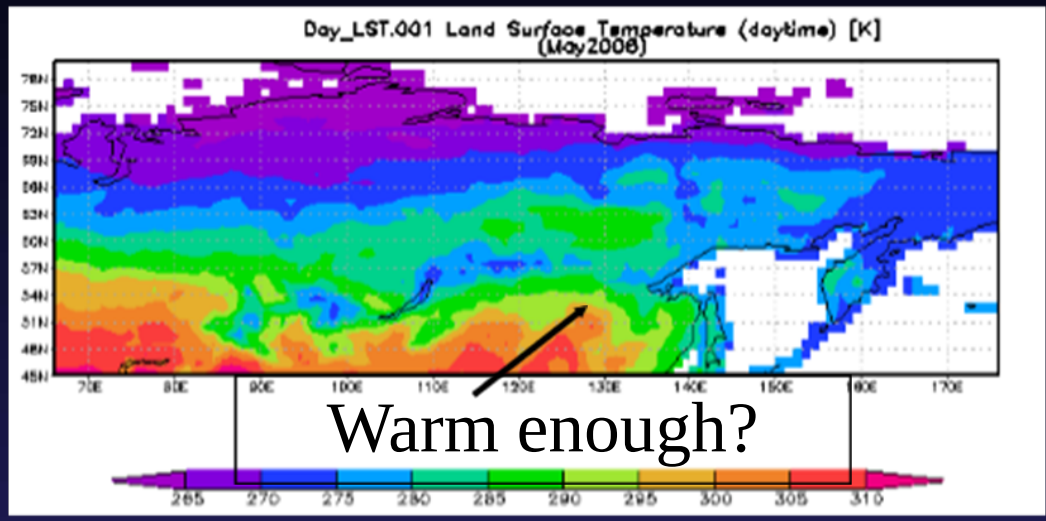




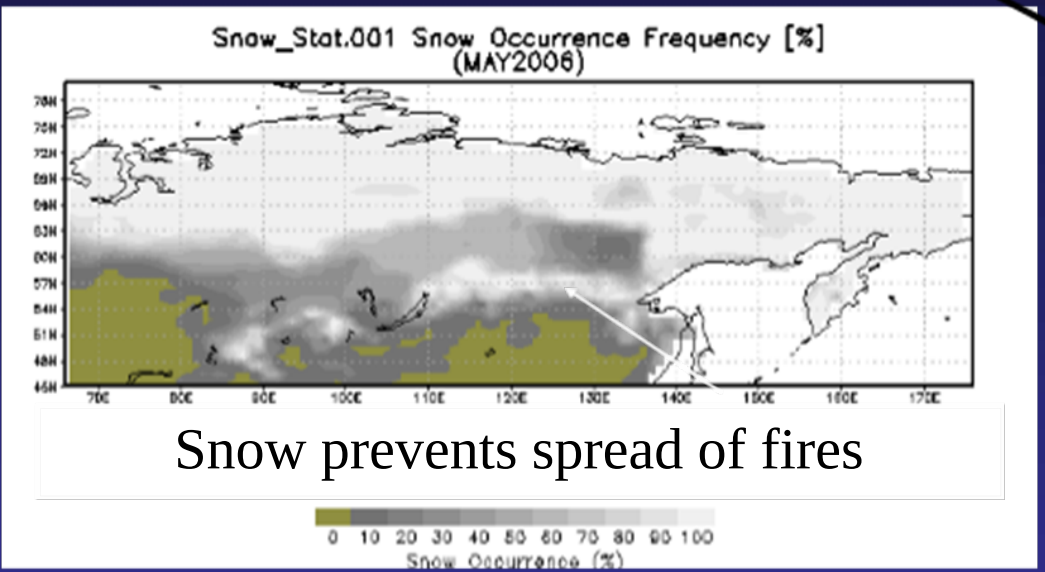
Fires, Temperature, Snow, AOT Maps



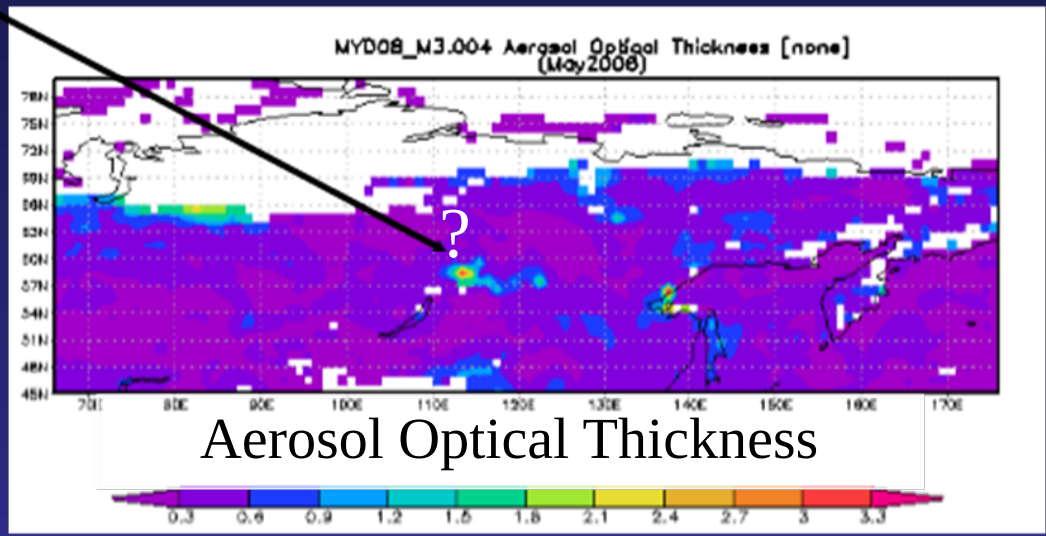
Fires



Warm enough?



