Issues and Solutions in Bringing Heterogeneous Water Cycle Data Sets Together

James Acker[1], Steven Kempler[2], William Teng[1], Deborah Belvedere[3], Zhong Liu[4], Gregory Leptoukh[2]

[1]NASA Goddard Space Flight Center/SESDA2, [2]NASA Goddard Space Flight Center, [3]UMBC/GEST, [4]GMU

Steven.J.Kempler@nasa.gov

Workshop Results (http://news.cisc.gmu.edu/cewisworkshop.htm)

Abstract

The water cycle research community has generated many regional-to-global -scale products using data from individual NASA missions or sensors (e.g., TRMM, AMSR-E); multiple ground- and space-based data sources (e.g., Global Precipitation Climatology Project [GPCP] products); and sophisticated data assimilation systems (e.g., Land Data Assimilation Systems [LDAS]).

However, it is often difficult to access, explore, merge, analyze, and intercompare these data in a coherent manner due to issues of data resolution, format, and structure.

These difficulties were substantiated at the recent Collaborative Energy and Water Cycle Information Services (CEWIS) Workshop, sponsored by NASA Energy and Water cycle Study (NEWS) Program Manager Jared Entin, where members of the NEWS community gave presentations, provided feedback, and developed scenarios which illustrated the difficulties and techniques for bringing together heterogeneous datasets.

This presentation reports on the findings of the workshop, thus defining the problems and challenges of multi-dataset research. In addition, the CEWIS prototype shown at the workshop will be presented to illustrate new technologies that can mitigate data access roadblocks encountered in multi-dataset research, including:

-Quick and easy search and access of selected NEWS data sets.
-Multi-parameter data subsetting, manipulation, analysis, and display tools.
-Access to input and derived water cycle data (data lineage).

It is hoped that this presentation will encourage community discussion and feedback on heterogeneous data analysis scenarios, issues, and remedies.

Challenges/Issues in Bringing Together and Utilizing Heterogeneous Data Sets

From CEWIS Workshop:

Steps taken to gather and prepare data for multi-data set inter-comparisons

***Responses*:**

Retrieving Data
- Identify the sources of data, found by searching data center archives and web
- Find relevant data with the desired data characteristics
- Get samples of data; sort out data 'quirks'; determine if data are readable/correct
- Check units, timestamp, quality control flags
- Understand data characteristics (format, time period and resolution)

Assembling Data
- Collocate from different instruments
- Bring data sets to common grid (interpolate, collocate)

Analyzing Data
- Acquire data read code
- Perform data subsetting
- Perform data intercomparisons
- Homogenize different data sets for objective comparisons

Roadblocks encountered when bringing heterogeneous data sets together

Responses:

Data Access
- Finding and gaining access to the data
- Data sets tend to be organized on a project-specific basis, making it difficult to know what other data sets might be applicable to a given problem
- Lack of a nice "search engine" to quickly locate the data

