SubX User's Guide

Chapter 1: Data Quick Reference

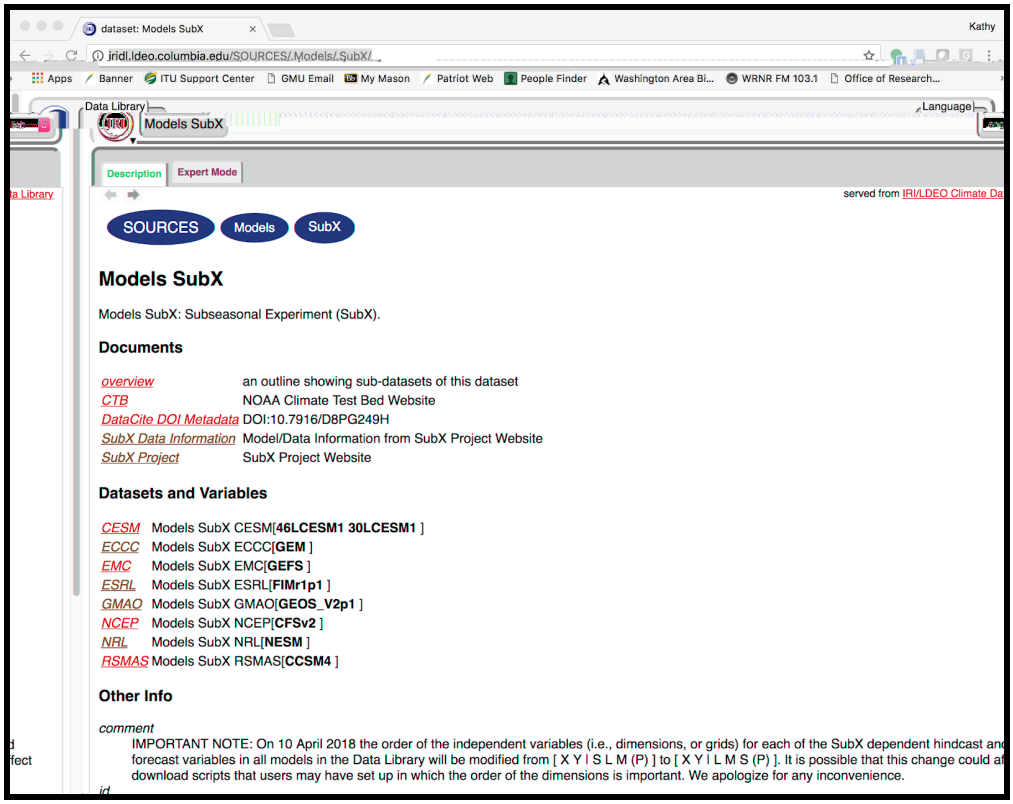
1. Where can I find the data?

SubX retrospective and real-time forecast data are located on the International Research Institute for Climate and Society Data Library (IRIDL). The URL for this data is: http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/

2. How do I know what data are available?

A brief summary of the data available can be found on the SubX website under "Data Description" (http://cola.gmu.edu/kpegion/subx/data/descr.html)

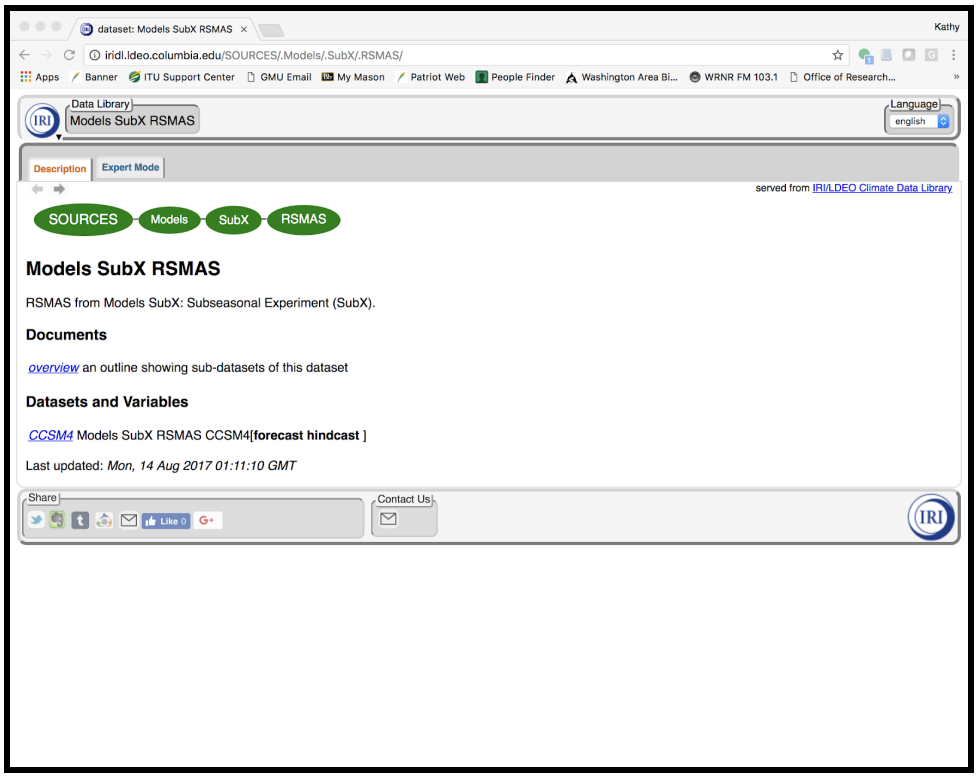
You can find more specific details about the data available by navigating to the SubX database at the IRIDL. The web interface allows you to click through the metadata and see what data is provided .



