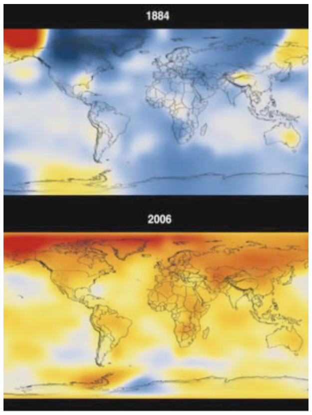

GROUP ON
EARTH OBSERVATIONS

GEOSS Workshop XLIII
Sharing Climate Information and Knowledge

Place: NCAR Center Green, Boulder Colorado, USA
Time: Friday, 23 September 2011, 8:00am-5:30pm
Fee: Free - all are welcome (registration requested, see below)

Organized by:
David Arctur, Ben Domenico, Glenn Rutledge, Nate Booth,
Al Gasiewski, Siri Jodha Singh Khalsa, Francoise Pearlman, Stefano Nativi

This one-day workshop is focused on how the Global Earth Observation System of Systems (GEOSS) can contribute to studying climate change, how GEOSS can enable broader community access to climate information and knowledge; and how scientists can manage and work with integrated data sources having different levels of quality.

Organized in association with:

OGC Making Location Count
USGS science for a changing world
NOAA National Oceanic and Atmospheric Administration U.S. Department of Commerce
CNR-IIA
IEEE
UCAR University Corporation for Atmospheric Research
ICEO IEEE Committee on Earth Observation
NCAR National Center for Atmospheric Research
unidata

Final Agenda

Morning Sessions	(Moderator: Al Gasiewski)	
Time	**Topic**	**Speaker**
8:00	*Registration Check-in*	
SESSION 1: Opening		
8:30	Welcome and Opening	**Sandy MacDonald, NOAA**
8:45	Logistics and Introduction to Workshop Objectives	**David Arctur, OGC**
SESSION 2: Bringing Climate Data to GIS and GEOSS		
9:00	Introduction to GEOSS and its Approach to Interoperability	**George Percivall, OGC**
9:25	GEOSS Portal and Service Oriented Architecture	**Siri Jodha Singh Khalsa, NSIDC**
9:50	Tools for Analyzing Societal Impacts of Climate Change	**Olga Wilhelmi / Jennifer Boehnert, NCAR GIS**
10:15	*Refreshment Break*	
10:45	Climate Data Networks & Portals	**Jay Hnilo, NOAA NCDC**
11:10	Issues in Working with Climate Data in GIS: Provenance, Gridding	**Dave Blodgett, USGS**
11:35	Open Discussions	
12:00	*Lunch – No Host*	
Afternoon Sessions	***(Moderator: Ben Domenico)***	
SESSION 3: Tools and Demonstrations		
13:30	THREDDS Data Server	**Ethan Davis, Unidata**
13:50	WCS 2.0 Extension for netCDF-CF; CSW and THREDDS Interoperability	**Stefano Nativi, CNR-IIA**
14:10	Climate Change Impacts on Biodiversity and Ecosystems	**Stefano Nativi, CNR-IIA**
14:20	Linked Data Practices for the Geospatial Community	**Stephan Zednik, RPI / Tetherless World Constellation**
SESSION 4: Breakout Groups		
14:45	Charter to Breakout Groups	**David Arctur, OGC**
14:55	*Coffee break and discussion to organize into breakout groups*	

SESSION 4: Breakout Groups		
15:15	**Group A**: How can the GEOSS infrastructure be evolved to better support climate research?	Group A Moderator: Ben Domenico
	Group B: How can the informatics community help scientists to better manage and work with disparate data sources?	Group B Moderator: David Arctur
	Group C: (1) What questions can the informatics community help climate scientists address? (2) How can open science improve broader societal awareness and reasoned discussion of climate change issues?	Group C Moderator: David Blodgett
16:40	*Coffee & tea break while breakout groups prepare summaries*	
SESSION 5: Closing	***(Moderator: David Arctur)***	
17:00	Reports of Breakouts, Conclusions, and Recommendations	
17:30	*Workshop Closes*	

Attendance: 65 registered; 54 attended

Notes from Presentations

Morning session: Al Gasiewski moderator & note taker

The workshop was opened by Prof. Albin Gasiewski of the IEEE, who welcomed attendees to Boulder and introduced Dr. Sandy MacDonald, Director of the NOAA Earth Systens Research Laboratory.

