A Data Library to Archive, Analyze, Visualize and Serve Online Datasets from Multiple Domains in an Interoperable Framework

Rémi Cousin[1], Sruti Devendran[1], and John del Corral[2]

[1]Columbia University of New York
[2]Columbia University

November 28, 2022

Abstract

The PRISM Data Library (DL) is designed to optimize the display, analysis, and retrieval of multiple domains datasets. Originally created for climate data, we aggregated data from agriculture and hydrology domains, as well as non-traditional domains for the DL such as ecology, finance, power outage and space weather data. These datasets range from simple geospatial point observations, to spatially gridded data products, to high-resolution satellite measurements, to GIS representation of administrative or domain-specific geographic entities. These datasets are represented in a consistent multi-dimensional (most often spatial and temporal) framework. As a result, dimension-wise comparisons are easily enabled through selection or transformation. Gridded data can be averaged over discrete geometrical entities (e.g. Counties, Bird Conservation Regions). The DL can be used in a browser, by connecting to servers at San Diego Supercomputing Center (SDSC) over the internet. Data selection, processing, and analysis are performed by the SDSC DL servers, and the resulting images or data files are sent back to the client's desktop. This model optimizes the use of internet bandwidth.

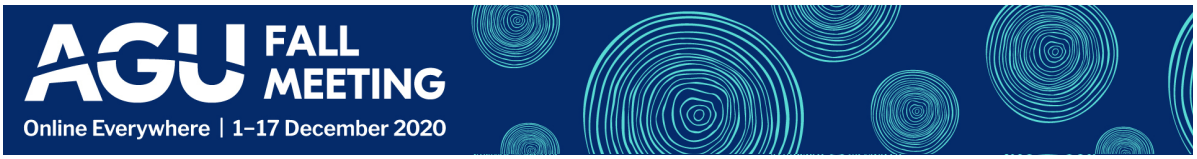
A Data Library to Archive, Analyze, Visualize and Serve Online Datasets from Multiple Domains in an Interoperable Framework

Rémi Cousin, Sruti Devendran, John del Corral

International Research Institute for Climate and Society, Earth Institute, Columbia University



PRESENTED AT:

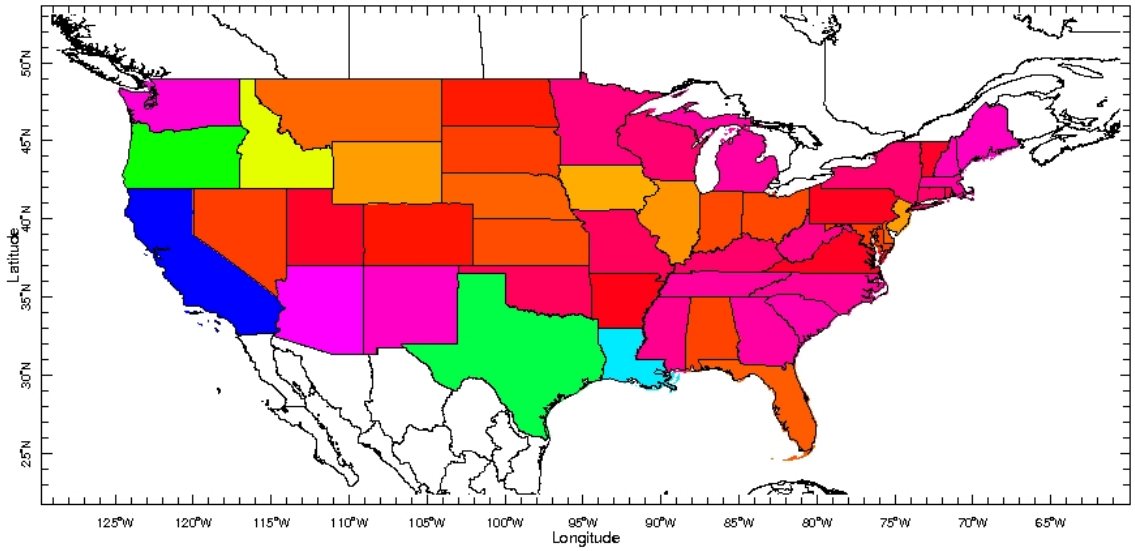


DATA REPOSITORY AND INTEGRATION

Illustration of different datasets in the SDSC DL (click maps to get to DL interface):

ECOLOGY: BBS (http://datalibrary.sdsc.edu/SOURCES/.USGS/.PWRC/.BBS/.State/.Indices/) bird Shannon diversity index

by state, yearly (2002 shown)