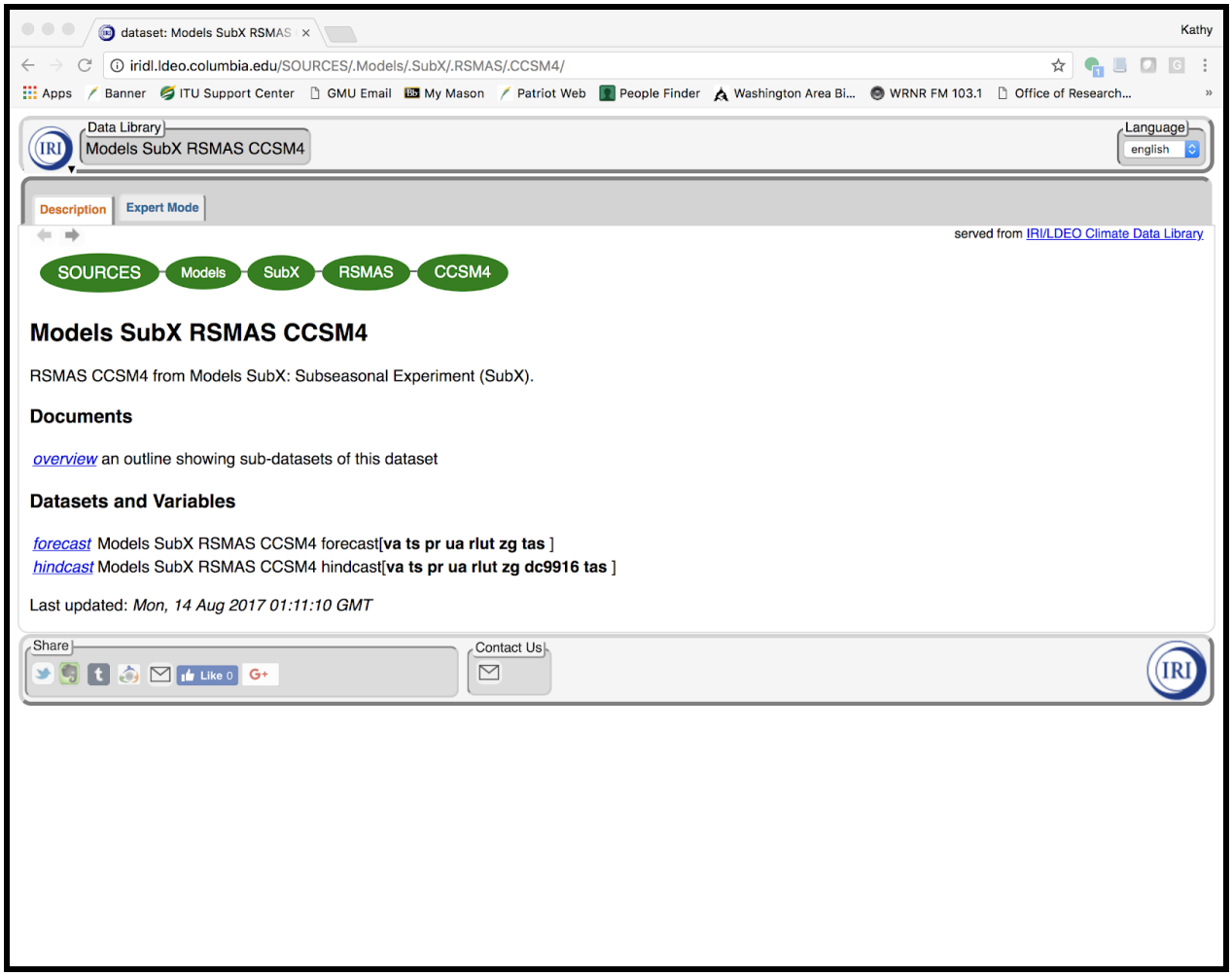
This page shows the modeling groups providing data to SubX. Here, you can see that there are 8 groups with available data (CESM, ECCC, EMC, ESRL, GMAO, NCEP, NRL, RSMAS).

The following steps will walk through an example of how to view more information about available data:

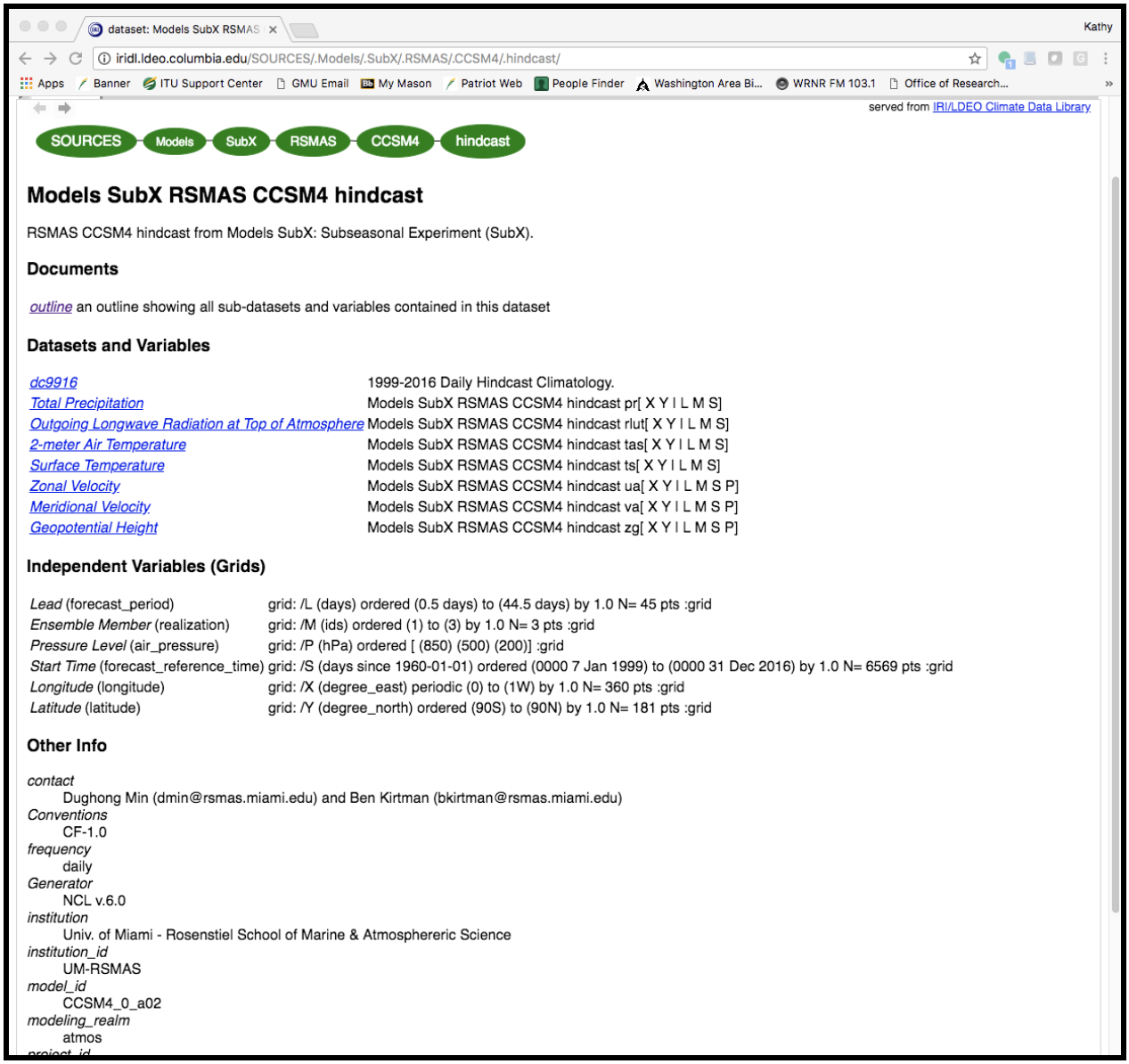
a) Click on one of the groups to see the next level of available data. After clicking on RSMAS, the page below appears. This page shows that there is one model provided by the RSMAS group, CCSM4. If the RSMAS group provides additional models in the future, there will be multiple selections listed here.



