List of Figures

Simplifying and accelerating model evaluation by NASA satellite data

Dr. Charles S. Zender
Department of Earth System Science, University of California, Irvine

Project Summary.

The fidelity of geoscientific model results are increasingly evaluated by comparison to products derived from NASA satellite measurements. The satellite data are archived in HDF-EOS format, which is now a superset of the netCDF format employed by most geoscientific models. Putting NASA-generated (HDF-EOS) data and model-generated (netCDF) data on a common grid, in the same format, for numerical comparison can be arduous because of data format incompatibilities. Moreover, some analysis tools for netCDF data have no counterparts or equivalents for HDF-EOS data. Many researchers desire a common toolkit for both HDF-EOS and netCDF data that would 1. simplify and accelerate the independent analysis of both data formats (HDF-EOS and netCDF), 2. exploit the strengths of netCDF's underlying HDF data format with easy-to-use netCDF tools, 3. ease evaluations of model predictions (in netCDF format) by NASA-generated data (in HDF-EOS format).

This project will simplify the workflow involved in intercomparing HDF-EOS format data to model results in netCDF format. It will do so in a user-friendly and transparent way, by improving the netCDF Operators (NCO) which are robust components of the scientific data analysis software stack already employed at most Earth science modeling centers. The key NCO improvement will be to support group hierarchies. Groups are nestable namespaces that allow for hierarchical storage (the "H" in HDF). Utilizing groups to store ensembles of observations and predictions would vastly simplify and accelerate the characterization, evaluation, and intercomparison of multiple geophysical observations and simulations. Until now this has been impossible since NCO supports only "flat" datasets. The proof of this claim will be demonstrated by applying the improved NCO to a prototypical, NASA-relevant, Earth System Science research problem: to characterize, evaluate, and intercompare Earth System Model-simulated and NASA-retrieved snow cover and albedo trends and variability in the CMIP5 models to be used in the IPCC AR5 climate assessment.

NCO is a robust element of the scientific software stack used by the community of Earth Science researchers inside and outside of NASA for over fifteen years. Researchers worldwide employ NCO's user-friendly commands, honed through years of open source, developer-user feedback, to process terascale model datasets (often in preparation for comparison to HDF data). However, there is not yet an NCO-equivalent for processing HDF-EOS data. This is partly because NCO does not yet understand all the powerful HDF capabilities now accessible through the netCDF API. The project will remediate much though not all of this deficiency. The primary outcome will be the applicability of NCO to ever-increasing sets of HDF-EOS data, and of netCDF data, that utilize groups to organize and contain data. The proposed work directly responds to the ACCESS call to increase use of EOS data by the climate modeling community analyzing and evaluating the CMIP5 simulations. Moreover, the improved NCO capabilities will apply to all geophysical data archived in HDF-EOS and netCDF formats.

The significance of the proposed work is expected to be greatest for applied science researchers wishing to more fully exploit NASA data to evaluate model simulations. The PI is a long-standing climate modeler, software developer, and NASA-funded researcher who understands many of the barriers to model evaluation and who has developed, in the form of NCO, an elegant solution to some of them. The PI participates in the relevant geoscientific communities, including as a past reviewer for the ESDS Standards Process Group, for the last two IPCC climate assessments, and in the development of ESG-supported models such as the Community Earth System Model.

1 Introduction

The fidelity of geoscientific model results are increasingly evaluated by comparison to products derived from NASA satellite measurements. The satellite data are archived in HDF-EOS format, which is now a superset of the netCDF format employed by most geoscientific models. Putting NASA-generated (HDF-EOS) data and model-generated (netCDF) data on a common grid, in the same format, for numerical comparison can be arduous because of data format incompatibilities. Moreover, some analysis tools for netCDF data have no counterparts or equivalents for HDF-EOS data. Many researchers desire a common toolkit for both HDF-EOS and netCDF data that would 1. simplify and accelerate the independent analysis of both data formats (HDF-EOS and netCDF), 2. exploit the strengths of netCDF's underlying HDF data format with easy-to-use netCDF tools, 3. ease evaluations of model predictions (in netCDF format) by NASA-generated data (in HDF-EOS format).

This project will simplify the workflow involved in intercomparing HDF-EOS format data to model results in netCDF format. It will do so in a user-friendly and transparent way, by improving the netCDF Operators (NCO) which are robust components of the scientific data analysis software stack already employed at most Earth science modeling centers. The key NCO improvement will be to support group hierarchies. Groups are nestable namespaces that allow for hierarchical storage (the "H" in HDF). Utilizing groups to store ensembles of observations and predictions would vastly simplify and accelerate the characterization, evaluation, and intercomparison of multiple geophysical observations and simulations. Until now this has been impossible since NCO supports only "flat" datasets. Furthermore, wrappers will be added to facilitate direct application of NCO to HDF-EOS datasets. The efficacy of these improvements will be demonstrated by applying the improved NCO to a prototypical, NASA-relevant, Earth System Science research problem: to characterize, evaluate, and intercompare Earth System Model-simulated and NASA-retrieved snow cover and albedo trends and variability in the CMIP5 models to be used in the IPCC AR5 climate assessment.

This proposal is organized as follows. The relation of HDF-EOS and netCDF data formats to NASA-archived data and to geoscientific model evaluation are described in Section 2. Section 3 describes the type of geoscientific application that motivates our software project, and that will serve as proof-of-accomplishment. Our specific software engineering objectives and methods to improve the workflow involved in characterizing and intercomparing NASA data and model output, along with the project timeline to accomplish these tasks, appear in Section 4. Section 5 describes the results of our relevant, prior research. Section 6 summarizes the technology readiness, robustness, and persistence of the software project, as per ACCESS requirements. Projects related to ours and potential broader scientific impacts are in Section 7. A list of acronyms and abbreviations appears at the end as a supplementary document.