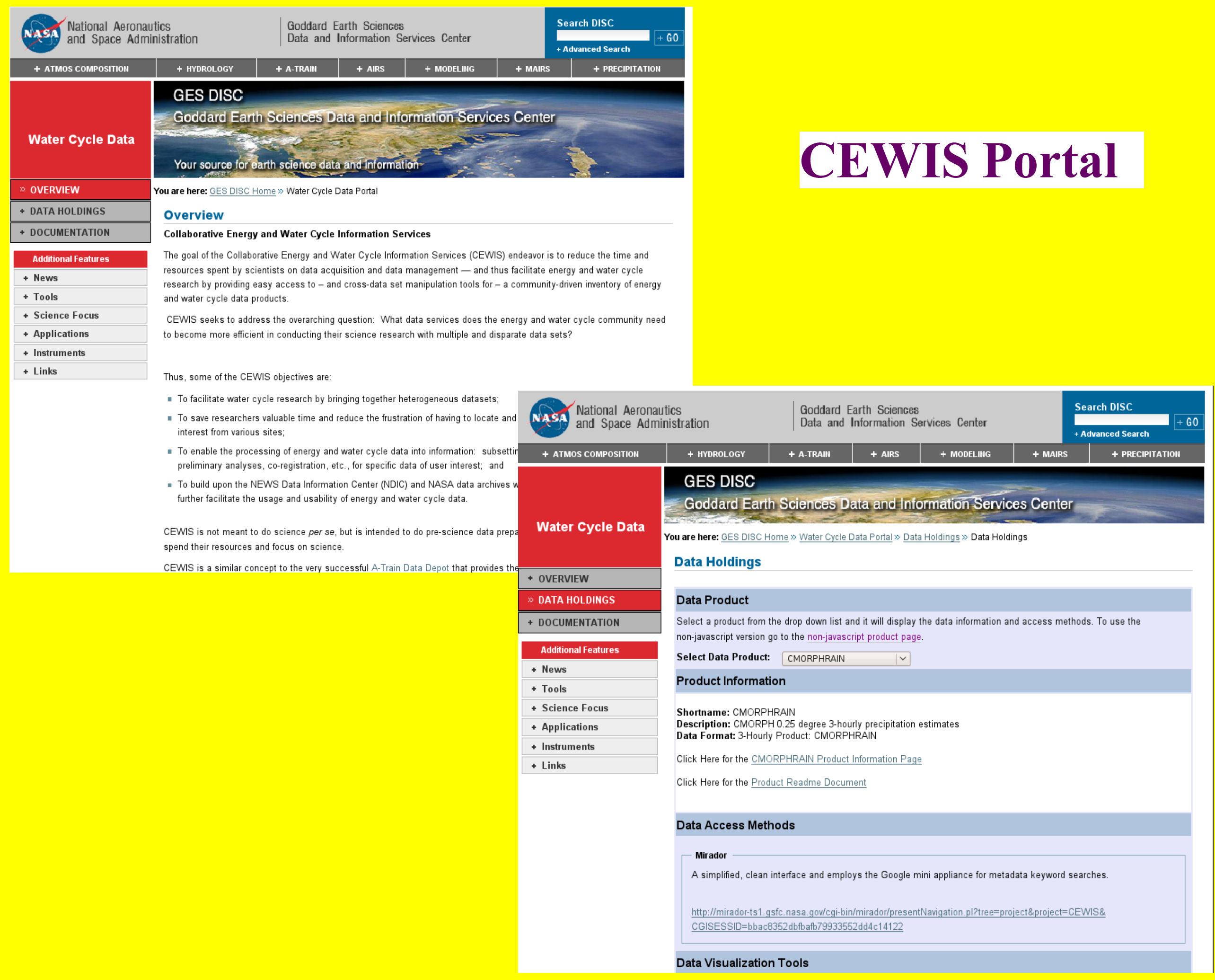
Data Characteristics
- Learning how to read data correctly
- Data volume → download interruption
- Spatially and temporally subsetting data

Combining Datasets
- Finding collocated data sets
- Converting data of different formats
- Properly converting data to common grid
- Users are expected to have knowledge of multiple sensors
- Dealing with different spatial/temporal resolutions, spanning periods, aspects of heterogeneous data sets, instrument fields of view

Verifying Combined Data sets
- Quantifying errors introduced during interpolation
- Distinguishing natural signal from systematic errors
- Time/space gridding mismatches across data sets; reconciling data that are not uniform in space and time

Data Documentation
- Undocumented features in data
- Inadequate quality control, error estimation, detailed documentation of data

CEWIS Portal



Demonstrating a Solution: CEWIS

The goal of the Collaborative Energy and Water Cycle Information Services (CEWIS) is to develop a community-based evolving set of data and information services that would facilitate users to locate, access, and bring together multiple distributed heterogeneous energy and water cycle datasets.