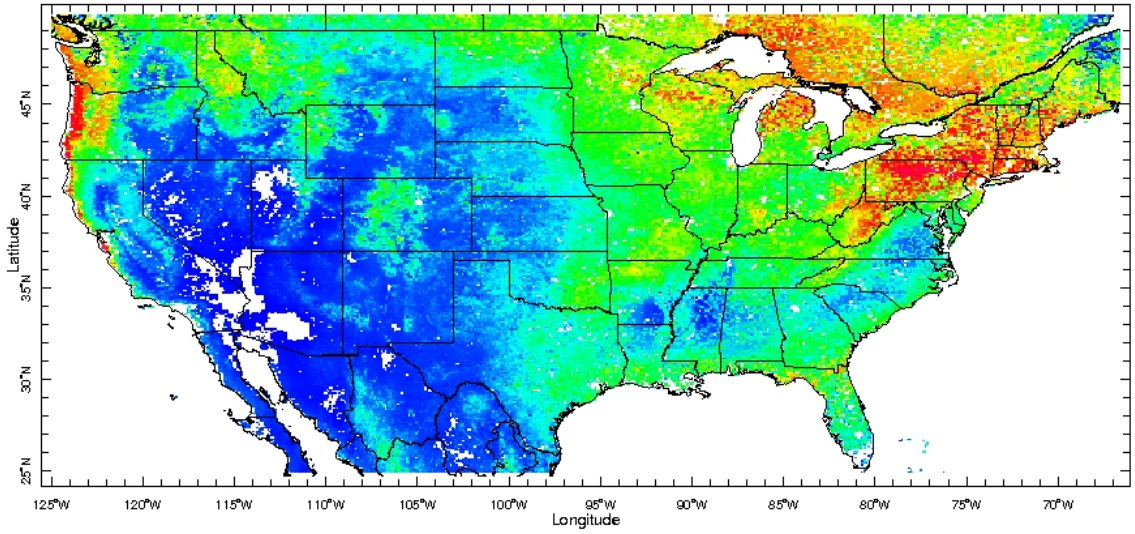


2002

(http://datalibrary.sdsc.edu/SOURCES/.USGS/.PWRC/.BBS/.State/.Indices/a-/.SOURCES/.WORLDBATH/.bath/-1/min/-a-/.the_geom/-a-/.shannon/-a-/.the_geom/-a/X/Y/fig%3A/white/mask/blue/fillby/black/stroke/countries/%3Afig//T/690.0/plotvalue//plotborder/72/psdef//plotaxislength/432/psdef/?bbox=bb%3A232.42%3A20.80%3A296.38%3A52.02%3Abb&T=2002#expert)

AGRICULTURE: MODIS GPP (http://datalibrary.sdsc.edu/SOURCES/.USGS/.LandDAAC/.MODIS/.MOD17A2H/.version_006/.CONUS0p1/)

0.1° res, 8-daily (4-11 Jul 2002 shown)

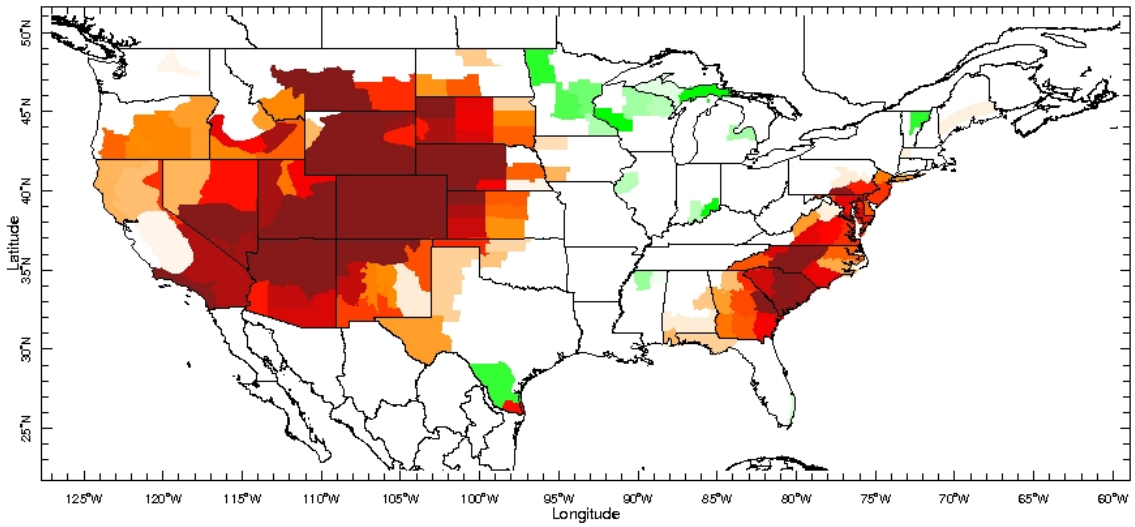


4-11 Jul 2002

(http://datalibrary.sdsc.edu/SOURCES/.USGS/.LandDAAC/.MODIS/.MOD17A2H/.version_006/.CONUS0p1/.Gpp/a-/-a/X/Y/fig-/colors/states/-fig//T/7302.5/plotvalue//plotborder/72/psdef//plotaxislength/432/psdef/?T=4-11%20Jul%202002#expert)

