Easy Access to and Analysis of NASA and Model Swath-like Data

Dr. Charles S. Zender[1], Dr. Christopher S. Lynnes[2], and Walter Baskin[3]
[1]Departments of Earth System Science and Computer Science, University of California, Irvine
[2]NASA Goddard Space Flight Center GES DISC, [3]NASA Langley Research Center ASDC

Project Summary.

Swath-like data (hereafter SLD) are defined by non-rectangular and/or time-varying spatial grids in which one or more coordinates are multi-dimensional. It is often challenging and time-consuming to work with SLD, including all NASA Level 2 satellite-retrieved data, non-rectangular subsets of Level 3 data, and model data on non-rectangular grids. Researchers and data centers would benefit from user-friendly, fast, and powerful methods to specify, extract, serve, manipulate, and thus analyze, SLD. This project addresses these needs by extending the functionality, user-base, and integration into NASA data services of the netCDF Operators (NCO), an open-source scientific data analysis software package which our current ACCESS project has augmented with HDF capabilities applicable to most archived and virtually all new NASA-distributed data.

The remote sensing and the weather and climate modeling and analysis communities face similar problems in handling SLD including how to easily: 1. Specify and mask (include/exclude) irregular regions such as ocean basins and political boundaries in SLD (and rectangular) grids. 2. Bin, interpolate, average, or re-map SLD to regular grids. 3. Derive secondary data from given quality levels of SLD. These common tasks require a data extraction and analysis toolkit that is SLD-friendly and, like NCO, is used in both communities. We will improve NCO to support these tasks so that users can 1. Analyze regions specified by familiar names (including Climate-Forecast standardized region names) or user-supplied boundaries. 2. Access sophisticated statistical and regridding functions that are robust to missing data. 3. Create, save, share, and re-use user-defined functions or "recipes". These capabilities will ease data specification, software-reuse, and, because they apply to SLD, minimize transmission, storage, and handling of unwanted data. Proof of these claims 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 surface energy budget trends and variability in Greenland, a cryosphereic region experiencing rapid darkening and nearly unprecedented melt.

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. We will coordinate infusion of the enhanced NCO into NASA Earth Science data services offered by the Goddard DISC and the LaRC ASDC. Collaboration will include working with data portals such as Giovanni (which already uses NCO) to understand, refine, and address data center needs for SLD capabilities. These centers will expand the community of users interested in their Level 2 data, while researchers will benefit from simplified processing of SLD with higher spatial and temporal resolution and information content than Level 3 data. The proposed work will improve users' ability to access and use EOS data, with methods seamlessly adopted by the modeling and model analysis communities (both are ACCESS 2013 goals). The improved NCO capabilities will apply to all geophysical data archived in HDF and netCDF formats.

The PI is a long-standing climate modeler, software developer, and NASA-funded researcher who understands obstacles to model evaluation by and use of NASA data and who has developed, in the form of NCO, an elegant solution to some. One Co-I is a long-standing information systems architect familiar with NASA data services' operations, community, and vision. The other is primary software engineer responsible for serving data from ASDC. We participate in relevant geoscience communities, including ESDS working groups, IPCC climate assessments, and community model development.

1 Introduction

Swath-like data (hereafter SLD) are defined by non-rectangular and/or time-varying spatial grids in which one or more coordinates are multi-dimensional. It is often challenging and time-consuming to work with SLD, which includes all NASA Level 2 satellite-retrieved data, non-rectangular subsets of Level 3 data, and model data on non-rectangular grids. Researchers and data distribution centers would benefit from user-friendly, fast, and powerful methods to specify, extract, serve, manipulate, and thus analyze, SLD. This project addresses these needs by extending the functionality, user-base, and integration into NASA data services of the netCDF Operators (NCO), an open-source scientific data analysis software package which our current ACCESS project has augmented with HDF capabilities applicable to most archived and virtually all new NASA-distributed data.

The remote sensing and the weather and climate modeling and analysis communities face similar problems in handling SLD including how to easily: 1. Specify and mask (include/exclude) irregular regions such as ocean basins and political boundaries in SLD (and rectangular) grids. 2. Bin, interpolate, average, or remap SLD to regular grids. 3. Derive secondary data from given quality levels of SLD. These common tasks require a data extraction and analysis toolkit that is SLD-friendly and, like NCO, is used in both communities. We will improve NCO to support these tasks so that users can 1. Analyze regions specified by familiar names (including Climate-Forecast standardized region names) or user-supplied boundaries. 2. Access sophisticated statistical and regridding functions that are robust to missing data. 3. Create, save, share, and re-use user-defined functions or "recipes". These capabilities will ease data specification, software re-use, and, because they apply to SLD, minimize transmission, storage, and handling of unwanted data. Proof of these claims 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 CMIP5 models used in the IPCC AR5 climate assessment.

This proposal is organized as follows. Section 2 introduces the software technologies that this project will link to address the SLD problem. 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 tasks and methods to address them, 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 reusability, lifecycle costs, dissemination, persistence, and our ESDSWG participation, as per ACCESS requirements. Projects related to ours and potential broader educational impacts are in Section 7. A list of acronyms and abbreviations appears at the end as a supplementary document.

2 Background

This project integrates three streams of proven Open Source software technologies: (1) HDF/netCDF/NCO. NCO is a toolkit that accesses data stored in either of the two dominant standards: HDF—the official storage standard for NASA EOS satellite retrievals and netCDF—the de facto format for weather/climate models. (2) WKT/GEOS/PostGIS—an Open Source format, geometry library, and database for map description and manipulation. (3) SCRIP—a robust and accurate library for spatial interpolation and regridding in spherical coordinates.

We will take advantage of the open source availability and convergence of the NASA data distribution formats (HDF and netCDF) with geospatial toolkits and map databases (WKT/GEOS/PostGIS, and SCRIP) to extend the existing NCO toolkit to work with orbital swath data and irregular model data. The improved NCO will give researchers an easy way to access, regrid, and manipulate SLD with intuitive and familiar commands, without having to learn any GIS programming. The result will be an indispensable software assistant to individual researchers and research centers interested in processing or distributing satellite, weather, and climate data.

2.1 HDF, netCDF, and NCO

Two formats dominate geoscience data archival of satellite- and model-derived data. 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).

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 is the most commonly used archival format for large scale geophysical models, such as climate and weather models. netCDF version 3 (netCDF3) is 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. netCDF version 4 (netCDF4) was introduced in 2006 as a backwards-compatible format that implements an enhanced data model containing many features long present in HDF (*Rew et al.*, 2006), including hierarchical storage. The back-end storage format of netCDF4 is HDF5.