2 Background

This project integrates four existing, proven, Open Source software technologies: (1) HDF-EOS—the official storage standard for NASA EOS satellite data. (2) netCDF—the de facto standard format for climate model data and many Earth-bound geophysical observational datasets. (3) netCDF Operators (NCO)—a commonly used toolkit for file manipulation and analysis of netCDF data. (4) OPeNDAP—a standard, approved by the Earth Science Data System Working Groups (ESDSWG), for network-transparent client/server access to geophysical data. We will take advantage of existing synergies and convergence among these standards to extend the existing NCO toolkit to work with HDF-EOS data, and to exploit the hierarchical storage ability of the HDF format as implemented in both HDF-EOS and netCDF4. The result be an indispensable software assistant to individual researchers and research centers for local and distributed processing of climate and satellite data.

2.1 HDF-EOS

Two formats currently dominate geosciences data archival. The first is the Hierarchical Data Format (HDF) (*NCSA*, 2004), and the subset called (HDF-EOS) adopted by NASA for archiving data from its Earth Observing System (EOS). Hence HDF-EOS is the dominant format served by the EOS Data Information System (EOSDIS).

This project will add support to NCO to exploit advanced features of the HDF storage format, including its eponymous hierarchical storage capabilities. Long used with HDF-EOS data, the capabilities and potential efficiency of hierarchical storage have not yet been fully exploited by either satellite or model datasets. This project will also develop “wrappers” to enable application of of NCO directly to HDF-EOS files. Automatic support of hierarchical storage in HDF-EOS and netCDF files by data analysis tools like NCO will eliminate a longstanding bottleneck in evaluation and intercomparison of models and observations.

HDF-EOS presents a large and complex Application Programmer Interface (API) to programmers. Partly as a result of this, some toolkits like NCO do not natively support HDF-EOS data. However, HDF-EOS files are now easily convertible to netCDF4 format (described below). Hence the time is right to extend and adapt NCO to apply to both HDF-EOS and netCDF datasets, to ease analysis and intercomparison of data in these formats.

2.2 netCDF

The second popular scientific data format is the Network Common Data Format (netCDF), developed by Unidata at the National Center for Atmospheric Research (NCAR) (*Unidata*, 2004; *Rew and Davis*, 1990). netCDF has become the most commonly used archival format for large scale geophysical models, such as climate models.

netCDF version 3 (netCDF3) was significantly less-powerful than HDF-EOS because it lacks features such as data compression, irregular grids, threading, and parallel I/O. However, netCDF has a simpler API than HDF, and has been widely used in the geophysical and climate modeling community by practicing scientists.

Recently, netCDF version 4 (netCDF4) was introduced as fully backwards-compatible format that implements an enhanced data model containing many features long present in HDF (*Rew et al.*, 2006), including hierarchical storage. In fact, and by design, the back-end storage format of netCDF4 is HDF.

netCDF is the storage format used by the Coupled Model Intercomparison Project (CMIP) multi-model datasets CMIP3 and CMIP5, based on the most recent Intergovernmental Panel on Climate Change (IPCC) climate assessments (*Meehl et al.*, 2007). These datasets figure prominently in the geoscience application described in Section 4.

2.3 NCO: netCDF Operators

Tools to manipulate and view netCDF files are relatively easy to write since the API is considerably smaller and simpler than, say, HDF. The netCDF Operators (NCO) (*Zender*, 2011) may be the best-known toolkit for numeric and metadata analysis and manipulation of netCDF data.

Traditional processing of scientific data works with an intra-file paradigm. Users open a file, read a variable from the file, and manipulate it. The intra-file paradigm works well in cases where all the pertinent data are stored in one or a few files. In some disciplines, however, data storage requirements dictate that relevant data be spread over multiple files. Satellite-derived information, for example, may be stored in a file-per-day or file-per-orbit format. Data produced by geophysical time-stepping models is output every timestep or averaged over many timesteps. Climate models, for example, archive data once-per simulated day or month, and simulate years or even centuries producing hundreds or thousands of large files in a single simulation. In such applications, the intra-file paradigm becomes unwieldy and the optimal tool for data reduction must support an inter-file paradigm.

We developed some guidelines based on our extensive experience with geophysical and climate data and implemented them in NCO. NCO assumes that processing large numbers of geophysical data-files is most efficient and intuitive when:

1. Files are the fundamental unit of data. NCO makes it easy to add, subtract, and manipulate entire files.
2. Files to be processed in a single step are homogeneous. NCO assumes the structure of each file (i.e., the fields present and their dimensions) are identical to the structure of the first file in the sequence. The two exceptions NCO allows are that the record variable (i.e., time dimension) length, and, in some cases the number of variables present, may change from file to file.
3. Distinctions between *dimensions*, *coordinates*, and *variables* are maintained.
4. Operators have defaults that make sense and may be over-ridden with a simple, mnemonic command line switch.
5. Operators must provide an **audit trail that tracks data provenance**
6. Operators must be as generic as possible, imposing no limitations on data dimensionality, size, or type.
7. Conformance to metadata conventions is paramount

Apparently NCO's guiding philosophy, "do what a sane user would want" has succeeded! NCO runs on all modern operating systems, and its use is fully detailed in the NCO User's Guide. To my knowledge, all established national and international climate modeling centers, including NASA, NOAA, NSF, and DOE centers install and maintain NCO for their system users. See, for example, NCO usage at DOE ARM, DKRZ, LMD, JISAO, NCAR, NOAA GFDL, NOAA CDC, and PRISM. In other words, NCO is widely used as middle-ware at geophysical institutions for data post-processing, hyper-slabbing and serving. The improvements and extensions to NCO proposed here will improved the shared-cyberinfrastructure that will include a whole new scientific community, the users of HDF-EOS products.