HYDROLOGY: NOAA PDSI (http://datalibrary.sdsc.edu/SOURCES/.NOAA/.NCDC/.CM/.Drought/.HPDI/)

by climate division, monthly (Jul 2002 shown)



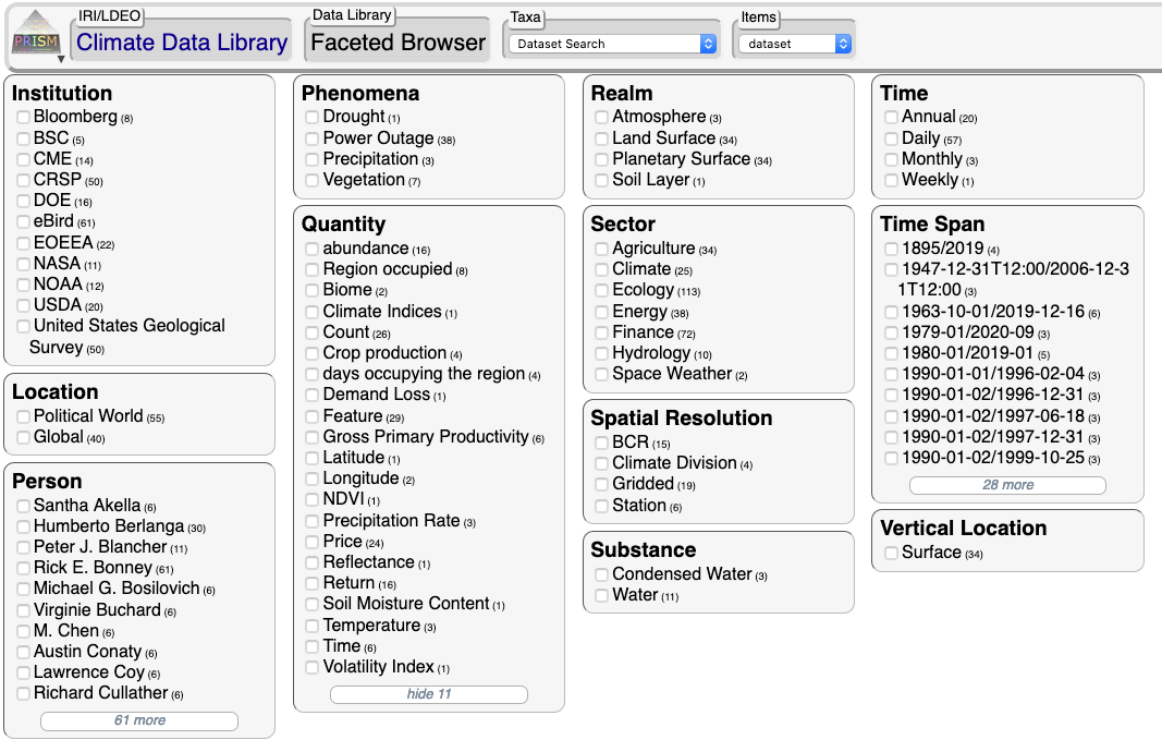
Jul 2002

(http://datalibrary.sdsc.edu/SOURCES/.NOAA/.NCDC/.CM/.Drought/.HPDI/a-/.SOURCES/.WORLDBATH/.bath/-1/min/-a-/.the_geom/-a-/.PDSI/-a/X/Y/fig%3A/white/mask/blue/fillby/black/states/%3Afig//T/-779.5/plotvalue//plotborder/72/psdef//plotaxislength/432/psdef/?T=Jul%202002&bbox=bb%3A233.83%3A22.66%3A295.20%3A49.91%3Abb#expert)

Current data files extensions read in the SDSC DL:

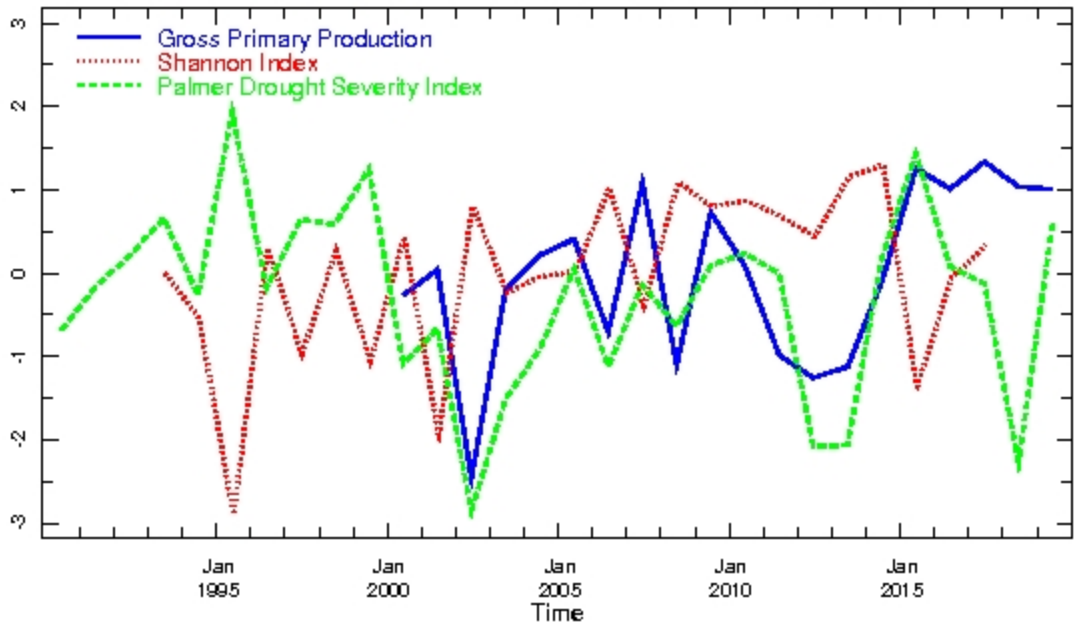
.txt, .tsv, .asc, .tif, .shp, .nc, .nc4, .hdf, .dat.gz

Facted search illustrates wealth of data domains covered (note that tagging is on-going and thus incomplete):