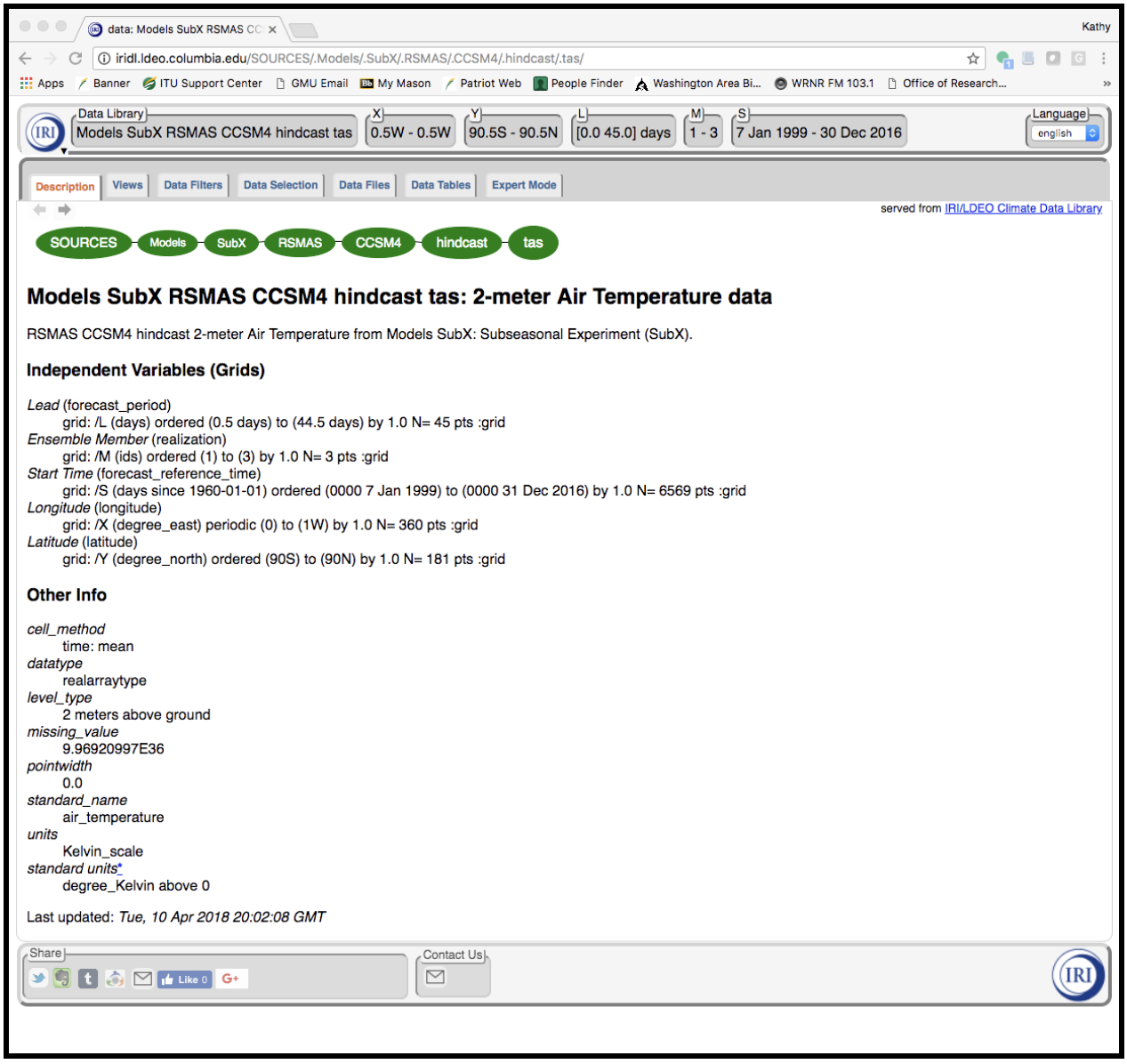
b) Click on the model ("CCSM4") to see the next level of available data. The results are shown in the figure below. This page shows that there are hindcast and forecast data available for this model.



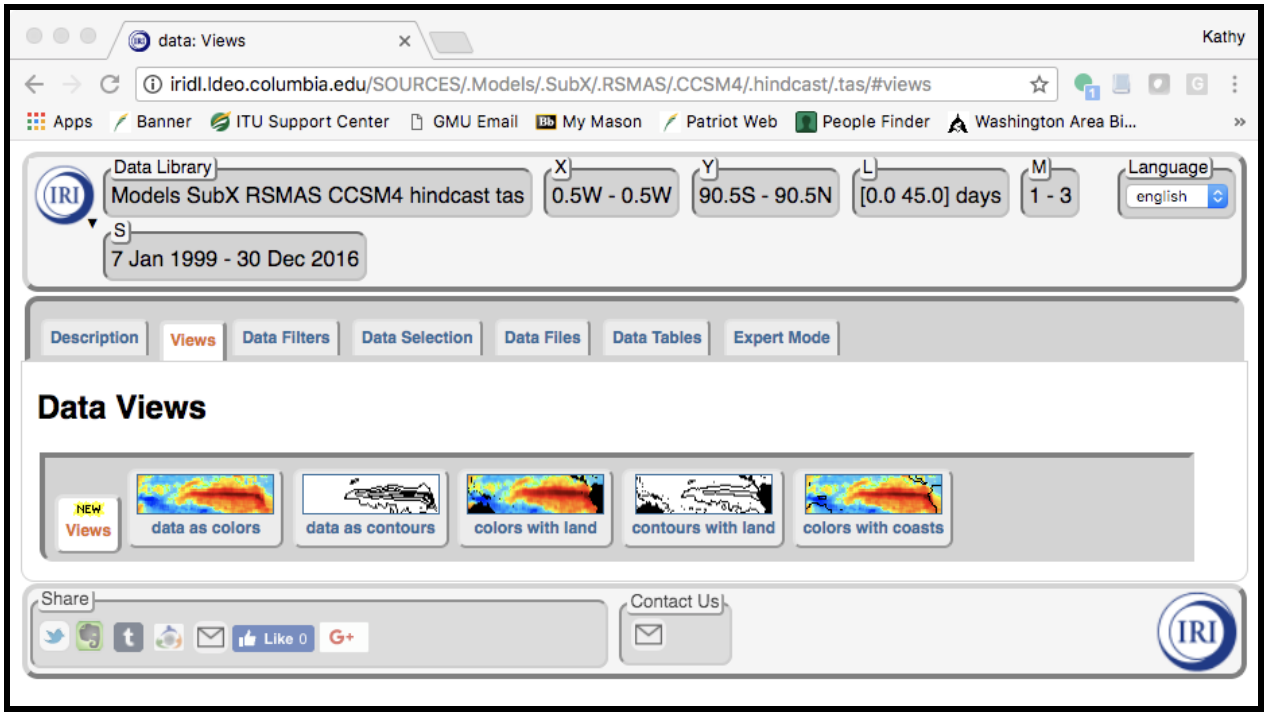
c) Click on “hindcasts” to see the next level of available data. The results are shown in the figure below. This page shows the available variables for this model (“Datasets and Variables”), the dimensions of the data (“Independent Variables (Grids)”), and additional metadata for this datasets (“Other Info”).



