Global Organization for Earth System Science Portals (GO-ESSP)

Collaboration designed to develop a new generation of software infrastructure that will provide distributed access to observed and simulated data from the climate and weather communities.

GO-ESSP 2011 Workshop Summary

The 9th Annual Workshop, May 9-13 2011
Asheville, NC
http://nomads.ncdc.noaa.gov/go-essp/

The 2011 Global Organization for Earth Systems Science Portals (GO-ESSP) conducted its 9th Workshop in Asheville NC, USA continuing its decade-long international collaboration for the coordination of services to the geosciences community

Overview
The Global Organization for Earth Systems Science Portals (GO-ESSP) is an active grass-roots unfunded community interested in leveraging emerging technologies in support of enabling climate and weather research. It is a self-organizing community that continues to develop and support many of today's earth sciences' most widely used technologies, vocabularies, and web-based applications with the goal of reducing costs and improving resource sharing by providing an open forum for the exchange of ideas and open source software and tools.

To take advantage of the confluence of individuals two other meetings were co-located with the GO-ESSP workshop (see Agenda in Appendix A): an Earth Systems Grid Federation (ESGF) meeting (see Agenda is Appendix B); and a Climate and Forecast (CF) metadata vocabulary meeting (Appendix C) each lasting a full day.

2011 Workshop Host
NOAA's National Climatic Data Center (NCDC) and the NOAA-National Operational Model Archive and Distribution System (NOMADS) Team hosted the week long workshop and this report captures both the presentations and notes thru web links to presentations and also it provides (mostly) unedited notes created by all participants via an on-line note-taking capability called "TitanPad" (http://titanpad.com/). Every effort has been made to preserve the original notes as scribed by participants therefore it is important to realize that only the presentations themselves represent the author's viewpoint. The associated notes herein represent participants' views. No attempt has been made to summarize the more "important" outcomes of the workshop since that is entirely dependent upon which aspect of user services individual participants represent. There were however many agreements that were reached for coordinated (non-duplicative) future work. Finally, this style of on-line note-taking contained in this report represents well the nature and spirit of GO-ESSP: resource sharing and community involvement to improve the quality of information for the better understanding of earth systems sciences.

For more information or to participle in GO-ESSP please consult http://go-essp.gfdl.noaa.gov/ or contact any of the GO-ESSP Principal Investigators listed on the cover page or in Appendix D.

Submitted May 27, 2011
Glenn K. Rutledge for the
GO-ESSP Workshop Organizing Committee

Global Organization for Earth System Science Portals (GO-ESSP)

Collaboration designed to develop a new generation of software infrastructure that will provide distributed access to observed and simulated data from the climate and weather communities.

GO-ESSP Day 1

NCDC's Glenn Rutledge opened the meeting and introduced the GO-ESSP Principal Investigator's (see Appendix D). Glenn then acknowledged Tom Karl (NCDC) in attendance and both Bryan Lawrence (BADC) and Karl Taylor (PCMDI) attending via WebEx and teleconference. Glenn then introduced Scott Hausman (acting- Director NCDC) who then welcomed the participants and acknowledged the outstanding efforts this grass roots effect sustains.

Glenn then reviewed the workshop focus. Since the t2011 workshop brought together over 100 of the worlds' top technologists and scientists (see *Participant List* in Appendix E) the focus of the workshop and the associated *Call for Participation and Abstracts* spanned the following areas of interest:

- Data preservation, data curation, and long-term access
  - Digital libraries, Digital Object Identifiers
  - Quality control
- Data provenance and usability
  - Model metadata, mark-up languages and semantic systems
  - Workflows and workspaces
- Serving data for climate science
  - Remotely sensed observations, *in-situ* observations and reanalysis data sets
  - Downscaled data
  - CF conventions and CF-OGC interoperability
- Next generation data services
  - Scaling data services to extremely high volumes
  - Automated systems
- Federation, interoperability, and security

The Principal Investigator's provided general guidance on the workshop structure and goals. The two day GO-ESSP workshop is designed to allow participants an opportunity to provide brief overview discussions rather than formal "presentations" regarding their data management activities that might be leveraged by others in the community. Open discussion and Q/A after each talk is highlighted below. Additionally GO-ESSP organizers stressed the importance of the afternoon sessions on Day Two that was a "demo-day" that allow participants to show their wares and discuss with their peers issues, concerns and most importantly advancing collaborations.

GO-ESSP Day 1

Tuesday May 10th, 2011 (Session 1)
Federation, interoperability, and security

Presentation 1

Luca Cinquini - Earth System Grid Federation: building a software framework of open source, modular components for analysis of large distributed scientific data

Presentation:
http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/GOESSP_2011_ESGF_Luca_Cinquini.pdf

Workshop Notes:

ESGF (http://esgf.org) is an open source project for unfunded groups that are wanting to share CMIP5 data. There is also a focus on observational data and use them to validate the models. Users want it to be more modular, more configurable. Big focus on web services, API for interoperability. Needs to be able to interoperate with existing organizations (NASA, NOAA, ESIP, etc). Tools in both Java and Python. Want the system to evolve towards a peer to peer architecture (p2p). The grid is populated with different types of data nodes. They are elastic and can leave and join without effecting the grid.

Q: what is the relationship between CF metadata, ESGF metadata and Metafor?
A: CF metadata covers physical quantities (fields and grids), Metafor Common Information Model covers also software entities (models, model components); ESGF metadata harvests both following CMIP5 requirements: vb)

Q: is the p2p-ness/pluggability/modularity being somewhat overstated?
A: There is no doubt some hidden dependencies all over the place! Second, is p2pness desirable?

Recent developments: metadata conventions for observations, modular security infrastucture, search service for data across nodes using Solr, web frontend, Live Access Server, exdpanded configurability, ESGF registry, ESGF dashboard, rich client access to ESGF services, and integration with OODT.

Security Services: The grid is composed of different authentication centers that allow for access controls to the data holdings for each of the centers. Uses: SSL, OpenID, PKI/X509, SAML (XML encoding for signing authentication/authorization). Implementers can now secure opendap servers whether Java based or Python based. Java Components for ESGF security: Identity provider: to register and auth users: Attribute and Auth Service: SAML Assertions about the users Used to add authentication to allow access to services.
Search: using Apache Solr, does full text searches quickly. The metadata does need to get encoded ESG is separated into two sides, backend and frontend.

ESGF wants to setup a "ESGP p2p Testbed" between the different groups in the federation
Ben D: EU is using a product Shibboleth, ESGF decided to not use it, yet.

Gavin Bell - ESGF: How to build an elastic distributed system over "Big Data"

Presentation:
http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/ESGF-P2P-GO-ESSP-2011.pdf

Notes:
Development: Different pieces of the ESGF have been broken into parts.
Node Architecture: 4 different configurations Index, IDP, Compute and Data. (set by --type in the install)

Node Manager: Consistent across all the node installations. All nodes deal with message passing in the same way. Uses the node manager to allow the process to be changed as it happens. Allows messages to be sent through nodes and by the nodes. Uses a gossip protocol. The communication allows for the elasticity. Lets nodes come in a and fall out of the grid. p2p is basicaly letting the nodes coordinate among themselves.

Core components: Monitoring, Metrics, Notification, Registry (keeps track of the nodes), Dashboard (gives a graphics interface for looking at what is going on and where data is flowing) Compute and visualization is via LAS: Ferret, CDAT, NCL. Data via OpenDAP, File download, and GridFTP.

Q: is LAS the compute and visualization service?
It can, but currently leverages CDAT as part of ESGF to do sub-setting and visualization, and a rich client to do science on the data.

ESGF is hosted at http://ESGF.org/ and is composed of 12 different projects.

Q: Search: will you chat the search across all the nodes.
A: Central search is done, not across the grid. Searches local copies for the metadata across the grid.

Plans for a map-reduce framework to look for best compute ready *machines* to do the work. The gossip allows the nodes to update their state and then pass it around to two others. It does this in log n. Security is maintained using the security layer even with the distributed p2p nature.

Feiyi Wang (ORNL), Galen Shipman, and John Harney - The Earth System Grid Federation User Interface

Presentation:
http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/harney_ESGF-P2P-UIV6.pdf

ORNL has had a lot of requests for searches across model and observations. The design for the interface is simple using spring, jQuery along with ajax-solr (MVC). The UI is customizable and looks like a standard site. It has standard *wigets* like login, search, etc. Has basic text search with auto-complete.

The ESGF UI allows you to store results also so you can save the results for later. Breadcrumbs to return back there where you are earlier. Facets for searches and are in the side bar accordion with counts. Also has a *facet browser* broken into category. Selections users make are stored in the current selection box. ESGF UI does temporal search with start and stop date range; and does bounding box and centroidal selections for doing spacial bounding. The current selection results can be removed by clicking on them. Search results have links for metadata, a feature to add the results to a shopping cart and to *visualize and analyze.*
Datasets added to the *data cart* and different files of the dataset can be selected. The items in the cart can be visualized and analyzed on LAS or another client.

Q: are key value pairs pulled from NetCDF attributes, or implemented in XML down the work flow chain somewhere?
A: I believe this is a TDS harvesting of netCDF attributes that are then converted to name value pairs and put into SOLR.

Q: are date ranges dynamically calculated from CF coordinate time variables?
A: […this will be…] future work: a Test and Feedback, vizGal for multiple datasets. Subsetting via OpenDAP. Currently download whole file via regular http.

Q: does query interface use OpenSearch [-Geo]?
A: this tool does not support sub-setting of data yet.

There will be different capabilities at different levels of "UI", some on the browser; some using CDAT/F-TDS/GDS; some using netCDF files on your local machine. The palette of options is going to be richer the closer to the metal you go. We probably need an agreement what services go at what level. e.g the Q about calendar support: should calendars, gridspecs, etc be supported at the browser level?

Nathan Wilhelmi - The ESG Gateway

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/wilhelmi-ESGF-gateway-v1.pdf

The front end portal for ESG [is called] the Gateway. ESG-G provides the gateway services currently serving 15TB/month through the NCAR gateway (http://earthsystemgrid.org)
The gateway is open source (apache 2). Uses Apache CLA to protect outside IP that gets added to the project. It is developed using Agile (2 week sprints). Uses Atlassian Suite and JIRA/Greenhopper, There is a strong emphasis on unit testing and peer code reviews. 1.3.0 has spring under the hood and give them several options like RESTful urls. RC2 is available on the website. 1.3.1 will offer REST for datasets, CIM 1.5.

Looking into developing model execution and LAS integration. As part of our ESG work, we have integrated the NCAR Command Language (NCL) as a backend engine for LAS. This is now a 64-bit version of NCL, which can handle much larger files in the analysis process. Working on providing parallel computation on the backend as well.

Improvements (currently working on them) SOLR integration for search, improved speed, metadata exchange and how that happens,

Future Plans: CMIP5 support, interfaces, high pri. issues, architecture and usability.

Q: what is Trackback interface about?
A: It is a display of model metadata.