1. Alexander MacDonald: Welcome and Opening

Dr. MacDonald discussed Earth System Information Systems and the value of data visualization in climate understanding. He emphasized through advanced graphical simulations the complexity of the atmospheric and oceanic systems and how global rendering of high-resolution models of Earth processes can be extremely useful. He noted that new models such as the Nonhydrostatic Icosahedral Model will be able to be routinely run at 4 km resolution globally for forecasting purposes. He described the computational complexities of producing a ten-day forecast with one minute time steps using this model. The key is to organize the data so that the resulting products are available over regular communication lines but based on complex large data sets. A question arose as to how it can be ensured that all the data is available for those interested user, along with the low-bandwidth products that are most commonly needed. This problem of geophysical data dissemination is currently being considered at NOAA.

2. David Arctur: Motivation / Workshop Agenda / Logistics

Dr. Gasiewski next introduced **Dr. David Arctur** of **OGC**, who discussed the workshop logistics and objectives. He noted that this is the 43rd GEOSS workshop, and discussed the importance of climate scientists working with the informatics community to help develop interoperable tools for dissemination of climate data. He noted that the Intergovernmental Panel on Climate Change (IPCC) will soon be developing their fifth assessment report (AR5), and the improvement of dissemination systems such as the GEOSS web portal will be extremely valuable in support of this activity. He recognized the concern expressed by Dr. MacDonald in being able to deliver visual images from extremely large data sets of extensive scope. He described the logistics of the breakout sessions along with draft questions that would be used to motivate the breakout group discussions.

3. George Percivall: Introduction to GEOSS and its Approach to Interoperability
The next speaker was **George Percivall** of **OGC**, who described an introduction to GEOSS and its approach to interoperability. He distinguished GEO (the Group on Earth Observations) and GEOSS (the Global Earth Observation System of Systems). He notes that GEOSS connects observations to decision making through an interoperable data management system. The GEO work plan, designed as part of a 10-year plan approved in Brussels in 2005, defines the set of tasks which are the means by which participating organizations and member countries develop the system of systems. He highlighted tasks IN-03, IN-05, ID-04, CL-01, and CL-02 – all relevant for GEOSS infrastructure, institutions and development, and information for societal benefits. CL-01 an CL-02 were detailed as relevant to workshop discussions. He introduced the GEOSS common infrastructure (GCI) as a means of allowing individual systems to meet their particular objectives but co-exist as part of the SoS. The GEOSS interoperability arrangements were described as fostering open interfaces through open standards and open intellectual property rights. The GCI is all about discovering data sources and information that are registered by with GEO. There is a GEO web portal that can be used to discover data sources and access registered (and some non-registered resources) through the GEOSS clearinghouse. The standards and interoperability forum (SIF) provides the ground rules for managing the registry. Refinements in the GCI are continuing to date. He noted that using the GCI can reduce the task of performing scientific studies using Earth data from months to minutes. He noted the analogy of current data-driven Earth science to on-line travel bookings, which have become far more efficient due to web-based search techniques. He described scenarios in which working with the scientists and policy makers have provided a more efficient means of defining GEOSS architecture. This Architecture Implementation Project (AIP) has provided a demonstration roadmap toward development of GEOSS applications. In this development process rendering .kml data is becoming more widespread, open and valuable. Two search examples, one on Arctic weather and another on habitat change, were presented. These pilot projects are rapidly becoming operational in the GCI.

4. Siri Jodha Singh Khalsa: GEOSS Portal and Service-Oriented Architecture
The next speaker was **Siri Jodha Singh Khalsa** of **CU/NSIDC**, who discussed the GEOSS portal and service oriented architecture. The GEOSS portal and clearinghouse are described in the GEOSS best practices wiki. Registered resources must declare their service endpoints and associated standards and other interoperability arrangements. This information is stored in the standards and interoperability registry (SIR). The SIF manages the SIR, and provides analysis, education, and advice on interoperability. Several resource categories are considered, from datasets, to monitoring and observing systems, computational models, initiatives, websites and documents, data services (analysis and visualization, alert,

RSS, and information feeds, and catalogues, inventories, and metadata collections), and software and applications. A number of issues have been discovered in implementing GEOSS, however, these are being identified and the CGI is being evolved to help accommodate these problems. This evolutionary plan includes a vocabulary of EO observations parameters, identification and priority assessing of data sets, assisting in online access, and development of tutorials. Challenges include lowering the entry barrier, and development of a brokered GCI.