2.4 DAP

The Open-source Project for a Network Data Access Protocol (OPeNDAP) Data Access Protocol (DAP) is a data transmission protocol designed specifically for science data. DAP libraries implement network-transparent access to data via a DAP server that in turn uses the HTTP protocol for communication. Although DAP-technology originated with oceanography data, it applies to virtually all scientific data. DAP 2.0 was adopted as an ESDS standard (ESDS-RFC-004). For simplicity, we use OPeNDAP and DAP interchangeably, concealing some nuances in their meaning (namely the organization vs. the protocol, and the specific implementor of the protocol). The NCO User's Guide and this OPeNDAP Presentation provide more details.

Any binary netCDF application (like NCO) may be DAP-enabled by linking to an DAP-enabled netCDF Client Library, such as that included in the standard netCDF distribution since version 4.1 (roughly 2010). Once NCO is DAP-enabled the operators are DAP clients. All DAP clients have network transparent access to any files controlled by a DAP server. All of the commands described in the following sections may be issued to files residing on a DAP server simply by replacing the filename arguments with the HTTP DAP server equivalents.

3 Geoscientific Domain Application

We will apply our improvements to the problem of analyzing HDF-EOS and netCDF datasets pertaining to Earth's water and energy cycle. In particular we will analyze observed and simulated snow cover (a form of surface water storage) and snow albedo (together with clouds, the primary determiner of shortwave energy redistribution in the cryosphere). The PI has active NASA and NSF projects on the energy balance of the cryosphere that motivate this domain application.

We choose the sub-discipline of climate data analysis for our application for two additional reasons. First, the NASA ACCESS call specifically mentions the CMIP5 multi-model dataset being assembled for use in the IPCC AR5. Second, our analysis involves tangible quantities (e.g., snow cover, temperature) and dimensions (latitude, longitude, height, time) which are familiar to all geophysicists regardless of their field. We emphasize that the improvements made to NCO will benefit any any domain where HDF-EOS or netCDF is the storage format of choice.

Simulations contributing to the CMIP5 dataset must be completed in 2011. Multiple climate simulations from each of approximately two dozen contributing models will be archived by, though not necessarily at, the Program for Climate Model Diagnosis and Intercomparison (PCMDI). URLs referencing the CMIP5 data files will have a site-dependent prefix followed by a Data Retrieval Syntax (DRS) -specified directory structure. Our domain application can apply to CMIP5 data stored remotely on the Earth System Grid (ESG) and accessed via DAP, as well as to CMIP5 data downloaded to the local directory structure.

Section 1.2.2 of the ACCESS call states that improving accessibility of the modeling analysis community to NASA data is a priority. We will illustrate the current state of accessibility and the goals of our project by asking three scientific questions, in increasing order of data analysis complexity:

1. What are the characteristic trends and variability of snow cover and albedo recorded by NASA MODIS or CERES observations?
2. To what extent do historical simulations of snow cover and albedo by model X agree with NASA MODIS or CERES observations?
3. To what extent do historical simulations of snow cover and albedo by all models agree with NASA MODIS or CERES observations?

The PI's group has written numerous paper on specific, science based hypotheses regarding snow cover and albedo evolution (*Flanner et al.*, 2007, 2009; *Wang and Zender*, 2010a,b, 2011; *Allen and Zender*, 2010, 2011a,b). These physical processes involved include feedbacks between dynamics, snow cover, temperature, and snow albedo. We are most interested in assessing the fidelity of model representation of snow-albedo feedback (i.e., albedo response to the seasonal cycle of temperature) because it is measurable by satellite and it explains most of the model differences in snow-albedo feedback to future climate change (*Hall and Qu*, 2006).

3.1 Accelerated Analysis of HDF-EOS data

Question 1 above demonstrates the problem of analyzing timeseries of HDF-EOS datafiles. *Zender and Mangalam* (2007) and *Zender* (2008) describe efficient solutions for determining the statistics (e.g., mean, trends, and variability) of data in netCDF classic format. This project must and will expand the applicability of those techniques to both HDF-EOS format data and to netCDF4 data employing hierarchical storage.

When one retrieves from the EOS ClearingHOuse (ECHO) the pertinent MODIS snow cover data on (for simplicity) the Climate Modeling Grid (CMG), for all months of 2007, one obtains twelve HDF-EOS files:

```
MOD10CM.A2007001.005.2007108111758.hdf
MOD10CM.A2007032.005.2007108111758.hdf
.
.
MOD10CM.A2007335.005.2008003152659.hdf
```

To proceed, many users (like this one) convert these HDF-EOS files to netCDF format. This is not strictly necessary since many data processing languages such as the Interactive Data Language (IDL), the NCAR Command Language (NCL), and Matlab understand HDF-EOS formatted data. One reason why many users convert HDF-EOS to netCDF formats is to facilitate processing and file manipulation of large numbers of large files with NCO.