Demonstration Purpose

- To show data services for NEWS data sets that facilitate multi dataset research
- To provide some starting points of discussion on NEWS multi-dataset analysis requirements

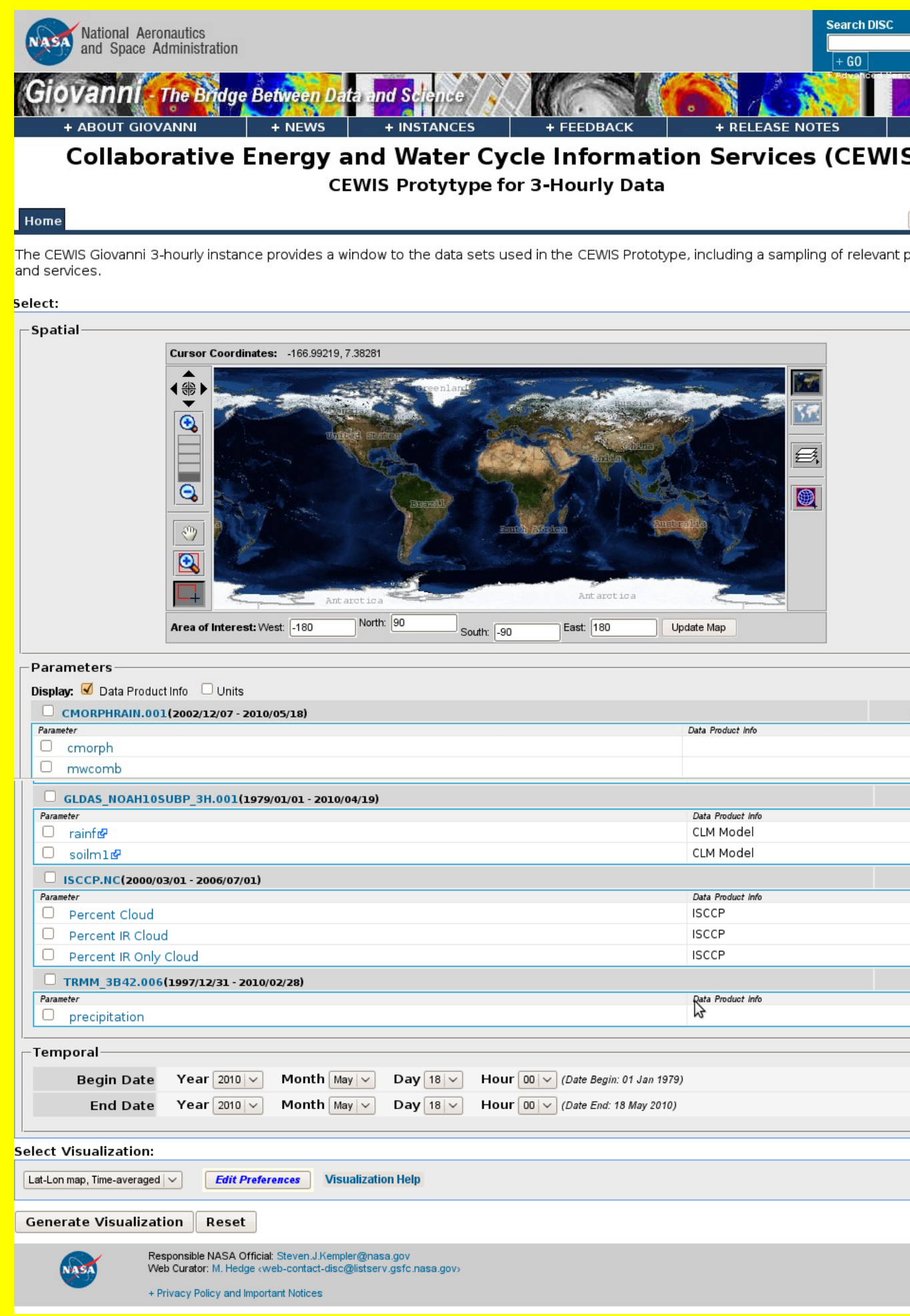
Demo Components

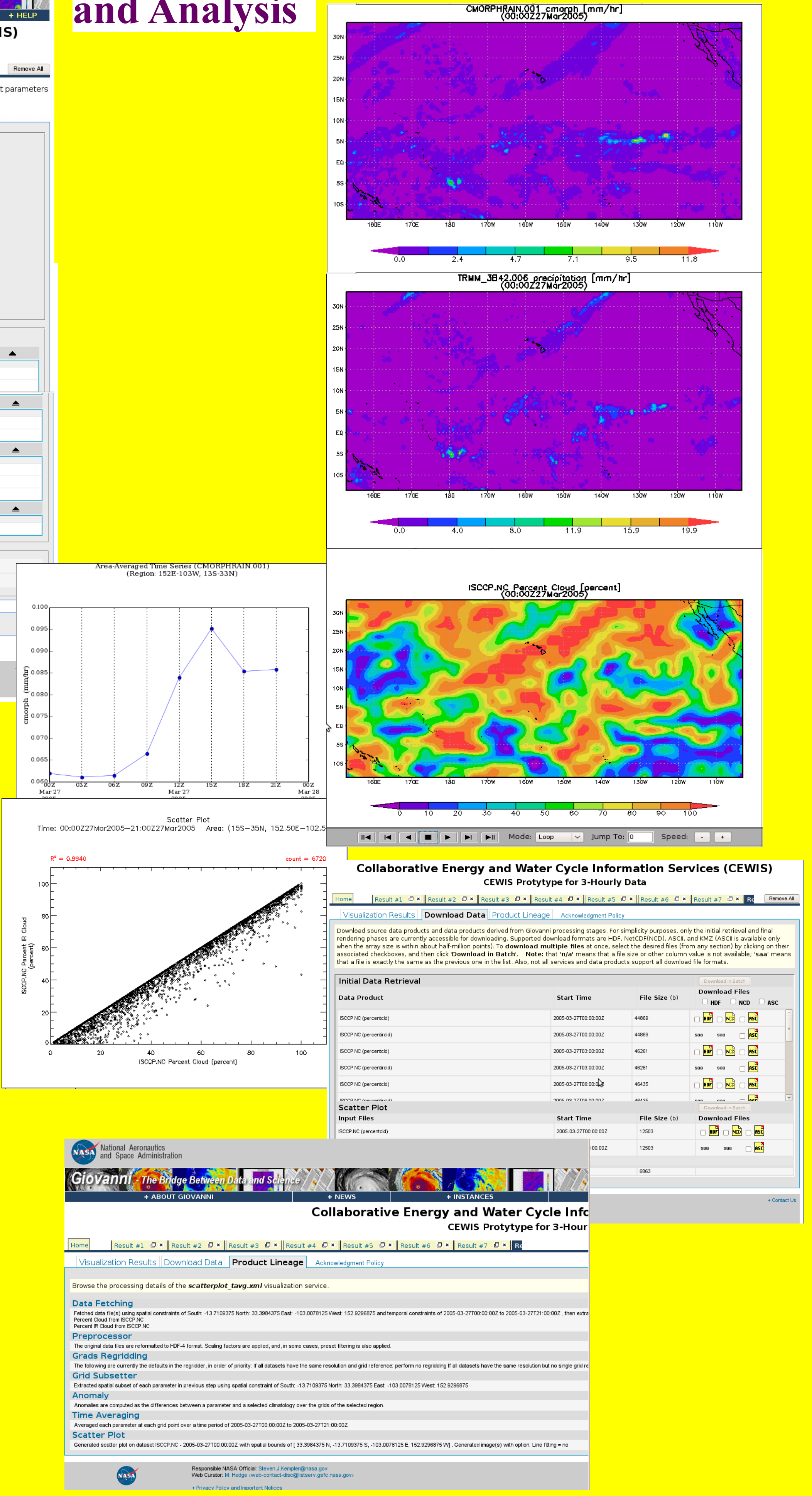
- CEWIS Portal
- Data search and access (Mirador)
  - "Manual" option
  - OpenSearch options
    - Data provider has their own search engine
    - Data is published to ECHO
    - Data provider installed provided search engine
- Data visualization and analysis (Giovanni)
  - 3 instances (monthly, daily, 3-hourly)

Thanks to our NEWS PIs Demo Collaborators

- Eric Fetzer - Merged Atmospheric Water Data Set from the A-Train
- Bill Rossow – ISCCP
- Robert Joyce – CMORPH
- George Huffman – GPCP2
- Matt Rodell – GLDAS