5. Olga Wilhelmi / Jennifer Boehnert, NCAR GIS: Tools for Analyzing Societal Impacts of Climate Change

The next speakers were Drs. Olga Wilhelmi and Jennifer Boenhert of the **NCAR GIS Project** Team. **Dr. Wilhelmi** discussed the pressing importance of understanding the impacts of climate change, and the special importance of the geography of climate change. Society and climate are co-evolving in a manner that could place more vulnerable populations at risk. What is "Usable Science"? This is a science that meets the changing needs of decision makers, and goes beyond data formats into data content and data quality. It addresses the relevant spatial and temporal scales and perception of usefulness to foster usability. An example of the use of climate data in vulnerability assessment was presented. A matrix of climate data supply and demand was discussed as a means of addressing missed opportunities. A variety of obstacles prevent relevant users from benefiting from available climate data. A broad range of users exists, and their needs are being identified. They require familiar, interoperable formats, diverse access means, documentation and expert support, spatial analysis tools, and use cases. A variety of means of providing information was discussed.

Dr. Boehnert spoke next on the details of the THREDDS data server, illustrating the means by which users can access climate data. The resulting data can be imaged using a standard geophysical browser. WMS can also be accessed through ArcMap. Integration of different data domains and formats can lead to important new data discovery opportunities, for example, overlaying severe weather and hospital facilities would suggest where medical facilities would be in demand during a severe weather event. She also discussed the utility of netCDF operators and Climate Data Operators (CDOs), and showed an example of the use of the CDO in an extreme temperature event. Extracting and appending methods are particularly important in implementing rendering that serve the user's specific needs. She also mentioned the use and dependency on 3rd party software, and described educational resources, for example, the Earth Exploration Toolkit.

6. Jay Hnilo: Climate Data Networks & Portals

Dr. Jay Hnilo, Chief Scientist for NOMADS at NCDC spoke next on climate data network and portals at NOAA NCDC. He indicated the critical needs to address he spatial and temporal resolution of current climate data, which are going to soon exceed 300 PB (petabytes). In contrast the current capabilities are of order 1 PB. There are a number of NOAA data portals addressing the IPCC reports, which are increasing their data needs. The NOMADS goals are to establish a unified climate and weather data model and promote model evaluation. The GO-ESSP goals are supportive to developing data delivery services. The NOAA climate service portal goals and objectives further promote data discovery ad delivery. He also spoke of NOAA's global interoperability program, and the National Climate Predictions and Projections platform. The mission and strategy of NCPP were described. NOAA and USGS are jointly (under an MOU) applying scientific tools to increased understanding of climate change. The Earth System Grid Federation (ESGF) is an open consortium of institutions, laboratories, and centers around the world that are dedicated to sharing global climate data. Through NOMADS, OPeNDAP is a practical means of getting remote scientific data and metadata. OPeNDAP accesses NWP analyses and provides an ensemble probability tool.

7. Dave Blodgett: Issues in Working with Climate Data in GIS: Provenance, Gridding
The last speaker of the morning session was **Dr. Dave Blodgett** of USGS, who spoke on issues in coordinate reference systems in geo-coordinate gridding. The USGS Geo Data portal Project is a new effort to develop common tools to integrate climate, weather, and geosciences. For example, the impact of weather may be different within different recognized eco-regions within the U.S. A datum is a coordinate system based on a spheroid, ellipsoid, etc…, and includes a measurement system. In GIS systems, conversion from a spherical earth model to a higher resolution version is a significant issue to be studied. OSGeo.org's PROJ.4 is often presented with this problem. Climate sciences use terrestrial data. The difference between the WGS84 and GCM coarse spherical earth datum can be as much as 20 km in the mid-latitudes. Errors are thus usually small, but can be substantial for newer regional applications and especially in mountainous terrain. If no datum is provided, users waste much time trying to find the right one. Low spatial resolution should be explicitly indicated. The takeaway is that latitude and longitude are not unique and depend on the specific datum. As we move toward more high-resolution data sets such issues become more important.

Afternoon session: Ben Domenico moderator, David Arctur note taker

8. Ethan Davis, THREDDS Data Server

- THREDDS supports various data formats netCDF, GRIB, HDF, …
- And recognizes various scientific feature types: gridded, discrete sampling, …
- Support standards-based catalogs, data access services: WMS 1.1, 1.3, WCS 1.0, NetCDF subset service
- CF Climate and Forecasting convention
- Q: potential for using database IO/SP for improved performance?
- A: probably not; however, see RAMADDA: indexes all data in a thredds server; includes a twiki CMS

9. Stefano Nativi, WCS 2.0 Extension for netCDF-CF

- Italian CNR / GEOSS research programs: EuroGEOSS
- ISO 19123 WCS 2.0: written by stakeholders
- Discrete / Sampling Geometries (feature collections)