Snow prevents spread of fires

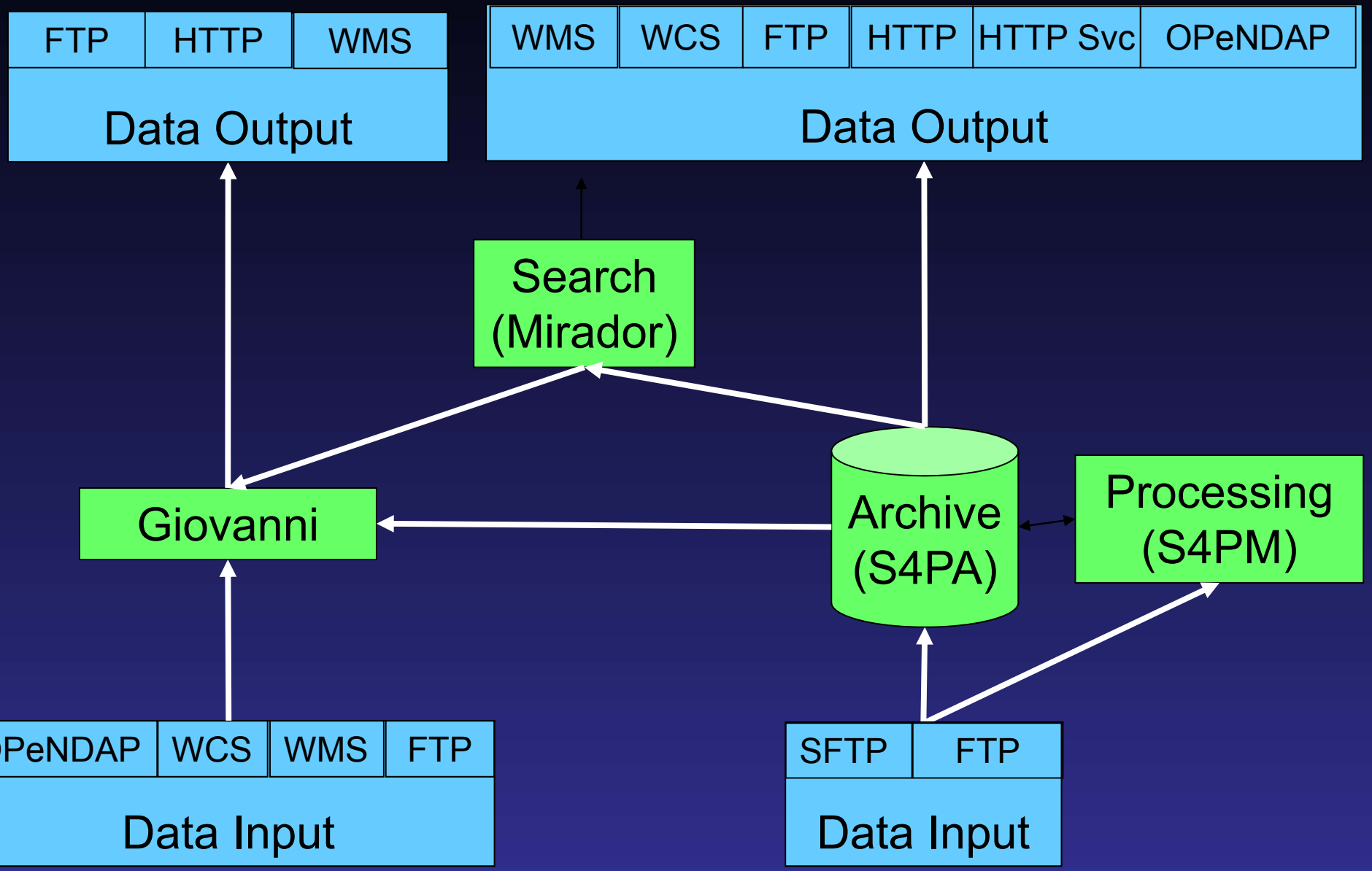


?

Aerosol Optical Thickness

Technicalities: servers, protocols, interoperability, etc.

GES DISC Interoperability Architecture



Processing with S4PM

- Simple, Scalable, Script-based Science Processor for Measurements
- Robust, fully automated (partially autonomous) processing system
- Runs science-team-provided science algorithms
- Open Source, reused within NASA and EOSDIS

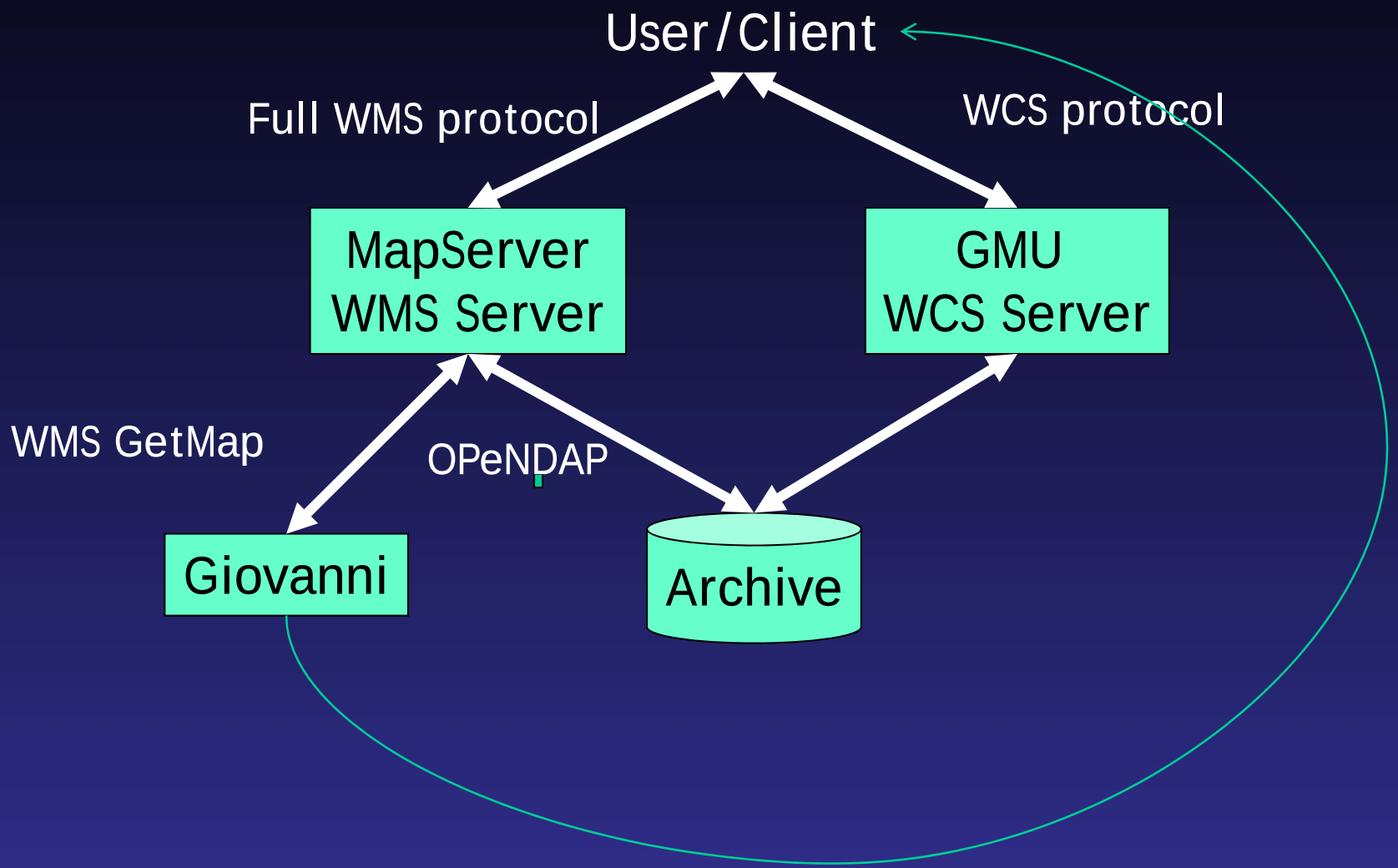
Archive in S4PA

- Simple, Scalable, Script-based Science Product Archive (S4PA)
- Disk-based archive on anonymous FTP
  - Data can be available through HTTP as well
    - Data can be restricted-access through HTTP
  - Data are distributed amongst multiple hosts
- Data organization
  - Assigned to host based on mission/instrument and processing level
  - Stored in directory tree based on dataset and Start date/time of data

S4PA Metadata

- File-level metadata are stored as companion XML files
- Dataset-level metadata are in the Global Change Master Directory (GCMD) Direct Interchange Format (DIF)
- Metadata are translated to EOS Clearinghouse schema and published
  - XSLT used for mapping