NCO is explicitly designed around the paradigm of the file being the unit of data (*Zender*, 2008) so that the 2007 yearly mean MODIS CMG snow cover data is easily obtained by

```
ncra MOD10CM.A2007*.nc  MOD10CM_2007_average.nc  # Works now (flat files)
hera MOD10CM.A2007*.hdf MOD10CM_2007_average.hdf # This project task
```

The first line shows how the NCO operator `ncra` (netCDF Running Averager) uses shell wildcarding to simplify multi-file netCDF data manipulation (averaging, in this case) (*Zender*, 2008). The second line shows the invocation of `hera`, the proposed HDF-EOS Running Averager to be developed in year one by this project.

To answer questions like Question 1 above, requires this project to complete two significant tasks called Task 1 and Task 2 in Section 4. Briefly, Task 1 refers to the implementation of support in NCO for hierarchical groups using the netCDF4 API (which, recall, is itself implemented with the HDF5 API). This hierarchical group feature will allow analysis of files that contain not only a single, top-level group (the netCDF3, aka netCDF classic model, default) but also arbitrary number of nested groups with their own namespaces.

For example, EOSDIS archives NASA High Resolution Dynamics Limb Sounder (HIRDLS) datasets in HDF5-EOS (`*.he5`) format. These datasets are not "flat"—the top level group in each file is named `HIRDLS` and it contains two subgroups, `Data_Fields` and `Geolocation_Fields`. The `Data_Fields` group contains the data fields that interest most users, e.g., `CFC11`, the CFC-11 mixing ratio.

In the future it is likely that model and satellite datasets will further exploit the hierarchical group capabilities of HDF to store multiple fields in groups. In the single-model examples discussed below, we envision each model realization of an ensemble of runs from a single model being stored as a first level group within the file. In the multi-model examples discussed below, each model might form a top-level group and its ensemble of realizations would be stored as second level groups within the file. Task 1 will implement in NCO the ability to traverse an arbitrarily-deep hierarchy of nested data. Hence arithmetic properties (e.g., averages, differences and derived variables) of nested data files will be as easy to compute as the arithmetic properties of flat (no groups) files currently are with NCO.

Task 2 is the implementation of wrappers (e.g., `hera`) that transparently (to the user) convert HDF-EOS data to netCDF4 data so that the hierarchical operators written with the netCDF API in completing Task 1, can be applied to HDF-EOS data.

We may call, by analogy, the HDF-EOS wrappers for the NCO operators the HDF-EOS Operators (HEO). Then Question 1 is completely answered by HEO statistical operations analogous to those already employed by NCO users and described in *Zender* (2008) and in the NCO User's Guide (*Zender*, 2011). We will assume that entire HEO toolkit will be developed (as wrappers) based upon the corresponding operators of the NCO toolkit. Section 4 describes the details of how will build these wrappers. The HEO toolkit will output netCDF4 files that are a valid subset of the HDF-EOS format. Hence no conversion of output from netCDF4 to HDF-EOS will be necessary, since netCDF4 files are valid HDF-EOS files.

3.2 Accelerated evaluation of single model instances by HDF-EOS data

For concreteness in discussing Question 2, let us consider a single model, the Community Earth System Model (CESM) from the National Center for Atmospheric Research (NCAR). ESG nodes will host tens, if not hundreds, of terabytes of CESM data. CESM, like all CMIP5 models, will be used in historical simulations (aka climate hindcasts, based on observed forcings and emissions) as well as future simulations. For robust comparison of CESM simulations with NASA observations of, say, snow cover and albedo, one would like account for the internal variability (aka, deterministic chaos) of the model by integrating it forward (with time-evolving forcings) from slightly different initial conditions. This results in an *ensemble* of model datasets.

One key advantage of hierarchical storage is that it naturally expands to the dimensionality of the experimental design, including ensemble experiments. In the CMIP3 dataset for AR4, many NCAR Community Climate System Model (CCSM, the immediate predecessor of the CESM) simulations were repeated eight times with perturbed initial conditions. Currently a single, eight-member ensemble of CESM simulations for CMIP5 would be evaluated against 2007 MODIS snow cover data using a meta-script such as

```
for run in '1 2 3 4 5 6 7 8'; do
  ncdiff CCSM_2007_${run}.nc MOD10CM.A2007.nc CCSM_${run}_minus_MODIS_2007.nc
done
```

With the support for hierarchical storage that this project will implement in NCO, it would make more sense to store all eight ensemble members in the same file, each in its own top-level group. The same data analysis would then be accomplished with a much simpler NCO invocation, i.e.,

```
ncdiff CCSM_2007_all.nc MOD10CM.A2007.nc CCSM_minus_MODIS_2007.nc
```

The explicit loop over each file in the ensemble has been replaced, automagically, by subtraction within the single `ncdiff` invocation as it recursively-descends through the input files. Hierarchical storage of ensembles will simplify file management for data providers by reducing the depth of the filesystem hierarchy (by eliminating the "run" level for ensembles), and by reducing the overall number of files (by a factor equal to the number of members of the ensemble).

3.3 Accelerated evaluation of model ensembles by HDF-EOS data

Each historical IPCC AR5 emissions scenario will be simulated by models representing all major international climate modeling groups. Each model in this ensemble will perform an ensemble of simulations. Hence the full CMIP5 multi-model dataset for a given historical scenario comprises and ensemble of ensembles. Currently such "grand ensembles" would usually be evaluated in a double loop (over models, and simulations of each model, respectively). The barriers that currently face researchers who wish to analyze the "grand ensemble" of CMIP5 data include aggregating the climate data from remote servers to the local host (which may have relatively limited storage), organizing the pertinent files, and writing scripts to perform the looping.