  - So data & storage models could support field-based (WMS) AND feature-based (WFS) AND sampling-based (SOS) data
  - Ragged Arrays – new
- WCS 2.0 can be expressed using binary (netCDF-CF) or GML content
- TDS + broker (GI-cat) accessible via OGC services (CSW OpenSearch-GEO and OAI-PMH)
-

10. Stefano Nativi, Climate Change Impacts on Biodiversity and Ecosystems

- Studying impact on species distribution using Ecological Niche Model
- Determining habitat uniqueness and what will change
- GEO Biodiversity Observation Network – see implementation plan from GEOSS
- GEO AIP-2: studied the role of climate in the decline of the American pika
  - Very sensitive to climate conditions; will tend to move up the mountains with climate change
- GEO AIP-3: EuroGEOSS/GENESIS "e-Habitat" project to assess environmental changes due to anthropogenic activities.
  - Bringing Climate Change study together with Bio-diversity research
- Q: is the community broker previously mentioned a human or computational entity? How is it done?
- A: broker described was primarily software components but requires institutional coordination and skilled development. The discovery broker is mediating queries across multiple data models. Works okay as long as service types are very similar. If service types differ greatly, a semantics broker should be considered.
- Users should be involved in decisions for its application

11. Stephan Zednik, Linked Data Practices for the Geospatial Community

- Consider the value proposition for linked data:
  - To enable easy discovery of resources
  - Leverage http, easy to implement
  - By using RDF the relationships should be easily specified
- 4 principles of linked data (Tim Berners-Lee):
  - Use URIs as names for things
  - Use HTTP URIs so people can look up those names
  - When one looks up a URI, provide useful info via standards (RDF, SPARQL)
  - Includes links to other URIs so they can discover more things
- Linked data web: data references data
- Linked data languages: SKOS, RDF Schema (RDFS), OWL, RDFS+ subset of OWL
- Reuse concerns: Usage & uptake, maintenance & governance, coverage, expressivity, modularity
  - Vocabularies should be reused when/where possible, but there's so far no definitive directory of these
- There's a basic RDF vocabulary for WGS84-based location terms, and it has wide usage, but it's too simple for GIS tools
- OGC GeoSPARQL provides vocabulary for spatial information, and a set of computation functions and query transformation rules.

- Discovery of linked data is typically via other linked data
    - OpenSearch most recommended interface
- Querying linked data typically via SPARQL endpoint, or faceted browsing tools
- SOAP web services not complementary with linked data (SOAP ignores HTTP access semantics). RESTful services can be complementary with linked data, but requires design.

Dr. David Arctur used the period before lunch to develop the set of draft questions for the breakout sessions.

Open science was defined as publications beyond the normal channels of publication, including rough lab notes, raw data, and other less refined sources of information. For example, how can publications properly reference the data that was used to develop the conclusions. Linked data may represent a paradigm shift in providing pathways to the essential data sources. It was commented that "…open data supports technocracies, and open science supports the public." Provenance management or informatics tools maybe an important issue to discuss. Data quality is also important as a means of inferring provenance. Using web processing services (WPS) to publish the processing that data goes through so that users may understand and work with them would be valuable. Other suggestions included the need for projection/mining techniques for customization and size reduction of massive data sets for specific user needs. Closely related to this was the concept of the virtual source librarian who may use a catalogue of interoperability services along with metadata catalogues to help identify specific data sets. How can a broker be used to support data access? Perhaps the equivalent of a Google search service for earth data is needed.

The draft breakout questions were accordingly modified to address the above comments and issues.

Breakout Questions and Reports

Group A: How can the GEOSS infrastructure be evolved to support climate research? (Identify gaps in current capabilities)

1. **Are there rich enough metadata standards to describe climate models so that end users can compare the output of different models?**
    - A European METAFOR project is addressing this. ESMF has metadata requirements. OpenMI
    - Subtask of GEO-GEOSS (Model-Web) also addresses question of needed metadata. Chair is Gary Geller of NASA/JPL
    - See PCMDI requirements for model metadata
2. **Are existing OGC services sufficient for serving climate data?**
    - Time issues are difficult
        - 360-day calendars, averages, baseline for anomaly calculations