Roland S.: Has the open source CLA changed?
A: No.
Q: Roland S. How is peer review done?
A: It uses Crucible

Q: John Caron: What is whitelisting of services?
A: This allows them to share particular data to the node/gateway users.
Q: Are the CMIP5 datanodes the same as the new data nodes that are coming online?
A: There are 8 gateways that access CMIP5 and other datasets as well. Going after both model data and obs. Using Cadis().
Q: What is the difference between this and Luca's?
A: The gateway has been developed over 4 years and has over 20K users. Managing over a Petebyte of data.
Q: Is it expected that it will interoperate with existing other systems that are being developed now.
A: Luca: Yes.

Philip Kershaw - Security Mash-up with the MashMyData Project: Delegation and workflows with OPeNDAP and OGC based services

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/KershawPhilipMashMyDataGO-ESSP2011.pdf

Workshop Notes:
MMD (MashMyData) uses OGC Web Processing Services, Pydap, SAML. OAuth
Proxy Certificates are easier to add for ESGF. MyProxy on ESGF is used differently than on normal grid systems because ESGF does not do delegation. Can authenticate an entire workflow (delegation) using a certificate. But if user comes in with OpenID, it needs to translate to a certificate. MyProxy Online CA does the credential translation service. The WPS retrieves a proxy certificate to access OpenDAP. All the user needs is their OpenID -- everything else happens behind the scenes.

Further work needed with OAuth, provides sublte differences in delegation that could be important.

Q: Luca: What are the paths that are not solid enough? [if] they are not sure about the passing the authentication

A: Gavin: Granting access to the OGC webservice? Need a mech. for discover like how Shibboleth is used at BADC.
This (discovery of user certificates) would enable security through service chains. Presently requires a priori registration. EGI and OpenID? ESGF needs CAs for centralized trust infrastructure.

Q: Caron: Does the OPeNDAP server need to know and auth all the users come in or can certain other centers be blanket access.
A: Roland: Are there roles? Yes, but they are left out of the presentation because of complexity.
A: Roland: Doesn't that take away the authorization per user. There is an extra step that is left out.

Benno Blumenthal - Using OpenID/OAuth to access federated data

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/BLUMENTHAL.pdf

Workshop Notes:
CMIP3 had a pydap server that had OPeNDAP access with basic authentication.
http://esgcet.llnl.gov/dap/ipcc4/?thredds

One flaw with the system is that you can't do mashup authentication. It can only authenticate one set of credentials. Basic authentication schemes (and even OpenID?) are susceptible to man-in-the-middle attack vectors. This also removes the ability to authenticate with a third-party. OAuth allows third party authentication. It is token based so you can pass a bearer token in one channel and MAC token over an open channel. Oauth passes the tokens around to make sure everyone is OK in the communication.

OAuth 2.0 will get installed in parallel with basic and digest authentication to allow unauthorized response gives information for authentication. With 2.0 callbacks go straight to the authentication service without the initial loop in OAuth 1.0. OAuth will join basic and digest auth, will that change HTTP. IT does have facilities to authenticate in different ways.
Cache Access is difficult with also allowing anonymous access as well.

Session 2
Data preservation, data curation, and long-term access

Eric Nienhouse - Data Management and digital preservation for arctic science: CADIS and Chronopolis

Presentation:
http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/Nienhouse-GO-ESSP-2011.pdf

Workshop notes:
A case study of arctic data using the SGS system and is a great case for digital preservation. "Arctic science meets digital preservation". Some datasets are unique and cannot be replaced, digital information loss is a real problem AON projects had varied needs, standards, and interoperability was difficult. Interviewing user groups was important for developing use cases and metadata profiles for CADIS.

Metadata profiles and editors are a highlight of the project. Used teams to curate and validate the metadata in the project. CADIS was an ideal candidate for a preservation pilot since it was smaller than 1TB.

Q: Is this a metadata editor in which users input information or is a display of data metadata harvested from files?
A: Eric says it is a GUI interface where users fill in about 40 attributes.

Chronopolis uses iRODS

Q: What is long-term plan for iRODS support (maintenance and continued support for use of iRODS framework in general)?
Community engagement had a powerful effect on contributing to the success of the project due to interviews to create strong metadata profiles.

There is capability to extend the facets and define what is harvested from the underlying catalogs. Formally created object rep. of the metadata then translated them to RDF. Solr has been a factor in helping with the metadata implementation. When they started the project Solr was not readily available. They are moving toward solr to work with metadata.
We used THREDDS metadata specification to help organize the metadata. To help map wildly varied data used CDM classed and IDV for examples. We also had to make sure there are lat/lon
---

Kyle Olivo - FREMeta: Efficient and flexible metadata re-writing

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/OLIVO_FREMetar_Efficient_and_Flexible_Metadata_Rewr.pdf

CMOR, for the uninitiated, is the PCMDI-supplied tool for converting netCDF files to the CMIP specification. FREMetar has a command line interface that will break things into chunks. The user provides a source directory. Those get moved to HPC, check to see if data is needed, gathers stats about missing values, compares metadata on the files, records the stats and copies the file back to the destination. This is primarily used with AR5 data.

Q: Does support multiple specifications (other than CMIP?)
Q: Jerry Potter: Are you going to make this available to other sites?
A: The system in its current state is not transportable. May be possible for version 2.0. Main reason is that the backend is a GFDL-internal RDBMS which is probably not portable.

Q: How are you verifying that you make no CMOR compatible metadata,
A: There are checks on the source data to make sure the data hasn't changed. It doesn't change the data only missing values to meet CMIP standards.
Q: Are you making any use of NCML?
A: No.

Russ Rew - Updates on Unidata Technologies for Data Access

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/rew-2011-goessp.pdf

Originally there was not a NetCDF standard. There was a lot of movement in the standards area. NetCDF was endorsed as FGDC (US federal) standard. Also an OGC core binary encoding standard. NetCDF/OPeNDAP allows subset access using DAP and faster than whole file access like FTP. netCDF uses a dispatch layer to isolate the lower format layers and piping them into the dispatch layer then to netCDF then to the application.

Q: Is the Jira instances open to everyone or closed to project developers only:
A: [response missing]

Oliver Clements - Production of a search and browse interface for an environmental science thesaurus

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/clements_go-essp-nvs.pdf

Workshop Notes:
This is a very active site. Results are given in an HTML table, very simple. It does have a modified date. The terms are links, but no browse ability, you get an XML snippet. Good for machine usage. Visited users with the problems and documented their complaints.
Other organizations decided to stop using them because of the difficulties with using the thesaurus with their thesauri.

New version need for end user consultation, Focus on NERC initially. A simple search, like Google is needed. The new results page is paged and loads quicker for smaller devices (ie. mobile). New results contain the source thesaurus. Now shows the term metadata and it has a key, preferred label, description and last modified version which is now correct. Future versions may have changes for the term like a repository. There is also quick links to related thesauri. It now has a cross-walk, and within a few clicks it allows easier and faster browsing of terms. The new version has caching to speed it up.

Future Work: Search filters (pre and post), visualization of concept relationships. We want to add a layer on top using SPARQL backend to speed up queries. They also in the future want to release example client code for the API.

Stephan Kinderman - Next generation data services: c3-INAD goes ESGF

Presentation:
http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/kinderman_go-essp-2011-c3inad.pdf

Workshop Notes:
Plans for integrating the national infrastructure into ESGF. C3Grid received climate community specific funding. Developed an integration pilot of the climate centers. The project agreed to use ISO 19139 for metadata. Next C3-INAD Grid = C3Grid with ESGF. Need to redesign their C3 code infrastructure because of its use of Globus. The middleware manages the clients, and decides when and where it is to be done. By transferring slices in the DSpace or Data Space. C3Grid Portal uses MyProxy to authenticate. The metadata from THREDDS is harvested in the C3 portal. Database is used for search.

Next we plan to add C3 INAD data stager to fulfill request and creates wget scripts. Data staging involves sub-setting and composition of data from the datanode and optionally format conversion. Initially this is done in C3-INAD, later also using OPeNDAP to data nodes. Caching is supported by the GNDMS data management middleware of C3-INAD. GNDMS also manages data lifetime (old data ages off..)

Next steps: multimodel and multiensemble workflows, climate scientist to do use case driven workflows and ESGF data integration.
Q: Will this be Shibboleth?

A: No. Main use case is for a portal and not rich client integration.

Stephen Pascoe - Maximizing the utility of OPeNDAP datasets through the NetCDF API

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/GO-ESSP_2011_spascoe.pdf

Workshop notes:
These are two ideas he has about the future and how the way that OPeNDAP is being deployed with the NetCDF. The applications can talk to ESGF Security, local files, etc. The users may not have direct file system access. Performance for client access to NetCDF resources is being further developed for greater throughput.

The OPeNDAP Test Framework will allow tools to grab OPeNDAP request, load testing, benchmarking. We tried it with several OPeNDAP servers and measure the results. For example: ask it for 45x45degree frame from a single file with a single variable:
CDAT had too many request and needs to be fixed CDO has a lot less but still more suited for the file system. Tested on servers: PyDAP and TDS and similar, but Hyrax is much higher and depends on the platform. The machines are able to talk at close to the limit of the pipe they are connected to, with TDS performing the best.

DAP response tiling/chunking: it uses tile cache for DAP request similar to Google maps tile cache for Google Maps like in WMS. OPeNDAP is in theory cacheable and could have a big impact on performance. Doing this on dynamic datasets this could make a large improvement because the server could cache things like means for the dataset.

Q: Who decides how tiling is done? Is the tile size/shape optimized by the server? client? hints in data?
A: There is no set way. Not all server-side processing can be expressed as a URL, but how do we keep them RESTful and in the spirit of OPeNDAP? OODT is a project that could have ideas, but is not RESTful.Web Processing Service (WPS) could help but too has some standards overhead. There are steps being made toward integrating WPS with OPeNDAP.

Q: Don't WPS servers die under polling requests?
A: It depends on implementation of server, polling frequency, maybe other factors. Good designs should not fall over.

Session Three

Serving data for environmental sciences

Giri Palanisamy - Metadata Standards for in-situ Observational Datasets

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/palanisamy_metadata_stds_GO-ESSP_May_2011.pdf

Workshop notes:
Many agencies had data and wanted to add it to the assessments. The goal is to increase the number of observations and make them similar to models. The community has been working on metadata conventions for obs/in-situ. All must pass CF and CMOR checkers; use the DRS specification for files and their file system structure.

NASA selected the best products to be propagated through the ESGF. ARM is to review the interactions of areosol and cloud, etc with various research sites and measurements collecting over 200 measurements. From this large collection of data they have selected products for ESGF based on CMIP5 categories (cloud diag. and monthly mean atmos fields and surface fields). We created NetCDF files and WaterML from these.
The AmeriFlux network consists of volunteers at over 142 sites in 5 countries. The network datasets have unrestrictive rules to be active with AmeriFlux. However, this creates a large variety of data that they are producing. This creates a huge problem for processing the data and bringing them into a common format. Sites used to change their formats year to year and takes a lot of time to readjust tools for the data. For met data they have 4 levels of data:
Level 1: native data,
Level 2: has QA
Level 3: like EU network Level4:

The data gets pushed into something very similar to DRS, and uses CMIP5 or CMOR variables. This is station based data. The file level naming includes instrument. Identified global attributes for the in-situ datasets and those included are in NetCDF file format.