The hierarchical data storage support described above reduces this complexity considerably. Instead of meta-code loops that resemble

```
for model in 'CESM GISS ECHAM ...'; do
    for run in '1 2 3 4 5 6 7 8'; do
        ncdiff ${model}_2007_${run}.nc MOD10CM.A2007.nc \
               ${model}_${run}_minus_MODIS_2007.nc
    done
done
```

storage of the grand ensemble results in a hierarchical manner collapses both dimensions of the loop, which NCO handles by the recursive descent traversal of the hierarchy within the single model file that contains the entire grand ensemble

```
ncdiff CMIP5_2007_all.nc MOD10CM.A2007.nc CMIP5_minus_MODIS_2007.nc
```

In other words, the analysis "script" is virtually the same for multi-model as for single-model ensembles. This illustrates, we believe, only some of the untapped power of hierarchical data storage to simplify and accelerate evaluation of models by NASA HDF-EOS datasets.

As part of the PI's previous NSF-funded project, NCO was optimized for lengthy, script-based data analysis workflows. The Script Workflow Analysis for MultiProcessing (SWAMP) actually compiles shell-scripts of NCO commands, identifying basic blocks, intermediate files, and results (*Wang et al.*, 2007a,b). Execution of basic blocks is scheduled optimally (*Wang et al.*, 2008, 2009), intermediate files are automatically removed, and the user receives only the results. Hence, SWAMP and NCO together on a DAAC server would give users a transparent system to perform server-side (e.g., at the DAAC) data reduction and analysis. The benefit of server-side analysis is, of course, that it significantly reduces the total amount of data that must be sent over the network to the user. However, SWAMP has never been used operationally and is only TRL 4, insufficiently mature to complete for this ACCESS proposal. This proposal aims to put an improved NCO squarely in the hands of users and researchers at their desks, so it would be beyond the scope of this proposal to focus on server side data reduction at NASA and/or ESG centers. We mention SWAMP mainly to demonstrate that NASA could support NCO-based workflows allowing server-side data analysis and intercomparison of HDF-EOS and netCDF data in some later project.

4 Tasks: Software Engineering and Configuration

As described in Section 3, this project must accomplish two main software engineering tasks:

1. Implement support for hierarchical groups
2. Create the wrappers called HEO, i.e., HDF-EOS wrappers for NCO that may be applied directly to HDF-EOS files.

4.1 Task 1: Hierarchical Groups

The most important, and difficult, software engineering task in this project is the refactoring of the NCO codebase to incorporate support for group hierarchies. Currently NCO supports only "flat" data files, i.e., those comprising a single top-level group, also known as a global group. To support arbitrarily nested groups, each with its own parent, children, and namespace, the NCO code must be refactored to:

1. Recursively descend through all groups when processing files
2. Track the relationships between groups, each of which may have a single parent, multiple siblings, and multiple children.
3. Implement namespaces so that variable, attribute and dimension names in one space are visible to children yet invisible to siblings (and their children).

The crux of the software engineering in the project will be transforming the code from simple single- or doubly-nested loops (usually over files and the variables in each file, respectively) to a recursive-descent code. This will be implemented in software with a tree structure to track group relationships in the hierarchy. Command line syntax will be chosen so that users may restrict which groups an operation (e.g., averaging) applies to. The current and future forms allowed for selecting variables from a file, for example, will be something like

```
ncra -v temperature        in.nc  out.nc # Current for netCDF files
ncra -v group1:temperature in.nc  out.nc # Future  for netCDF files
hera -v temperature        in.he5 out.nc # Future  for HDFEOS files
hera -v group1:temperature in.he5 out.nc # Future  for HDFEOS files
```

Note that the .nc suffix on the output files simply indicates that out.nc is a valid netCDF4 file. All netCDF4 files are, by definition, valid HDF5 files. The HDF Group maintains the h4toh5 and h5toh4 programs which automatically convert between output HDF formats, should the netCDF4 format not be desirable. This would be accomplished in the same manner as the wrappers described next in Section 4.2.

The operators will also need to be equipped with a more forgiving algorithm to determine which variables in one file match those in a comparison file. Consider, for example, the difference between a model-generated and satellite-observed dataset such as

```
ncdiff -v snow_cover CESM.nc MODIS.hdf difference.nc
```

Currently, NCO will only subtract variables of the identical name in the second file from the first file and place the difference in the third file. With the possibility that either or both the first and second files may contain an ensemble of simulations (stored as distinct groups), each of which may contain a snow_cover variable, NCO must be taught to "broadcast" the arithmetic operation (here subtraction) from the first dataset throughout all conforming groups in the second dataset. Broadcasting is described in more detail in *Zender and Mangalam* (2007). In other words, if CESM.nc contains eight different snow_cover variables (one per top-level group) and MODIS.hdf contains only one snow_cover variable (presumably representing the best observed snow cover estimate), then difference.nc will also contain eight top-level groups each containing the snow_cover field corresponding to the ensemble in CESM.nc differenced from MODIS.nc.

4.2 Task 2: HDF-EOS Wrappers for NCO

As mentioned above, we will create wrappers (e.g., hera to wrap ncra) that transparently (to the user) convert HDF-EOS data to netCDF4 data so that the hierarchical operators can be applied. The goal of the conversion is to produce netCDF4 files with which NCO already works (with the exception of hierarchical groups to be implemented in Task 1). It is not strictly necessary to implement the conversion functionality as wrappers. The NCO executables could be modified to automatically detect the format of each input file. Those with HDF suffixes (i.e., .hdf, .he2, .he4, .hdfeos, .he5), or with HDF "magic numbers" as determined by an octal dump (i.e., od -An -c -N4 input results in 211 H D F) would be converted to netCDF4 prior to processing.