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*, 2013) may be the best-known toolkit for numeric and metadata analysis and manipulation of netCDF data. NCO comprises a dozen operators that perform the data and metadata manipulation and analysis functions most desired by researchers. Our ACCESS 2011 project has added support to NCO to exploit advanced features of the HDF storage format, including its eponymous hierarchical storage capabilities. NCO now makes it easy to perform powerful manipulation of netCDF files *and* netCDF-compliant HDF files. Thanks to ACCESS 2011 the name NCO is now somewhat of a misnomer since NCO works well with all netCDF-compliant HDF files, including HDF5 and HDF-EOS5 (now, directly) and HDF4 and HDF-EOS2 (via wrappers in development). Hence this ACCESS 2013 project will benefit users of nearly all NASA-distributed HDF data.

Also relevant to this project is that 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.

2.2 NCO Philosophy

Traditional processing of scientific data works with an intra-file paradigm: users open a file, read a variable from the file, then 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. SLD from satellite, for example, is often stored as one file-per-orbit, for thousands of orbits. Data produced by geophysical time-stepping models (weather and climate models) is output every timestep or averaged over many timesteps, for, again thousands of files per simulation. In such applications, the intra-file paradigm becomes unwieldy and optimal tools 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 processed in a single step are usually homogeneous: they contain the same variables, usually but not necessarily of the same type, rank, and shape.

3. Distinctions between *dimensions*, *coordinates*, and *variables* are maintained.
4. Operators defaults make sense and may be over-ridden with simple, mnemonic command line switches.
5. Operators provide an **audit trail that tracks data provenance**
6. Operators are as generic as possible, imposing no limitations on data dimensionality, size, or type.
7. Conformance to metadata conventions (like Climate Forecast, CF) is paramount (*Gregory*, 2003).

Apparently NCO's guiding philosophy, "do what a sane user would want" has succeeded! NCO runs on all modern desktop and server operating systems, and its use is fully detailed in the NCO User's Guide. To our 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, sites where NCO is used as backend middleware or is promoted as a user-solution, including AeroCom, DOE ARM, DKRZ, EU ENES, EU PRISM. NASA GES DISC, NASA LARC, NCAR, and NOAA CDC. These institutions find consistent advantages to using NCO. For example, speed. NCO works at the speed of C, its native language. Co-I Lynnes and colleagues at GES DISC improved one workflow by a factor of seven by replacing Python time-averaging code with NCO. Also, NCO provides easy access to the underlying structure and contents (dimensions, attributes, variables) of data without requiring programming, compiling, or debugging. NCO enables many operations that are otherwise tedious to code. These examples illustrate why NCO is currently used at NASA, and give insight into how it may be better exploited.

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 attract new users who are interested in working with swath-like data in NASA observations and models, but who are not necessarily skilled or familiar with GIS or regridding (or who wish quicker methods for accomplishing these tasks).

Note that NCO supports the Open-source Project for a Network Data Access Protocol (OPeNDAP, aka DAP) protocol, a data transmission protocol designed specifically for science data. DAP 2.0 is an ESDS standard (ESDS-RFC-004). The NCO User's Guide and this OPeNDAP Presentation provide more details. NCO (and all DAP clients) have network-transparent access to any files controlled by a DAP server. All NCO commands described in the following sections will work with files residing on a DAP server simply by replacing the filename arguments with the HTTP DAP server equivalents.

2.3 WKT, GEOS, and PostGIS

This project will leverage three established open source geographic information system (GIS) technologies to simplify access to SLD regions in NASA and model datasets. The first technology is a markup language for describing boundaries and areas. This technology, called Well-Known Text (WKT), is an Open Source standard. WKT and its binary representation WKB provide the equivalent functionality of Shapefiles (`.shp` and `.shx` files) defined and used by GIS industry leader Environmental Systems Research Institute (ESRI). Henceforth "shapefile" (lowercase "s") will denote a generic description of a boundary or region. Whether it is in WKT/WKB or Shapefile format is immaterial since the PostGIS back-end (below) easily converts between them. NCO will use WKT internally for representation of non-rectangular regions and boundaries. No user-knowledge of WKT will be required, though users may supply WKT/WKB files to describe regions not in the PostGIS region database to which NCO will have access.

The second GIS technology NCO will leverage is the Geometry Engine Open Source (GEOS) library which implements powerful methods for manipulating spatial geometric regions. GEOS is a complete C/C++ port of the Java Topology Suite (JTS), and will provide NCO with powerful spatial operators, predicate operations (e.g., intersections, overlaps), as well as WKT/WKB encoding and decoding. Third, NCO will link to the PostgreSQL Geographic Information System (PostGIS) database to store, access, query region and boundary information. Users will not need to know any internals to exploit the powerful GEOS/PostGIS features NCO will support.

2.4 SCRIP

The final ingredient necessary to make NCO a “killer app” for analyzing SLD is a regridding library which understands regional and global map grids of arbitrary complexity. The Spherical Coordinate Remapping and Interpolation Package (*Jones*, 2001) (SCRIP) is unsurpassed at providing this functionality. SCRIP is employed for similar purposes by the NCAR Command Language (NCL) and by the Climate Data Operators (CDO) (*Schulzweida*, 2013). CDO does provide a flexible and intuitive interface to SCRIP, some of which we will emulate because it is well-done. The advantages of our implementation will be improved handling of missing values (not handled by NCL as of version 6.1.1), seamless application to multiple files with differing input geometries (not handled by NCL or CDO), and integration with NCO’s new SLD capabilities. This prototype command conveys our ultimate goals—it regrids two Greenland swaths to a common 1° × 1° output grid, averages those internally, and outputs the result:

```
ncra --rgn=greenland --grid=1x1 sld1.h5 sld2.h5 1x1.h5
```

NCO will incorporate GIS and regridding technology unobtrusively so many (most?) investigators can conduct their weather/climate/geoscience research on NASA or model SLD without becoming GIS experts.

3 Geoscientific Domain Application

We will apply our improvements to the problem of analyzing HDF and netCDF datasets pertaining to Earth’s cryospheric water and energy cycles. In particular we will analyze factors contributing to the observed and simulated surface energy budget (SEB) of Greenland. Greenland experienced unexpected and unprecedented melt covering 97% of its surface in July, 2012. More troubling is Greenland’s long-term trend of darkening where mean July albedo has decreased monotonically by 8% absolute since 2000 (*Box et al.*, 2012). Explanatory hypotheses include long-term trends in Greenland snow thermodynamics (*Box et al.*, 2012; *Zender*, 2012), column water vapor (*Kapsch et al.*, 2013), low-level clouds (*Bennartz et al.*, 2013), and heat transport (*Fettweis et al.*, 2013; *Kapsch et al.*, 2013). This project supports the thesis research of UCI graduate student (Wenshan Wang) who studies the relative roles of these processes and their implications for other cryospheric regions like Antarctica (*Zender*, 2012).