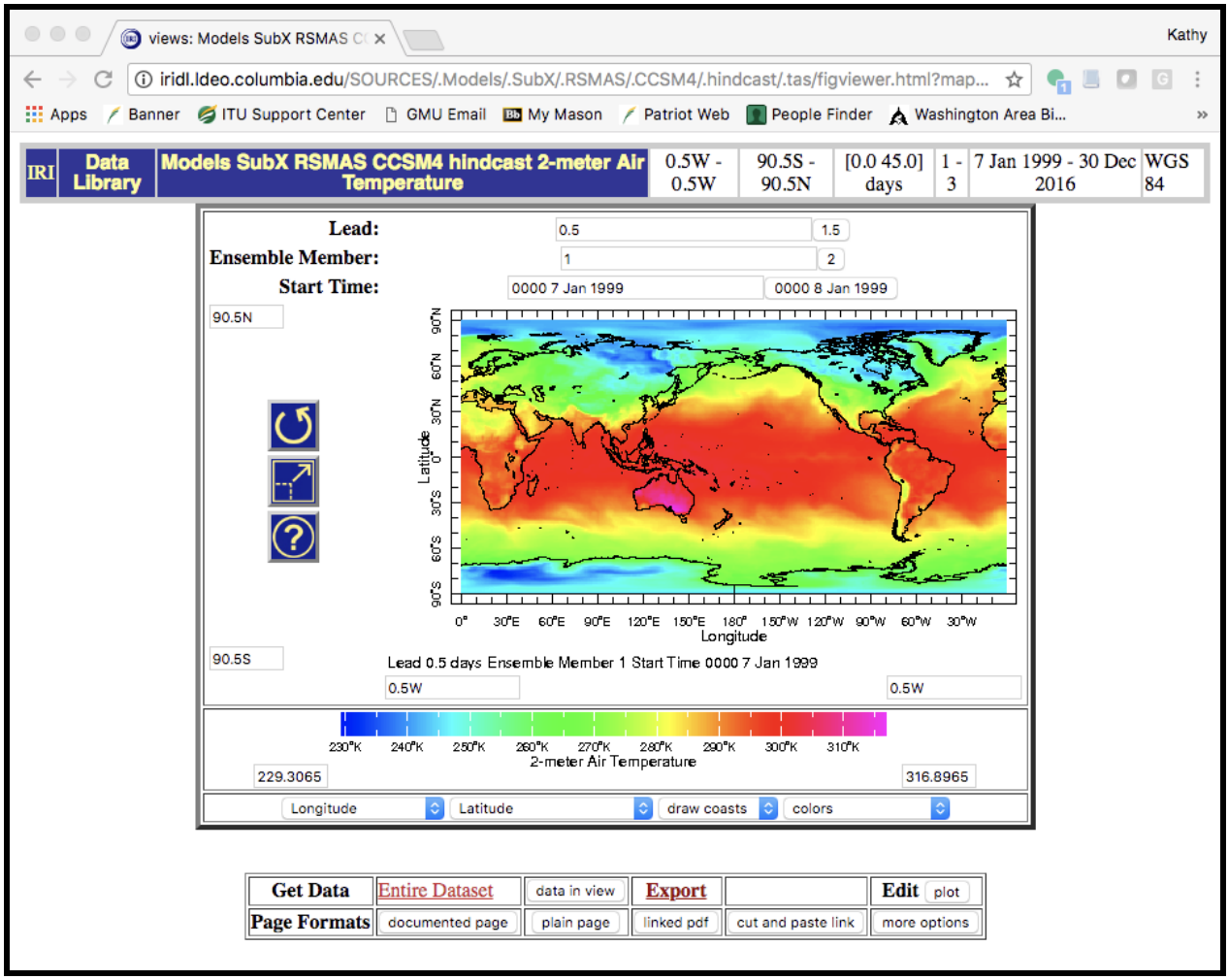
d) Next, select a variable (e.g. 2-meter Air Temperature) to view the available data for that variable. The results are shown in the figure below. The dimension information is shown ("Independent Variables (Grids)" as well as other metadata ("Other Info").



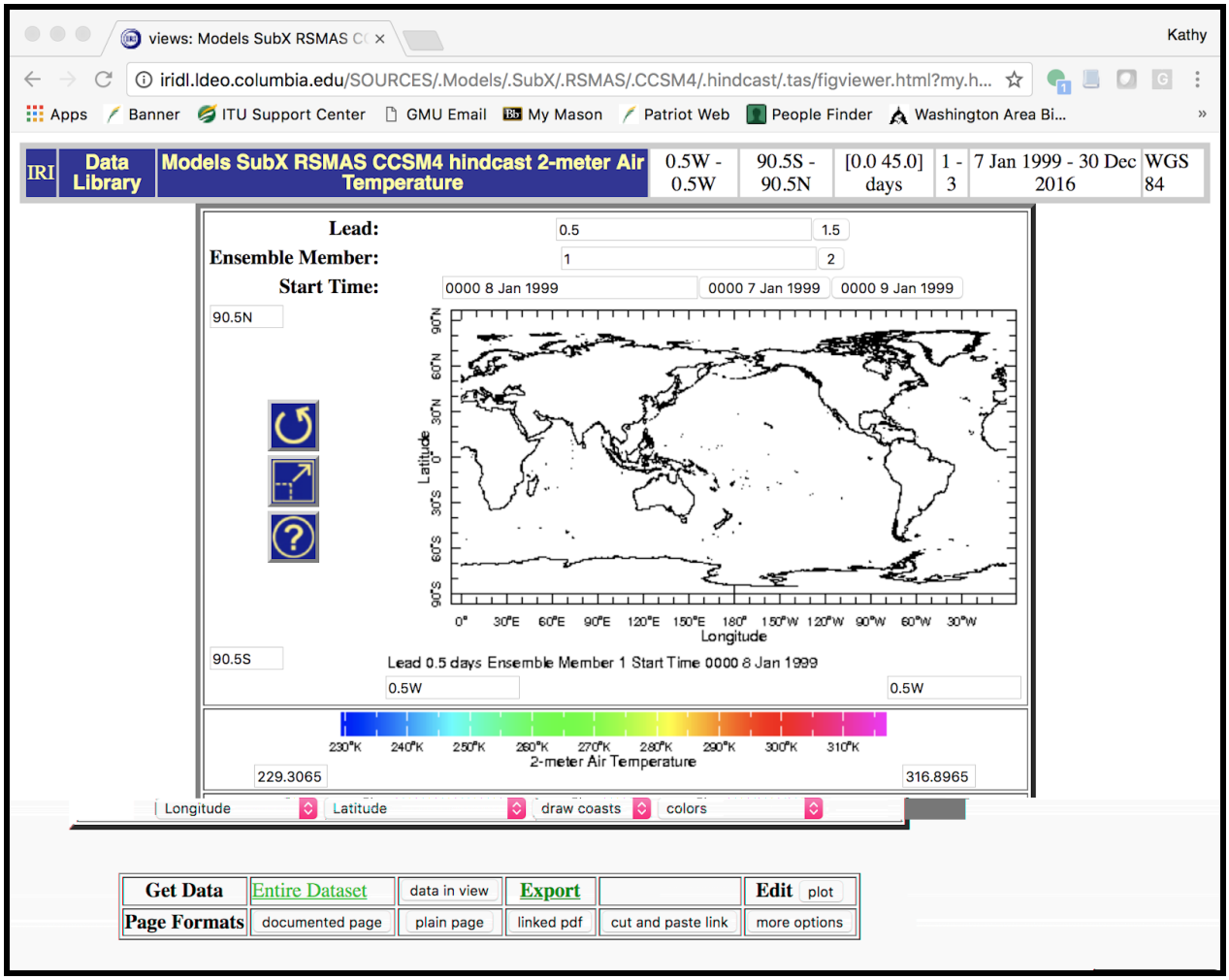
e) The IRIDL provides a data viewer that allows the user to see a graph of the data as well as various other options for working with the dataset. Near the top of the page, there are 7 tabs, one of which is labelled “Views”. Selecting this tab will bring up options for how you wish to see the data in the data viewer (see figure below).