Observational datasets have many metadata standards used. (FGDC, DC, ISO 19115, 19139 and DIF) Using these standards offer provenance, quality information , keywords, citations Service level information and hierarchical metadata fields and support for data discovery

Comment: Steve Hankin: NOAA PMEL is using NetCDF standards also for their metadata and wants to work together.
Q: Are you working on profiles?
A: Yes, they are defining the core metadata fields.

Aparna Radhakrishnan - NOAA/GFDL - Model Development Database Interface (MDBI)

Presentation:
http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/GO-ESSP-Aparna-GFDL-pdf.pdf

We are using CM2.0 and CM2.1 for AR4 but with AR5 there are a lot of additional models. Moving from 12TB to no less than 300TB. AR4 used CMOR AR5 will be using FREMor. Before teh tracking process was used with phones and email, Now MDBI is being used to track QC with AR5. MDBI is a transparent view of the curator database in a user friendly way. Uses ExtJS javascript framework with JSP on the back end with MySQL for DB. Allows for an administrative view with the ability to hide data that you are testing.

The curator role: This change in the different levels of model output, experiment information, ARr5 variable mappings, FREMetarized files, the QC service is complete.

QC: login is via NEMS/LDAP, then navigate to the correct experiment and check the global attributes. Users can categorize the variables in the CMIP5 tables. There are two levels of QC that are asked for, file level (checking DRS, variable name, CMOR min and max, mapping to the CF name, conversion of units (ie. deg. C to deg. K)) They can exchange comments on variables, but not read anyone else's comments. You can see who entered the comment. You can back-track comments and variable QC by u-nchecking. They want all the variables "checked and green". Want to enforce doing QC in the interface rather than another way of tracking it.

Q: Where do the max and min come from?
A: It is compared to CMOR Trying to say that the data in the model is what it says it is, not measuring the quality of the model.

Jianfu Pan - Continuously Enhancing Usability of Remote Sensing Data for Climate Models

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/PAN_GOESSP2011.pdf

Workshop notes:
Three 'user' groups: User, Data and Technology.
Data: NetCDF (most used), HDF (most common at their site), Custom Binary, and ASCII The data is access using services and has some preparation task: sub-setting, re-gridding and projection/interpolation (harder than it may seem). Quality filtering and format conversion

There are many packages available to review data, but you often have to use special data formats to use those tools. On-the-fly web services: server-side data preparations, Rest-like url’s, format conversion (OTF conversions).

Data Quality Screening Service (Quality Filtering) is for level 2 Satellite data that often comes with qc flags, users used to have to write their own qc to filter.

Other technologies are being integrated, IDV, Panoply, Pomegranate.

Giovanni is an system that integrates data prep, anal and viz int services and workflows with simple interfaces for the user. Handles everything for the users: feftching, etc. Coming to Giovanni allows a workflow for fetching, sub-settting, re-gridding, etc. Can come from external data sources. Giovanni is being re-engineered: true service oriented, community based, interoperable with other services such as data download, customized climos (user constructed), provenance and advisory aspect (part of Giovanni to help users to anal the data based on the know knowledge of the data and help the user analyze it.

Jean-Yves Peterschmitt - Paleoclimate Modelling Intercomparison Project (PMIP) Phase 3

Presentation:
http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/PMIP3_GO-ESSP_2011_peterschmitt_v3.pdf

Workshop notes:
We have CMIP5 data from paleoclimate periods. This is the third phase. 1st phase is just atmosphere models with AMIP variables names, and FTP. 3rd phase will be fully CMIP5 compliant. Paleo Experiments: we will check how well models perform with unusual boundary conditions and long term experiments Can compare these.

{…we have to…} change the boundary conditions of the models drastically, along w/ orbital parameters, trace gases, veg. etc. (Larger ice sheet than today as high as 3K meters). PMIP3 DB: paleo model data will be in the official cmip5 db and a subset will be mirrored at IPSL. Some of the PMIP3 participants are some of the CMIP5 participants, but the others are not in the CMIP5 archive and store them at IPSL.

We also have tools for non-CMIP5 experiments, but are in the PMIP3 database at IPSL. Most of the data is already available, but not “CMORized” (CF, NetCDF compliant). The new data will arrive in the DB mid 2012.

We have deployed a ESGF datanode to distribute our CMIP5 data. There is 4 Petabytes of data. We will deploy an ESGF datanode at the CEA HPC. Those centers will have the core data (CMIP5, PMIP4 AND *MIP) along with operational data. Working currently on OpenID, a contribution for the ESGF stack is mirroring a subset, non-cmip5 models/experiments documenting with METAFOR. At IPSL they are in charge of one of the METAFOR packages. Customizing LAS for distributing model data to non-programmers (climate proxy data community)

Q: How much has been done on distributing to the 14 groups and how much data?
A: There will be a lot, and it may be reduced because you don't need all the data. You may only need monthly means or pre-computing.

Eric Stephan - Leveraging the Earth System Grid for Integrated Regional Earth System Modeling (iRESM) Research

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/tues/stephan_go-essp_51011.pdf

Workshop notes:
Developing re-gridding of data for the region. Then build it for agr, hydro and other data models. How can this be leveraged for earth system grid? How do we save ag or hydro data and what if it doesn't fit in the netCDF model? There are significant challenges in spatial scale, variability, temporal scale, etc. Needs server side processing (regional processing), programmatic access for ad-hoc searches from an API combined with data retreival. Provenance in existing data and how to capture the intermediate results and convey how much we trust the results is being developed. We want to use standards with community buy-in. Since we capture the intermediate results then take the RDF mapping and tie them into ontologies and building them into rules that are already established. We are 18 months underway and are just getting started. We want to advance data and meta data standards to the climotological community.

Q: What are you thinking about using SWAP(?)) for?
A: Hoping that a lot of the efforts we are building are for understanding the relationships with models. Some model integration has to do with synchronizing time steps, or building new models that algorithmically simulate better. Semantic web services as a way to describe the data, self describing.

Day 2: Wednesday May 11th, 2011
Session Four: Next generation data services

Jeff Daily - Parallel Analysis of GeOscience Data: Status and Future

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/DAILY_GO-ESSP-2011.pdf

Workshop notes:
Motivation for the work is big data reaching PB size. GCRM using global geodesic grid. Large sizes take up to 40 days to reach off of a disk at 300MB/s. They use parallel IO (Parallel NetCDF, NetCDF4/HDF5).

Sub-setting the Geodesic Grid is unordered so they need to be indexed. The subsets are masked based and all the edges have an index(?) Patterned after NetCDF operators (NCO), but created own parallelized command line tools (Pagoda command line tools) for unstructured grid subsetting. The files are large so they don't concatenate, but use aggregations. With 19 files at 8.5 GB each they noticed with a 4 core version of Pagoda that they have much better performance. Going from 15 variables to just 4 it scales. Scalability depends on dimension order and data distribution. Cautioned to be aware of your dimensions. Some cores threw out data and when those cores got turned off they had better performance. Using TAU profiler to look at I/O and it was 2/3 I/O.

Plans for a Python version of the libraries using Cython. Plans to hide I/O latency by mixing IO and computation. Need more users and user input. NCAR is using it for a nightly script.

Q: Does it work on other grids besides Geodesic?
A: We have tried it, but feel it would not be too difficult to support additional grids. The tools originated with the geodesic grid, but they do have some elements of other ugrid efforts.
A: (follow-up by author): I think the Argonne data was cubed-sphere, but I will double-check. I have tested against the GCRM, of course, but also the sample regular grids provided on the netcdf website.
http://svn.pnl.gov/gcrm/wiki/pagoda
Discussion Group: **http://groups.google**.com/**group/pagoda-dev**
They try to mix parrallelism, but IO is a problem.

Don't use OpenMP, but MPI rather with Global Arrays from PNNL.

Luca Cinquini - A Scientific Workspace environment for collaborative analysis of climate data

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/GOESSP_2011_COG_Luca_Cinquini.pdf

Workshop notes:
COG - is a 3 year project: Research, experiment and report on web applications.
Want to mix services, social communication to enhance collaboration.
Have a data workspace to do analysis (LAS, etc) and then a web environment to share information. This services acts as an indexing layer so projects can discover other projects.

History: Metadata infrastructure was created for an NCAR workshop in 2008. This workshop compared atmospheric dynamical cores. The workshop was supported for NCAR. It allowed them to compare results in a simple way.

Present: The CoG workspace is planned for use in a graduate and post-doc workshop in 2012 at the University of Michigan similar in structure to the 2008 workshop. CoG can be used for Generic Model Inter-comparison, collaborative data analysis, and hosting of applications that generate derived data products, coordinate development of multi-component models, etc. None of the existing applications combines data and metadata service with collaborative tools and project governance. The software is built upon a community framework, Django. Django Includes ORM API, RDB and is WSGI compliant. Development started 4-5 months ago and was focused on collaborative tools.

Currently, the software has the capability to host projects and to represent the formal relationships between projects. There is a project browser and a structured layout for governance (these will eventually become standard templates). Also has newscast to send messages between projects and pages that can be commented on, facilitating discussions. Each project can create arbitrary pages using a backend wiki (with standard mediawiki hooks). There will be configurable templates and automatic menu creation already exists. A faceted data search has been integrated to a resident ESGF node. This capability should work with any resident data service.
Future work: 1) Complete the integration with data services for search, (etc) and integrate with LAS, TDS, ESGF etc. and 2) explore for work with National Climate Predictions and Projections project (NCPP) and OpenClimateGIS initiative (based on geo-django) and develop a metadata processing pipeline and support the DyCore workshop in summer 2012.

Q: Does it integrate with the repositories?
A: Currently it is linking to a local repository located on the same server. Ideally we would place this layer on top of an ESGF data node. If the search on that node can see other nodes, then we should be able to link to other repositories. This is an active area for development.

Martin Juckes - ExArch: Climate analytics on distributed exascale data archives
Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/juckes_exarch_goessp_may2011.pdf

Workshop notes:
Funded by the G8-exa-scale research initiative we are working with partners in Canada, Japan, Germany, France and others. It is a research project, but will support some the development of some infrastructure which will be done under the GO-ESSP/ESGF framework. We are working to take calculations to the data. Exa-Flop computers using GPUs is rapidly increasing, but data movement is not. With CMIP5 (5PB) -> CMI7 (16xxPB's. It uses Climate Data Operators (CDO) behind web services.

Frederic Laliberte - Exascale Climate data analysis from the INSIDE out
Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/laliberte.pdf

ExArch work package3 Cutting edge climate diagnostics. For users they need to download the data and it takes time and bandwidth. Requires perfect data and well represented numerics. UofT will create diagnostics using simple server-side processing framework. Will monitor OPeNDAP with CDOs. Ideally we would like to the some query, process it and reduce the size. If that takes too long the user will decide to just download the data and do the reductions locally. If you have 6hr lat/lon hybrid and need a large size reduction down to lat-theda will be inaccurate if you are not hi-res. 160x320x60x1500x#years --> 160x128x128
Other diagnostics are based on EOFs, Tropical diagnostics of inter-seasonal variability that relies on the analysis of space-time spectra. Both methods requires long times series over ...
With server-side processing modeling groups would make the development of diagnostics easier and timely. Providing derived data from the native grid will also reduce numerical errors and improve inter-comparison.