The PI has had previous NASA and NSF projects on the energy balance of the cryosphere that led to this domain application. We have analyzed Greenland snow cover and snow albedo on climatological, annual, monthly, daily, and diurnal timescales (*Flanner et al.*, 2007, 2009; *Wang and Zender*, 2010a,b, 2011). Climatological, annual, and monthly analyses are easy to perform since NASA aggregates and distributes the necessary products in monthly-mean form on the regular Climate Modeling Grid (CMG). By contrast, daily and diurnal analyses are difficult because we are often confronted with Swath-like data (SLD), single scenes, and non-rectangular grids. Understanding these energy budget trends and anomalies requires intercomparison of these NASA observations with model simulations *Flanner et al.* (2009); *Wang and Zender* (2010b); *Allen et al.* (2012). Hence we are motivated to ease access to and analysis of both NASA and model SLD.

The ACCESS call states as priorities tools that (Section 1.2.1) “improve users’ ability to find, access, and only download data meeting customizable criteria and reduce the volume of unwanted data downloaded” and that (Section 1.2.2) improve “usability of NASAs Earth Science observational data for the modeling and model Analysis communities”. We will illustrate the current state of accessibility and the goals of our project by asking three scientific questions, in increasing order of analysis complexity:

1. What are snow albedo and surface temperature in a rectangular region of Greenland in a single swath?
2. What instantaneous column water vapor resides over an irregular region (Greenland) in a single swath?

3. To what extent do historical simulations of snow albedo and temperature by CMIP5 models agree with NASA observations in Greenland?

Zender and Mangalam (2007) and *Zender* (2008) describe how and why NCO is an efficient solution for determining geophysical statistics (e.g., mean, trends, and variability) of data in netCDF classic format. Thanks to our ACCESS 2011 project these methods now apply without change to HDF data such as MODIS datasets, and data with hierarchical group structure. The sections below focus only on extending existing high-performance tools (NCO) to apply to SLD, i.e., to non-rectangular regions, and to regrid SLD to rectangular grids for intercomparison and evaluation purposes. These GIS features are orthogonal yet complementary to accomplishments of our ACCESS 2011 project.

3.1 Accessing and Analyzing Rectangular regions from Swath-like Data

Question 1 above illustrates the problem of analyzing rectangular regions (e.g., the state of Colorado) in non-rectangular grids like swaths. Level 2 swath data retrieved from the EOS ClearingHOuse typically contains two-dimensional longitude and latitude arrays, or equivalent information such as a geolocation pointer together with along- and across-swath coordinates. In the latter case geolocating the grid is more difficult (and is allocated substantial programming effort in Section 4.1 below). For now we assume the geolocated two-dimensional coordinates are available and named `lon2d(i,j)` and `lat2d(i,j)` where i and j represent (for simplicity) the along- and across-swath indices. The coordinates are two-dimensional because both the longitude and latitude at (i, j) depend on i *and* j. On a rectangular grid i and j represent east-west and north-south indices so the coordinates are one-dimensional, `lon1d(i)` and `lat1d(j)`.

Extracting rectangular hyperslabs from a rectangular grid file in HDF5 or netCDF4 format (`rct.h5`) is simple, while doing the same from SLD (`sld.h5`) requires much more effort. This is illustrated with the canonical NCO operator named `ncks` (netCDF Kitchen Sink) (*Zender*, 2013), although the general principles apply to all analysis languages of which we are aware:

```
ncks -d lat1d,0.,90. -d lon1d,0.,180. rct.h5 out.h5 # Currently works
ncks -d lat2d,0.,90. -d lon2d,0.,180. sld.h5 out.h5 # This project
```

The first command works as expected and the output dataset contains only requested rectangle and no extraneous points. If the input file is global, then `out.h5` will be one-fourth the size (containing half the global latitudes and half the global latitudes). Although the user's intention with the second command is the same as the first, the second command would currently fail because the operator lacks the algorithms to automatically mask-in the intended region and return the most compact array possible.

Instead users who wish to analyze rectangular regions of SLD currently employ masking:

```
# ncap2 works with single or multi-dimensional lat and lon
ncap2 -s 'where(lat2d > 0 && lat2d < 90) && (lon2d > 0 && lon2d < 180)
 temperature=sst; elsewhere temperature=sst@_FillValue;' sld.nc out.nc
# ncwa weighted averager works only with single-dimensional lat and lon
ncwa -B 'oro < 0.5' -w area -a lat1d,lon1d in.nc out.nc
```

The masking procedures NCO now employs do keep values outside the Region of Interest (ROI) from contaminating computed statistics, yet do not reduce the output filesize. An ACCESS call priority (Section 1.2.1) is to "reduce the volume of unwanted data downloaded".

Task 1 described in Section 4 is designed to answer questions like Question 1 above, in a user-friendly manner that avoids unnecessary data transfer in accord with ACCESS guidelines. Briefly, Task 1 refers to the implementation of support for NCO to automatically mask and extract the smallest possible amount of data in specified rectangular region of an SLD dataset.

3.2 Accessing and Analyzing Irregularly Shaped Regions from Swath-like Data

Question 2 above illustrates the problem of extracting a non-rectangular shape from SLD. NCO currently supports specification of non-rectangular regions via three different, and by the standards of what we propose to implement, relatively crude methods:

1. NCO supports some of the numerous CF Conventions for Coordinate Systems. In particular, NCO supports Reduced Horizontal Grids and Unstructured grids in which the horizontal coordinates that describe a non-Cartesian grid are linked to and accessed via the `coordinates` attribute. NCO accepts rectangular region specifications as lat/lon bounding boxes, such as `-X 0.,180.,-30.,30. -X 270.,315.,45.,90.`. Any number can be daisy-chained to specify an arbitrarily complex region. This formalism is mainly used for time-constant irregular grids such as reduced latitude-longitude grids (fewer longitudes at polar latitudes) or completely unstructured grids (e.g., geodesic grids stored as one dimensional arrays in which the index indicates the horizontal cell position). This infrastructure allows NCO to identify, interpret, and process (e.g., hyperslab, average) variables on Reduced Horizontal and unstructured grids as easily as it works with regular grids. The main shortcoming of this method is that it is only common for model (e.g., CMIP5) datasets. Few NASA datasets are stored with the structure and metadata of reduced or unstructured grids.
2. Multi-slabbing: NCO's Multi-slabbing Algorithm (MSA) options allow users to specify any number of 1-D hyperslabs such as `-d lon,-180,-150. -d lon,0.30. -d lon,150.,180.`. However, NCO always takes the union of all matching points. Since irregular regions are more often composed of uniquely specified two-dimensional (not one-dimensional) regions, the MSA option is in practice usually used only for one dimensional fields such as timeseries.
3. Masks: NCO's `ncap2` implements a `where` command to create or employ binary masks, e.g., `where(mask==1) temperature=sst elsewhere temperature=sst@_FillValue`. Since `where` accepts any number of conditionals that can define masks for arbitrarily complex regions. However, `where` is only available in `ncap2` and NCO requires hand-scripting to apply the mask to an entire file. As shown earlier, NCO's `ncwa` also performs statistics (e.g., averaging) based on masks provided by multi-dimensional fields as long as the coordinates are 1D. However, files rarely contain all the regional masks (e.g., Greenland) that researchers want.