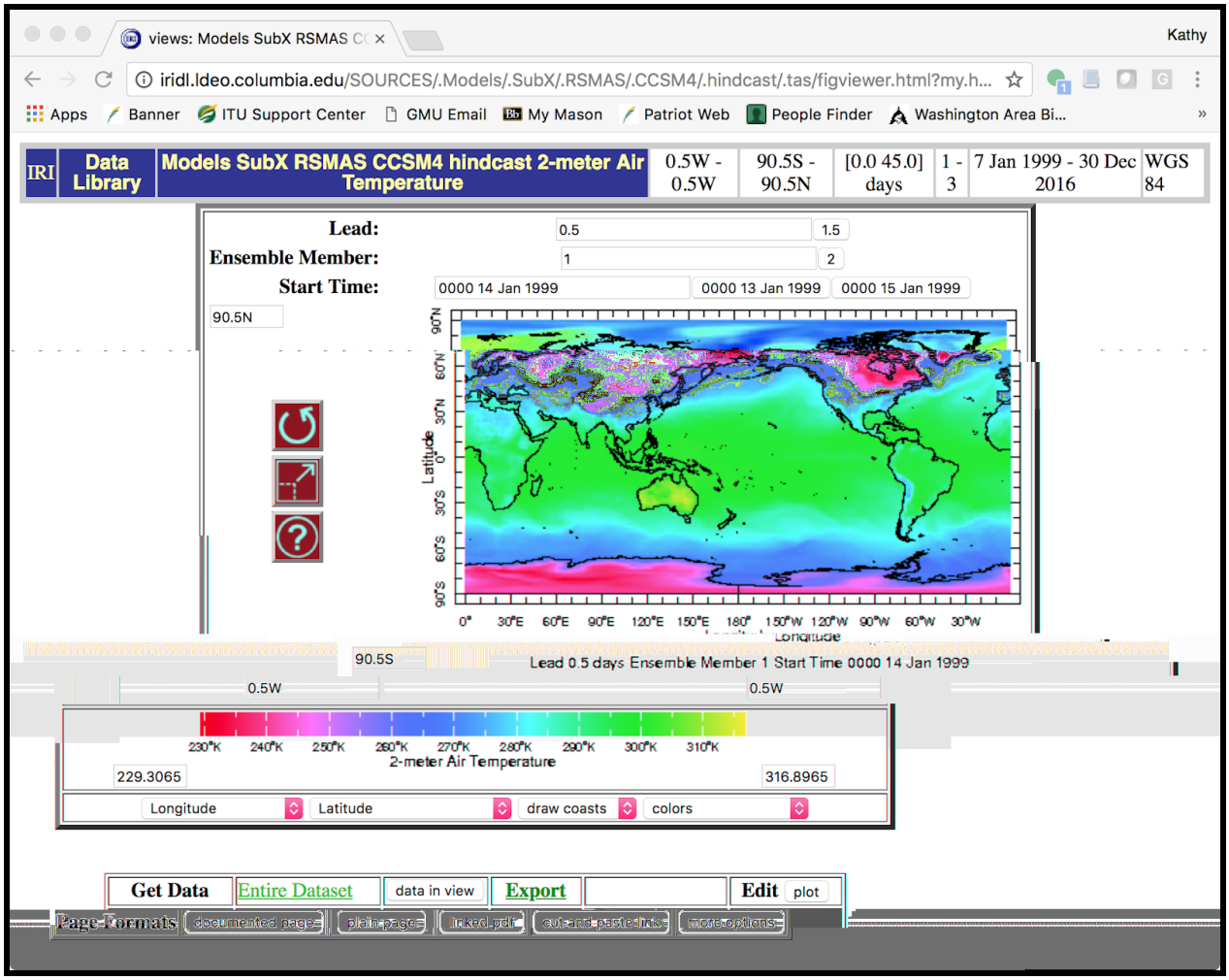
f) Choose the “colors with coasts option” to view the data on a map with coastlines drawn in (see figure below). This shows the 2m temperature data for the RSMAS, CCSM4 dataset for Lead=0.5, Ensemble Member=1, and Start Time: Jan 7, 1999 for the entire globe. In the data viewer, you can change the longitude and latitudes over which to view the data, the values over with to contour, and the lead time, ensemble member, or start time to view.



g) Since the SubX dataset contains data initialized weekly and the initialization data for the models can vary, it is helpful to use the data viewer to identify the start times for which there is data. Click the Start Time: 0000 8 Jan 1999 button to view the data for re-forecasts initialized on Jan 8, 1999. The results are shown below, which shows a map with no data. This is because no data exists for the RSMAS-CCSM4 model for start times of Jan 8, 1999.



h) The RSMAS-CCSM4 re-forecasts are initialized every 7 days (see table of this information for each model http://cola.gmu.edu/kpegion/subx/data/descr.html), so the next start date for which there should be available data is Jan 14, 1999. Click on the dates next to start time until Jan 14, 1999 has been selected. The results are shown below. The data viewer shows a contoured map of 2m Temperature for the RSMAS-CCSM4 model for Lead=0.5, Ensemble Member=1, and Start Time=Jan 14, 1999. All of the data for RSMAS-CCSM4 can be viewed in this way using the data viewer.



3. How do I download the data?

Overview

The IRI Data Library is an OPenDAP Server. Therefore various languages and environments can be used to access the SubX data via OPenDAP. Once an OPenDAP connection has been established, the data are stored as netCDF. To establish an OPenDAP connection, you supply the application a URL and the data can be directly downloaded.

To determine the URL for the data you wish to download, you can click through the available data until you arrive at a dataset you wish to download (see section 2). Following the example from section 2 and arriving at Figure 5, select the tab “Data Files” and “OPeNDAP”. The URL for this dataset is provided as:
http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/.RSMAS/.CCSM4/.hindcast/.tas/dods (Figure 10)