Rachana Anathakrishnan - Globus Online (GO): A hosted data transfer solution for climate scientists

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/Rachana-GlobusOnline-GO-ESSP.pdf

Focus is on distributed and open systems design/access/distribution using the Globus toolkit (ie ESGF). They have build-a-grid. Uses GridFTP and provides fast secure extensible standar and robust FTP based data movement. There are a lot of challenges for end-users (firewalls, configurations, mutliple providers/authentication. To overcome these they have created Globus Online. Has "fire-and-forget" data movement. Has 3rd party transfers and downloads. When moving multiple TB Globus online will keep track of transfers and can recover if there are faults. Performance optimized for you and autodetects the types of data you are moving and tunes it for you. When dealing with multiple security domains globus will help with that. There is even support for expert operations and offers support by looking at the transfers that are going on. There are 3 interfaces (Web, CLI, and HTTP RESTful interface). The CLI is a custom version of an ssh client.

Offers Endpoint Management for public endpoints with logical names, uses default cred. service. and Transfer Management: recursive transfers; Levels of synconization...
They have a very lightweight install that can be installed in 2 clicks. Ideal for laptop setups where you don't want to setup a gridFTP server, but don't want to go through all the trouble.

GlobusOnline does not have certificates, but ideally made for laptops. Will be released with ESGF Gateway, integrates with other ESGF tools.
Has two transf. paths http and ftp. Supports login via Shibboleth and OAuth.
With GO you can use your ESGF creds to login to GO. Focuses on transfer and sharing data

---

Reagan W. Moore - Policy Based Data Management (iRODS)
Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/MOORE_GO-ESSP-iRODS.pdf

Integrated Rule Oriented Data Service - Allows policies for data at each collection site. Each policy controls the execution of a workflow. The output of the policy give a state and that is stored in a metadata catalog. Each of the providers of the data have a archives where they have assembled the data. The properties of the stores is located at each site and get to decide what goes into each of the shared collections. Requires a consensus of the providers on what is shared. Have many PB size grids, NOAO, CyberSKA, etc. iRODS is implemented by putting middleware at each of the storage sites. Clients (48 so far) that access the datagrid will be redirected to where the data actually resides. The results of this is stored in a metadata catalog. They provide multiple levels of virtualization so they can offer the services to many different clients and the clients are independent of the data. Can be stored across many different types of storage and file systems. There are 71 policy enforcement points where policies are applied. Can be used to check for errors and is highly controlled. Local rules control access to local storage. iRODS is highly extensible: selection of clients, policies and procedures for the type of data that is being stored and controlled. It is open source software via BSD.
iRODS can be used with other grids that do not have iRODS.

Robert Oehmke - ESMF Fast parallel grid remapping for unstructured and structured grids
Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/pres_1105_goessp_oehmkev2.pdf

ESMF regridding - flexible, accurate, portable, parallel and fast, community developed. The regridding utilities supports SCRIP from grid files or custom ESMF unstructured format. generates NetCDF weight file format, comes with source. Can either generate interpolation weights from NetCDF files or during model run. Currently handles regridding of mosaics of grids via unstructured grid, but will ultimately use gridspec. Supports global 2D logical rec. grids, Regional 2d logical rectangular. Support for cartesian x,y.

ESMF unstructured format describes the connections between the elements where SCRIP format does not. Interpolation types: Bilinear, higher order patch recovery and first order conservative. Supports masking (only logical rectangular grids). Runs test on 20+ platforms a day and they check the interpolation error and the compilation error. Checks the weights for accuracy. Performance is good but as cores are increased the performance flattens out (is still much faster than serial solutions). There may be an issue with parallel IO. Takes about a minute to do 10800x5400 lat lon to 1440x1440x6 nasa cubed sphere with 96 cores. Plans for support of GridSpec

---

Jay Hnilo - NOMADS and the National Climate Model Portal (NCMP): Science and Data Management Services

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/HNILO_GO-ESSP_NOAA_National_Climate_Model_Portal_Final.pdf

The primary goals of NOMADS/NCMP - have been consistent over the last 10 years: distributed format neutral access to high volume models data input and out. Now there is a priority to NOAA's Reanalysis output--- CFSR and Reforecast (Saha et al); ESRL's 20th Century Reanalysis Project (Compo et al)), and derived subsets of GFDL and other IPCC AR5 contributions as coordinated w/ the US CMIP5 archive at PCMDI (Williams/Taylor). NOMADS is a founding member of GO-ESSP and NCMP will be constructed as a service within the NOMADS framework.

NOMADS is built upon (mostly) open source libraries and tools, for distributed data access using community software and leverages resources wherever possible. Several key data application servers are at the heart of NOMADS: Unidata's THREDDS Data Server (TDS), PMEL's Live Access Server (LAS), IGES/COLA's GrADS Data Server (GDFS). New (updated) GRID technologies will also be implemented as a service in NCMP to include DOE's Earth System Grid Federation (ESGF) and GridFTP services. Prototype data management tools are also being tested such as UNC's DICE program's iRODS under the direction of Regan Moore. Data staging long time series some of the very high volume data is routinely performed under NOMADS, and must continue given the extraordinary data volume growth estimates. Data reduction policies are currently being explored to remove old fcst data; and saving only analysis or model restart files to overcome the costs of high volume data such as these. NOMADS OPeNDAP saved 80% based on a study by the NWS (National Weather Service - NOAA). CFSR (NCEP's Climate Forecast System Reanalysis) continues to be an extremely heavily access data set. Last year NOMADS serviced over 125million downloads w/ 87,000 unique IP's and a one day record serving 4.7TB- this includes subsetting operations therefore reducing the actual bytes of data users actually needed to download. Also- Grib2 has a 10:1 compression. Access statistics are already being exceeded given the addition of the NCEP reforecast products. NOMADS supports reanalysis.org, UAF (United Access Framework) and GIP (Global Interoperability Program) Emphasis for NCMP is water resource management, and the energy community . One dataset (CFSR) is .5 PB. NOMADS is the storage facility and NCMP is the data discovery mechanism. Created Flash components for THREDDS for WMS, WCS along with navigation of THREDDS catalogs and Multigraph (http://multigraph.org) Will offer online climate model analytical engines using LLNL' developed "Climate Data Analysis Tools" (CDAT) Studies of variability, on-line pre-computed indices, and diagnostics will also be part of NCMP.

NCMP will be offering information on climate variability. Work with GIS users to NetCDF using tile information and representing them in NetCDF. Many of the datasets have no datums so mapping features can be off by 5-50km.

NCMP is working with USGS Center for Integrated Data Statistics (CIDA) who have implemented upload of shapefile, computation of stats on gridded data as a OGC web processing service, using TDS, OpenDAP, and custom components.

Steve Hankin - Unified Access Framework (a pretentious name for a simple idea)

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/Hankin-UAFatGOESSP2011-v2.pdf

UAD is all about enterprise wide integration of data and it's a very difficult problem. Sharing helps, but the people who are doing this are making solutions for themselves. NOAA has many different viewpoints on data and a solution is to make a system of systems. Later became GEO-IDE. Seed funding finally came last year. The traditional approach didn't seem to be working. Rather than repeat that traditional approach a more agile approach is being taken. "Don't solve problems, copy success." UAF decided to copy gridded data products to create a powerful interoperable platform. Not all CF, and often unaggregated. Metadata is weak or minimal. Often "trash" files end up being served in catalogs. UAF has the concept of a clean catalog that is well formatted, has metadata and aggregations. Reaching users with their pref. tools (MatLab, ArcGIS, IDV, Ferret, LAS, Google Earth, Godiva2, ERDDAP, R) ERDDAP is strongly RESTful and allows for easy access to R, MatLab, etc. Have created a website with a set of "How-To's" for using the different tools. There are also ways to access the data in THREDDS via views and also access the metadata via ncISO. Currently: evaluating *mature* discovery tools (RAMADDA, Geo-Portal, GI-CAT) All ways to crawl the UAF clean catalog. Roland wrote the catalog cleaner (another THREDDS crawler). Working to make it more automated or highly automated with less hand holding. Latest version re-creates the entire THREDDS tree and allows access to other viewers via OPeNDAP calls to the underlying catalog. Working on in-situ obs collections with CF discrete Geometries spec, ncStream (cdmRemote), ERDDAP, LAS, IOSP for data base access, NCMP aggregations of 1d file collections.

UAF approach: a way of organizing integration that is simple, open, cheap, compatible (ESGF, NOMADS, IOOS, Ingrid, Giovani, OGC) Should be a broader topic than just NOAA. Many OPeNDAP servers have little or no documentation. Request for docs to host on your CF app. With rapidly changing data where sets come and go then UAF may not be a great a solution, but through working together we can find a way to work together to create clean sets.

Antonio S. Cofino - The ENSEMBLES / Unican Downscaling Portal

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/Hankin-UAFatGOESSP2011-v2.pdf

The Ensembles downscaling portal allows friendly statistical downscaling. These needs to be defined: Predictors (large scale reanal fields), Predictands (local vars).

They have daily observations, Reanalysis (4d global coverage), GCM scenarios (climate change) Orignially for season simulations. Some projects supported are forest fires, health, impacts, integration (impact on hydrology, crops, economy), metadata for GCMs (metafor) They have created several web services including downscaling. You can select the downscaling method (regression, analogs, weather typing, etc.) obtains cross-validation in present climate. Should not be used a black box so correct software is used with the data produced. Very flexible and will do more than just GCM output. Integrating METAFOR services for downscaling metadata. Plans to incorporate as many possible downscaling techniques. MERRA is not available yet, but is being tested and will be available soon.

Alison Pamment (BADC) - CF Standard Names

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/pamment_go-essp_2011.pdf

Large growth in standard names since 2006. CMIP5 has requested a large number of these -- has contributed to a large amount of the growth in # of std. names.
CEDA Vocabulary Editor. Keeping subversion repository of xml.

PI note: extensive discussion lead by Alison occurred during "CF-Day" that occurred at the end of the week.

Rich Signell - The US-IOOS Modeling Testbed Cyber-infastructure: Unstructured Grid Standards and Standards Based Tools for Analysis of Ocean, Atmosphere & Climate Model Data

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/Signell-2011-05-11_GO-ESSP_Asheville_v2.pdf

The US IOOS is across the US and has federal and state government, academic institutions working on ocean observing and modeling. The IOOS Modeling Testbed groups are broken into three groups Chesapeake Bay (estuarine hypoxia), Gulf of Mexico (shelf hypoxia) and Coastal Inundation on the Gulf and East US Coasts. Focus on toolkits for the scientist, that are flexible and powerful for analysis. Use common scientific analysis environments: Python, Matlab, R, IDL, etc. Focus first on Matlab because used by 80% of oceanographic community.

Takes different model output and change them using NcML through the Unidata Common Data Model in NetCDF-Java and out through web services and finally into standard clients.
NCTOOLBOX for Matlab is a google code project: http://code.google.com/p/nctoolbox
Able to do comparisons with 5 different models for Deep Water Horizon. Works well for the structured grid data, but want to be able to handle unstructured grid data.