Thus the methods NCO already supports are not what we would call intuitive, complete, uniform, or file-level. They are also clumsy and not well-suited for integration into GUIs.

This project proposes to implement SLD-access methods that are 1) Intuitive, 2) Complete, 3) Uniform, and 4) File-level. Such access methods will, by construction and intent, share the most important attribute of any new feature: user-friendliness. In this context "Intuitive" means that regions may be accessed by their familiar (and sometimes standardized) names. For example, non-Intuitive methods to specify the Greenland region would include those that require the user to provide the (lat,lon) coordinate-pairs that described Greenland's perimeter in some specific format, while Intuitive methods include those that allow the user to specify the region name, e.g., `--region=Greenland`. This is practical for the most common regions, yet users must be able to specify perimeters for regions that do not have standardized names.

In this context "Complete" means that the access methods work on any desired geographic region, regardless of its shape or contiguity with other regions. For example, the Multi-slabbing algorithm already implemented is not Complete because it cannot describe regions that require disjoint multi-slabs in both latitude and longitude, such as all the "red states" or "blue states". Our new method guarantees Completeness by allowing irregular regions to be specified as any collection of irregular sub-regions. For example, `--rgn=England,France`.

"Uniform" means that the access methods work in the same predictable way for all operators. For example, simple (rectangular) hyperslabbing in NCO is Uniform because it works with the same switches and syntax across all applicable operators, e.g., `-d lat,0.,90. -d lon,0.,180.`. The new hyperslabbin gmethods will likewise be uniformly incorporated into NCO.

Finally, “File-Level” means that, by default, the methods apply to the entire dataset and need not be hand-coded to apply to multiple variables. Users usually want to extract the same regions for all variables in a given dataset and NCO (rectangular) hyperslabbing, multi-slabbing, and auxiliary coordinates are all Uniform methods. However the *where* syntax only works in `ncap2` so it is not a File-Level feature.

Task 2 described in Section 4 is designed to answer questions like Question 2 above. Briefly, Task 2 refers to the implementation of backend GIS technology (GEOS/PostGIS) that the user accesses with intuitive region names (or shapefiles) and geometry commands.

3.3 Regridding Observations and Models for Analysis and Intercomparison

Question 3 above illustrates the problem of regridding data from SLD (and rectangular) grids to a common grid, which is usually but not always rectangular, for purposes of intercomparison or statistical analysis. Creation of gridded monthly averages, for example, involves regridding and combining scores of swaths (approximately one per day). This is typically done with meta-code loops that resemble

```
for fl in `ls sld*.nc`; do
    regrid --region=greenland --grd_out=1x1 ${fl} 1x1_${fl}
done
ncra 1x1_*.nc greenland_monthly_average_1x1.nc
```

If the user only desires the final monthly average, then regridded intermediate files (`1x1_sld01.nc`, `1x1_sld02.nc`) are not of interest. Our solution will avoid storing intermediate results in on disk, and instead aggregate each input file onto the output grid in RAM, and only write out the final file:

```
ncra --rgn=greenland --grd_out=1x1 sld*.nc modis_greenland_avg_1x1.nc
```

Likewise statistics in irregular regions from source files at different resolutions, such as the CMIP5 multi-model ensemble, become a simple one-line command:

```
ncea --rgn=greenland --grd_out=1x1 ncar.nc ecmwf.nc ukmo.nc out_1x1.nc
```

Here `ncra` and `ncea` are time and ensemble operators, respectively (*Zender*, 2013).

Weather and climate analysis often involves intercomparing NASA with model datasets. Often a rate-limiting step in this procedure is conversion of both datasets to a common grid. Our proposed solution replaces that procedure with

```
for model in ’cesm giss echam ...’; do
    ncdiff ${model}_YYYYMMDD.nc modis_L2_YYYYMMDD.nc \
           --grid=1x1 ${model}_minus_MODIS_YYYYMMDD.nc
done
```

If the model datasets are stored as a hierarchical ensemble of groups in the same file, as made possible by our ACCESS 2011 project, then the outer loop over model names may be omitted because differencer, `ncdiff`, automatically loops over groups (a feature we call group broadcasting).

4 Tasks: Software Engineering and Configuration

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

1. Implement masking, extraction, and then geolocation of rectangular regions in SLD grids
2. Implement these same capabilities for non-rectangular regions
3. Support regridding from SLD and rectangular grids to SLD and rectangular grids

4.1 Task 1: Region Masking

The first software engineering task for this project is to refactor the NCO codebase to automatically mask and extract the smallest possible amount of data in specified rectangular region of an SLD dataset. Two NCO operators (ncap2 and ncwa) already support masking by conditions placed on multi-dimensional variables (cf. where example in Section 3.2). This task will be accomplished by generalizing and extending that (ncap2) functionality to file-level operations performed by the rest of NCO. The key step is to implement routines that automatically parse rectangular region specifications placed on multi-dimensional coordinate variables (e.g., -d lat2d,0.,90. -d lon2d,0.,180.) into the conditional clauses which are already implemented. Once the condition is parsed, points inside or outside a ROI can easily be set (i.e., masked) to a user-selected value including _FillValue.

Many NASA SLD datasets will require that the coordinate system underlying the swath be geolocated into absolute latitude and longitude arrays (i.e., lat2d and lon2d) before further operations like masking can occur. HDF-EOS datasets provide geolocation fields with the necessary information that can be extracted through the netCDF4 API. In some cases it may be necessary to use the HDF API directly. By contrast nearly all model SLD contains geolocated coordinates. Our work plan is therefore to implement masking first (Year 1) on datasets which provide geolocated coordinates. In Year 2 we will implement internal NCO geolocation support to handle the more general case.