OGC Architecture



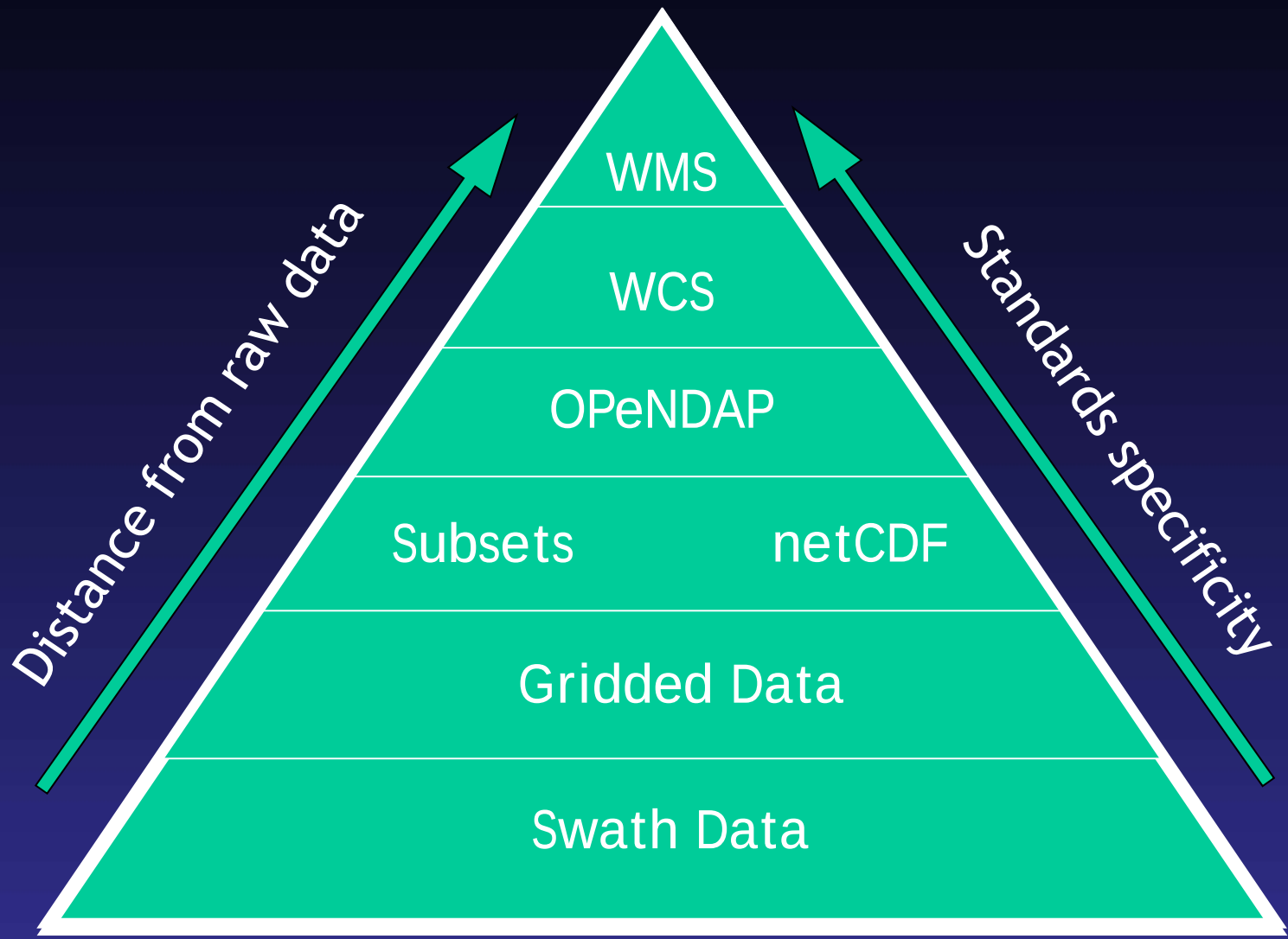
HTTP Services

- REST-RPC Hybrid service
  - URL-addressable (GET requests)
  - URLs generated in Mirador
- File returned from server as MIME attachment
- Services:
  - NetCDF Conversion for AIRS grids and swaths (save radiance products), TRMM grids
  - Variable subsetting for OMI L2G
  - Spatial subsetting for MERRA model output

OPeNDAP

- All gridded data available
- AIRS Swath data available
- OMI data support NetCDF conversion through OPeNDAP

Available Forms of Data



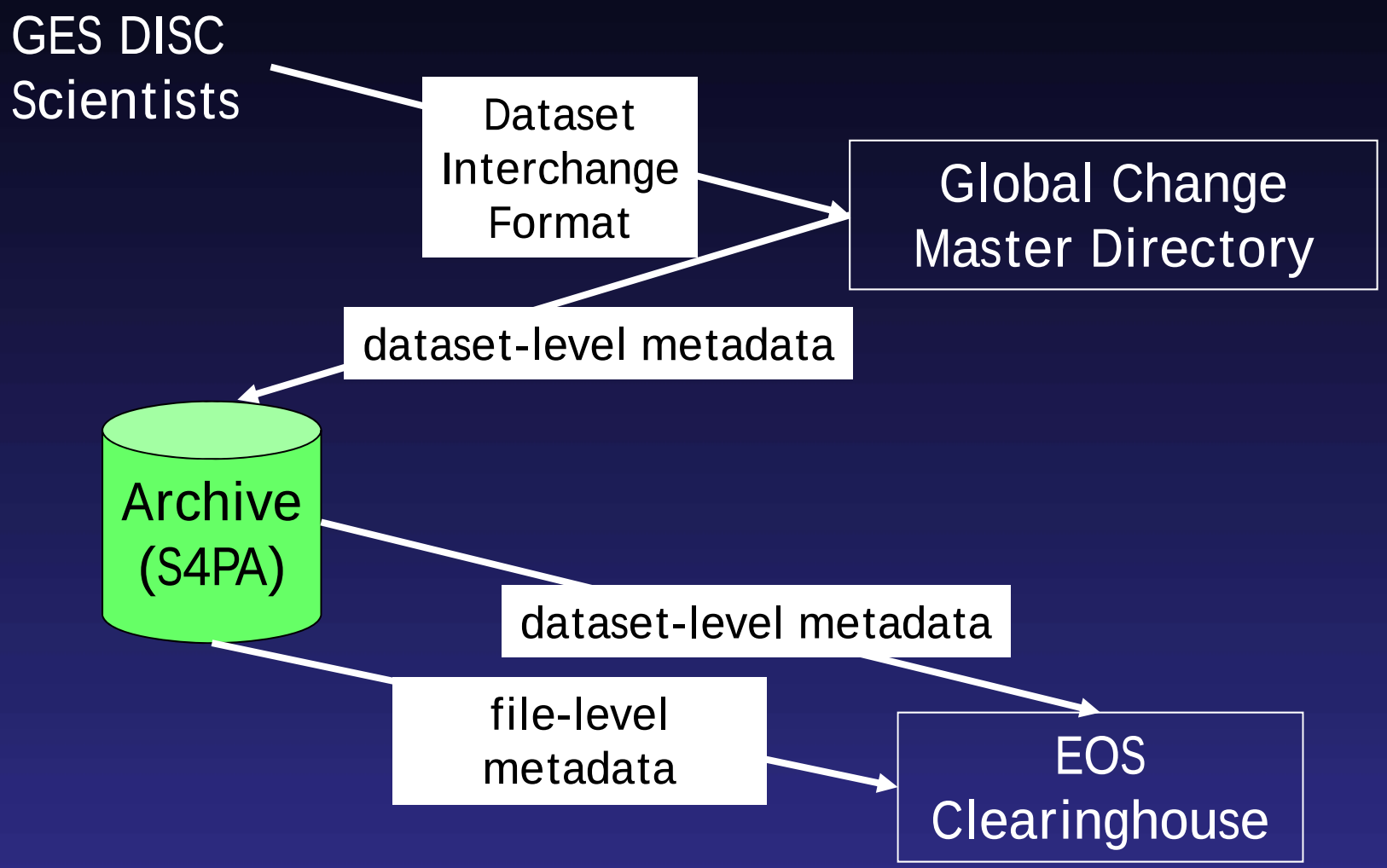
Mirador

- Interactive search tool
  - Keyword - space - time
  - Hierarchical navigation
  - Search by geophysical event
- Machine-accessible search tool
  - Follows OpenSearch convention with Date and Geo extensions: http://www.opensearch.org
  - Available as RSS or Atom response
  - Returns URLs to data files
    - Could be enhanced to return URLs to services
- Now includes ontology