        - Refer to NCAR GIS group for summary of time issues they have encountered.
        - CF addresses these items, but not many clients can use those conventions
    - Climatological statistics have similar problems and lack clients who can use the CF constructs
3. **Lower the entry barrier for scientific users**
    - e.g., use brokers to hide the complexity of the underlying data services
    - enable easy discovery of resources: virtual librarian; community information brokers (eg, EuroGEOSS)
4. **These resources have to be registered in GEOSS**
5. **Much simpler registration system is needed. Approach should be geared toward scientist users rather than IT people.**
6. **GEOSS should be linked with similar efforts: GCOS, GO-ESSP**
7. **Compliance tools for screening for adequate metadata**
8. **Ensure datasets are accessible via web service rather than just static web site URL.**
9. **Need to get more people to register their data.**
    - Provide incentives (e.g., publication credit for academics)
    - Make one's data more visible
10. **Need mechanism for contributor to specify scope and use of data. Provide tools for data provider to specify quality.**
11. **Provide Amazon-like ranking for users to comment on datasets.**
12. **Any formal review mechanism would require even more incentives for both contributors and reviewers.**

Group B: How can the informatics community help scientists to better manage and work with disparate data sources?

1. **GIS tools should <u>require minimally acceptable metadata</u>** for provenance mgt/analysis, data quality, etc., and <u>preserve these fields through stages of processing</u>. NOTE: OGC WMS, WFS, WCS, SOS, etc. DO NOT exchange these fields.
2. **Additional tools are needed for evaluating aspects or properties of fitness for use**: scale of data, uncertainty, accuracy, completeness, representativeness, consistency, bias, trust of source, trust of decisions made in processing, etc.
3. **GEO should track W3 Provenance WG (w3 prov)**
   www.w3.org/2011/prov/wiki/Main_Page
    - General approach; orthogonal but complementary to domains
    - Provenance represents relationships among contextualized entities and events that have occurred (working definition). Conceptual model (UML) is being defined, and a formal model (RDF)
    - Stakeholder questionnaire now being circulated (google doc)
        - Stakeholders: GEOSS, OGC, NCAR GIS, GeoViQua, EuroGEOSS, …
        - What are the use cases: simple reuse, processing, evaluating for fitness
4. **For GEOSS: Need strong business case: what capability could be added that justifies the time & effort? How will lives be saved or lost as a result of these measures?**

Group C-1: What questions can the broader geosciences informatics community help climate scientists address? Where are the misunderstandings? Who do we need to discuss this with? How do we, as geoscience professionals present value to the climate science community (motivation to change)?

1. **We need to <u>propose operational, vetted, and official solutions to get adoption</u>.**
   - Defensible, best available, trusted.
2. **Operational Data: Respected publisher, known procedures, reliability vs quality depends on purpose.**
3. **Conservative vs. Accurate systems.**
   - Multiple tiers of data. Quantity vs. Quality (crowd source vs official source)
   - Ensembles help to ensure quality is expressed clearly.
   - Difficulty with immediately available data and high quality documented data.
   - Trade offs between timeliness and quality. Automated systems can solve this.
4. **Data for Science, Archive: Quality is implied via review of systems and documented use.**
   - Data citation could lead to greater trust of data resources. (linked data?)
   - Data citation may be a pre-condition to tracking provenance. (DOIs) CF, GEO-GEOSS, Data citation guideline.
   - A formal system for feedback on datasets would be a good addition.
5. **Uncertainty: We have OK ways of expressing uncertainty but not communicating it.**
   - An area of expertise that is difficult to put into general formats.
   - Ensembles solve this problem roughly, allow people to interpret (not always good).
6. **Provenance tracking: Critical to open science, tracking provenance "openly".**

Which leads us to the next major question:

Group C-2: How can open science improve broader societal awareness and reasoned discussion of climate change issues?

7. **WPS implications**:
   - A user can see the input output and methods of the "black box"
   - Little control of what people do with outputs or how they use the service.
   - Do we have a responsibility to control how processing and output is used?
   - Deliberate mis-use can be less of a problem than ignorant use.
   - Open data and processing "could" be the best tool to build credibility.
   - Automatic "open" systems take away the contact and research involved in "closed" access to data and processing algorithms. The human interaction allows human-to-human caveats.
   - <u>Social networking concepts could be used to get the "human touch". Forums and bug tracking for data and services.</u>
   - <u>GEOSS could provide this system either in the GEO portal or in the data providers'.</u>
   - Difficulty in deferring to data provider's feedback system. Provide mechanism to link to data providers'. Keep it simple and lightweight.

8. **How to encourage "reasoned" discussion:**
   - Transparency is important.
   - Partition issues of objective data issues and interpretation.
   - Is there a place for "summary" or atlas level information in GEOSS?
   - <u>Outreach information needs to be evident with data</u>.
   - Need to provide mechanisms to express applicable scale and general application.

Summary Recommendations to GEO/GEOSS

To be continued (derived from above notes)…