Another difficult part of Task 1 is to extract the most compact portion of an SLD grid that meets the rectangular region specification. In general the result will be a (highly) non-rectangular grid in which each along- and across-track row and column has different number of points, i.e., the extracted lat2d and lon2d arrays will be ragged. Thus there are at least two options: 1) The extracted coordinates can be stored in rectangular arrays where the extraneous points are set to a predefined value, e.g., _FillValue. And 2) The ragged extracted coordinates can compressed to consume minimal storage using a convention such as the CF convention for compression by gathering. This convention essentially encodes ragged multi-dimensional data into a 1-D array on disk, and indicates to tools (like NCO) the intended multi-dimensionality for decompression using the compress attribute. Compressing coordinates (as opposed to fields like SST) could lead to a host of portability problems, i.e., downstream tools may not understand the CF convention. Hence we will start with Option 1 and implement Option 2 if time permits.

4.2 Task 2: Integration with GIS backend

The second software engineering task for this project is the refactoring of NCO to support user specifications of non-rectangular regions using intuitive regions names (or shapefiles) that access backend GIS technology (GEOS/PostGIS) and geometry commands. A key design goal is that NCO users need not know GIS internals. Our solution does *not* add full GIS functionality to NCO. The most useful GIS features we can add in the 2 year scope of this ACCESS project are intuitive, non-GIS commands to combine (union and intersection), mask, and extract regions specified by intuitive names or user-supplied shapefiles. This functionality will be implemented extensibly so that more powerful features of GEOS/PostGIS can be added in future years.

The plan for implementing the GIS technology (introduced in Section 2.3) is best illustrated by describing the flow of code executed during a prototypical command, say, to extract Greenland from SLD data:

```
ncks --region=greenland modis_level2_swath.hdf modis_greenland_masked.nc
```

In this prototypical command, the output is identical to the input (including the same SLD grid) except all fields are set to the missing value (aka _FillValue) outside the boundaries of Greenland. The user need specify only an intuitive region name ("greenland") and the input file to mask and NCO will mask all fields in that file. NCO will use the librx regular-expression library to map user-specified text and regular expressions to the names of region objects (shapefiles) stored in the PostGIS database to be installed with NCO. This default NCO database will contain boundaries of CF-standard regions. User-supplied shapefiles will be supported via a command-line switch (e.g., `--shapefile=my_database`).

After verifying the requested shapefile is available, NCO will retrieve the map grid geometry from the MODIS file and format it internally as a geometry understandable by GEOS. Then NCO retrieves the selected region (Greenland) as a WKT region object from PostGIS. With both the shape and the map grid represented as WKT objects in RAM, NCO calls the GEOS toolkit to mask all points in the map outside Greenland. Other GEOS geometry manipulation options will be accessed with commandline options. We will initially support inclusion/exclusion of single regions, and then implement support for multiple regions via lists, e.g., `--rgn=greenland,iceland`. The default region combination method will be the union, with options use to specify intersections and overlaps, e.g., `--rgn_op=intersection`.

Finally, NCO stores the masked or extracted region in the output file. This step presents the same two storage options described in Section 4.1, i.e., storing rectangular arrays with extraneous values or storing ragged arrays using compression-by-gathering. We will leverage the experience gained from and code used in Task 1 to write the output file.

This solution augments NCO's current, clumsy non-rectangular region specification methods (Section 3.2) with a clean, intuitive interface that can easily be populated by a GUI front-end. Co-I Lynnes at the Goddard Earth Science (GES) Data and Information Services Center (DISC) will provide guidance and support on linking this capability to NASA data access and delivery systems/GUIs such as Giovanni and the Simple Subset Wizard (SSW).

An neat application of this technology which DISC is particularly interested in is subsetting all points within a given radius of a target:

```
# Radius from Center: Within 1000 km of Boulder
ncks --rgn_ctr=40.,105. --rgn_rds=1000 sld.h5 out.h5
```

We will implement this capability into the SSW in Year 2. Depending on subsequent demand, we may add masking options for analytic shapes besides circles too.

4.3 Task 3: Regridding

As important and challenging as Region Specification is the task of Regridding or Remapping (we do not distinguish the two). Regridding is perhaps the most common manipulation users wish to perform on SLD. It is complex owing to the wide variety of input and output grids, regridding choices, and treatment of missing data. These elements combine to make regridding a formidable barrier to analysis for novice researchers. Alleviating these difficulties could substantially increase the utility and value of SLD to researchers outside the experienced remote-sensing and GIS communities.

We eschew "reinventing the regridding wheel" by adopting the time-tested Spherical Coordinate Remapping and Interpolation Package (SCRIP) (*Jones*, 1999, 2001) as the back-end tool for remapping. SCRIP is flexible, accurate, and robust. "Flexible" in this context means the algorithm supports the most common input and output grid types. *Jones* (2001) describes how SCRIP supports the most common gridtypes: rectilinear, curvilinear, and unstructured. "Accurate" regridding refers to the mathematical properties of the procedure. Accuracy metrics of the transformation algorithm include its conservation properties (i.e., are area-weighted integrals conserved?) and gradient properties (i.e., how smooth are first derivatives?). SCRIP supports five types of remapping: second-order accurate conservative, bilinear and bi-cubic interpolation, distance-weighted, and particle-based. "Robust" refers to the reliability of the remapping algorithm under non-ideal circumstances. SCRIP robustly handles the "Pole problem" in spherical geometry.

One difficulty is with the problem of missing (i.e., undefined) values. The application must ensure that undefined (i.e., missing) source cell values do not contribute to any partially overlapping destination grid cells. This requires that NCO re-search and tag the SCRIP-provided re-mapping weight matrix for every field re-mapped. Essentially valid source cells that contribute to a partially missing destination cell must be re-weighted. Other implementations of SCRIP (e.g., NCL as of version 6.1.1) do not do this, and instead set the entire destination cell to the missing value. NCO's treatment of regridded missing values will avoid this pitfall.

The crux of the software engineering for Task 3 is creating an optionally invoked SCRIP-abstraction between NCO's input and output. NCO currently copies the input grid to the output file and the abstraction layer will re-direct input to the regridding routines where necessary. Currently NCO is unaware of the map grid descriptors that are (necessarily) demanded by generic regridding libraries like SCRIP. To support such flexible remappings with minimal user-intervention, the NCO code must be refactored to:

1. Point to existing SCRIP-compliant grid transformation specification files (aka "SCRIP-files") and transformation options (e.g., regridding algorithm) for frequently used grids. A SCRIP-files is a simple netCDF file with variables and attributes that describe a single grid (*Jones*, 2001).
2. Create SCRIP-files from a template data file. Often users wish to regrid SLD data to the grid used by another data file they have. The task here is to read the grid information from the data file, supply any missing information with sensible defaults, and convert it to a SCRIP-file.
3. Call SCRIP with the pre-existing or NCO-generated SCRIP-files. Given the input and output grid SCRIP-files, SCRIP generates a "remap matrix" of weights to remap the input to the output grid.
4. Use the same remap matrix as many times as necessary to re-grid the entire input file.