Giovanni Data sources and their access protocols

Data sources	Protocol	Data
NASA GES DISC	Local access	AIRS, TRMM, OMI, MLS, HIRDLS
NASA MODIS DAAC	FTP	MODIS
NASA Ocean Color DAAC	FTP	SeaWiFS, MODIS
NASA Langley DAAC	OPeNDAP	CALIPSO, MISR, TES, CERES
NSIDC	FTP	AMSR-E
NOAA	FTP	Snow, Ice, NCEP
Univ. of Maryland	FTP	MODIS fire, NDVI
Colorado State Univ.	FTP	CloudSat
CIESIN Columbia University	FTP	Population
JPL	FTP	QuickSat
EPA via DataFed	WCS	PM2.5
Lille, France	FTP	Parasol
ESA	FTP	MERIS
Juelich, Germany	FTP → WCS	*HTAP*
Paris, France	OPeNDAP	*AEROCOM*

NASA-level Catalog Interoperability



Formats

- Hierarchical Data Format (HDF)
  - Versions 4 and 5
  - “Standard” format but wide variation in data structures and semantics
- HDF-EOS2, HDF-EOS5
  - EOSDIS Standard structures and limited semantics
- Network Common Form (netCDF)
  - Standard format with wide variation in structures and semantics
- COARDS
  - netCDF convention on dataset dimensions
- CF1
  - COARDS successor with controlled vocabulary for variable names
- Binary, ASCII
  - Non-standard formats

Available Protocols & Services

- Anonymous FTP
  - Available for all public data
- Public HTTP
  - alternative to anonymous FTP, for most data
- Open-Source Network for Data Access Protocol (OPeNDAP)
  - Supports subsetting, ASCII download
  - Supports netCDF conversion for HDF5 data
  - Available for many datasets
- OGC Web Coverage Service
  - Typically implies interpolation / reprojection
  - NetCDF/CF-1 profile
  - Does not support vertical profiles
  - Offered for a few datasets
- OGC Web Map Service
  - Reprojected, mapped visualization
  - Offered for TRMM AIRS, AIRS Near-real-time

Availability Table

	Storage							NetCDF Conversion		High-level Protocols			
Type of Data	*ASCII*	*Binary*	*GRIB*	*HDF4*	*HDF-EOS2*	*HDF5*	*HDF-EOS5*	*HTTP Svc*	*OPeN DAP*	*OPeN DAP*	*WCS*	*WMS*	*Air Quality Variables*
AIRS Grid				X	X			X		X	x	x	CO, CH4, temp profile
AIRS Near-Real-Time Swath				X	X							x	CO, Volc SO2
AIRS Radiance				X	X					X			Volc SO2
AIRS Swath				X	X			X		X			CO, CH4, temp profile
GLDAS Model Output			X					X		X			land surface
HIRDLS Swath						X	X			X			O3, CFC, aerosol, temp profile, HNO3
LIMS Profiles	X												O3, NO2, HNO3, temp profile
MERRA Model Output				X	X			X		X			winds, trace gases
Microwave Sounding Unit		X		X									air temp
MLS Swath						X	X			X			trace gases, temp profile
OMI Bin						X	X			X			aerosol, HCHO, NO2, SO2, O3
OMI Grid						X	X		X	X	x		aerosol, O3
OMI Swath						X	X						aerosol, BrO, ClO, HCHO, NO2, SO2, O3
SORCE	X												solar irradiance
TOMS Grid				X	X			X		X			aerosol, O3
TOMS Swath						X	X						aerosol, O3
TRMM Grid				X						X			rainfall
TRMM Merged 3-Hr Grid				X				X		X		X	rainfall
TRMM Raw Data		X											rainfall
TRMM Swath				X									rainfall
UARS Grid		X											NOx, O3, HNO3, SO2, CO, CH4

Metadata - Transport

- Anonymous FTP and Public HTTP
  - Internal to file
  - Alongside of file
    - S4PA metadata files
    - OAI/PMH not yet supported
- Within protocol
  - WMS - GetCapabilities doc
  - WCS
    - DescribeCoverage doc
    - Internal to transmitted file
  - OPeNDAP
    - Separate protocol requests: DDS, DAS, DDX

Metadata Format & Semantics

- Format
  - Depends on transport
  - XML
    - Alongside of file
    - In OGC protocols
    - OPeNDAP DDX
  - ODL (Object Definition Language)
    - Internal to HDF-EOS files
- Structure and Semantics
  - Limited by format and transport
  - ISO 19115, Part 2? Not yet supported
    - But could be mapped with XSLT
  - WDC-RSAT?
  - EOS Clearinghouse?

Key Issue: Quality Control Info

- Complex info for most products
  - In data file
  - External files
  - External documentation
- How to handle in WCS / WMS?
  - Prescreen (replace suspect pixels with fill)?
    - See Provenance…
  - Bundle with data?
    - Not possible with WMS

Key Issue: Provenance

- Key aspects
  - Original source of data
    - Data attribution
  - Processing
    - Quality screening
    - Reprojection
    - Aggregation
- How to transport?
  - WMS - no mechanism
  - WCS - use netCDF global attribute?
  - Develop extension to protocols?

Key Issue: Ancillary Materials

- Types of ancillary materials
  - Supporting documentation
  - Supporting tools
- Which links need to be preserved?
- How do we preserve them?

Science quality of the multi-sensor analyses

Data harmonization

- This is not only about data formats and protocols
- This is about science knowledge and provenance

Main challenge:

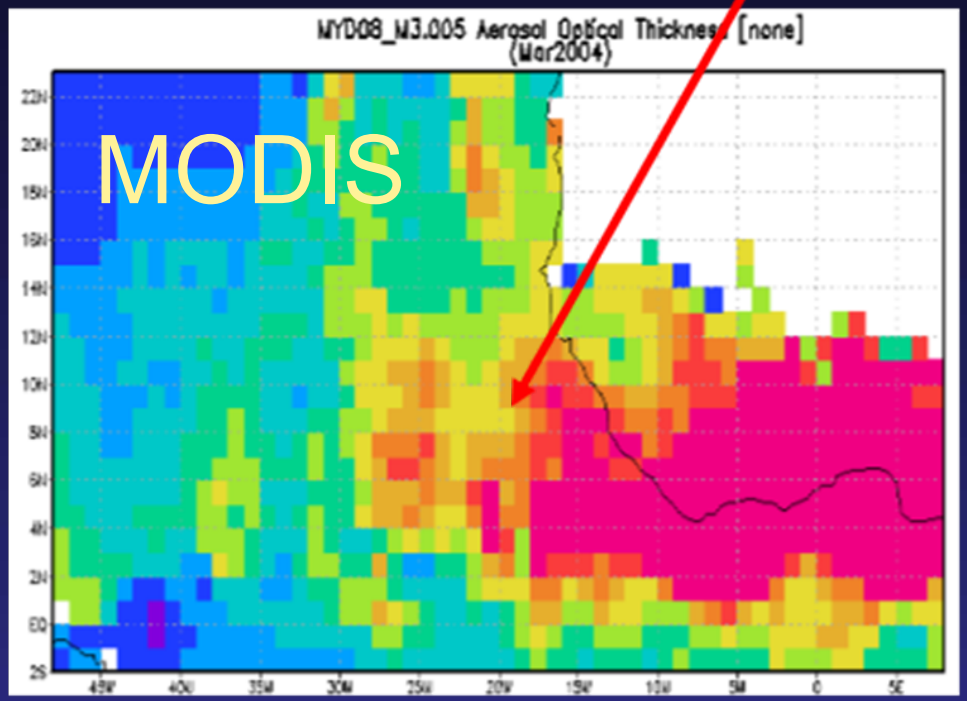
Even after all the data are read, colocated and coregistered on the same grid, how can we assure that the data quality and provenance are used, and used correctly?