Want to handle the ugrids with the same workflow as the other gridded data into THREDDS. Had NOAA/Unidata workshop back in 2006 where people said the really need a standard for ugrid and now they have a google group http://bit.ly/ugrid_group, http://bit.ly/ugrid_cf (netcdf java), http://bit.ly/ugrid_git (git repo), and http://bit.ly/ugrid_m (Matlab toolbox). One driving factor on the standard was to design it so that existing grids could be modified to work with the standard using NcML. A UGRID class has now been added to NetCDF-Java, and also a UGRID class for NCTOOLBOX. Searching: You can harvest the THREDDS metadata using ncISO or GI-CAT and then plug it into matlab and pull out links to get DAP links, etc for use in Matlab.Plans to do server-side subsetting more ugrid methods for matlab, ESMF, subsetting with THREDDS.

Rob Raskin - Mapping CF Standard Names to the SWEET Ontology

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/RaskinSWEET.pdf

Developed by NASA but includes a lot of earth sciences, units, space-time, and quality. http://sweet.jpl.nasa.gov/ Has 8 high level:Representation, Process, Phenomena, Realm, State, Matter, Human Activities and Quantity. Now there is sweet 2.1 (where state was added) Added Roles, color. size, equilibrium, type activity, level, connectedness, impact, substance etc. Has 4400 classes, 2200 individuals and 600 relations. CF names are long strings that are joined of all the different attributes of the parameter. In SWEET it is broken into multiple attributes or semantic representation: Quantity, transformation, state, substance, medium, process.
Work is going on to map the SWEET ontology with CF (2150 done thus far).
Satellite Observational Data is not well defined in CF (Spectral ranges, Source)

Alexander Pletzer, Ed Hartnett - Progress on LibCF including support for Gridspec

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/PletzerGOESSP-May2011.pdf

CF 1.5 wants all the data stored in a single NetCDF file. Mostly covers lat/lon grid, each var has assigned attributes such as standard names and units, grid and data live in the same file. WIth traditional lon/lat grids there are issues with lat/lon spacing goint to 0 at the poles and therefore numerical stability with explicit schemes. Over resolution at the poles is a waste of resources. Mosaics share a tile and fold in funny ways, but don't always need to be cube grids. Mosaics have more flexibility in indexing than curvilinear and more regularity than ugrid and geodesic. Mosaic files contain the connectivity to the different sides of the square grid. libCF is written in C to allow it to work closer with NetCDF Issues with field staggering not in CF, CF assumes fields are nodal. Can it use cell methods for curvilinear grids: possibilities: super grid, rely on the dimensions and dual grids.

Presentations provided as background for demo's later in the day include:

1) Jennifer Miletta Adams & Brian Doty (IGES/COLA): "New GIS Interfaces in GrADS"
Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/ADAMS_DEMO_GO-ESSP_2011_JADAMS.pdf

2) Roland Schweitzer et.al: "F-TDS: A General Purpose System for Server-side Analysis of Earth Science Data"
Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/roland_F-TDS-DEMO.pdf

3) Roland Schweitzer et. al: The Interactive Earth Science Data Visualization Gallery (vizGal)
Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/roland_vizGal-DEMO.pdf

4) Steve Ansari (NCDC)) The Weather and Climate Toolkit
5) Sylvia Murphy (CIRES), et al. Implementation of CMIP5 Metadata in the ESG Gateway
6) Luca Cinquini (NOAA/JPL) A Scientific Workspace
7) Rachana Ananthakrishnan (DOE/ANL) Globus On-Line
8) Antonio S. Cofiño (University of Cantabria,)UC downscaling inter-comparison Project
9) Reagan Moore (UNC/DICE) iRODS
10) OpENDAP James Gallagher

End of GO-ESSP Workshop

Earth System Grid Federation "Day"
Monday May 9th, 2011

8:30 am - 5:00 pm Earth System Grid Federation "Day"

Presentation:
http://nomads.ncdc.noaa.gov/go-essp/presentations/esgf/ESGF%20Proposal-2.pdf

Most of the discussion time centered around the need for resource sharing and coordination of efforts now that ESGF is an open source Federation. The group emphasized the need for tightly specified interfaces -- and profiles (such as data publishing profiles for THREDDS) that are well defined as well as defined file formats and API's.

Using version control for different interfaces. Release schedule and deployment facility for the different interfaces. Open source is a key tenant, but ESGF needs a development process. Now that many of the ESGF nodes and gateways are going operational it is important to keep a more operational mindset. The nodes and gateways need to stay up or have failover.
The development process is key, but it must be well documented. There is a need for institutional collaborations. We need to think about money sources for supporting resources, but the institutions need to know what they are getting. Security has well defined documentation using the Interface Control Document (ICD). This is to be used for a template to help the documentation of the other interfaces by creating a questionnaire.

The different organizations have different needs, but need to have strong documentation for the interfaces. Interfaces having a much stronger commitment rather than delivering software. *Make the software support the interface*.

Does there need to be a working group for Implementation? Are existing processes that will work? OGC process: structured, made for collaboration across institutions (1 annual testbed has 10 national agencies sponsoring activities by 30+ orgs) OGC testbed process could be adapted to make things work for non-OGC groups like ESG (similar to OPeNDAP, NetCDF) look to OGC for maintenance

There does need to be a balance between specification and software. Spending a lot of time on details for specifications may not produce a lot of software, but will gain stronger integration. Documenting the interface is documenting how it talks to the other side. ICD is a good interface, but only recently due to stronger documentation. ICD is a good example of this it needs to be replicated. The documentation process needs to part of the release process. Documentation leads the team to create a more holistic view, but how can this be done? More physical meetings and the need for creating drafts? OpenID and ICD are successful examples, but need to be

extracted. There were two different implementations so they had to work it out. Build the use cases around the needs for interfaces; sometimes you don't have a big enough picture.
Define the interfaces and functionality

Look at OGC Process as a framework or guide. Concerns about OGC being able to be quick. Pluses and minuses here. Disparate groups got together and produced waterML; other standards originating outside OGC are being brought into OGC to address long-term maintenance (GeoSciML, SoilsML, OpenMI, netCDF/CF, OPeNDAP). WMO larger possibly older systems, but they well coordinated are standards and tend to work. There needs to be a careful open process with documentation. Create specifications that we can produce and support. Need to consider issues like intellectual property rights (IPR) protection, specification life cycle, certification - these are tough for developers/programmers to be concerned with, but should not be ignored. (OGC addresses these).

Agree on tools and services. Team up on semantic web. HPC center has issues with installing and it becomes difficult to ask to re-install a system with the install script fails Have a session on how to improve the software stack and it needs to be painless. How can we help Gavin and take some work off his back. Is there a need for training sessions, documenting experiences. Virtual machines(still there?). Aim is to deploy a federated archive, not just an FTP site. Originally developed for CMIP5, since then has grown into other datasets. Was originally just for scientist, but its scope is broadening. Upgrades, testing and deployment, can this be automated? Very difficult to test nightly builds. Could have a simple release process with a red light and green light for stability. Emphasis on up-to-date documentation ... w/o having to e-mail someone to look for it

CF Metadata

NetCDF Climate and Forecast Metadata Convention

Thursday May 12th, 2011

CF Day

Steve Hankin: - "CF-R-Us"
Introduction: CF workshop will include several sessions then break out and close with a governance meeting. Reminder that CF is really OUR standard, not something "they" do to us!

Topics: http://nomads.ncdc.noaa.gov/go-essp/presentations/CF/CF_Topics.pdf

---

John Caron - CF Chapter 9: Discrete sampling geometries times series, vertical proejections

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/CF/caron_CFdiscrete.pdf

Originally called the point observation convention. An encoding standard for netCDF classic files (represent ragged arrays). Classifies data according to the connectedness of time/space coordinates. Defines netCDF data structures that represent features. Makes it easy to store and extract features from a file and subset on space and time.

Feature Types: point (single data point), timeSeries: a series of point at the same location with monotonically increasing time, trajectory (trajectory feature): a series of data points along a path through space with monotonically increasing times, profile: an ordered set of data points along a verticle line at a fixed horizontal position and fixed time, timeSeriesProfile: a collection of profile features, but at the same space location (28 or 43 vertical sounders scattered around the US sampling the atmospherics every 15 minutes, trajectoryProfile: a collection of profiles except rather than being at the same point they follow along a trajectory (a ship traveling around an ocean taking soundings, and you may interpolate around those soundings).

Closed polygons are not being addressed in this proposal. The USGS CIDA group (Nate Booth, Dave Blodgett) are working on adding GIS features to NetCDF/CF. This new CF standard (approved yesterday) in Chapter 9 is all point data.

Feature Instances: instance variables: only the instance variables have the station variable, but are important to pick out when looking at larger sets of data.
instance dimension = station; instance variables = lat, lon, alt, station_name, desc, stuff (Provide the metadata that define these variables)

Representations: putting ragged arrays into single dimension arrays with pointers to the separations. Orthogonal multidimensional array (2D with lat lon altitude and set station to the station ID) float humidity(station, time); float lon(station); float lat(station); float alt(station); float time(time) where time and station are constant across the variables. Each point in index space is a coordinate in.

Started to think about lagrangian tracers. Benno said that you end up with tracers at each time. There is a lead time and start time. It is noted in the track site for CF. Tracers end up in the base body of CF, just need discussion. There are incomplete multi-dimensional arrays. Each timeSeries can have its own set of coordinates. This allows for missing values in the ragged array. Because it's multi-dimensional and it has to be squared off and leaves missing values for stations that are missing data. Contiguous ragged array: Turns the multi-dimen. array into a single dim. array. There is a variable called row size and each station records the number of samples for each station. Station 0 goes to 0 -> n-1 and Station 1 goes from n-1 to.. Index ragged arrays are a varient of the 1d array by recording with each variable the station it belongs to. "station_index=i" Requires you to read the whole thing to find out about a single station. Sample is the index for the data from the station and indexed by the station.

Status: approved **42**-page standard. implemented in netCDF java library: nested tables, point feature dataset APIs, TDS has feature collections (aggregations v2). Creating data query services in TDS (alpha), Starting to replace old APIs in IDV (2 or three done, but more to do) CDM is currently based on an older spec, the convention is more general than CDM. John has created implementation notes also (see slides).
Issues: time it takes for a complex proposal to get approved. Relationship of the specification to the implementation. Is the implementation necessary?, Does CDM have a special role? Also, generality vs. specifity: stick to the files or answer the question. Few are willing to go through the proposal and approval process. Innovation in the standard is irresistible. The process took from 2007/09 until 2011/05 to create, vet and get approved total. CF approval started in 2008/10. Other complex CF proposals: (RADAR/LIDAR) from Mike Dixon (trac #59), If people get involved CF for RADAR/LIDAR could get approved, but it is a long time to wait and is needed. Currently working on version 2. CF-Satellite: currently and email group and has a proposal with SSEC due June 3rd. Unstructured Grids: There is a google group who are discussing this, but there are a lot to discuss with this group and there are many efforts.

David Arctur: Q: how does this work relate to the OGC O&M work?
John Caron: OGC tends to work from top down, CF works from bottom up, hope to meet in the middle. Can we map CDM to OGC?
Bryan Lawrence: Thinks CDM and O&M activities are complementary.
John: Yes, OGC and Unidata CF/CDM have worked a lot together discussing point data conventions.
Ben Demenico: Unidata has been making sure that CSML is harmonized, and the CSML folks are very aware of the OGC O&M.
Bryan: CSML 3 is compatible with O&M.
John: Important to distinguish between encoding format and data model
Bryan: there is some overlap in OGC SWE, however
Jennifer: Is this new point feature type consistent with opendap point data, like Dapper?

