Workflow Group

Plenary Tutorial:
Framework and Tools for Supporting Model Integrations and Offline Analysis
Overview

- **Process Flow**
  - Task Leads: Ben Mayer and Kerstin Kleese-Van Dam

- **Data Management**
  - Task Leads = Sasha Ames and Rachana Ananthakrishnan

- **Diagnostics & Metrics**
  - Task Leads = Jeff Painter and Brian Smith

- **UV-CDAT**
  - Task Leads = Charles Doutriaux and Aashish Chaudhary

- **User Interface**
  - Task Lead = Matthew Harris
Task Leads = Ben Mayer and Kerstin Kleese-Van Dam
Overview of Current Data Management

- Model Run Setup
- Run Management
- Many sites
  - Model runs
  - Resource availability
- Post-Process
  - Transfer
  - ESGF Publish
  - Interpolation
  - Diagnostics
  - Archive to HPSS
- Data Logistics
Model Configuration

- **Required** to **Enable proper testing of scientific question**
- Query collaborators for output fields
- Configuring Input files, model parameters and output fields
- Perform 5 day run
  - Check for ability to run
- Perform 25 day run
  - Check outputs are as expected
Run Management

• Typical time to complete production run 6-9 months;
  – time = money
• Need to be **constantly running model** or progressing towards running to **minimize time to solution**
• Status reports are generated to track progress
DOE Computing Diversity

- DOE sites have specific strengths (compute, disk, archival storage)
- We can and do leverage these unique capabilities
- It does introduce complexities with data location and security/accounts
Data Transfer

- Allows for leveraging multi-site strengths
- ESNet and Globus enable high speed transfers
ESGF Publication

- Allows organized and searchable data distribution to collaborators across sites
- Publishing of v0.1 data in progress (See Dean’s Talk)
Interpolation

- **Functional Requirement** to compare models and observations or intermodel (v0 to v1), and intermediate step between climatologies and diagnostics
- Performing several types of interpolation
  - Converting resolutions (higher -> bilinear, lower -> conservative remapping)
  - Native grid to lat/lon
Diagnostic Analysis

- Required to diagnose how case is progressing and evaluating model output
- We are using both UV-CDAT and NCL based diagnostic tools to examine case progress
- Creating new diagnostics framework using UV-CDAT to integrate into larger workflow
- Atmosphere and Land diagnostics are almost complete
Archive Generated Output

- Ability to, at a later date, perform analysis or recreate a case configuration
  - Long term availability
- Use HPSS for increased reliability over disk
  - Two copy or RAIT for even higher reliability
  - Also cost effectiveness
  - Automatically Computed CRC for bit error detection
Capability Summary

• Proper Configuration to test Hypothesis
• Simulation Progress and Status Reports
• Take best advantage of resources
  – Globus Transfers
  – ESGF Publications
• Critical functions for ability to compare and evaluate how model is progressing
• Provide final output and configuration information for later analysis via archiving
Data Logistics

- Production Compute
- Disk: Need 1-5TB
  - Model 6hr Run Request per day (on average)
  - 5 hour of run per 6 hr of request
  - ~1 hour of writing
  - 10x to 100MB/s
- Network drain:
  - In practice 550Mbps
  - (70MB/s)
  - Can happen constantly
- HPSS Archive: Very large storage
  - Tens of PB to ExaB
  - 200 MB/s
  - Nearly Constant
- Analysis of entire case
- Disk: Maximum Case size
  - Currently 130TB
  - Need ~150TB
  - Single Node of compute

ACME Accelerated Climate Modeling for Energy

U.S. Department of Energy
Task Leads = Sasha Ames and Rachana Ananthakrishnan
ACME Data Publication to ESGF

• The ESGF enterprise system is a framework for the management, dissemination and analysis of model output and observational data (http://esgf.llnl.gov)
  – Over 40 projects use ESGF (CMIP5, CORDEX, etc.)
  – Tens of PBs of data archived (ACME ➔ 20 datasets, ~50TB)
  – Easily accessible via http or Globus
• We’re publishing model output and climo data sets now
  – Goal to integrate automated publication with managed workflow
• ESGF nodes online at LLNL, ORNL, LANL, ANL, PNNL
  – NERSC coming soon
  – ANL upgrades coming soon
• Need to publish?
  – Now: contact publication “Liaisons” for ACME
  – Future: use Publication Service web user interface
  – Know where your data is located and values for search categories
• Usage tutorial
  – https://acme-climate.atlassian.net/wiki/display/WORKFLOW/Development+of+documentation+to+access+data+on+ESGF
Earth System Grid Federation (ESGF)