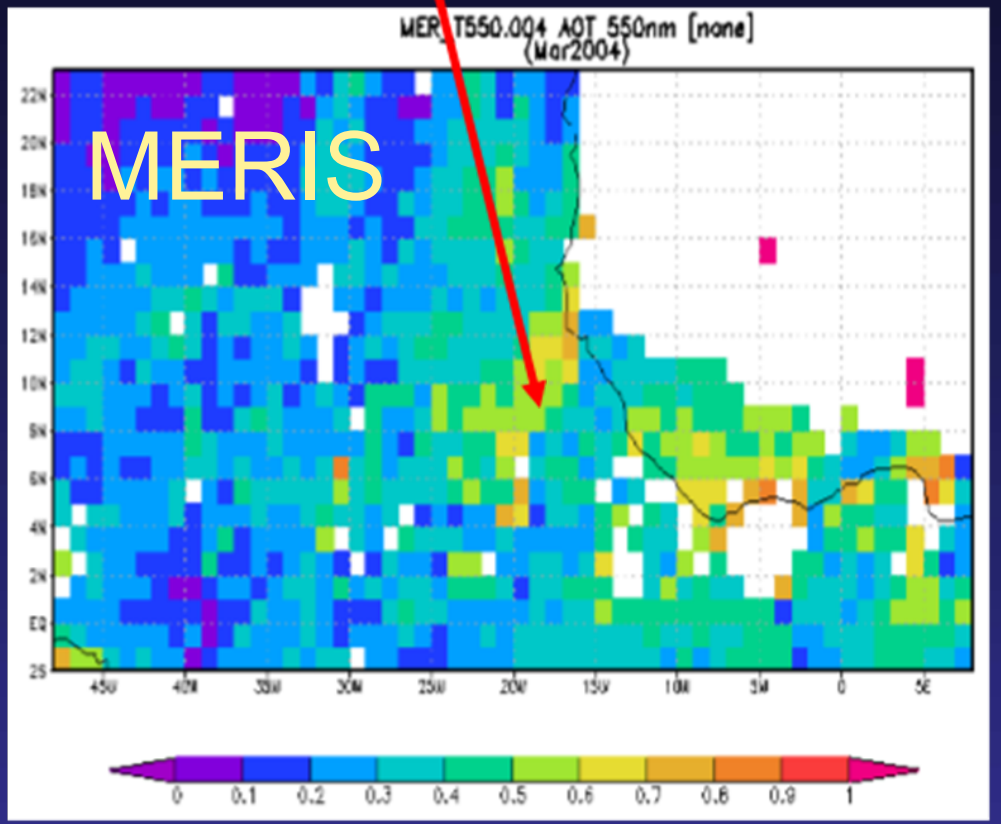
Science Quality and Data provenance

***Data Provenance**: the source of data, including the execution history of the processes that produced them*

Same parameter

Same space & time





Different provenance → Different results – why?

Data Provenance and Science Quality

- We can save time by providing convenient services to scientists but…
- Science quality of our results is imperative for scientists to be able actually trust and use them
- Documenting all the steps leading to the final product is paramount
- Also, providing assessment of sensitivity of the results to variations in processing algorithms/steps… published in peer-reviewed papers and presented to users in convenient, easy-to-find-and-read fashion
- Only working closely with scientists can guarantee science quality

Science Quality of Giovanni Results

- Giovanni operates mostly on the standard data products
- Giovanni results are the same as produced using the standard data out-side of Giovanni
- Data can be misused in Giovanni as well as (and may be more so than) without Giovanni
- We implement Science Team recommendations
- We provide warnings and caveats
- We perform sensitivity studies together with scientists in the corresponding fields