John: Yes. It's likely you would access data into these NJ classes via opendap connections like Dapper, and could save to NetCDF files with these conventions.

Balaji: on subject of innovation, it's important to test out to see if things are useful before suggesting as a standard.
John: this should be done by clear versioning, alpha, beta releases, etc.
Steve: let's make sure to pick this up later today -- worthy of 20 min of discussion.

Jonathan Gregory: The CF Data Model

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/CF/gregory.pdf

CF can really be considered an abstract data model, independent of NetCDF file format or particular language liike Java or C. This abstract model can be described in plain language, or with UML. (Used Enterprise Architect to generate the UML)

A data model is centered on a space object.

Space: dimension, aux coords, cell methods, cell measures, transformations
If space has no data, it's just a grid.

Space = Data + Grid

Rich Signell: It would be interesting to compare the CF UML diagram to Bill Howe's GridFields data model.

The space may or may not contain data. The most important part of the grid is the dimensions.

CF should include:
- a document defining its data model
- a dcoument explaining how this is implemented in NetCDF
- a reference software implementation (Python, Java?)

Ben: how does this relate to Stefano Nativi's UML model?
Brian: Cell bounds and cell methods are very important to us, not so visible in Stefano's UML.
Ben: ISO 19123 data model is very relevant to this data modeling activity, also, and should be considered. Also a GML specification.

Art Burden - C-RDR Case Study: Fun with Metadata Conformance and netCDF-4.1

Presentation: http://nomads.ncdc.noaa.gov/go-essp/presentations/CF/art_The%20C-RDR%20Case%20Study.pdf

Climate Raw Data Project - VIIRS, CrIS, ATMS, OMPS inst. on NPP sat. bird Will deliver level0 C-RDRs = RDR converted to level1a data: reconstructed, unprocessed, packed with support data needed to calibrate and geolocate in netCDF4, will simplify access to the raw data. Archive guidelines require 19115 but using 19115-2 (remote sensed). Provides a unique case for metadata conventions. The approach taken was a hybrid. Follow CF where applicable, contain ACDD (NetCDF Attribute Convention for Dataset Discovery) and includes metadata that map to relevant to NPP ... They are mapping everything to ISO 19115.

Q: Rich Signell: How does this relate to ncISO ?
A: CF conventions for sat. are not fully established. Swath data coordinates, band, sample, scan. Bounding box attributes: G-Rings (Geographic min/.max vlaues are useless for long swaths) You would have a grin lat/lon that gets stored. When you process you want to know the area that those are in. With one orbit of data your bounding box will be identical. Bounding box for discovery of data will nto be useful. Engineering data: performed periodically, 500 scans, but only 20 sets of engineering data in a file. keep the NPP names. Need two unlimited dimensions: scan and ?, so NetCDF 4 or bust.

NPP fill values fall within data range.. not recommended by CF, no practical solution available. C-RDR Data Format: netCDF4: classic vs. enhanced: Raw data really lends itself to be stored in groups. Wanted to use the nc_string data type, multiple unlimited dimensions

Most users are power users; want to improve calibration, etc. but want to use the classic model. Small hurdles alienate data users. Cause problems with Matlab if they don't want to work with the data some. Need to make netCDF-4 transparent to end user. Support for Viz and anal. of data with current software is piecemeal currently. Latest version of Matlab bombs with enhanced netCDF-4 built-in routines. 2001a just released but too does not support enhanced netCDF-4. Wrote read C-RDR for Matlab. can spit out ncdump version of data. IDL 8.0 does support some netCDF-4.1 features using the built-in hdf_browser function Accourding to ITT there will be support for netCDF-4.1 (?)

Software developers, data providers, users: who goes first? End up paying a penalty for pushing the envelope. Support lags in COTS and customers get upset. The priority is long-term stewardship, and not short term gain to appease end-users currenlty. Endusers offten don't care about extensive metadata, just want to get working.

Big problems for reading into various packages were: multiple unlimited dimensions, NC_STRING and structures.

Russ Rew: we could build these factors into NC_COPY so that you could flatten and "fix", converting from NetCDF4 to NetCDF3.

CF TOPIC DISCUSSION

<u>Satellite Data</u>: Rew, Alison, Raskin, Jianfu. Russ is leading a proposal to get funding for people to work on this as their day job. There is a CF Sat mailing list. BNL agreed to write a letter of

support. Bryan also agrees to assign a person on his staff (Victoria) to setup an initial telecon (and the possibly a series of telecons) to get the ball rolling and harvest the low-hanging fruit (e.g. a coherent proposal for pixel descriptions and simple swaths). Need to consider the LTDP angle too (e.g. L0, L1) and how these possibly different applications (LTDP versus science/vis) cience/vis) may proceed. What about SAFE?

Unstructured Grid: Jeff Daily, Bob Oehmke, Rich Signell, Alex Pletzer, Kyle Wilcox, Bert Jagers.
Action item: Rich will organize regular telecons with this group -- folks who are actually writing code to work with unstructured grid data in NetCDF. The goal is to get the small differences between the ESMF unstructured grid, the Deltares unstructured grid, and the Karen Schuchardt/Jeff Daily geodesic grid format worked out.

LibCF: John Caron points out that a API is a binding of a data model to a particular language, and that a clear data model design would be important and useful activity that would imform people working in different languages. Jonathan agrees that this would be very valuable, and will start a ticket.

CD: Multiple APIs are likely to be used for writing out files conformant to CF grid conventions. Resource issue is that since LibCF is new, writing out conformant files, especially for complex/unstructured grids, may be faster and funded if it does not go through LibCF. Development of a clear data model design would be useful since then there can be API bindings that use community-developed grid representations for modeling, including ESMF. Here it may also be useful to experiment with a grid API that does not just encompass mosaics but includes unstructured grids in the same context.

JC: TDS and CDM have feature collections. Method for handling (millions) of files. Partition time series by time (aggregate over time). Alpha s/w available now. (Confusion over distinction between aggregation by time and collections). Allow coordinate space access (as opposed to index space, because index space is impossibly slow). Ala WCS/WFS??
Need to go beyond NetCDF/OpenDAP API (e.g. Dapper reasonable example).

Significant number of input formats supported (e.g. BUFR etc, possibly including SWE) into these collections. (See JC's earlier presentation).

Nate's Booths group at USGS and IOOS Modeling Testbed are working on SOS services for THREDDS Data Server, utilizing the new point featureType constructs.

Balaji: there will be a need to store iceberg model output, and icebergs come and go.

Caron: this is similar to oil spill and other particle tracking modeling applications. This was discussed quite a bit on the CF list, and progress was made, but beyond the scope of the current proposal. I'm willing to pick this up again now that the point proposal has been approved.

Datum and Coordinate Issues:

David Arctur: Lon/Lat datum matters in many domains (not global modeling), and is required to move data from CF datasets into GIS tools (e.g. ArcGIS) and OGC standards.
EPSG is the defacto standard for coordinate systems, but is not an open database, and some datums have been rejected by EPSG. How about hosting such a database at OGC?
Action item: David will lead, Rich will moderate.

Caron: We need at CF Best Practices statement -- what users should do when datums are not supplied, issues, etc. This is a bit different than the conformance document.

Common Concept:

Trac ticket 27 has lots of discussion on URNs but not resolved. Ticket 29: How do we describe what we mean by common concept, e.g. how fuzzy? Ticket 24 ?

Action item: elevate ticket 27 (last discussion 2.5 years ago) and try to finish it.

Caron: Standard names group should take this on.

Separate out the issue of external URIs. Finalize that and close #27.

The original use cases are still valid:

bundle of attributes that are all needed to uniquely ID a variable

high_cloud_amount = stdname=cloud_amount_in_atmospheric_layer + layer_bounds=(z1,z2)

Perhaps keywords identify ways to use regular expressions on combinations of attributes?

Other example (MPI): surface temperature is "tas" for CMIP5 purposes... how to announce that this is the "CMIP5 name" for surface temp? We wanted to use common_concept.

Perhaps these use cases can be solved without the full machinery of common_concept.

Action items: Alison is the moderator for track ticket 24 (though not listed on the trac site), and she will get telecons going to elevate this ticket. Alison says that issues on 27 and 29 need to get resolved in order to finish 24. Steve Hankin agreed to write text to update tickets with current discussions.

Standard Names:
Actual Min/Max: Caron, Rew
GridSpec: Antonio, Daily
Using NaN: Caron, Antonio
Calendar Time: Benno, Jainfu

External Metadata Linkages: Giri, Hankin
Discovery Metadata: Baird, Hankin
Scalar Auxiliarly Coords
David Arctur: NASA supported project OWS-8: WCS 2.0 JPEG2000, HDF4-EOS, HDF5-EOS2, NetCDF encodings

CF Governance

Action item: form small groups with more telecons around key issues, with people who have contributed a lot asked to participate. Moderator should be able to decide whether telecons are public or go off line if necessary to make more progress. We should also seek to document what the moderator's role is.

David Arctur: Consider "core & extensions" model for CF evolution: have widely-agreed core functionality versioned separately from optional, thematic, application-specific, or contentious draft functionality.

Ben Domenico: Note there will be an OGC meeting hosted by UCAR at Center Green in Boulder, Sept 19-23. It would be fine to schedule a 2-4 hour "CF ad hoc" session as part of the Met-Ocean Domain Working Group (DWG) agenda. Contact Ben if interested. Need to coordinate with Met-Ocean DWG co-chairs Chris Little <chris.little@metoffice.gov.uk> and Marie-Francoise Voidrot <marie-francoise.voidrot@meteo.fr>.

CF-OGC Relationship
Ben Domenico and David Arctur: about coordination between CF community and OGC process. CF team could continue grassroots development, and submit updates to standards to the OGC when ready. If OGC adopts without change, keep the same version# on both. Loose coupling, with some chances for isolated parallel development on tasks of mutual interest.

Alternatively, CF community could use OGC convening support directly: quarterly meeting planning, collaboration portal/twiki, working group process, etc. Most if not all CF team are already OGC members through their org's (UK NERC/BADC, USGS, NOAA, NASA, UCAR, and OPeNDAP org is planning to join), so membership requirement for portal access may be a minor issue. CF team could participate through Met-Ocean DWG, and just ask for time in OGC meeting agendas as whenever it makes sense (don't have to meet quarterly just because the meeting is being held; depends on status of work, location of meeting, etc).

Benefits to using OGC meetings in addition to GO-ESSP meetings for CF coordination:
- Met-Ocean DWG has WMO coordination status, with co-chairs Chris Little and Marie-Francoise Voidrot representing both OGC and WMO interests & issues.
  - There is also direct connection to WMO IPET-MDI (Inter-programme expert team for metadata & data interoperability) through Jeremy Tandy, UK Met Office, current chair of that team.
- just having more frequent opportunities to meet can help advance projects more quickly;
- raises awareness of CF progress and issues to broader, more diverse audience;

- great way to have extended, informal talks with experienced OGC programmers and technical managers beyond just the CF team -- this can sometimes slow down standards process but almost always for good reasons (consider other stakeholders' needs; avoid duplication of effort or divergent approaches, etc)
- document management system for specifications, best practices & other doc types;
- OGC staff take care of meeting venue logistics and collaboration portal admin

End of CF Workshop

Appendix A

GO-ESSP Workshop Agenda

Global Organization for Earth System Science Portals (GO-ESSP)

Collaboration designed to develop a new generation of software infrastructure that will provide distributed access to simulated, observed, and reanalysis data from the climate and weather communities.