Conversion from HDF-EOS to netCDF4 could be done multiple ways. These include the NCL ncl_convert command, which handles HDF-SDS, HDF4-EOS, and HDF5-EOS formats. The HDF Group-supported eos52nc4 and aug_eos5 command can accomplish the equivalent when converting HDF5-EOS files to netCDF4 format. Operators which take multiple input files will parallelize conversion of those files (up to some reasonable maximum number in advance). Conversion would take place in, and the resulting netCDF4 file would be stored in, a secure directory which would default either to the current working directory or /tmp. Or a location configured with environment variables or switches.

NASA satellite data are currently archived in at least three versions of the HDF format (HDF-SDS, HDF4-EOS, and HDF5-EOS). The file conversion alone will not be sufficient to allow intercomparison of model with satellite data. Preliminary testing shows that EOSDIS-distributed HDF files often contain empty groups, field names spelled slightly differently than model and netCDF conventions (e.g., Time instead of time and MissingValue instead of missing_value). It would be onerous for the user to have to manually rename (e.g., with NCO's ncrename) all such minor discrepancies before the satellite and model data could be compared. Hence, in addition to converting the HDF data, NCO will be taught to recognize common paradigms and equivalancies between model conventions, such as the Climate Forecast (CF) metadata convention (*Gregory*, 2003), and NASA EOSDIS conventions. We will start with the most common use case for evaluating CMIP5 models, which is probably the analysis and comparison of the models to HDF-SDS or HDF4EOS data on the climate modeling grid (CMG).

4.3 Schedule and Milestones

Year 1. *Goals*: Implement group hierarchies

1. PI: Spin-up programmer on NCO API and train student on science issues
2. PI: Manage NCO releases and dissemination to community
3. Programmer: Refactor NCO codebase to support group hierarchies (Task 1)
4. Programmer: Participate in ESDSWG activities (Section 6.3)
5. Grad student: Develop CMIP5 analysis scripts to evaluate multi-model snow cover and albedo against MODIS and CERES data (Section 3)

Year 2. *Goals*: Document and harden releases

1. PI: Manage NCO releases and dissemination to community
2. Programmer: Develop wrappers to apply NCO directly to HDF-EOS files (Task 1)
3. Programmer: Document and illustrate NCO usage with HDF-EOS files
4. Programmer: Participate in ESDSWG activities (Section 6.3)
5. Grad student: Apply NCO improvements to CMIP5 analysis (Section 3)
6. PI and Grad student: Write paper on multi-model and observed snow trends
7. PI and Programmer: Write paper benchmarking workflow improvements

5 Results from Prior Funding on Related Projects

Zender was PI on two previous scientific data analysis related projects: First, NSF ATM-0321380, $1,105,000, 2003-2006, *MRI: Acquisition of an Earth System Modeling Facility (ESMF) for Coupled Climate, Chemistry, and Biogeochemistry Studies*. The ESMF was UCI's main supercomputer facility from 2004–2009. Over 50 students and researchers used ESMF in graduate modeling seminars and in terascale climate simulations. Also Used ESMF as an OptIPuter node to demonstrate high-bandwidth distributed data reduction and analysis. Second, NSF IIS-0431203, $594417, 2004–2008, *SEI(GEO): Scientific Data Operators Optimized for Distributed Interactive and Batch Analysis of Tera-Scale Geophysical Data*. Improved, invented, implemented, benchmarked, and distributed new capabilities for the netCDF Operators (NCO). Led to one Master's and one PhD (D. Wang) degree in Computer Science on "Compilation, Locality Optimization, and Managed Distributed Execution of Scientific Dataflows". Developed Script Workflow Analysis for Multi-Processing (SWAMP, `http://swamp.googlecode.com`). Resulted in four peer-reviewed papers and twelve conference abstracts (`http://nco.sf.net#pub`).

Zender has been PI on three NASA science projects which involved comparison of GCM simulations in netCDF format to NASA satellite data in HDF-EOS format: First, NASA New Investigator Program (NIP) (NAG5-10546), $330k, 2001–2004, *Influence of Mineral Dust Aerosol on the Chemical Composition of the Atmosphere*. Second, NASA Earth and Space Science Fellowship (ESSF) for Mark Flanner, $48000, 2005–2007, *Climate Sensitivity To Snow Radiative Processes: Improving Physical Representation And Understanding With MODIS/MISR*. Third, NASA International Polar Year (IPY06 NNX07AR23G), $607000, 2007–2011, *Black Carbon Impacts on Cryospheric Climate Sensitivity and Surface Hydrology*. HDF-EOS data analyzed in the course of these projects include products from MODIS (optical depth, snow cover, snow albedo), AMSR-E (soil moisture), and QuikSCAT (wind speed). These projects have produced a few dozen papers, and, more to the point, provided first-hand experience for understanding the needs of researchers in intercomparing netCDF and HDF-EOS data.

6 Technology Readiness, Robustness, and Persistence

6.1 Technology Readiness Level

This project adheres to the NASA ACCESS guidelines of focusing on near-term improvements in high Technology Readiness Level (TRL) technologies to improve the usability of NASA data. netCDF and HDF-EOS are, clearly, both high TRL components of the model geoscientific computing software stack. They both have complete user documentation, years of operational use in mission-critical environments, and dedicated engineering support teams in place.