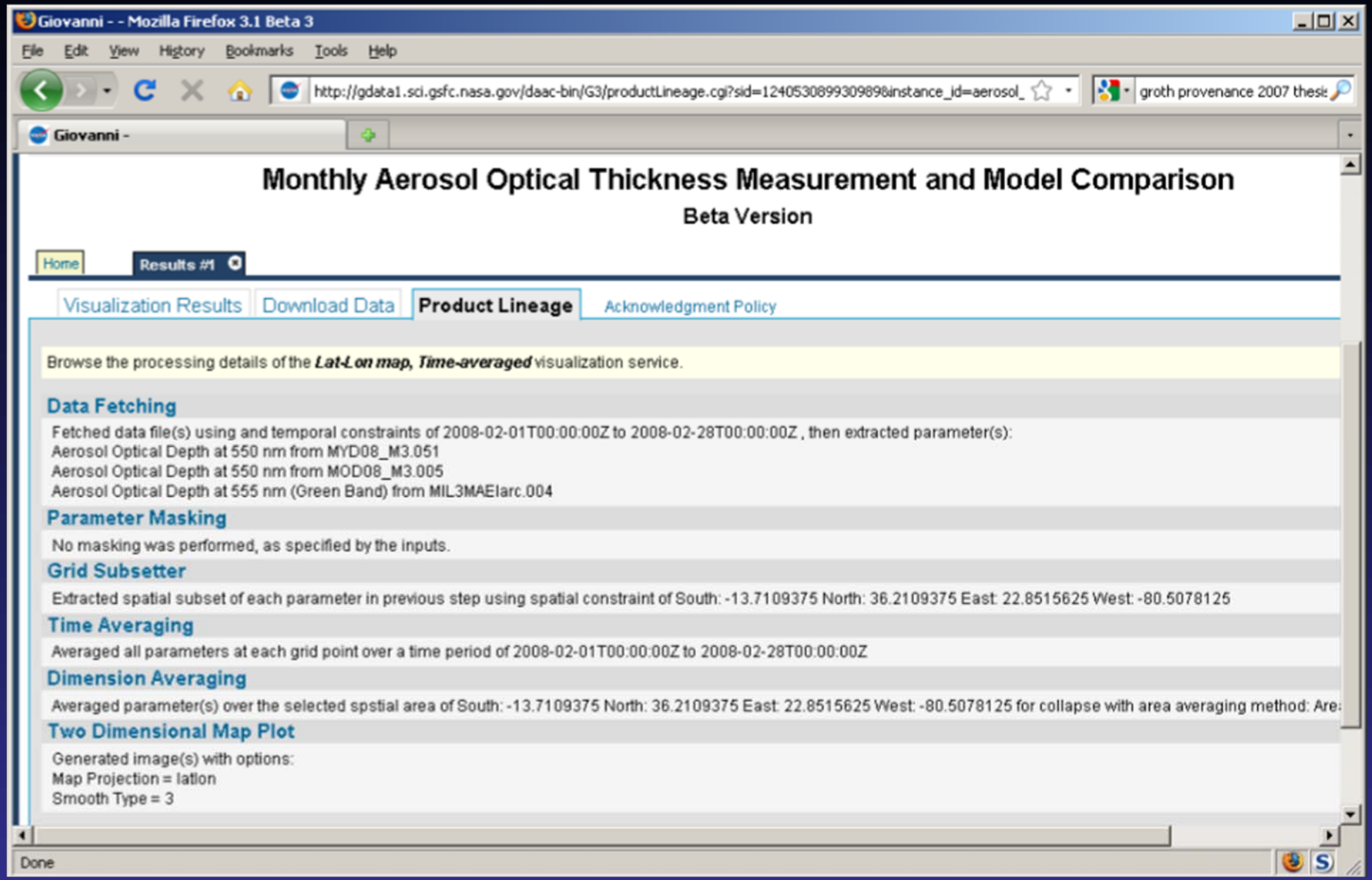
MLS Science Team recommendations

Parameter	Min. Altitude (hPa)	Max. Altitude (hPa)	Quality thresh (min)	Convergence thresh (max)
ClO	100	1.0	0.8	1.5
CO	215	0.0046	0.2 < 100 hPa 1.2 ≥ 100 hPa	1.8
Geopotential Height	316	0.001	0.6	1.2
H2O	316	0.002	0.9	N/A
HCl	100	0.15	1.0	1.5
HCN	10	0.1	0.2	2.0
HNO3	215	3.2	0.4	1.8
N2O	100	1.0	0.5	1.55
O3	215	0.022	0.4 < 100 hPa 1.2 ≥ 100 hPa	1.8
OH	32	0.0032	N/A	1.1
Relative Humidity with respect to Ice	316	0.002	0.9	N/A
Temperature	316	0.001	0.6	1.2

All parameters are screened where 'Status Flag' bit 0 is not set (i.e. odd value).
Data values where the precision is negative are excluded.

- Quality – flag where larger values indicate 'good' radiance fits.
- Status – various bit encoded flags, e.g. bit 0 indicates error, bit 1 questionable data, etc.
- Convergence – ratio of the fit from the retrieval algorithm to the estimated fit

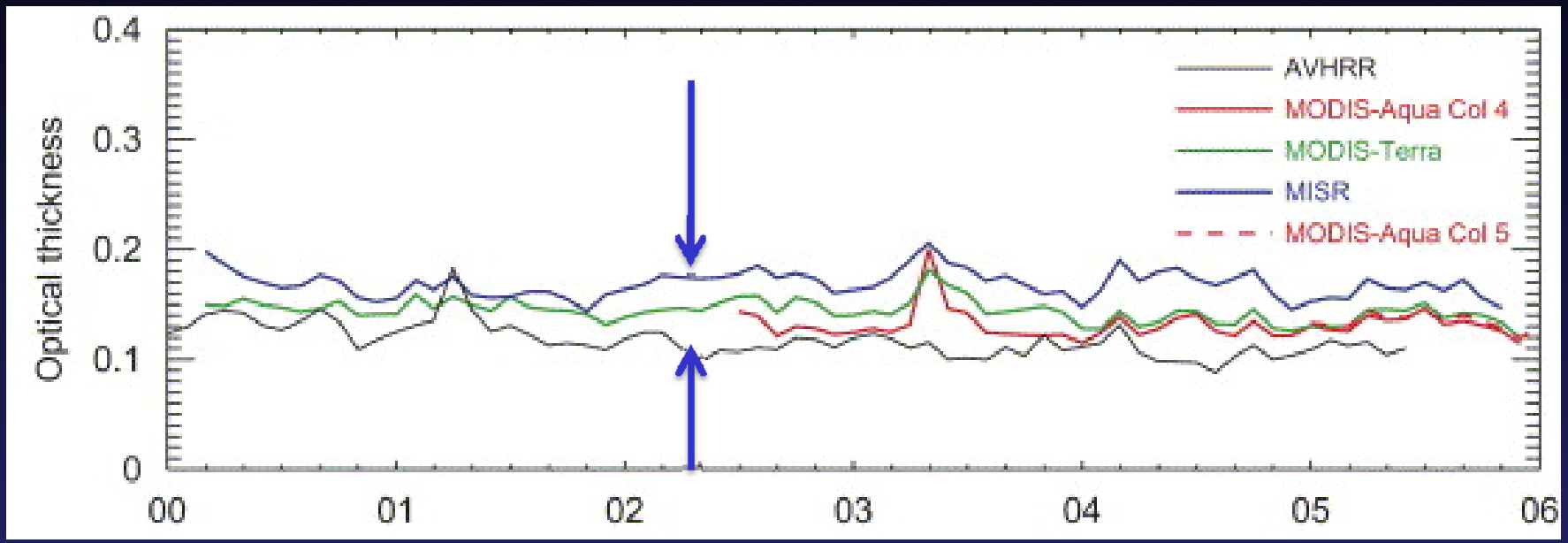
Product lineage in Giovanni



Provenance for Intercomparison

- *Automated or semi-automated intercomparison* of two apparently comparable parameters exposes a challenge in the proper consideration of the data provenance.
- Dealing with two or more provenance chains is much more difficult.
- Provenance should be described with enough *semantic richness* for users to assess and eventually *assure the scientific validity* of an intercomparison operation.
- Complicating this task is the dispersion of data and services to multiple sources, to be accessed via *heterogeneous workflows*.
- Persisting and transmitting the rich provenance requires *provenance interoperability* in addition to data interoperability.

Differences between sensors

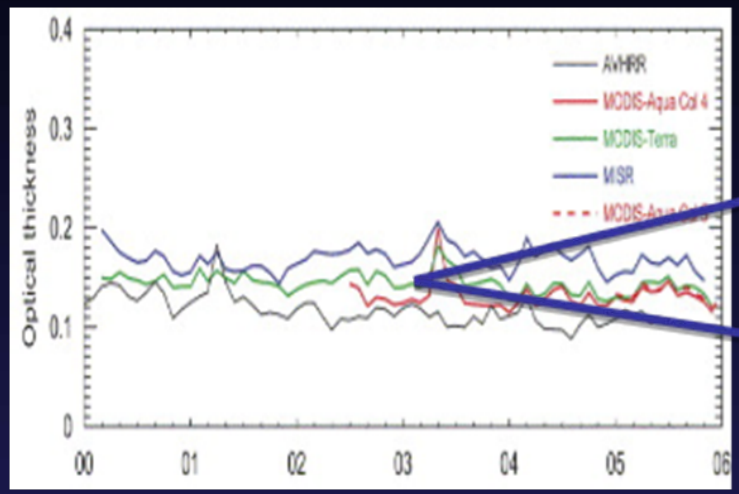


Time series of the global mean values of the AOT over the oceans from **Mishchenko** et al., J. Quant. Spectr. and Rad. Transfer, 2007, **106**, 325

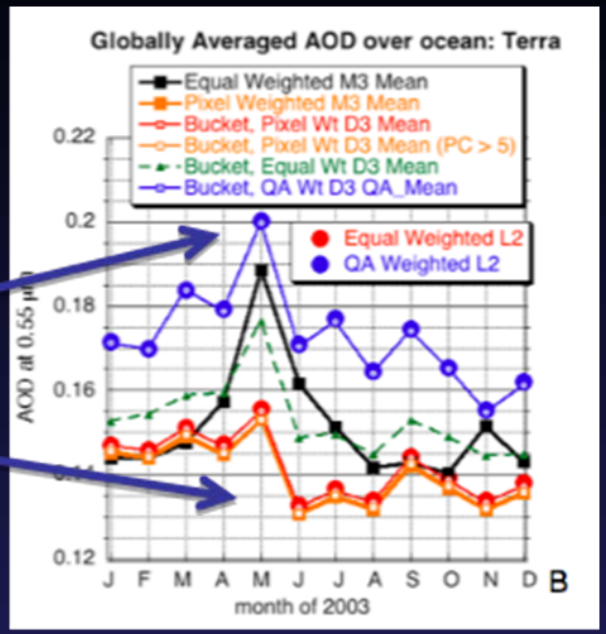
Differences in AOT between various sensors exceed reported accuracies of each sensor