GO-ESSP Public Note-Taking Site: http://go-essp.titanpad.com/asheville Password: go-essp

Day 1: Tuesday, 10 May

GO-ESSP Presentations

Session One: Federation, interoperability, and security

08:30 – 09:00	Arrival time	Glenn Rutledge (Host)	30 min
09:00 – 09:05	Welcome and logistics	Glenn Rutledge, (NCDC)	5 min
09:05 – 09:10	Opening remarks	Scott Hausman, Director (acting) (NCDC)	5 min
09:10 – 09:25	The Earth System Grid Federation: building a software framework of open source, modular components for analysis of large distributed scientific data	Luca Cinquini (NASA), *et al.*	15 min
09:30 – 09:45	ESGF: How to build an elastic distributed system over "Big Data"	Gavin Bell (LLNL)	15 min
09:50 – 10:05	The Earth System Grid Federation User Interface	Feiyi Wang, (ORNL), Galen Shipman, and John Harney	15 min
10:10 – 10:25	The ESGF Gateway	Nathan Wilhelmi (NCAR), *et al.*	15 min
10:30 – 10:45	**BREAK**		15 min
10:50 – 11:05	Security Mash-up with the MashMyData Project: Delegation and Workflows with OPeNDAP and OGC Based Services	Philip Kershaw (BADC), *et al.*	15 min
11:10 – 11:25	Using OpenID/OAuth to access Federated Data Services	Benno Blumenthal (IRI)	15 min

Session Two: Data preservation, data curation, and long-term access

11:30 – 11:45	Data Management and digital Preservation for Arctic Science: CADIS and Chronopolis	Eric Nienhouse (NCAR) *et al.*	15 min

11:50 – 12:05	FREMetar: Efficient and flexible metadata re-writing	Kyle Olivo (GFDL)	15 min
12:10 – 12:25	Updates on Unidata Technologies for Data Access	Russ Rew (Unidata)	15 min
12:30 – 13:40	**LUNCH**		70 min
13:45 – 14:00	Production of a search and browse interface for an environmental science thesaurus	Oliver Clements (BADC)	15 min
14:05 – 14:20	Next generation data services: C3-INAD goes ESGF	Stephan Kindermann (DKRZ)	15 min
14:25 – 14:40	Maximising the utility of OPeNDAP datasets through the NetCDF4 API	Stephen Pascoe, Science and Technology Facilities Council UK	15 min
14:45 – 15:00	**BREAK**		15 min

Session Three: Serving data for environmental sciences

15:00 – 15:15	Metadata Standards for *in-situ* Observational datasets	Giri Palanisamy (ORNL) *et al.*	15 min
15:20 – 15:35	NOAA/GFDL- Model Development Database Interface (MDBI)	Aparna Radhakrishnan (GFDL)	15 min
15:40 – 15:55	Continuously Enhancing Usability of Remote Sensing Data for Climate Models	Jianfu Pan (NASA)	15 min
16:00 – 16:15	Paleoclimate Modelling Intercomparison Project Phase 3	Jean-Yves Peterschmitt, National Center for Scientific Research, FR	15 min
16:20- 16:35	Leveraging the Earth System Grid for Integrated Regional Earth System Modeling (iRESM) Research	Eric Stephan (DOE/PNL)	15 min
16:40 – 16:45	**Closing comments and adjourn Day 1**		5 min
18:30 – 21:00	**Workshop Dinner: Self-organizing**		2-3 hrs

Day 2: Wednesday, 11 May

Presentations and Demonstrations

GO-ESSP Presentations

Session Four: Next generation data services

08:30 – 09:00	Arrival Time	Glenn Rutledge (Host)	30 min
09:00 – 09:15	Parallel Analysis of GeOscience Data: Status and Future	Jeff Daily (DOE/PNL)	15 min
09:20 – 09:35	A Scientific Workspace environment for collaborative analysis of climate data.	Luca Cinquini (NASA/JPL)	15 min
09:40 – 09:55	ExArch: Climate analytics on distributed exascale data archives	Martin Juckes, Science and Technology Facilities Council UK	15 min
10:00 – 10:15	Exascale Climate data analysis From	Frederic Laliberte, University of	15 min

	the INSIDE out	Toronto	
10:20 – 10:30	**BREAK**		10 min
10:30 – 10:45	Globus Online: A hosted data transfer solution for climate scientists	Rachana Ananthakrishnan (DOE/ANL)	15 min
10:50 – 11:05	Policy-based Data Management	Reagan W. Moore (RENCI)	15 min
11:10 – 11:25	Fast parallel grid remapping for unstructured and structured grids	Robert Oehmke	15 min
11:30 – 11:45	NOMADS and the National Climate Model Portal (NCMP): Science and Data Management Services	Jay Hnilo (NCDC, CICS-NC)	15 min
11:50 – 12:05	The Unified Access Framework (a pretentious name for a simple idea)	Steve Hankin (NOAA/PMEL)	15 min
12:10 – 12:25	The Unican Downscaling Portal	Antonio S. Cofiño (University of Cantabria, Spain)	15 min
12:30 – 13:40	**LUNCH**		70 min

Session Five: Areas of Major CF Enhancements

13:45 – 14:00	The US-IOOS Modeling Testbed Cyber-infastructure: Unstructured Grid Standards and Standards-Based Tools for Analysis of Ocean, Atmosphere & Climate Model Data	Rich Signell (USGS)	15 min
14:05 - 14:20	CF standard names: recent developments and a new vocabulary editor	Alison Pamment (BADC)	15 min
14:25 – 14:40	Mapping CF Standard Names to the SWEET Ontology	Rob Raskin (NASA/JPL)	15 min
14:45 – 15:00	Progress on LibCF including support for Gridspec	Alexander Pletzer, Ed Hartnett (Unidata)	15 min

Session Six: Presentation, demonstration and discussion of existing services

Those interested in providing a live demonstration of systems or services are invited to join in this session and provide a brief introduction prior to the demo session.

15:00-15:15	Demo Overviews	Rutledge (Host)	
	The New GrADS Interface for GIS Data Formats	Jennifer Adams, Brian Doty (COLA)	2-3 min
	Visualization Gallery (vizGal)	Roland Schweitzer, Weathertop Consulting, LLC	2-3 min
	The Weather and Climate Toolkit	Steve Ansari (NCDC)	2-3 min
	Implementation of CMIP5 Metadata in the ESG Gateway	Sylvia Murphy (CIRES), *et al.*	2-3 min
	A Scientific Workspace	Luca Cinquini (NOAA/JPL)	2-3 min
	Globus On-Line	Rachana Ananthakrishnan (DOE/ANL)	2-3 min
	UC downscaling inter-comparison Project	Antonio S. Cofiño (University of Cantabria, Spain)	2-3 min

	iRODS	Reagan Moore (RENCI)	2-3 min
	OpENDAP	James Gallagher	

GO-ESS Demonstrations

Session Seven: Demonstrations

15:15 – 17:00	Live Demonstrations of Tools and Software	105 min
17:00 – 17:30	Open Discussion: Feedback from Demos & Collaboration Opportunities	30 min
17:30 – 17:35	**Adjourn 2011 GO-ESSP Workshop**	5 min

Appendix B

2011 Earth System Grid Federation (ESGF) Day

Governance and Technical Meetings Agenda

Asheville Renaissance Hotel

Swannanoa Room, Monday, 9 May

Welcome and Opening Remarks

08:30 – 08:55	Arrival Time	Glenn Rutledge (Host)	25min
08:55 – 09:00	Welcome and Logistics	Glenn Rutledge (NCDC)	5 min
09:00 – 09:10	Opening Remarks	Dean Williams (PCMDI) Don Middleton (NCAR) Bryan Lawrence (BADC) (remotely) Michael Lautenschlager (DKRZ)	10 min

All of the ESGF team are invited to spend the entire day together for discussion themed around three, and depending upon the audience 4, sessions:

Open Discussion Session One: *(+/-)* ***2hrs.*** *How we work together*

* Define what ESGF is (not the resulting infrastructure but the "working-power").

Can it be managed? How? How are we dealing with project responsibility? Are we all on the same page? It is important to be clear what it is, as it is not a project. There are no resources allocated for ESGF directly nor anyone is directly responsible for it.]

How do we work together on:

- Daily to weekly time scales…

* How can we be more responsive to “bug fixing problems” at the federation level?
* current system, critics and improvements. (e.g. define guidelines for accessing the repositories, like anonymous read-only, etc)
- Weekly to monthly…
* How can we prioritize activities so we can get the best benefit of our joint activities?
* We can't force anyone to do anything... we can't even "incentive" people, so: How can we assure things get done?

- Monthly to yearly…
* What are the directions in which we are going?
* What does every group "expect" from ESGF. (Own goals vs ESGF's ones)
* What about other type of data? (aka. is there life after cmip5? :-)
* How does NCAR's opening of the Gateway's source code change things? What can we re-use, what can the gateway use from us. (eg. security: it would be nice that interfaces required for communication between data node and gateway are developed by the same team if possible)

~11:30 - 1:00 LUNCH

Session Two: Operations for CMIP5 1:00-2:30
- Operations check and status: how's everything running, critics and problems.

  - Status of Japan Gateway (G. Rutledge for Michio Kawamiya JAMSTEC) 5min
- Load balancing & streamlining. What can be done and how?
- Helpdesk current stand. problems, critics.
- User experiences gathered so far. (e.g. downloading files)
- Redundancy and fail-safe systems. What, where and how?

  (from repo to PCMDI's attribute service)
- QC. How's that working critics & problems.
- Replication and gridFTP. Speed and other problems. (scripts, clients & co.)
- DRS structure too rigid for other data. Is this really required?
- Registry status. How flexible is it (adding new nodes/gateways/services)? How flexibly should

  it preferably be?
- Security. How does the integration with other services work? Is any integration from ESGF to other systems already in place? Planned?

Session Three: What are the ESGF futures? 2:30-3:30

It might seem precipitate to discuss that before we are even properly operational, but it is clear that there are parallel activities going on (and they will continue). What is a sensible architectural philosophy that celebrates both our desire to work together, and our differences?
- What is the architectural philosophy we should prioritize?
- Where do we want ESGF to be aimed to after CMIP5?
- What can we do in the short/mid/long term?

Session Four: Technical session: Code-sprint.

If there's time and anyone is interested we could solve current configuration problems (or of any other kind) together to gain a better insight in the system.