- ESGF Web Front End (Current)

- ESGF CoG (Improved front end released now. Officially released in Sept. 2015)
Publication as a Service

- **Goals:**
  - Simple interface for scientist to publish data
  - Managed publication process
    - User provides some input,
    - Submits publication task
    - Task managed by service
    - User notified on progress/errors

- **Mappings:**
  - Community: ACME
  - Sub-communities: Atmosphere, Land, etc
  - Collection: Case
  - Dataset: Run
Choose a collection

A collection is created for each Case
Add description

Submit: Describe this Dataset

The name of the Institution that generated the data
Institution *
Computation Institute

The name of the user who generated the data
Run Owner *
Ananthakrishnan
Rachana

Description of this dataset
Description *
Test submission

Data Node
Oak Ridge

Metadata and ESGF node to store data
Assembles files to publish

Identify files to publish and transfer to ESGF data node

Submit: Assemble this Dataset

Click the button to assemble the files you would like to publish to this dataset. Files chosen will be transferred to ESGF data node. Once all your data is assembled choose next to publish.

For ESGF publication, the files should be transferred in this directory structure
'ACME/data_type/<experiment>/<versionnum>/<realm>/<regridding>/<range>
Globus will scan all transferred files, and create the needed directory structure in
Refresh the page to see status of your transfers. Once completed, you can

Transfer Files

Transfer Request Submitted Successfully, Task ID: af3ee7e-eed9d-11e4-b6b4-123139209a8

Endpoint: climate#acme
Path: /data/1/ACME/h0/b1850c5_m1a/V0_1

Endpoint: publish#ami
Path: /submission0_37/

select all | none  up one folder  refresh list refresh list
b1850c5_m1a.cam.h0.0001-01.nc  4.01 GB
b1850c5_m1a.cam.h0.0001-02.nc  4.01 GB
b1850c5_m1a.cam.h0.0001-03.nc  4.01 GB

more options  Label This Transfer
This will be displayed in your transfer activity.
Login to ESGF

Publish to ESGF
Your data files will be scanned and parsed to extract metadata, create a THREDDS catalog and publish to the ESGF service. This might take some time, and you can refresh to see status.

Click on Publish to start the publication process. If you want to add more files use Previous button to transfer additional files.

Progress of Publication

To be able to publish to ESGF, you need to authenticate with your ESGF OpenID and password.

ESGF OpenID
https://esgf.ccs.ornl.gov/esgf-idp/openid/rachana

Password
**********

Login to ESGF
Pick search categories

Optional step to rearrange files in directory structure needed by ESGF publish tool

Facets selection

3 files have been found but not all of them are in the directory structure required by ACME project: `ACME/data_type/experiment/version номер/experiment/range/`. Please, select appropriate facets from the dropdown lists below. If a required facet is missing, please contact Support before proceeding.

- **Project:** ACME
- **Data type:** climo
- **Experiment:** b1850c5_m1a
- **Version number:** v0.1
- **Realm:** atm
- **Regridding:** re30_g16
- **Range:** all

Submit

< Previous Cancel/Save Next >
Status updates

Manage remote metadata extraction, generation of THREDDS catalogs and push to ESGF search index.
Planned Work

- Deployment of next version with email notification
- Production deployment of service
- Support for other ESGF data nodes
- Programmatic interface for supporting automation and scripting
- Seamless integration with ESGF policies for publication
- Integration with publication tool configuration manager (developed at ORNL)
Task Leads = Jeff Painter and Brian Smith
Command Line Diagnostics

Specify:

- Model information - data location, optional filter specification, optional flag indicating climatology vs raw
- Observation information (optional)
- Package type - atmosphere / land / ocean
- Variables, seasons (optional), additional variable options (optional)
- "plot set"

```bash
diags --outputdir ~/diagout/ --modelpath=~/metrics_data/cam35_data/,climos=yes --obs path=~/metrics_data/obs_data_5.6/,filter="f_startswith('NCEP')",climos=yes --package AMWG --set 4 --vars T --seasons ANN
```
Command Line Diagnostics

Output is:

• One PNG image file for each plot
• Another PNG image file for all plots combined
• A NetCDF file with the data used to compute each plot.
  – suitable for further analysis or preparing specialized plots for publication
Command Line Diagnostics

metadiags.py:

- Wrapper around diags script to create all diagnostics collections with a single command line invocation
- Additional diagnostic “collections” are easy to add
- Example:
  ```python
diags_collection['so'][desc'] = 'Tier 1B Diagnostics (Southern Ocean)'
diags_collection['so'][regions'] = ['Southern_Extratropics']
diags_collection['so']['SHFLX'] = {'plottype': '3', 'obs': 'LARYEA_1'}
diags_collection['so']['QFLX'] = {'plottype': '5', 'obs': 'LARYEA_1'}
```

- Takes model location, obs location, and output directory as arguments

climatology.py (and new climatology2.py):

- Creates climatology files
- Supports unstructured native grid datasets
- Takes input data location and output directory as arguments
Click on “Set Path” to choose the model data location. Optionally, you can type in a filter or state that data is climo files.

Click on “Use Observations 1” and “Set Path” to choose the obs data location. Click on the “Obs set 1” pull-down menu to choose among the obs sets in that location.
UV-CDAT Integration

Choose a plot set

Choose a variable and a season. Sometimes more options are available – a level set in this case.

Finally, click on Apply!
UV-CDAT Integration

Generated Diagnostics

Diagnostics Controls
Task Leads = Charles Doutriaux and Aashish Chaudhary
Documentation ([http://uvcdat.llnl.gov](http://uvcdat.llnl.gov))

- **Input/Output: cdms2**

- **Arrays: NumPy/NumPy.ma/MV2**

- **Utilities: genutil and cdutil**

- **Visualization: VCS**

- **Tutorials**
Quick Script Example

- import sys, os, cdms2, vcs
- x = vcs.init()
- x.setcolormap("rainbow")
- gm = vcs.createboxfill()
- f = cdms2.open(os.path.join(sys.prefix, 'sample_data', 'clt.nc'))
- s = f("clt")
- x.plot(s, gm)
- #x.png('test_vcs_basic_boxfill.png')
Plot Interaction

- import sys, os
- Import cdms2, vcs
- x = vcs.init()
- x.setbgoutputdimensions(1200, 1091, units="pixels")
- x.setcolormap("rainbow")
- gm = vcs.createboxfill()
- f = cdms2.open(os.path.join(sys.prefix, 'sample_data', 'clt.nc'))
- s = f("clt")
- x.plot(s, gm)
- x.interact()
Graphical User Interface
Visualization Improvements

- VCS 2D and VCS 3D code sharing
- Sophisticated labeling of contours
- Customization of font face, color, background, etc for each isoline.
- Better vector plots
- Various bug fixes, rendering improvements.
Software Quality “Control”

- **Workflow**
  - **Git** branch based workflow
  - **Master** and **Release** branches
  - **Review** and then merge
  - Automated testing using CTest / BuildBot / Travis CI

- **Testing**
  - 400+ tests
  - Tests algorithms, state changes etc.
  - Includes diagnostics, IO, visualization, and interactions
  - **Buildbot** is used to monitor the git repositories and schedule builds and tests when pull requests are created or the integration branches (master/release) are changed.
  - **CMake/CTest** deals with the actual building and testing, and then posting the results on dashboards hosted by **CDash**
Software Quality “Control”

**GIT Dashboard**

**BuildBot**

**Travis CI**
Task Lead = Matthew Harris

User Interface
All users must authenticate.
Easy access to every ESGF node
Convenient node search

Real-time node information and availability
Results from selected node with desired search criterion
Dashboard panel selector

Save and load your favorite layouts
User color scheme
Selecting a variable to display a plot
Classic Viewer

Menu

Output Display

Controls

Diagnostic Selection

ACME
Accelerated Climate Modeling for Energy