Statistical aspects of spatio-temporal aggregation of MODIS aerosol data

AOD difference between sensors



Mishchenko et al., 2007



Levy, Leptoukh, et al., 2009

MODIS Terra only AOD: difference between diff. aggregations

Q.: How sensitive are AOT time-series to different aggregations to monthly products?
A: Very sensitive. For MODIS-Terra alone, AOD difference can be up to 40%

- Pixel count weighting (correctly) applied to L3 data represents L2 sampling, and leads to spatial and temporal (mostly clear sky) bias in the result.
- Applying Confidence weighting to L3 leads to a different Confidence-biased L2 sampling result
- Grid number weighting (correctly) applied to L3 data during spatial or temporal aggregation represents L3 sampling and a lesser spatial and temporal (mostly clear sky) bias.
- To compare data from different sensors, it is important to use the same statistical aggregation

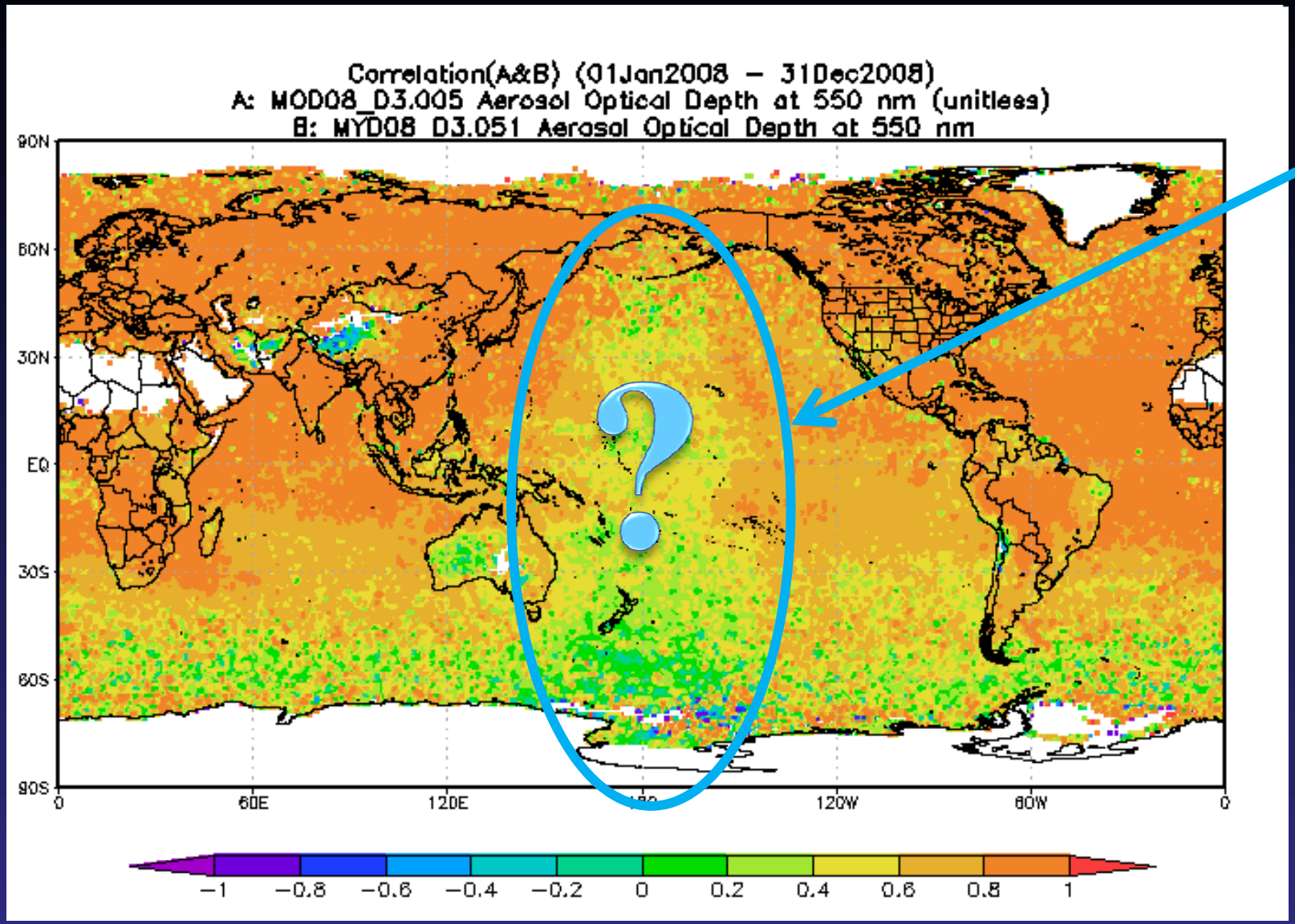
Intercomparison Questions

- Are those differences due to real differences between different measurement techniques?
- What about AOT differences for the same sensor due to different aggregation to Level 3 monthlies?
- If computed using Level 2, would these big differences still be present?

Major question:

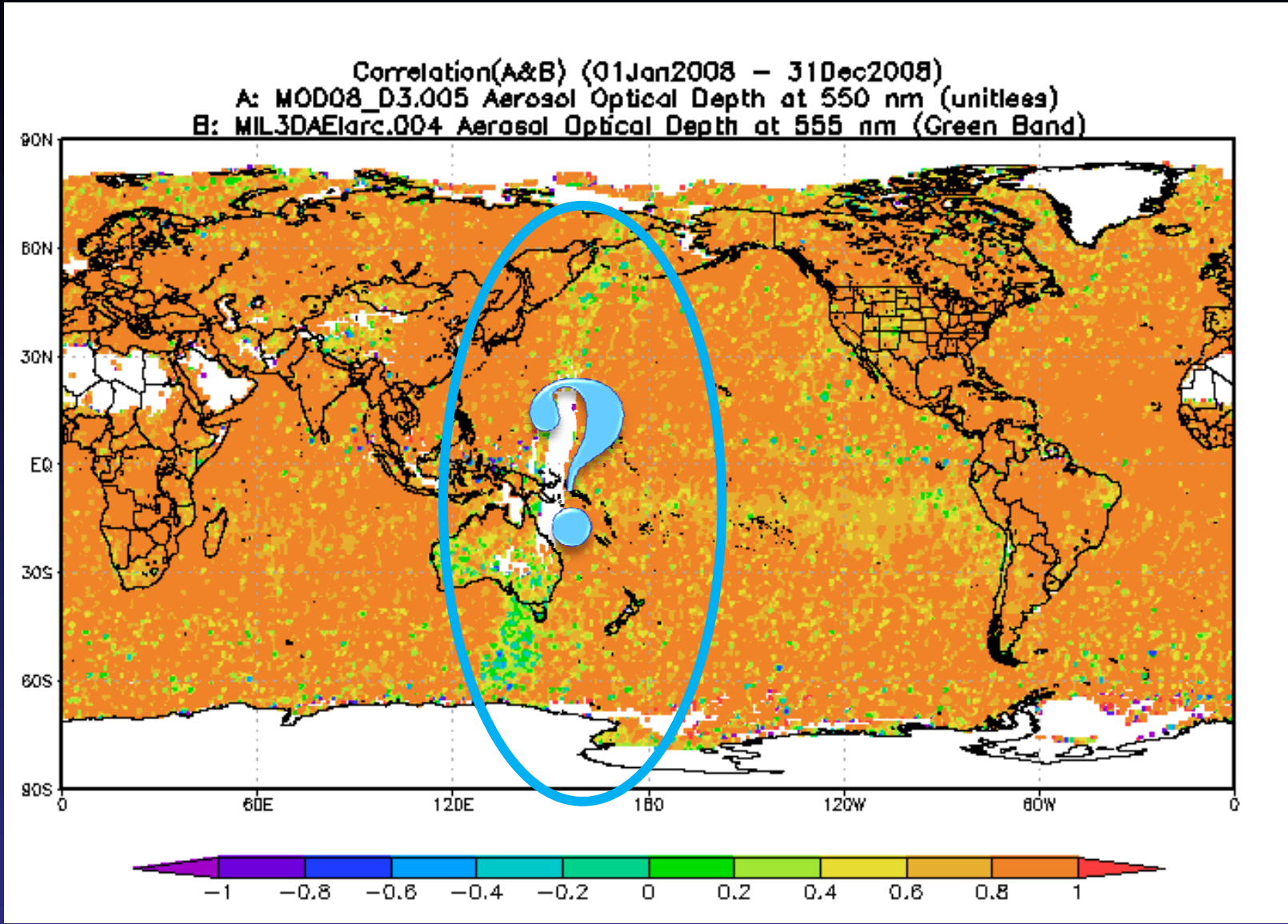
How sensitive are AOT time-series to different aggregations to monthly products?