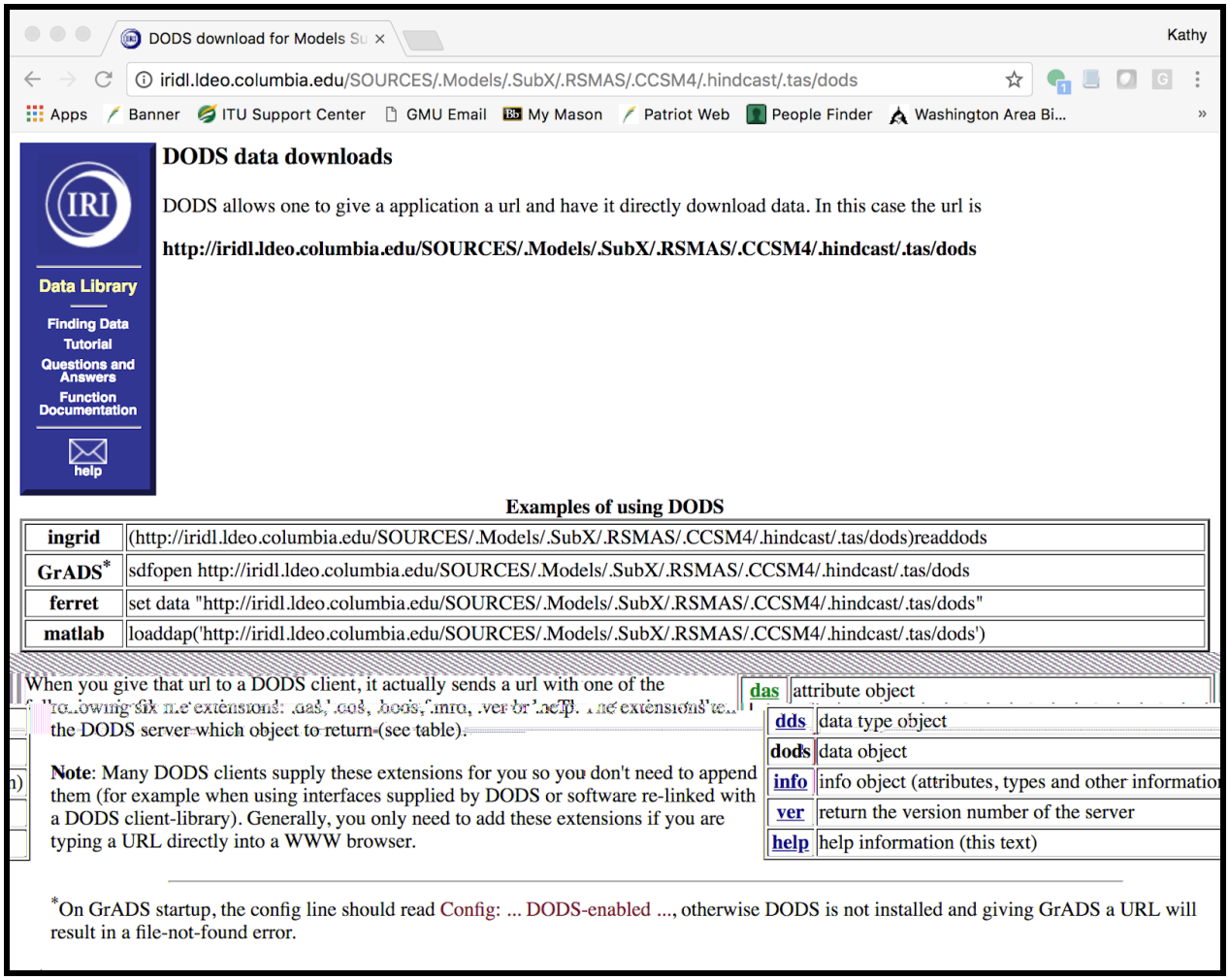


Figure 10

Sample Programs & Tutorial

Sample programs to download data are available in Matlab, Python, NCL, and GrADS. These codes are available on Github (https://github.com/kpegion/SubX). The step-by-step instructions below show how to get the data using the Python sample codes.

a) Download the Python code from Github

The quickest way to download codes from GitHub for users who are not experienced with git is to navigate to the github page for Version 1 of the codes (https://github.com/kpegion/SubX/releases) and then “Source Code (tar.gz) (see below). The file SubX-1.0.tar.gz will be dowloaded. You will need to place the file on the correct computer and in the location you want it. For example, I use scp to transfer the file to my local servers used for data processing.

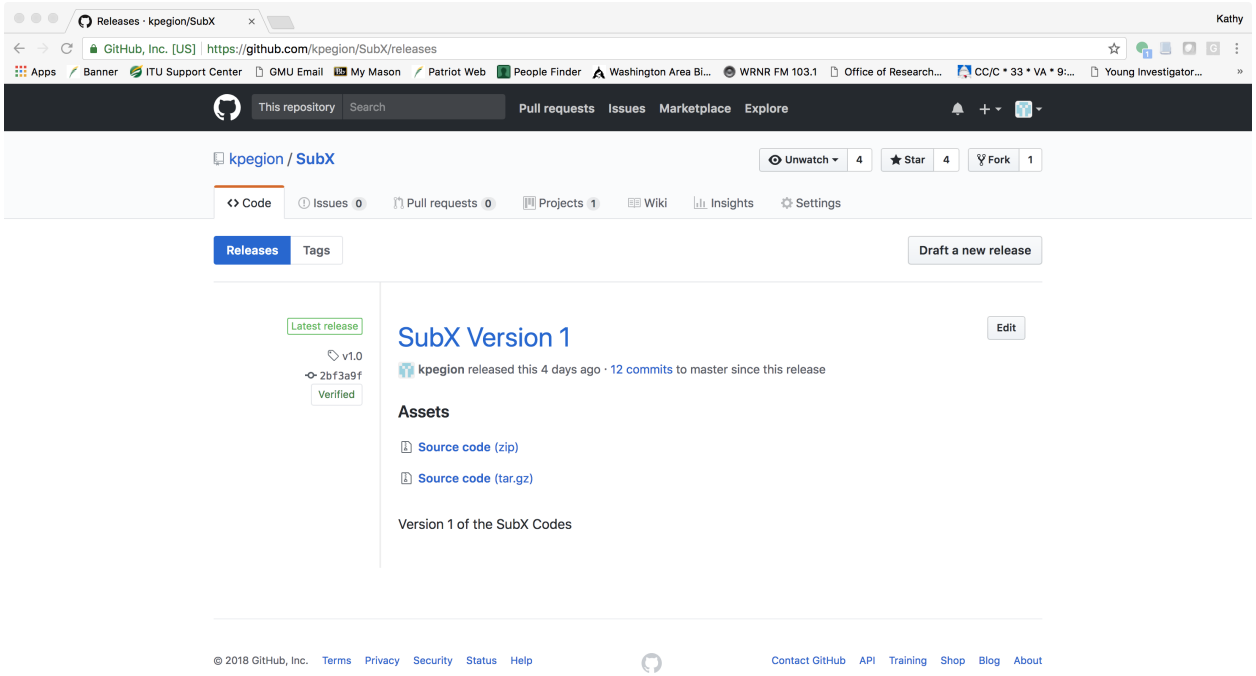


Figure 11

b) Once the file has been moved to the correct location, unzip and untar the file. You will now have a directory called SubX-1.0 that contains Version 1 of the SubX codes from GitHub.

c) cd to the SubX-1.0 directory where you will all of the GitHub SubX codes

d) cd to the Python directory, where you will see a file called getSubXFieldsIRIRXYLMSP.py

e) View the file getSubXFieldsIRIXYLMSP.py in your favorite text editor.

In the section labelled “Variables to be modified by user”, modify the following:

outPath - change to the location you wish to download the data
varnames - change to the variables you wish to download
plevstrs - change to the variables/levels you wish to download (these must match with the varnames)
groups - change to the modelling group for data you wish to download
models - change to the models for data you wish to download (these must match with the group names)

As an example, to download the RSMAS, CCSM4, 2m temperature data, set the variables to the following:

```
varnames=['tas']
plevstrs=['2m']
groups=['RSMAS']
models=['CCSM4']
```

f) Run the program.