Giovanni Data Visualization and Analysis



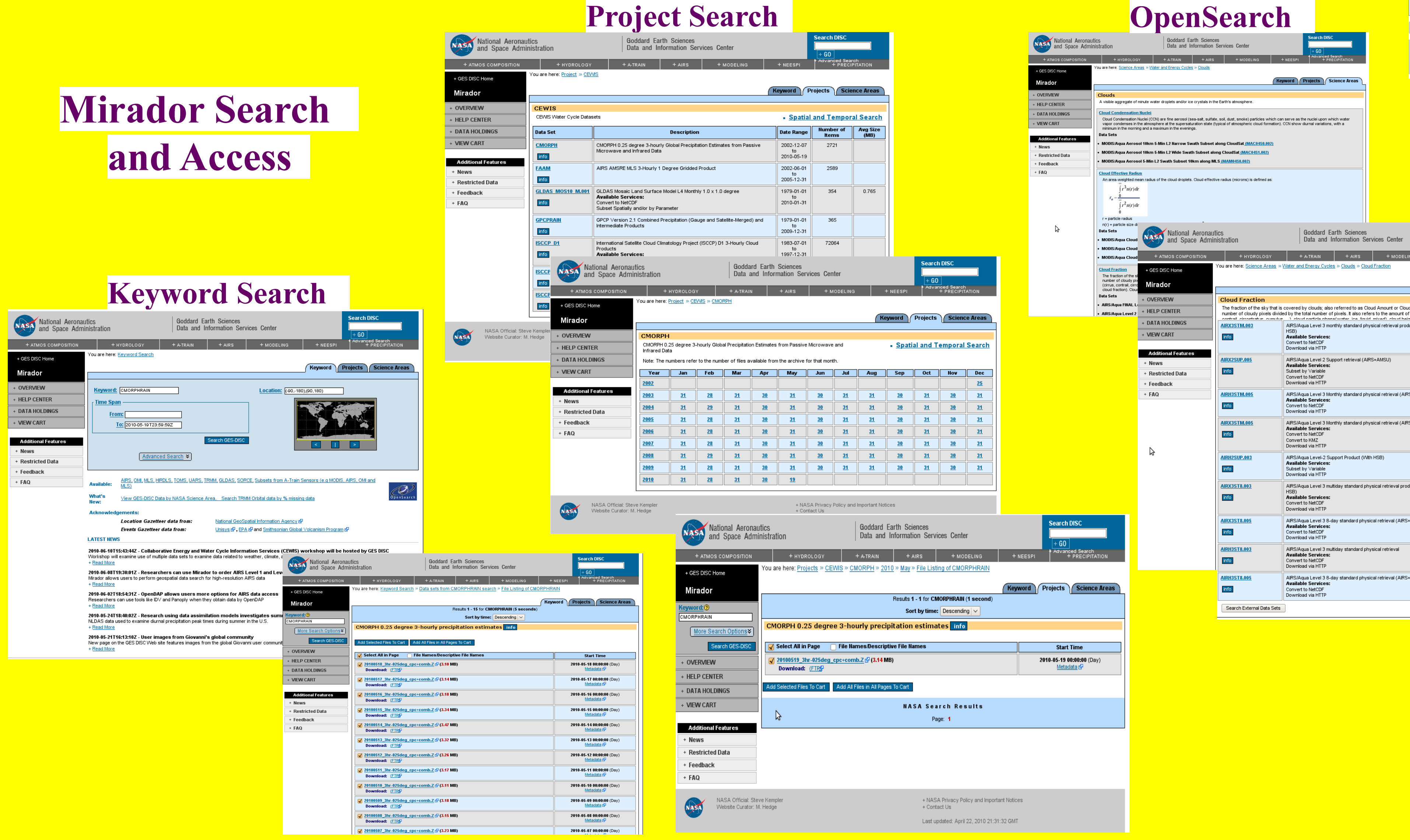


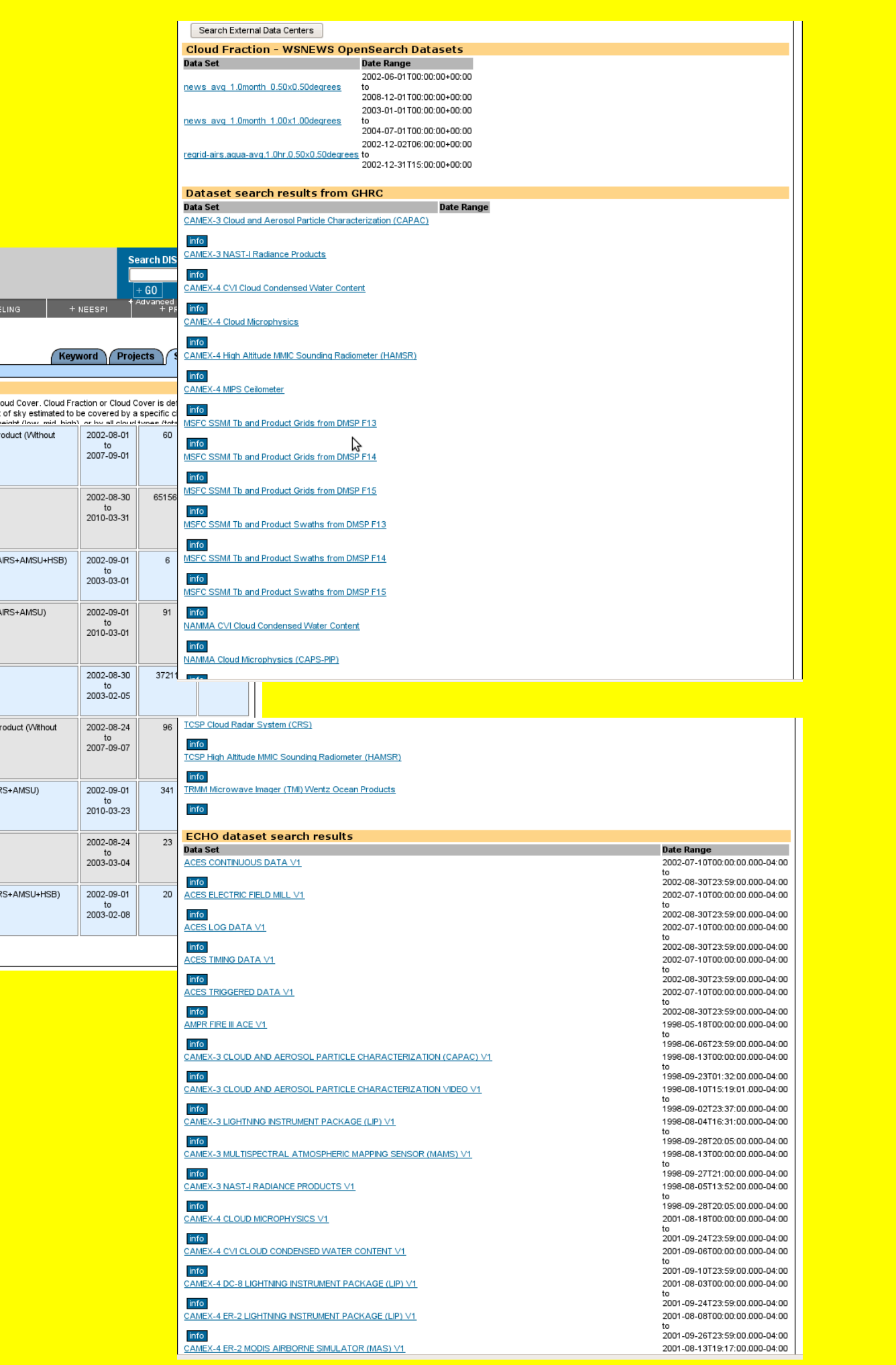
Project Search

OpenSearch

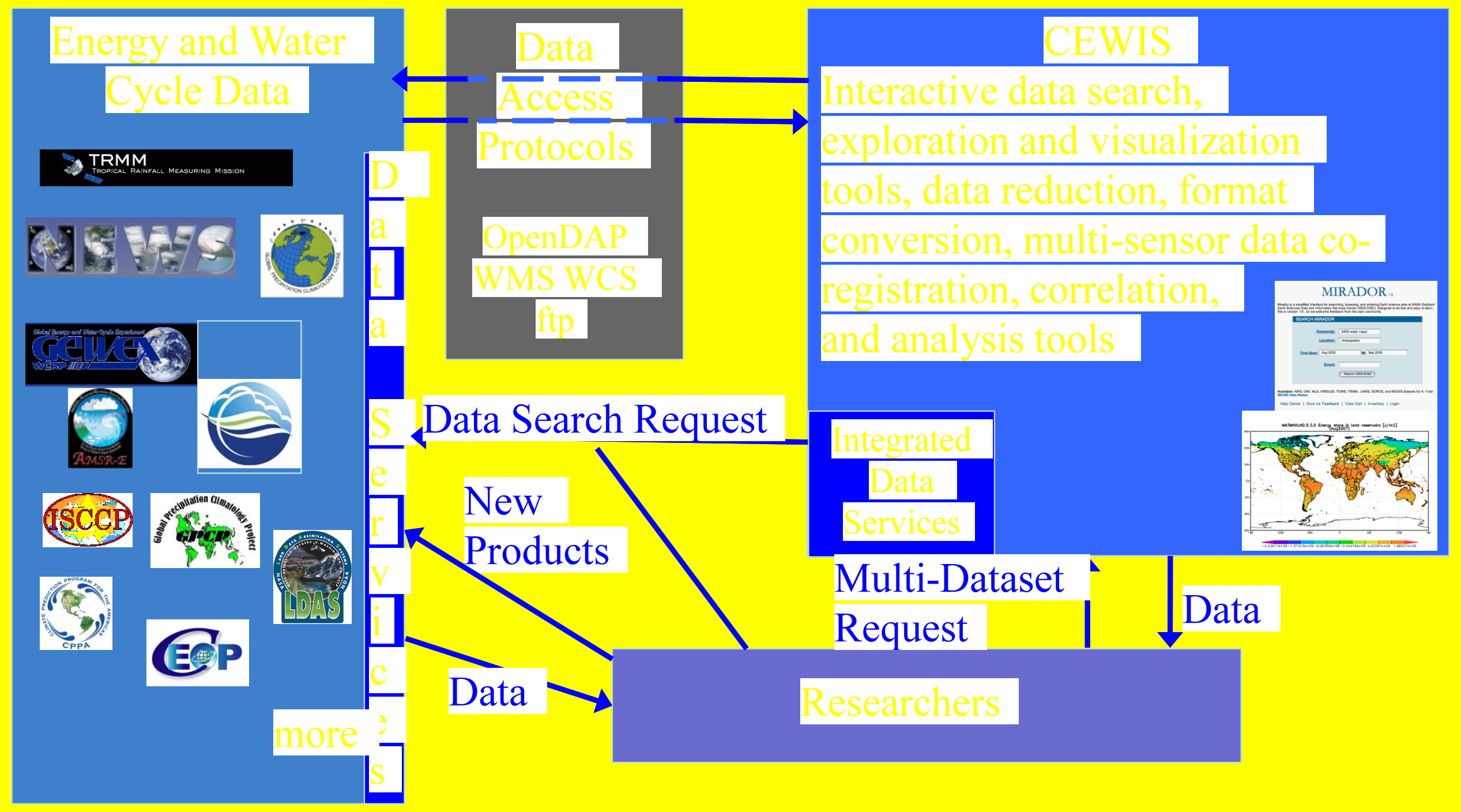
Mirador Search and Access

Keyword Search





The Collaborative Effort Model


Energy and Water Cycle Data
TRMM Tropical Rainfall Measuring Mission
NEWS
GEWEX
AMSR-E
ISCCP
GPCP
LDAS
CPPA
ECP
more
Data Services
Data Access Protocols
OpenDAP WMS WCS ftp
CEWIS
Interactive data search, exploration and visualization tools, data reduction, format conversion, multi-sensor data co-registration, correlation, and analysis tools
MIRADOR
Integrated Data Services
Data Search Request
New Products
Data
Multi-Dataset Request
Data
Researchers