Level 3 caveats: Data Day



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

Level 3 caveats: Data Day



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

Multi-Sensor Data Synergy Advisor (MDSA)

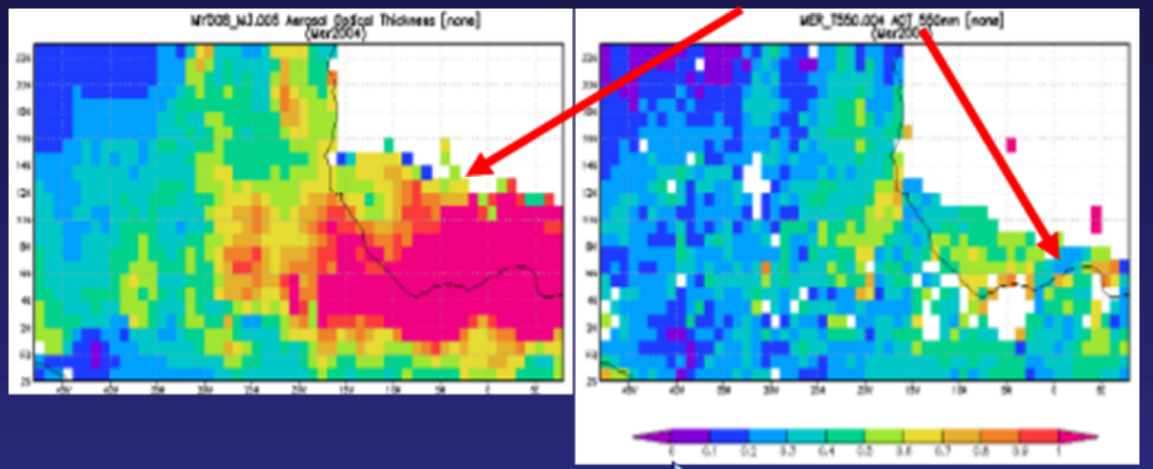
Expand Giovanni to include semantic web ontology system that captures scientist knowledge & data quality characteristics, and to encode this knowledge so the Advisor can assist user in multi-sensor data analysis.

Identify and present the caveats for comparisons.

Funding : ESTO

Same Parameter

Same Location and Time



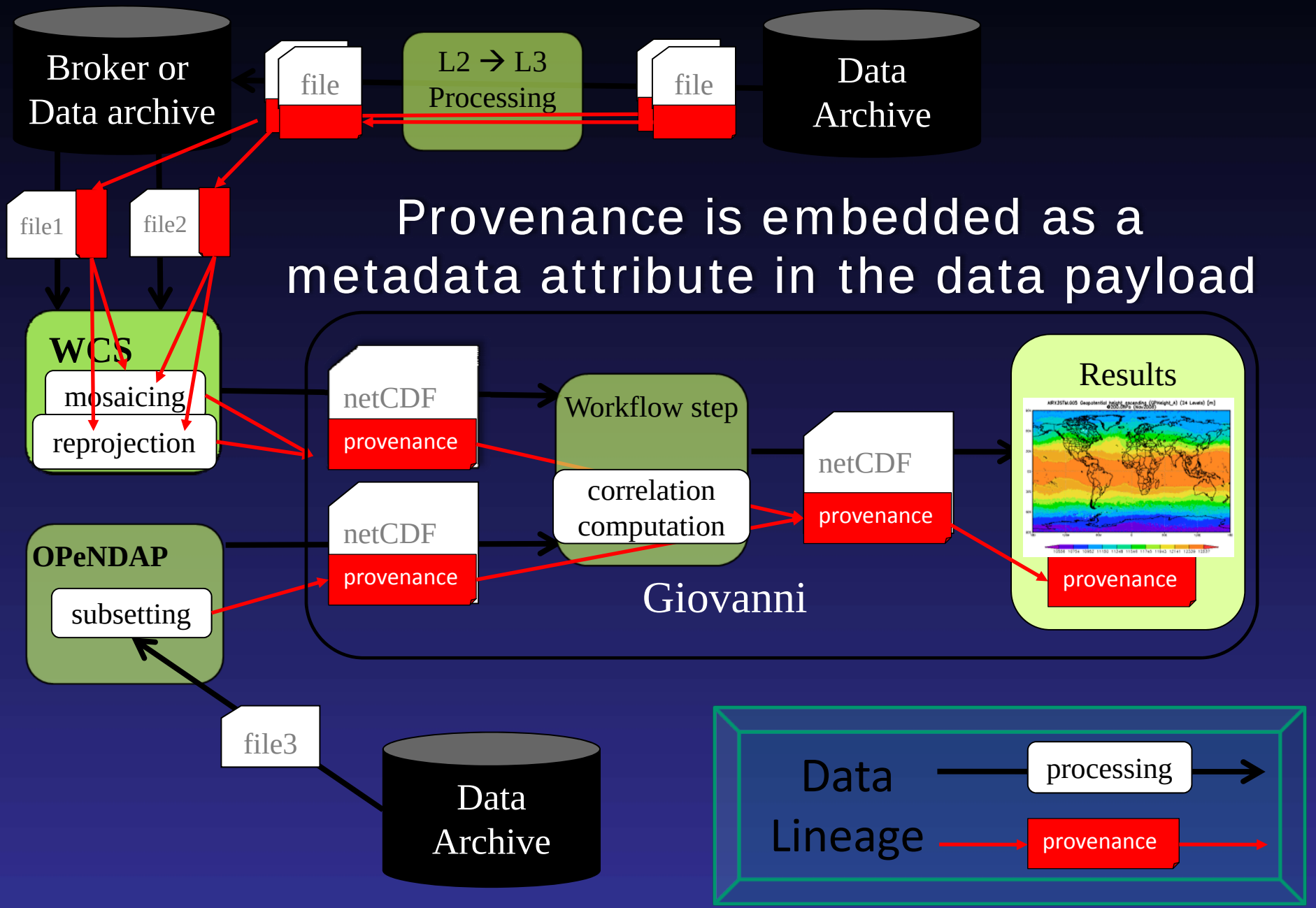
Different Provenance → Different Results

Importance of capturing and using provenance

International and inter-institutional data sharing projects involving Giovanni

- Atmospheric Composition Portal – a joint project between NASA GSFC and DLR serving as a prototype (to include more partners next year). WMS and WCS extended protocols.
- DataFed works with many organizations. WMS and WCS.
- Giovanni accessing data via WCS and OPeNDAP from various data centers.

Chaining provenances:



Chaining provenances: Out-of-Band Approach