~ 3:30-4:30/5:00 Closing Comments and Adjourn

Appendix C

Climate and Forecast Convention Governance and Technical Meetings Agenda

Thursday 12 May 2011 Discussion and Decision of Challenges

--Presentations with extended discussions --

08:00 – 08:30	Arrival Time	Glenn Rutledge (Host)	30 min
08:30 – 08:35	Welcome and Logistics	Taylor / Hankin	5 min
08:35 – 08:45	Individual Introductions	All	10 min
08:45 – 09:00	Discrete Sampling Proposal to the CF meeting	John Caron (Unidata)	15 min
09:05 – 09:20	CF Data Model Discussion	Jonathan Gregory (WebEx)	15 min
09:25 – 09:40	Climate Data Record CF attributes	Art Burden	15 min
9:45 - 10:00	BREAK		15 min
10:00 – 10:30	LibCF priorities from users perspective	All	30 min
10:30- 12:00	Group discussion: topics below		90 min

Some of the identified candidate CF topics for group discussion are below. The topics will be priority-ranked based upon input from the group. Each participant will be asked to indicate his/her top 3 priorities – please consider your priorities in advance. Discussions will occur based upon that ordering.

1. Particular elements of discrete geometry ("point data") conventions that generally effect CF (#37)
2. Particular elements of unstructured grid conventions that generally effect CF
3. Particular elements of gridspec conventions that generally effect CF (trac #63)
4. Particular elements of satellite conventions that generally effect CF
5. General considerations for the standard name conventions
6. Concerns with CF checker and/or compliance testing generally (trac #43,50,53, 54,60)
7. Use of NaN as a _FillValue (trac #52)
8. Common Concept -- external namespaces embedded within CF (trac #24, 27)
9. Proposal for standard attributes actual_min and actual_max (trac #31)
10. Handling of calendar time data – Gregorian (irregular) monthly axes
11. Cell Methods – proposed new methods; concerns over harmonizing CF units and cell_methods (trac #61, 65)
12. Scalar Auxiliary Coordinate Clarifications -- harmonized with other coordinate variables (trac #62)
13. The need for a formal UML data model
14. Other discussion topics

12:00 - 12:30	Open discussion: Improving the CF standards and approval process	30 min
12:30	**CF MEETING ADJOURN**	
12:30 - 14:30	**Steering Committee Working Lunch (2 Hours)**	2 hrs

Appendix D

GO-ESSP Principal Investigators

British Atmospheric Data Centre
Bryan Lawrence
B.N.Lawrence@rl.ac.uk

German Climate Computing Centre
Michael Lautenschlager
lautenschlager@dkrz.de

Lawrence Livermore National Laboratory
Dean Williams
williams13@llnl.gov

National Center for Atmospheric Research
Don Middleton
don@ucar.edu

NOAA/Earth Systems Research Laboratory
Cecelia Deluca
Cecelia.Deluca@noaa.gov

NOAA/Geophysical Fluid Dynamics Laboratory
V. Balaji
V.Balaji@noaa.gov

NOAA/National Climatic Data Center
Glenn Rutledge
Glenn.Rutledge@noaa.gov

NOAA/Pacific Marine Environmental Laboratory
Steve Hankin
Steven.C.Hankin@noaa.gov

Appendix E

List of Participants

Last Name	First	Affiliation	Email
Adams	Jennifer	IGES/COLA	jma@cola.iges.org
Ananthakris	Rachana	Argonne National Laboratory	ranantha@mcs.anl.gov
Ansari	Steve	NOAA / NCDC	steve.ansari@noaa.gov
Arctur	David	Open Geospatial Consortium (OGC)	darctur@opengeospatial.org
Arnfield	Jeff	NOAA/NCDC	Jeff.Arnfield@noaa.gov
Balaji	V.	NOAA/GFDL (GO-ESSP PI)	V.Balaji@noaa.gov
Austin	Matthew	NOAA/NESDIS	matthew.austin@noaa.gov
Barkstrom	Bruce	Barkstrom Enterprises	brbarkstrom@gmail.com
Bell	Gavin M.	PCMDI / ESGF.org	gavin@llnl.gov
Bhardwaj	Ashis	IPSL	abhipsl@ipsl.jussieu.fr
Biard	Jim	NOAA/NCDC/STG	jim.biard@noaa.gov
Blumenthal	Benno	IRI of Columbia University	benno@iri.columbia.edu
Brinegar	Danny	NOAA/NCDC	danny.brinegar@noaa.gov
Budai	Jeff	NOAA/NCDC (STG, Inc.)	jeff.budai@noaa.gov
Burden	Art	NOAA/NCDC	art.burden@noaa.gov
Caron	John	Unidata	caron@unidata.ucar.edu
Carter	Dean	NOAA/NCDC	dean.carter@noaa.gov
Cinquini	Luca	NOAA/ESRL	Luca.Cinquini@noaa.gov
Clements	Oliver	British Oceanographic Data Centre	daol@bodc.ac.uk
Cofino	Antonio	University of Cantabria	antonio.cofino@unican.es
Cogbill	Phil	NOAA - National Climate Model Portal	phil.cogbill@noaa.gov
Cross	Scott	NOAA/NODC	scott.cross@noaa.gov
Daily	Jeff	Pacific Northwest National Laboratory	jeff.daily@pnl.gov
Deluca	Cecelia	NOAA/ESRL (GO-ESSP PI)	Cecelia.Deluca@noaa.gov
Denvil	Sebastien	CNRS/IPSL	sebastien.denvil@ipsl.jussieu.fr
Devine	Gerard	NCAS, University of Reading, UK	g.m.devine@reading.ac.uk
Domenico	Ben	Unidata/UCAR	ben@unidata.ucar.edu
Gallagher	James	OPeNDAP	jgallagher@opendap.org
Galluppi	Kenneth	University of NC - RENCI	galluppi@renci.org
Gardiner	Ned	Climate Program Office	ned.gardiner@noaa.gov
Gonzalez	Estanislo	Max-Planck-Institute for meteorology	estanislao.gonzalez@zmaw.de
Grogan	Michael	NOAA/National Climatic Data Center (STG, Inc.)	mike.grogan@noaa.gov
Hall	Alan	NOAA/NCDC	alan.hall@noaa.gov
Hankin	Steve	NOAA / PMEL (GO-ESSP PI)	steven.c.hankin@noaa.gov

Harney	John	Oak Ridge National Laboratory	jfharney@ornl.gov
Hnilo	Jay	NCDC NOMADS / NCMP /CICS	jay.hnilo@noaa.gov
Idaszak	Ray	RENCI/UNC Chapel Hill	rayi@renci.org
Juckes	Martin	STFC	martin.juckes@stfc.ac.uk
Kawamiya	Michio	JAMSTEC	kawamiya@jamstec.go.jp
Karl	Thomas	Director NOAA NCDC	Thomas.R.Karl@noaa.gov
Kershaw	Philip	NCAS/British Atmospheric Data Centre	philip.kershaw@stfc.ac.uk
Kindermann	Stephan	German Climate Computing Center, DKRZ	kindermann@dkrz.de
Kostov	Georgi	NOAA / NCDC NOMADS	georgi.kostov@noaa.gov
Krassovski	Misha	CDIAC, Oak Ridge National Laboratory	krassovskimb@ornl.gov
Laliberte	Frederic	University of Toronto	frederic.laliberte@utoronto.ca
Lautenschlager	Michael	DKRZ (GO-ESSP PI)	lautenschlager@dkrz.de
Lawrence	Bryan	British Atmospheric Data Centre	B.N.Lawrence@rl.ac.uk
Loomis	Timothy	NOAA/NESDIS/STAR/EVL	tim.loomis@noaa.gov
McGuirk	Marjorie	NOAA/NCDC	marjorie.mcguirk@noaa.gov
Middleton	Don	NSF/NCAR (GO-ESSP PI)	don@ucar.edu
Moore	Reagan W.	University of North Carolina at Chapel Hill DICE/iRODS	rwmoore@renci.org
Morgan	Mark	Institut Pierre Simon Laplace	momipsl@ipsl.jussieu.fr
Murphy	Sylvia	NOAA/CIRES	sylvia.murphy@noaa.gov
Nienhouse	Eric	NCAR	ejn@ucar.edu
Oehmke	Robert	NOAA/CIRES	robert.oehmke@noaa.gov
Ogochi	Koji	Japan Agency for Marine-Earth Science and Technology (JAMSTEC)	ogochi@jamstec.go.jp
Olivo	Kyle	NOAA GFDL	kyle.olivo@noaa.gov
Palanisamy	Giri	ORNL	palanisamyg@ornl.gov
Pamment	Alison	NERC Centre for Atmospheric Science/British Atmospheric Data Centre	Alison.Pamment@stfc.ac.uk
Pan	Jianfu	Adnet Systems, Inc. / NASA	jianfu.pan@nasa.gov
Pascoe	Stephen	Science and Technology Facilities Council, UK	Stephen.Pascoe@stfc.ac.uk
Peng	Ge	NOAA/NCDC	ge.peng@noaa.gov
Peterschm	Jean-Yves	LSCE / IPSL	Jean-Yves.Peterschmitt@lsce.ipsl.fr
Pisut	Dan	NOAA Visualization Lab	dan.pisut@noaa.gov
Pletzer	Alexander	Tech-X Corp.	pletzer@txcorp.com
Potter	Jerry	NASA GSFC	gerald.potter@nasa.gov
Radhakrishna	Aparna	NOAA/GFDL	aparna.radhakrishnan@noaa.gov
Raskin	Rob	Jet Propulsion Laboratory	raskin@jpl.nasa.gov

Rew	Russ	UCAR Unidata	russ@unidata.ucar.edu
Rutledge	Glenn	NOAA NCDC (GO-ESSP PI)	Glenn.Rutledge@noaa.gov
Sanders	Drew	NOAA/NCDC	drew.sanders@noaa.gov
Santha	Suresh Kumar	ORNL	santhanavans@ornl.gov
Schweitzer	Roland	Weathertop Consulting, LLC	Roland.Schweitzer@noaa.gov
Shipman	Galen	ORNL	gshipman@ornl.gov
Shipman	Galen	Oak Ridge National Laboratory	gshipman@ornl.gov
Signell	Rich	USGS	rsignell@usgs.gov
Sim	Alex	Lawrence Berkeley National Laboratory	asim@lbl.gov
Stephan	Eric	Pacific Northwest National Laboratory	eric.stephan@pnl.gov
Stevens	Laura	CICS-NC/NCDC	laura.stevens@noaa.gov
Stockhause	Martina	German Climate Computing Centre (DKRZ)	martina.stockhause@zmaw.de
Swank	Dan	NCDC/NOMADS/STG	dan.swank@noaa.gov
Symonds	Jason	NOAA/NCDC/NIDIS Drought Portal	Jason.Symonds@noaa.gov
Tandy	Jeremy	(UK) Met Office	jeremy.tandy@metoffice.gov.uk
Taylor	Karl	Director PCMDI	taylor13@llnl.gov
Torrey	Cristen	University of Michigan	ctorrey@umich.edu
Wang	Feiyi	Oak Ridge National Laboratory	fwang2@ornl.gov
Webster	Phil	NASA Goddard	Phil.Webster@NASA.gov
Wei	Yaxing	Oak Ridge National Laboratory	weiy@ornl.gov
Wilhelmi	Nathan	NCAR	wilhelmi@ucar.edu
Williams	Dean	PCMDI (GO-ESSP PI)	williams13@llnl.gov