Our implementation of SCRIP will be complete once NCO accepts and parses intuitive command-line specifications for simple destination grid geometries, e.g., `--grd_out=1x1`, `--grd_out=T42`. In the spirit of interoperability we will adopt as synonyms where possible the SCRIP options supported by CDO (*Schulzweida*, 2013). Constructing such rectangular grids requires computing regularly spaced gridpoints and spectral coefficient for Gaussian quadrature. The PI is well-versed in these methods and has coded all this in Fortran. The result will be a SCRIP implementation that supports intuitive grid specifications and remapping options as well as pre-defined SCRIP files:

```
# Regridding using existing remap matrix weights in rgr_wgt.nc
ncks --rgr_wgt=rgr_wgt.nc in.h5 out.h5
# Conservative Re-gridding of in.nc to destination grid
ncks --rgr_dst=scrip.nc in.h5 out.h5
# Bicubic re-gridding of in.nc to template grid
ncks --rgr_mth=bicubic in.h5 out.h5
```

4.4 Schedule and Milestones

Dated entries indicate milestones to be reached within preceding 3 months. Activities denoted by *italics* are ongoing, and are not milestones. Personnel with direct responsibility indicated by **PI** (C. Zender, UCI), **CI1** (C. Lynnes, GSFC), **CI2** (W. Baskin, LaRC), **SE1** (Software Engineer/Scientific Programmer P. Vicente, UCI), **SE2** (Software Engineer/Scientific Programmer TBD, GSFC), and **GS** (Grad Student W. Wang, UCI). Hyperlinked Sections and Tasks refer to ACCESS project proposal.

Year 1. *Goals*: Masks, Single Irregular Shapes

20140208	ACCESS 2013 Project commences day after ACCESS 2011 Project ends
20140208	(**M1**) NCO 4.4.x (last pre-ACCESS 2013 NCO release) (**PI**)
	Refactor NCO to support SCRIP (Task 3) (**PI**)
	Refactor NCO to support JTS/GEOS (Task 2) (**SE1**)
	Manage NCO releases and disseminate to community (**PI**)
	Identify and iterate on GES DISC use cases for NCO (**CI1**)
	Identify and iterate on LaRC ASDC use cases for NCO (**CI2**)
	Infuse NCO recipes into GES DISC data distribution services (**SE2**)
	Analyze Greenland SEB using MODIS, MERRA, MISR, CERES (**GS**)
20140308	(**M2**) Visit GES DISC, collaborate on NCO recipes (**PI**, (**CI1**)
20140508	(**M3**) NCO 4.5.0: Link to JTS/GEOS (Section 4.2) (**SE1**)
20140508	(**M4**) Finalize internal masking API (Section 4.1) (**SE1**)
	Quantify cloud, water vapor, heat transport roles in Greenland SEB trends (Section 3) (**GS**)
20140808	(**M5**) Single shape mask works on geolocated grids (**SE1**)
20140808	(**M6**) CF standard region demonstrated (**SE1**)
	Compare observed with modeled (CMIP5) Greenland SEB (**GS**)
20141108	(**M7**) Rectangular extraction of geolocated SLD (Section 4.1) (**SE1**)
	Improve regression tests for RRL 8 (Section 6.1) (**SE1**)
20141108	(**M8**) Internal links to SCRIP (Section 4.3) (**PI**)
	Draft and implement recommended citation to complete RRL IP Level 8 (Section 6.1) (**PI**)
	Start implementing NCO geolocation framework (**SE1**)
20141108	(**M9**) Union of multiple shape masks works (Section 4.2) (**SE1**)
20141108	(**M10**) Greenland SEB trend (Section 3) (**GS**)
20141205	(**M11**) Present at AGU Earth and Space Science Informatics session (**PI**)
	Document regridding capabilities in NCO User's Guide (**GS**)
20150208	(**M12**) Regridding works on global files (Section 4.3) (**PI**)
	Paper on relative roles of clouds, vapor, heat transport in Greenland SEB trends (Section 3) (**GS**, **PI**)

Year 2. *Overall Goals*: Geolocation, Regridding, Multiple Shapes, NASA Infusion

	Integrate NCO SLD feature into Giovanni, SSW GUIs (**SE2**)
	Integrate NCO regrid feature into Giovanni, SSW GUIs (**SE2**)
	Infuse NCO recipes into LaRC ASDC data distribution services (**CI2**)
20150508	(**M13**) Geolocation by NCO works (Section 4.1) (**SE1**)
20150508	(**M14**) Regridding works on rectangular hyperslabs (Section 4.3) (**PI**)
20150808	(**M15**) Regridding works on non-rectangular regions (Section 4.3) (**PI**)
20150808	(**M16**) GEOS overlaps work (Section 4.2) (**SE1**)
	Paper on global long-term darkening trends (Section 3) (**PI**), (**GS**)
	Standardize/document internal APIs for RRL extensibility Level 6 (Section 6.1) (**SE1**)
20151108	(**M17**) Regridding works in multi-file operators (Section 4.3) (**PI**)
20151208	(**M18**) Present ACCESS accomplishments at AGU Earth and Space Science Informatics session (**PI**)
20160208	(**M19**) Multi-file regridding on non-rectangular regions (Section 4.2–4.3) (**PI**)
20160208	(**M20**) NCO 5.0.0: Finalize GEOS/PostGIS/SCRIP NCO (**SE**)
20160208	ACCESS 2013 Project ends

5 Results from Prior Funding on Related Projects

Zender is/was PI on two previous scientific data analysis related projects: First, 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 MultiProcessing (SWAMP, `http://swamp.googlecode.com`). Resulted in four peer-reviewed papers and twelve conference abstracts (`http://nco.sf.net#pub`). Second, NASA NNX12AF48A, $495840, 201202–201401, *Simplifying and accelerating model evaluation by NASA satellite data*. In progress. Project consists of improving existing netCDF software to analyze and manipulate data stored in the EOS-standard HDF format. Specifically, we are adding group support to NCO for HDF5 and netCDF4 files, and creating NCO wrappers for HDF-EOS2 files. Supporting one PhD (W. Wang) degree in Earth System Science on factors causing darkening of Greenland, and their implications for other cryospheric regions. Resulted in one publication (*Zender*, 2012) and three conference abstracts so far (`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 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 datasets 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 accessing non-rectangular regions and swath-like data.