In my estimation netCDF, HDF-EOS, and DAP are adults (TRL 9) whereas NCO is a teenager (TRL 8). NCO was first adopted by a wide user community (the NCAR Climate and Global Dynamics division) in 1996. In the fifteen years since, NCO has been deployed at all known geoscientific computing centers, and on thousands of desktops. It has a fairly complete User's Guide and regression test. NCO is also often taught as part of both the NCL and netCDF data analysis and provider workshops.

6.2 Improvements to NCO Software Robustness

If past is prologue then the most frequently asked question on the NCO forums as this project progresses will be "How do I build NCO version W on platform X with netCDF version Y and HDF version Z". In the course of performing the tasks in Section 4, the project team, and especially the Scientific Programmer, must improve the NCO build procedure until a straightforward `./configure;make;make test;make install` works (or fails with a useful error) on most platforms.

We will pursue two specific strategies to improve software robustness. First, the NCO configuration and regression testing mechanisms will be expanded to test the new HDF-EOS capabilities. Second, binary releases of NCO will be made in both RedHat Package Manager (RPM) format (used by RedHat Enterprise, Fedora, and CentOS flavors of Linux) and in Debian format (used by Debian and Ubuntu distributions). These binary release formats ensure stringent cross-platform portability and documentation requirements are met, as well as correctness of dependencies on shared libraries. They are critical to ensuring users can easily install and utilize NCO on their personal workstations.

6.3 Participation in Earth Science Data Systems Working Groups

The PI participates in the relevant geoscientific and IT communities. He is a past reviewer for the ESDS Standards Process Group (SPG), having commented on (to the best of my recollection) the Data Access Protocol (DAP), HDF 5, and netCDF Classic standards.

The project requests funds for a Scientific Programmer, 25% of whose time is reserved for participation in Earth Science Data System Working Group (ESDSWG) activities. The most like groups are the SPG and the Technology Infusion Working Group (TIWG). The PI recognizes that participation in such groups requires a combination of NASA scientific data and IT experience that not all scientific programmers possess. Should the programmer begin with insufficient experience for ESDSWG participation, the PI will take-over this function until the programmer is adequately trained. The PI will commit to this role indefinitely since he is very interested in infusing new software technology into the perennial problems encountered by researchers in managing large datasets.

6.4 Persistence of ACCESS Tools

There are at least three reasons why the NCO project will persist after ACCESS funding—its license and history, current trends in geophysical IT, and the scientific needs of its developers. The PI holds the copyright to all NCO source code and has released it under the terms of the Free Software Foundation's GPL for over fifteen years. Hosting the code on SourceForge.net and opening it to contributors (there have been seven active developers throughout the project) under the conditions of the GPL3 license ensures that no code is ever lost to the community. Development is relatively transparent. Bugs are tracked and comments promptly responded to.

Extending NCO to apply to HDF-EOS datasets is expected to significantly increase the user-base, and, subsequently help requests. The SourceForge infrastructure is known to scale to much larger projects than NCO. NCO user forum activity is roughly 1.5 messages per day (over the last year), a level we have maintained for about the last seven years (statistics since 2000 available here). Part of the Scientific Programmer's responsibility will be to assist helping these new users on the NCO forums.

netCDF and HDF-EOS have secured their positions as the dominant geoscientific data formats and their convergence has made NCO more relevant than ever. The PI has developed, ported, maintained, and secured support (only once, from NSF) for NCO over fifteen years. Hence NCO has been supported organically (by volunteer developers, code contributors and users) for eleven of its fifteen years. As a tenured professor, with life-long job security in climate science, the PI has every incentive to continue improving the utility of NCO for his own and his students' research, and to release NCO as free and open-source software since it is a fun and rewarding form of scientific community service.

7 Related Projects, Impacts, and Education

7.1 Related Projects

PI Zender has long collaborated informally with the netCDF development team at Unidata. He provides feedback and testing of the netCDF library, and Unidata reciprocates with software support and advice on exploiting netCDF features. Zender and netCDF developer Russ Rew have submitted one proposal (declined) and intend to submit another on improvements and extensions to the netCDF API.

Zender and NCAR Earth System Grid (ESG) PI Don Middleton are developing a proposal to integrate the advanced scheduling features of SWAMP (*Wang et al.*, 2009) into the ESG.

7.2 Dissemination to NASA Centers

NCO does not yet handle HDF-EOS files and so NCO may not already be installed at NASA Centers that primarily serve as repositories for HDF-EOS files, such as EOSDIS Distributed Active Data Archive (DAAC) sites. Our team would ascertain the extent to which interactive data analysis and processing takes place at the DAACs. (NCO is used in the back-end of many GUI front-ends that let users select data hyperslabs from repository data.) Should the need and opportunity for NCO to play similar roles at NASA DAACs seem apparent, we would consult with the ESDS Technology Infusion Working Group about possible next steps.

Larger geoscientific computing centers that are repositories for model data (e.g., ESG nodes) usually already have NCO installed. Dissemination of NCO to such centers will follow the normal path as system administrators upgrade to newer NCO version on a 1–2 year cycle.

7.3 Education

This project trains one graduate student in software engineering for geoscience data analysis. UC Irvine is a US Department of Education Minority Serving Institution with large pools of under-represented minorities (URMs) potentially interested in pursuing undergraduate research projects. The UCI ESS department where Zender teaches has three programs that pipeline URMs to ESS research opportunities: (1) CAMP (Campus

Alliance for Minority Participation) in Science, Engineering and Math, (2) an NSF REU in ESS (2011–2013, PI E. Saltzman of ESS), and (3) the Undergraduate Research Opportunities Program (UROP). With these resources Zender will open a year-round undergraduate research position in his group at no additional cost to NASA.