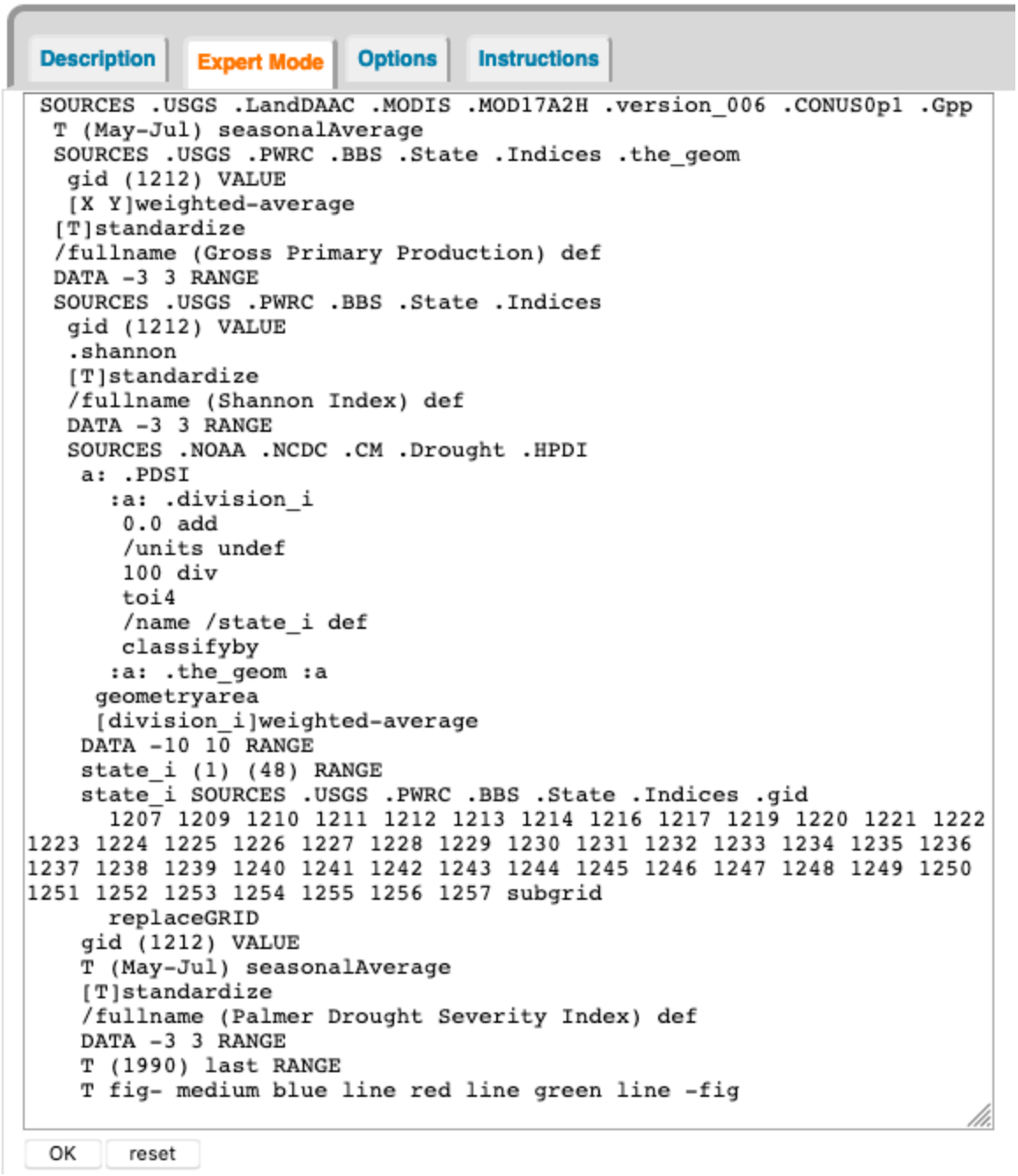
DATA ANALYSIS AND VISUALIZATION

Via clickable UI or programming in a text box, users can perform data analysis (e.g. regridding, time series analysis, spatial averaging, inter/extrapolation; EOFs, correlatin, clustering) as illustrated by the time series graph below that takes data from the 3 maps on left pannel averaged over CA for May-July & standardized:



1212

Programmed online (http://datalibrary.sdsc.edu/SOURCES/.USGS/.LandDAAC/.MODIS/.MOD17A2H/.version_006 /.CONUS0p1/.Gpp/T/(May-Jul)/seasonalAverage/SOURCES/.USGS/.PWRC/.BBS/.State/.Indices/.the_geom/gid/(1212) /VALUE/[X/Y]weighted-average/[T]standardize//fullname/(Gross%20Primary%20Production)/def/DATA/-3/3/RANGE /SOURCES/.USGS/.PWRC/.BBS/.State/.Indices/gid/(1212)/VALUE/.shannon/[T]standardize//fullname/(Shannon%20Index) /def/DATA/-3/3/RANGE/SOURCES/.NOAA/.NCDC/.CM/.Drought/.HPDI/a:/.PDSI/:a:/.division_i/0.0/add//units/undef/100/div /toi4//name//state_i/def/classifyby/:a:/.the_geom/:a/geometryarea/[division_i]weighted-average/DATA/-10/10/RANGE/state_i /(1)/(48)/RANGE/state_i/SOURCES/.USGS/.PWRC/.BBS/.State/.Indices/.gid/1207/1209/1210/1211/1212/1213/1214/1216/1217 /1219/1220/1221/1222/1223/1224/1225/1226/1227/1228/1229/1230/1231/1232/1233/1234/1235/1236/1237/1238/1239/1240 /1241/1242/1243/1244/1245/1246/1247/1248/1249/1250/1251/1252/1253/1254/1255/1256/1257/subgrid/replaceGRID /gid/(1212)/VALUE/T/(May-Jul)/seasonalAverage/[T]standardize//fullname/(Palmer%20Drought%20Severity%20Index) /def/DATA/-3/3/RANGE/T/(1990)/last/RANGE/T/fig-/medium/blue/line/red/line/green/line/-fig/#expert):



A breadth of words allow to program any type of analysis, as illustrated by the Function Documentation (http://datalibrary.sdsc.edu/dochelp/Documentation/funcmenu.html):

average Calculates the average

B

beginLoop marks the beginning of a loop

bias_mean calculates the **mean bias** for deterministic forecasts **fcst** from observations **obs**.

BofA=C Converts a variable A to a variable B using a table B(C=A) and linear interpolation. Out of range values beyond half a grid step are NaN

BofA=C-bounded Converts a variable A to a variable B using a table B(C=A) and linear interpolation. Out of range values are pegged to the extreme values

boxAverage Calcuates the box average. Commonly used for creating seasonal averages. Note: function should only be used with continuous data domain (see example below).

butt_design
designs a Butterworth filter

C

cca computes canonical correlation analysis between leftVar and rightVar.

changetruncation Changes variable truncations to nsp

classify Classifies data into categories, i.e. labels ranges of values.

classifyby Classifies var1 by var2

CofA=B Converts a variable A to a variable B using a table B(C) and linear interpolation. Out of range values beyond half a grid step are NaN