6 Technology Issues

NCO is Technology Readiness Level TRL 8. This exceeds the minimum ACCESS standard of TRL 7. A programmer dedicated to the task for a year or two could improve NCO to TRL 9, though in our opinion those resources would be better spent on improving NCO's features (as proposed) and its reusability.

6.1 Reuse Readiness Level

This project adheres to the NASA ACCESS 2013 guidelines of achieving a high Reuse Readiness Level (RRL) to ensure sustainment and infusion of NCO in Earth science data systems. We estimate that overall NCO currently meets the ESDS definition of the following topic area RRL levels:

1. *Documentation Level 6*: The NCO User's Guide is and will be kept fairly complete. NCO is often taught in workshops that provide their own documentation (e.g., NCL's and netCDF's workshops).
2. *Extensibility Level 5*: Dozens of contributors have, with minimal help, extended NCO through the years with features including auxiliary coordinates, package management, and regular expression support. Nevertheless NCO's internal API documentation is a weakpoint. The UCI SE1 will be tasked to make it cleaner, more self-consistent, and document extensibility capabilities.
3. *IP Issues Level 7*: NCO is distributed under the GPL3. All copyrights are held by the PI who is therefore the only necessary point of contact for IP issues. This is indicated atop each source code

file, in the documentation, and with the `--version` switch of all executables, as per GNU coding standards. The UCI PI is tasked to draft and package recommended citation and developer list to fully implement RRL IP Level 8 or 9.

4. *Modularity Level 7*: NCO already utilizes separate modules, configurable with GNU Autotools, to include/exclude five optional libraries and their capabilities including ANTLR, GSL, netCDF4/HDF, OPeNDAP, OpenMP, `librx`, and UDUnits. This project will add modules for SCRIP and (both or neither) GEOS/PostGIS
5. *Packaging Level 8*: NCO supports both GNU Autotools and Qt project as build systems. RedHat and Debian-based Linux distributions have distributed NCO RPMs and .debs for over a decade. UCI distributes the Windows version as an InstallShield package. NCO has been deployed at all known geoscientific computing centers, and on tens of thousands of desktops (Linux, MacOS, and Windows).
6. *Portability Level 7/8*: NCO has been ported to all modern scientific computing platforms.
7. *Standards Compliance Level 7*: NCO is C99 and C++ compliant, supports UDUnits dimensional units, and strives to support the most useful CF metadata standards.
8. *Support Level 6*: NCO accepts and responds to support requests on its SourceForge site. New packages are typically distributed every month or two and incorporate fixes to all known problems.
9. *Verification and Testing Level 6*: NCO performs over one hundred self tests (with `make check`) when built. Nevertheless the UCI SE will be tasked to make NCO's regression tests more numerous, thorough, and self-explanatory.

By improving the weaker areas of Extensibility and Verification, this project will move NCO to at least a strong RRL 6 in all areas, and better in many.

6.2 Dissemination to NASA, NSF, and DOE Centers

Thanks to ACCESS 2011 funding, NCO now handles HDF-EOS5 files and by February 2014 will handle HDF-EOS2 files. NCO is therefore a natural addition to NASA data centers that serve HDF and netCDF files, such as EOSDIS Distributed Active Data Archive (DAAC) sites. Co-I's Lynnes and Baskin are ideally suited to ascertain the extent of SLD access and analysis at GES DISC and ASDC. NCO is already used in the back-end of many GUI front-ends that let users select data hyperslabs from repository data, including Giovanni4. This is because NCO is faster than many alternatives. With the SLD features described herein implemented, NCO will give users and DAACs unprecedented options (e.g., regridding) to ease access to and analysis of L2 data.

Infusing and iterating these capabilities with the collaboration of the Co-Is at the GES DISC, Goddard DAAC, and ASDC is a preliminary step to infusing this technology at other DAACs. We note that handling SLD and regridding are the most common requests that the PI received for NCO improvements from informally polling attendees of the 2012 ESDSWG meeting in Annapolis and of Fall AGU. Larger geoscientific computing centers that are repositories for model data (e.g., ESG nodes like NCAR, ORNL, PCMDI) usually already have NCO installed. Dissemination of NCO to such centers will follow the normal path as system administrators typically upgrade to newer NCO version on a 1–2 year cycle.

6.3 Lifecycle Costs and Continued Maintenance

This ACCESS project will continue to update and distribute .debs for Linux, tarballed executables for MacOS, and a Windows-native self-extracting InstallShield GUI installer. Other pre-compiled formats (RPM for Linux, Fink and MacPorts for MacOS, Cygwin for Windows, and native AIX executables) will continue to be made by external contributors invested in those formats. The wide variety of external ports the community provides at no charge to this project or NASA testifies to the effectiveness and low life-cycle maintenance costs of Open Source projects that generate and retain satisfied users.

PI Zender will apply for NCO-related support from other agencies to continue NCO development and maintenance to continue after the ACCESS 2013 project funding ends. NSF has supported NCO once before (IIS-0431203). NCO may be competitive in the NSF SI2 and EarthCube programs. Should external funding from NASA and NSF not be secured, NCO maintenance and low-level development will be conducted by volunteers, as it as been for most of its 17-year history. NASA data services will always be able to freely update to the latest NCO versions long after this ACCESS 2013 project ends.

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 seventeen years. Hosting the code on SourceForge.net and opening it to contributors (there have been eight 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 datasets and to native Windows users (both accomplished with ACCESS 2011 funding) has significantly increased NCO's user-base. netCDF and HDF 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 supported NCO over seventeen years, with external NCO support (from NASA and NSF) for only six years. NCO was supported completely organically (by volunteer developers, code contributors and users) for eleven of its seventeen years. Of course, feature growth slows without dedicated programmer support. As a tenured professor, with life-long job security in climate science and in computer 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.

6.4 Participation in Earth Science Data Systems Working Groups

PI Zender participates in the relevant geoscientific and IT communities. He is a past reviewer for the ESDS Standards Process Group (SPG) where he reviewed the Data Access Protocol (DAP), HDF5, and netCDF Classic standards. Zender has actively participated in Earth Science Data System Working Group (ESDSWG) activities since 2012. He currently Co-Chairs (with Peter Leonard) the NASA ESDS Data Stewardship Interest Area (DSIA) working group "NASA ESDS Conventions for HDF5" (HDF5WG). He lurks on the Open Source and Technology Infusion WGs and will join the Geospatial WG in Fall, 2013.