Moreover, PI Zender helps train Orange County K–12 teachers in Earth Science curricula. An NSF Math Science Partnership project called FOCUS (Faculty Outreach Collaborations Uniting Scientists, Students and Schools) brings the teachers to UCI. Zender incorporates NASA materials into his FOCUS seminars. his outreach seminars to students (of all ages) of the Osher Lifelong Learning Institute (OLLI). Finally, Zender is member of the Long Beach Aquarium of the Pacific Science on a Sphere (SOS) team that brainstorms new ways to demonstrate to the general public the value of NASA satellite-retrieved data in understanding the Earth and its oceans.

References

Bibliography

Allen, R. J., and C. S. Zender (2010), The effects of continental-scale snow albedo anomalies on the wintertime Arctic Oscillation, *J. Geophys. Res. Atm.*, *115*(D23105), doi:10.1029/2010JD014,490. 3

Allen, R. J., and C. S. Zender (2011a), Forcing of the Arctic Oscillation by Eurasian snow cover, *J. Climate*, *24*(24), 6528–6539, doi:10.1175/2011JCLI4157.1. 3

Allen, R. J., and C. S. Zender (2011b), The role of eastern Siberian snow and soil moisture anomalies in quasi-biennial persistence of the Arctic and North Atlantic Oscillations, *J. Geophys. Res. Atm.*, *116*(D16125), doi:10.1029/2010JD015,311. 3

Flanner, M. G., C. S. Zender, J. T. Randerson, and P. J. Rasch (2007), Present-day climate forcing and response from black carbon in snow, *J. Geophys. Res.*, *112*, D11,202, doi:10.1029/2006JD008,003. 3

Flanner, M. G., C. S. Zender, P. G. Hess, N. M. Mahowald, T. H. Painter, V. Ramanathan, and P. J. Rasch (2009), Springtime warming and reduced snow cover from carbonaceous particles, *Atmos. Chem. Phys.*, *9*(7), 2481–2497, doi:10.5194/acp-9-2481-2009. 3

Gregory, J. (2003), The CF metadata standard, *CLIVAR Exchanges*, *8*(4), 4. 4.2

Hall, A., and X. Qu (2006), Using the current seasonal cycle to constrain snow albedo feedback in future climate change, *Geophys. Res. Lett.*, *33*(L03502), doi:10.1029/2005GL025127. 3

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007), The WCRP CMIP3 multimodel dataset: A new era in climate change research, *Bull. Am. Meteorol. Soc.*, *88*(9), 1383–1394, doi:10.1175/BAMS–88–9–1383. 2.2

NCSA (2004), *Hierarchical Data Format*, National Center for Supercomputing Applications, Champaign-Urbana, IL, `http://hdf.ncsa.uiuc.edu`. 2.1

Rew, R., and G. Davis (1990), NetCDF: an interface for scientific data access, *IEEE Comput. Graph. Appl.*, *10*(4), 76–82, doi:10.1109/38.56302. 2.2

Rew, R., E. Hartnett, and J. Caron (2006), NetCDF-4: Software implementing an enhanced data model for the geosciences, in *Proceedings of the 22nd AMS Conference on Interactive Information and Processing Systems for Meteorology*, p. 6.6, American Meteorological Society, AMS Press, Boston, MA. 2.2

Unidata (2004), *Network Common Data Format*, University Corporation for Atmospheric Research, Boulder, CO, `http://www.unidata.ucar.edu/packages/netcdf`. 2.2

Wang, D. L., C. S. Zender, and S. F. Jenks (2007a), Server-side parallel data reduction and analysis, in *Advances in Grid and Pervasive Computing, Second International Conference, GPC 2007*, *IEEE Lecture Notes in Computer Science*, vol. 4459, edited by C. Cérin and K.-C. Li, pp. 744–750, doi:10.1007/978–3–540–72,360–8_67, Springer-Verlag, Berlin/Heidelberg, doi:10.1007/978-3-540-72360-8_67. 3.3

Wang, D. L., C. S. Zender, and S. F. Jenks (2007b), A system for scripted data analysis at remote data centers, *Eos Trans. AGU*, *88*(52), fall Meet. Suppl., Abstract IN11B-0469. 3.3

Wang, D. L., C. S. Zender, and S. F. Jenks (2008), Cluster workflow execution of retargeted data analysis scripts, pp. 449–458, 10.1109/CCGRID.2008.69. 3.3

Wang, D. L., C. S. Zender, and S. F. Jenks (2009), Efficient clustered server-side data analysis workflows using SWAMP, *Earth Sci. Inform.*, *2*(3), 141–155, doi:10.1007/s12,145–009–0021–z. 3.3, 7.1

Wang, X., and C. S. Zender (2010a), MODIS snow albedo bias at high solar zenith angle relative to theory and to *in situ* observations in Greenland, *Rem. Sens. Environ.*, *114*(3), 563–575, doi:10.1016/j.rse.2009.10.014. 3

Wang, X., and C. S. Zender (2010b), Constraining MODIS snow albedo at large solar zenith angles: Implications for the surface energy budget in Greenland, *J. Geophys. Res. Earth Surf.*, *115*, F04,015, doi:10.1029/2009JF001,436. 3

Wang, X., and C. S. Zender (2011), Arctic and Antarctic diurnal and seasonal variations of snow albedo from