CofA=B-bounded Converts a variable A to a variable B using a table B(C) and linear interpolation. Out of range values are pegged to the extreme values

constantdata Returns constant data

correlate Calculates the Pearson Product-Moment Correlation coefficient of two variables over specified grids (i.e., independent variables)

cos Calculates the cosine of a number or variable (given in radians)

cosd Calculates the cosine of a number or variable (given in degrees)

cptv10 creates a CPT v10 file from the input

(http://datalibrary.sdsc.edu/dochelp/Documentation/funcmenu.html)

N

normalize Divides var1 by var2. Points less than minimum in var2 become NaN in the output

normalizeddistrib1D Returns the normalized frequency distribution of a set of data for a specified range and step interval.

normalizeddistrib2D Computes the distribution of A vs B (see distrib2D) and then renormalizes by the integral along B. This new variable has the property that the integral along B is 1

O

onsetDate Computes yearly onset dates from daily rainfall data

openquery Opens a query

openqueryby Opens a query indexed by indexvar

P

pad0 pads begining and end of stream along ordered grid with a length of pl steps of grid with mean value, ie the zeroth derivative order boundary constraint.

pad1 pads begining and end of stream along ordered grid with a length of pl steps of grid by y-axis reflection symmetry at each extremity point stream, ie the first derivative order boundary constraint.

pad2 pads begining and end of stream along ordered grid with a length of pl steps of grid by pi-rotation symmetry at each extremity point of stream, ie the second derivative order boundary constraint.

pairsums Pairwise sums along independent variable of variable. This is the variable equivalent of integralgrid

partial Takes partial derivative of variable along grid

partialeast Applies zonal derivative to variable in spectral coordinates

partialnorth Applies meridional derivative to variable in spectral coordinates

partitiongrid splits an independent variable into two parts: a coarse scale grid and a fine-scale subgrid. The two ivars point to each other with sophisticates and isSophisticatedBy.

pentad3Q Third quartile (by pentad of year) of multiple calendar years of values on a

DATA DISSEMINATION (DOWNLOADS)

Besides being able to download all the images generated through the DL, users can also download the raw data or any virtual results of their expert mode programs in a variety of data formats, or send to third party applications.

Illustration of image formats:

Illustration of data file formats:

ingrid	The Postscript-based software on which the Data Library is built.
CPT	Climate Predictability Tool More information
ferret	Interactive computer visualization and analysis software. More information
GrADS	Grid Analysis and Display System More information
matlab	Data analysis and visualization software. More information
NCL	NCAR Command Language More information
WinDisp	A public domain software package for the display and analysis of satellite images, maps and associated databases, with an emphasis on early warning for food security. More information

Other Available File Formats

Full Information Formats These files contain all of the available metadata.	
OPeNDAP	A system which downloads data directly to software, such as matlab, Ferret, GrADS, etc. Specific instructions are available in the table above. Note: OPeNDAP was formerly known as DODS (Distributed Oceanographic Data System). More Information
netCDF (network Common Data Form)	A commonly supported self-describing data format. More Information

Partial Information Formats These files contain only some of the available metadata.	
For the remaining data formats, the following information may be helpfulthe **scale_factor** is 1. and the **add_offset** is 0, i.e. the data is already properly scaled. The **missing_value** (flag for missing data) is NaN	
Columnar Table	A table with separate columns of numbers for each independent variable (i.e., grids) and for the data. This is an inefficient format, so you would have gotten a **HUGE** file for dataset of this size. This file will be approximately -1090676848 bytes, with 3 columns of 267024204 numbers.
2-Dimensional Tab-Separated Tables	Tab-separated-values (tsv) file with information about the independent variables (i.e., grids). The list to the left allows you to specify the format of the table. Note: The variable running across the top of the table (identifing columns) is listed first and the variable running down the side of the table (identifing rows) is listed second.
GIS-Compatible Formats There are three GIS-compatible formats available.	
2-Dimensional Table	A 2-dimensional ascii file that includes an ArcInfo Header.
IDA Image	File(s) in the Image Display and Analysis format. Typically used with WinDisp.
LAN Image	File(s) in the ERDAS LAN format. Typically used with various GIS programs, including ArcView and HealthMapper.
GeoTIFF Image	File in GeoTIFF format. Typically used with various GIS programs, including ArcView and ENVI.