PI Zender will continue to reserve up to 25% of his time for participation ESDSWG activities. He commits to this role since he is very interested in infusing new software technology into the perennial problems encountered by researchers in managing large datasets. The last two years have been a spin-up period in which he has learned more about NASA culture and niches where NCO can streamline NASA data services. The ACCESS and ESDSWG communities have been very welcoming to the PI and he is eager to improve his tools to help serve their needs.

7 Related Projects, Impacts, Education, and Public Outreach

7.1 Related Projects

PI Zender has long collaborated informally with the netCDF development team at Unidata. The NCO project is perhaps the most demanding test of the netCDF library and Unidata is no stranger to our bug reports. Unidata reciprocates with software support and advice on exploiting netCDF features. Zender and netCDF developer Russ Rew submitted one proposal (declined) on improvements and extensions to the netCDF API. Zender plans to submit an NSF proposal with Rew and Earth System Grid (ESG) and NASA

Co-PIs to help data users to exploit hierarchical groups by getting data providers to annotate and distribute data that way. Zender has a student intern studying more efficient chunking defaults for various access patterns of netCDF/HDF datasets. Co-I Lynnes architects data services at GES DISC which distributes data not only to single users through Giovanni and SSW, but also to other organizations such as ESG. Formatting and massaging such data into particular formats is a neverending task to which NCO is well-adapted, and in which Lynnes plans to leverage his understanding of NCO. NCAR, DISC, GISS, and GFDL all use NCO (according to personal communications with, respectively, G. Strand, C. Lynnes, R. Miller, and K. Dixon) to massage datasets for compatibility with CMIP5 standards. Thus improvements to and accelerations of NCO will continue to contribute to international assessments.

7.2 Education

This project trains one graduate student in understanding cryospheric climate change using advanced geoscientific data analysis techniques. 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 (2011–2014, PI E. Saltzman of ESS), and (3) the Undergraduate Research Opportunities Program (UROP). Zender has opened a paid summer undergraduate UROP position on the NCO project at no additional cost to NASA and will do so again in Years 1 and 2.

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. Our most recent project is the NASA Competitive Program for Science Museums and Planetariums (CP4SMP) grant "Our Instrumented Earth: Understanding Global Systems and Local Impacts through the El Niño Story". Zender is helping to help train dozens of Title I (underprivileged) K-12 teachers for this project.

Every year Zender teaches ESS 172/272 "Science Communication & Outreach" which sends UCI undergraduate and graduate students to teach global change lessons, including NASA material, at local (Irvine) and disadvantaged (Santa Ana) K–12 schools. If awarded this grant, Zender will apply for supplemental NASA Education & Public Outreach (EPO) support. This would modestly support support a student intern to coordinate our ESS graduate student K–12 outreach seminars for a Climate Literacy project called CLEAN, and it would support a student intern to help integrate NASA climate data into the atmospheric science section of the new educational loop at Crystal Cove State Park (CCSP). CLEAN serves hundreds of Orange County (OC) K–12 students annually, while CCSP's educational loop serves 50,000+ visitors annually.

References

Bibliography

Allen, R. J., S. C. Sherwood, J. R. Norris, and C. S. Zender (2012), Recent Northern Hemisphere tropical expansion primarily driven by black carbon and tropospheric ozone, *Nature*, *485*(7398), 350–354, doi:10.1038/nature11097. 3

Bennartz, R., M. D. Shupe, D. D. Turner, V. P. Walden, K. Steffen, C. J. Cox, M. S. Kulie, N. B. Miller, and C. Pettersen (2013), July 2012 Greenland melt extent enhanced by low-level liquid clouds, *Nature*, *496*, 83–86, doi:10.1038/nature12002. 3

Box, J. E., X. Fettweis, J. C. Stroeve, M. Tedesco, D. K. Hall, and K. Steffen (2012), Greenland ice sheet albedo feedback: thermodynamics and atmospheric drivers, *The Cryosphere*, *6*, 821–839, doi:10.5194/tc-6-821-2012. 3

Fettweis, X., E. Hanna, C. Lang, A. Belleflamme, M. Erpicum, and H. Gallée (2013), Brief communication "important role of the mid-tropospheric atmospheric circulation in the recent surface melt increase over the Greenland ice sheet", *The Cryosphere*, *7*(1), 241–248, doi:10.5194/tc-7-241-2013. 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. 7

Jones, P. W. (1999), First- and second-order conservative remapping schemes for grids in spherical coordinates, *Month. Weather Rev.*, *127*, 2204–2210. 4.3

Jones, P. W. (2001), *A User's Guide for SCRIP: A Spherical Coordinate Remapping and Interpolation Package*, Los Alamos National Laboratory, Los Alamos, NM, `http://climate.lanl.gov/Software/SCRIP`. 2.4, 4.3, 1

Kapsch, M.-L., R. G. Graversen, and M. Tjernstr om (2013), Springtime atmospheric energy transport and the control of Arctic summer sea-ice extent, *Nature Clim. Change*, *3*(5), –, doi:10.1038/nclimate1884. 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. 2.1

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. 2.1

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. 2.1

Schulzweida, U. (2013), *Climate Data Operators User's Guide, Version 1.6.0*, Max Planck Institute for Meteorology, Hamburg, Germany. 2.4, 4.3

Unidata (2004), *Network Common Data Format*, University Corporation for Atmospheric Research, Boulder, CO, `http://www.unidata.ucar.edu/packages/netcdf`. 2.1

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 multiyear Baseline Surface Radiation Network measurements, *J. Geophys. Res. Earth Surf.*, *116*(F03008), doi:10.1029/2010JF001864. 3

Zender, C. S. (2008), Analysis of self-describing gridded geoscience data with netCDF Operators (NCO), *Environ. Modell. Softw.*, *23*(10), 1338–1342, doi:10.1016/j.envsoft.2008.03.004, available from `http://dust.ess.uci.edu/ppr/ppr_Zen08.pdf`. 3

Zender, C. S. (2012), Snowfall brightens antarctic future, *Nature Clim. Change*, *2*(11), 770–771, doi:10.1038/nclimate1730. 3, 5

Zender, C. S. (2013), NCO User's Guide, version 4.3.2, `http://nco.sf.net/nco.pdf`. 2.1, 3.1, 3.3

Zender, C. S., and H. J. Mangalam (2007), Scaling properties of common statistical operators for gridded datasets, *Int. J. High Perform. Comput. Appl.*, *21*(4), 458–498, doi:10.1177/1094342007083,802. 3