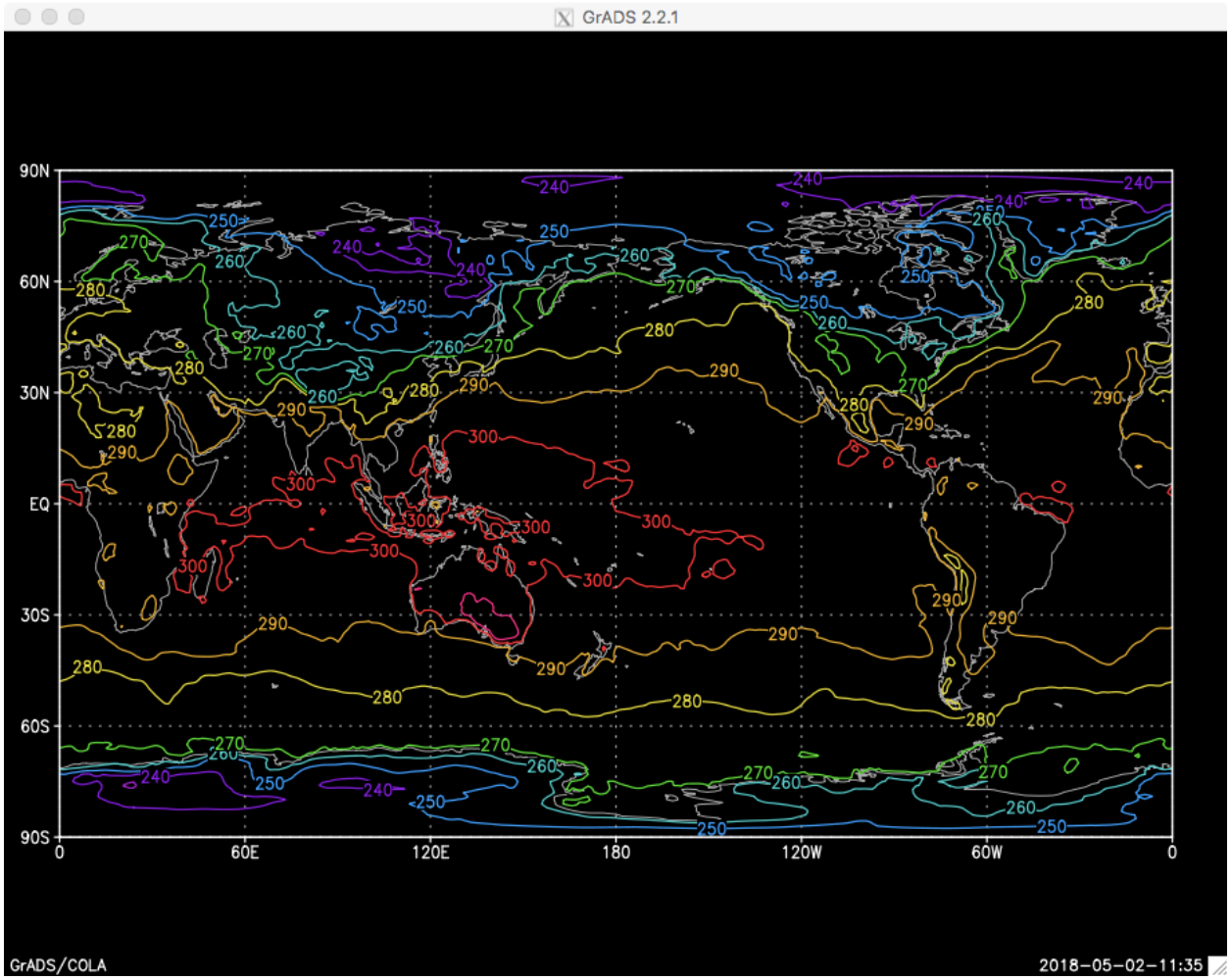
When you run the program, it may appear that nothing is happening. This is because the program is busy downloading the data. Since this is a large dataset, it will take a long time for the program to run.

g) Check to see that the data is being downloaded correctly. The data will be located in:

<outPath>/tas2m/daily/full/RSMAS-CCSM4/

```
[kpegion@atlas1 RSMAS-CCSM4]$ ls -la
total 31580
drwxr-xr-x 2 kpegion users       12 May  2 11:27 .
drwxr-xr-x 4 kpegion users        4 May  2 11:26 ..
-rw-r--r-- 1 kpegion users 11743938 May  2 11:26 tas_2m_RSMAS-CCSM4_19990107.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:26 tas_2m_RSMAS-CCSM4_19990108.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:26 tas_2m_RSMAS-CCSM4_19990109.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:26 tas_2m_RSMAS-CCSM4_19990110.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:26 tas_2m_RSMAS-CCSM4_19990111.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:26 tas_2m_RSMAS-CCSM4_19990112.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:27 tas_2m_RSMAS-CCSM4_19990113.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:27 tas_2m_RSMAS-CCSM4_19990114.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:27 tas_2m_RSMAS-CCSM4_19990115.e1.daily.nc
-rw-r--r-- 1 kpegion users 11743938 May  2 11:27 tas_2m_RSMAS-CCSM4_19990116.e1.daily.nc
```

h) Since the data are in netcdf, they are easy to view using your favorite method for plotting data. For example, viewing the data in tas_2m_RSMAS-CCSM4_19990107.e1.daily.nc using GrADS for the first lead time, produces the map below. I can confirm that these data are correct by comparing it with a map of the data in the IRIDL data viewer (see section 2).