Data Only Formats These files contain just the data without any of the available metadata.	
Binary direct access	A big-endian, ieee single-precision file in floating-point format. Also known as a binary random access file. This is a random-access file; it is purely data with no record-structuring information. The data is structured to correspond to the independent variables (i.e., grids) in X Y order, with the first grid varying the fastest.
DEC ALPHA direct access	Same as the binary random/direct access format above except that it is byte-swapped for DEC ALPHA's and PC's (little-endian).
Binary FORTRAN sequential access	A big-endian, ieee, single-precision file in floating-point format. This is a sequential-access file with each record containing all the X Y points. It **must** be read using FORTRAN sequential access. There is only one record for the data you have selected.
DEC ALPHA sequential access	Same as the binary sequential access format above except that it is byte-swapped for DEC ALPHA's and PC's (little-endian).
Text with tab-separated-values	Text file where data values corresponding to different **X** are separated by tabs and data values corresponding to different **Y** are on different lines. This is readable by most programs, including spreadsheets, but will be about four times larger than the binary or netCDF/HDF files noted above.
Text	Text file where data is arranged in chunks of X Y. There are five values per line and each chunk starts on a new line. This will be about four times larger than the binary or netCDF/HDF files.

Illustration of OPeNDAP workflow:

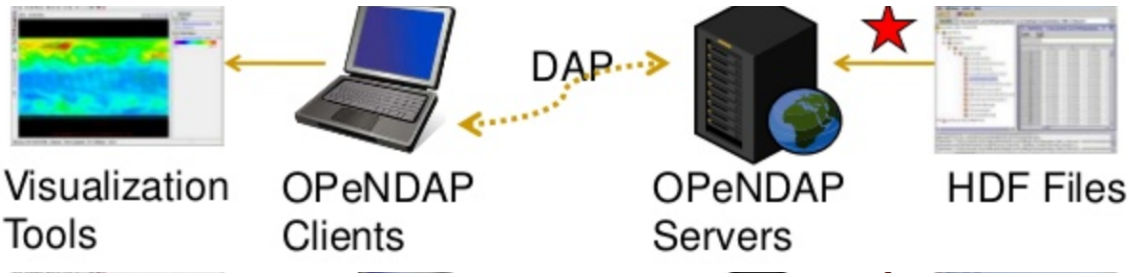


Illustration of sending to 3rd party app:

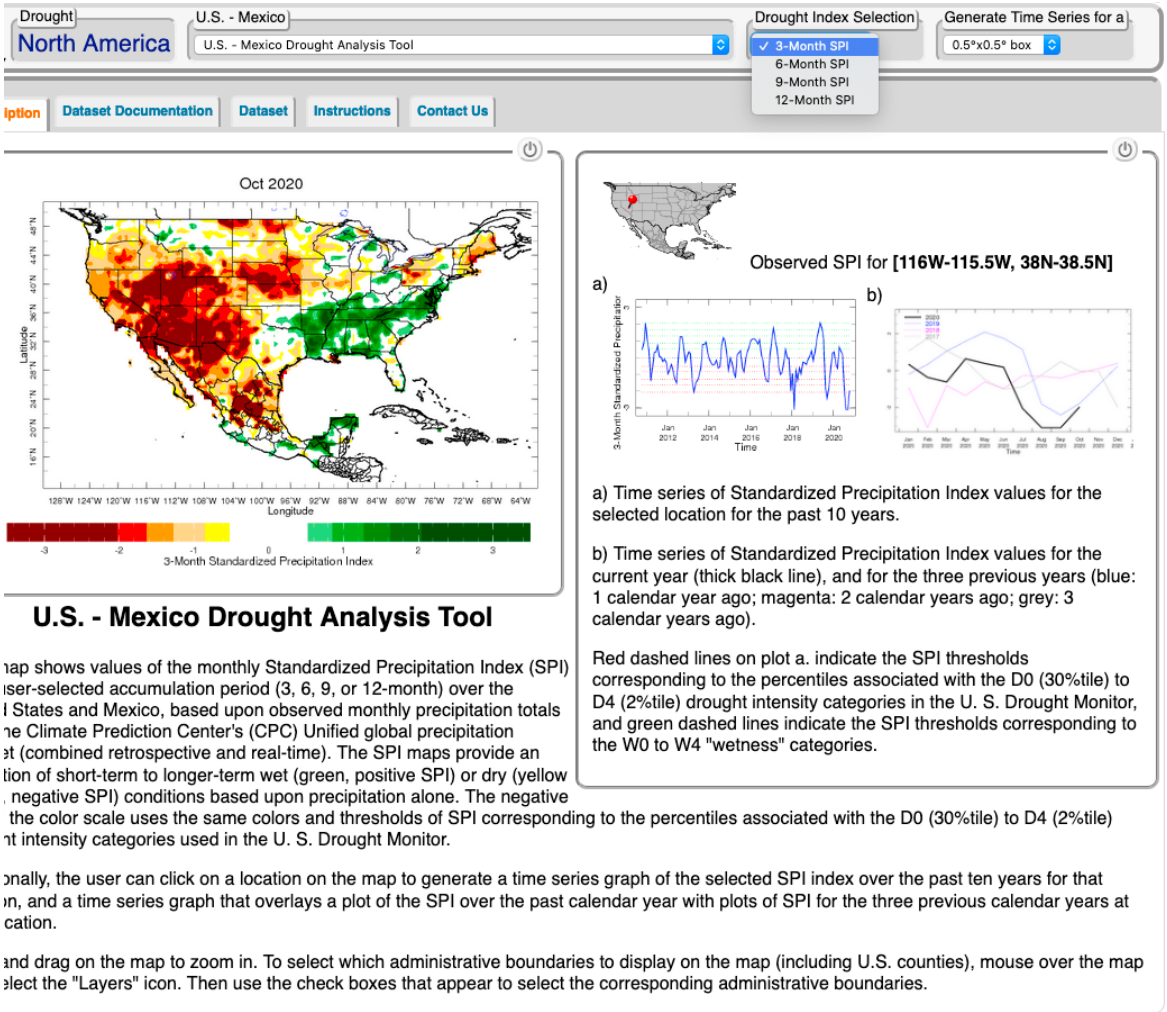
[VIDEO] https://player.vimeo.com/video/479978448?byline=0&badge=0&portrait=0&title=0

WEB INTERFACE & "MAPROOMS"

The Data Library is a web service and the preceding 3 panels illustrate what users can accomplish through it.

A request for data, data analysis, or image is assembled in a virtual task tree as the user refines the details of their request and is only executed when actual images or datafiles are requested by the user. The computations for a request are done on the server side where the data resides.

This framework facilitates the collaboration between researchers or any other kind of users. It also allows to build websites, dedicated at making potentially complex queries to the DL via simple html controls. We call these Maprooms, e.g. (http://iridl.ldeo.columbia.edu/maproom/Global/Drought/N_America/US_Mexico/Analysis.html?region=bb%3A-117.5%3A36.5%3A-117%3A37%3Abb):



(http://iridl.ldeo.columbia.edu/maproom/Global/Drought/N_America/US_Mexico/Analysis.html?region=bb%3A-117.5%3A36.5%3A-117%3A37%3Abb)

VISIT the PRISM SDSC DL (http://datalibrary.sdsc.edu/).

DISCLOSURES

This work was funded by the National Science Foundation, Award Number 1940276: 'Predictive Risk Investigation SysteM (PRISM) for Multi-layer Dynamic Interconnection Analysis'

AUTHOR INFORMATION

Rémi Cousin

Senior Staff Associate
Data Library Manager

Since 2008, Cousin has been a member of the IRI Data Library, which disseminates science-based climate information to its user community. To fulfill this task, Cousin assists in research activities, and engages with user communities through collaborative work or training workshops.

Sruti Devendran

Staff Associate
Data Library

Devendran was a member of the IRI Data Library in 2019-2020, which disseminates science-based climate information to its user community. To fulfill this task, Devendran assists in research activities, and engages with user communities through collaborative work or training workshops.

John del Corral

Senior Staff Associate
Database, GIS, Semantic Technology

He is interested in the role of computers and computational science in multi-disciplinary areas of geophysical research. This includes high performance computing, graphical techniques, geographical information systems, database technology, semantic technology, and learning the basic science in the areas of research. del Corral is actively involved in the maintenance of and upgrades for the IRI Data Library. is projects for the Data Library involve the use of GIS, databases, and semantic technology.

ABSTRACT

The PRISM Data Library (DL) is designed to optimize the display, analysis, and retrieval of multiple domains datasets. Originally created for climate data, we aggregated data from agriculture and hydrology domains, as well as non-traditional domains for the DL such as ecology, finance, power outage and space weather data. These datasets range from simple geospatial point observations, to spatially gridded data products, to high-resolution satellite measurements, to GIS representation of administrative or domain-specific geographic entities. These datasets are represented in a consistent multi-dimensional (most often spatial and temporal) framework. As a result, dimension-wise comparisons are easily enabled through selection or transformation. Gridded data can be averaged over discrete geometrical entities (e.g. Counties, Bird Conservation Regions). The DL can be used in a browser, by connecting to servers at San Diego Supercomputing Center (SDSC) over the internet. Data selection, processing, and analysis are performed by the SDSC DL servers, and the resulting images or data files are sent back to the client's desktop. This model optimizes the use of internet bandwidth.

REFERENCES

Blumenthal, M.B., Bell, M., del Corral, J., Cousin, R., Khomyakov, I., 2014. IRI Data Library: enhancing accessibility of climate knowledge. Earth Perspectives 1, 19.