i) Additional information

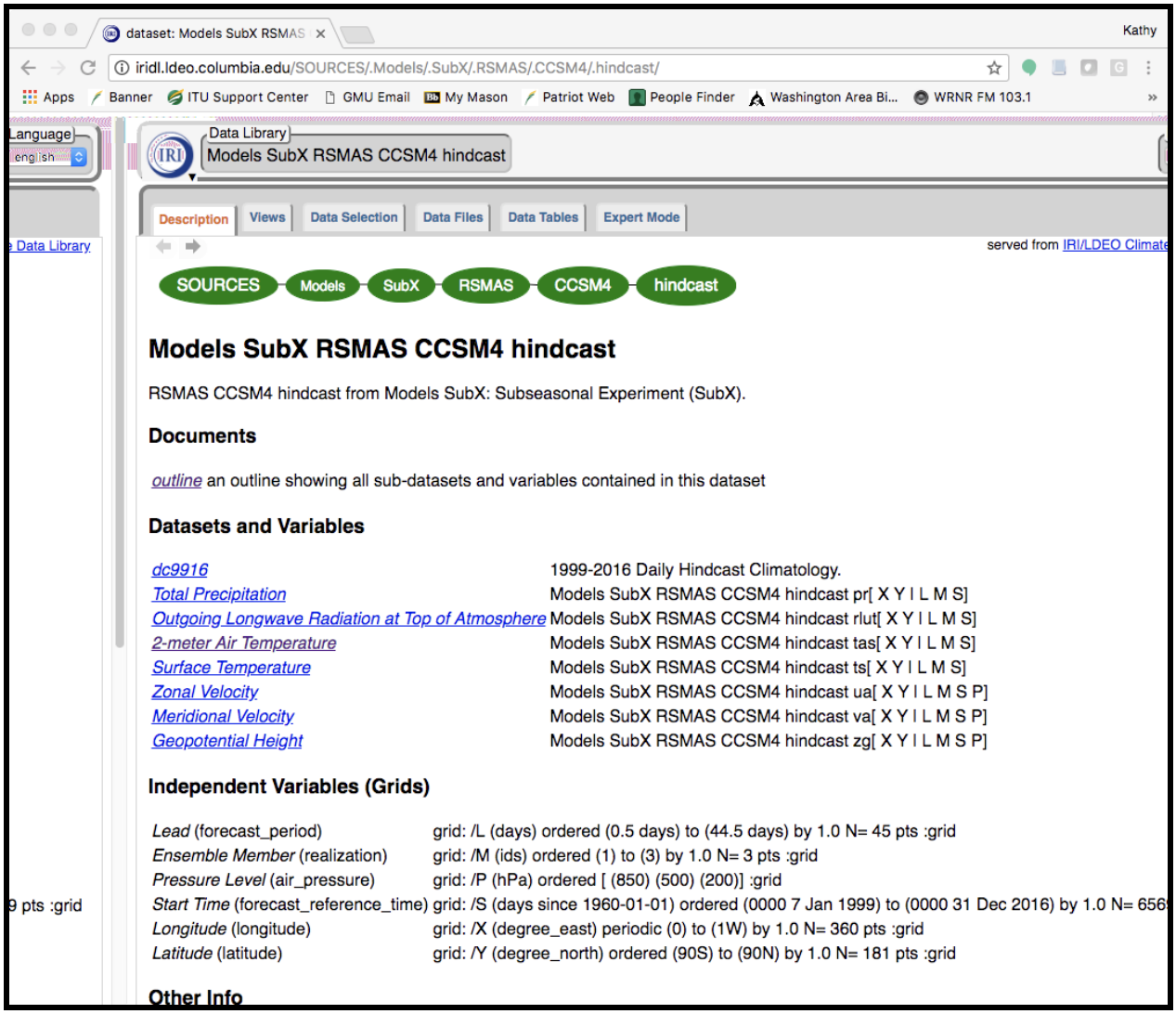
Data files will be downloaded for all start dates, this includes start dates for which there is all missing data. This means that some files will contain all missing data. As an example, the figure below shows the map for RSMAS-CCSM4, 2m Temperature for the file tas_2m_RSMAS-CCSM4_19990108.e1.daily.nc. As explained in section 2, RSMAS-CCSM4 does not have data for start dates on Jan 8, 1999. Therefore, a map of this data in GrADS looks like:

Chapter 2: Climatologies

Calculation of climatologies is necessary for producing forecast anomalies as well as for removing the model mean (systematic) biases due to model drift.

1. Getting climatologies from IRI Data Library

The climatologies for the 10 Priority 1 variables (http://cola.gmu.edu/kpegion/subx/data/priority1.html) are available from the IRI Data Library. Navigating to the IRI Data Library to the list of “Datasets and Variables” variable list for a given model will allow the user to see the available climatology data, which is called “dc9916” which stands for “Daily Climatology 1999-2016”. Clicking on “dc9916” will show the list of variables for which a daily climatology has already been computed. Codes for downloading climatology have not been provided, but it is relatively simple to modify the SubX download codes to download climatology data.



2. Calculating climatologies

If you wish to calculate climatologies for other variables not provided or want to understand how the climatology is calculated, the methodology is described below. Sample codes in Matlab are available at https://github.com/kpegion/SubX.

The methodology described here is used by NCEP/Climate Prediction Center for producing SubX real-time forecasts for use by CPC forecasters. This is also the methodology used for the real-time forecast anomaly maps posted on the SubX Website and for SubX model evaluation.

The methodology is applied to each model individually. The methodology is described for a given lon, lat point and can easily be applied to a full global field. It is applied to each lead-time separately to account for model drift.

a) *Calculate the ensemble mean for each forecast initialization*

The number of ensemble members (N) for each model from the hindcasts is shown in the table below.

Model	**# Ensemble Members**
ECCC-GEM	4
EMC-GEFS	11
ESRL-FIM	4
GMAO-GEFS	4
NRL-NESM	1
RSMAS-CCSM4	3
NCEP-CFSv2	4

b) For each initialization day of the year (1-366), average over all the years.

i. This produces what is called a “noisy” or “naive” climatology
Ii. Given the interval of hindcasts for different models, this may not produce a climatology for all days of the year for some models.

c) Apply a 31-day triangular smoothing to the noisy climatology.

The running mean is applied to the noisy climatology with a window of 31 days (+/- 15 days). It is applied to the noisy climatology in a periodic fashion such that Dec running means include Jan values and Jan running means include Dec values. A triangular window is applied such points in the window that are further away from the center point have less weight than those closest to the center point. The smoothing function applied here, called nanfastsmooth, has been optimized to run relatively quickly in Matlab. It was obtained from the Matlab Central database.

d) Lead-time dependence of Climatologies

Due to model drift from the initialized, observed state, to the model's own internal state, the climatology is expected to be different for different lead times. Therefore, the above procedure is performed for each lead-time separately.

Chapter 3: Verification Datasets

The verification datasets used for SubX Model evaluation are:

a) 2m Temp over land
CPC Global Daily Temperature (0.5x0.5)
Calculated as: $(T_{max}+T_{min})/2$
https://www.esrl.noaa.gov/psd/data/gridded/data.cpc.globaltemp.html

b) Precipitation over land
CPC Global Daily Precipitation (0.5x0.5)
https://www.esrl.noaa.gov/psd/data/gridded/data.cpc.globalprecip.html

c) Sea Surface Temperature
NOAA High Resolution Daily OISST (0.25x0.25)
https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.highres.html

d) Precipitation over Ocean
NASA Version 1.2 GPCP Daily Precipitation (1x1)
http://iridl.ldeo.columbia.edu/SOURCES/.NASA/.GPCP/.V1DD/.V1p2/.prcp/

e) Real-time Multivariate MJO Indices
from Matt Wheeler, Bureau of Meteorology, Australia
http://poama.bom.gov.au/project/maproom/RMM

Chapter 4: Skill Evaluation

1. Anomalies and Bias Correction

Anomalies are made by removing the lead dependent climatology (described in Chapter 2). This effectively performs a mean bias correction (see Becker et al. 2014). Currently, no further bias correction is applied to SubX re-forecasts.

2. How are weeks defined?

Evaluation of the retrospective forecasts uses two methods for defining the weeks. This is done because different models have different initial condition dates.

a) By Lead
Using this method each week is determined by each individual model's initial start date, meaning:
Week 1: Average of Days 1-7
Week 2: Average of Days 8-14
Week 3: Average of Days 15-21
Week 4: Average of Days 22-28
Because different models have different start dates, a multi-model ensemble cannot be produced using this method.

b) By Target
This method mimics the real-time forecast setup in which the start of the forecast period is Saturday and all forecasts initialized over the previous week are used (See Chapter 5 for real-time forecast details).

The algorithm is as follows:

- Identify all dates in the hindcast period (specified by startdate and enddate)
- Select the dates that are dow=7 (Saturdays)
- Identify the dates for the week prior to the given Sat (i.e. previous Sat through Fri)
- Loop through each Sat dates (referred to as basedate) for each model and determine if valid hindcast data exists this is done by checking if the entire field (lon,lat,lead) is NaN
- If a valid hindcast exists for a given day, determine how far it is from basedate and use this information to select the appropriate lead times for the given hindcast data to align its weeks 1-4 forecast target dates with the basedate forecast target dates.

Using this method, a multi-model ensemble can be calculated by averaging the models re-forecasts for a target date.