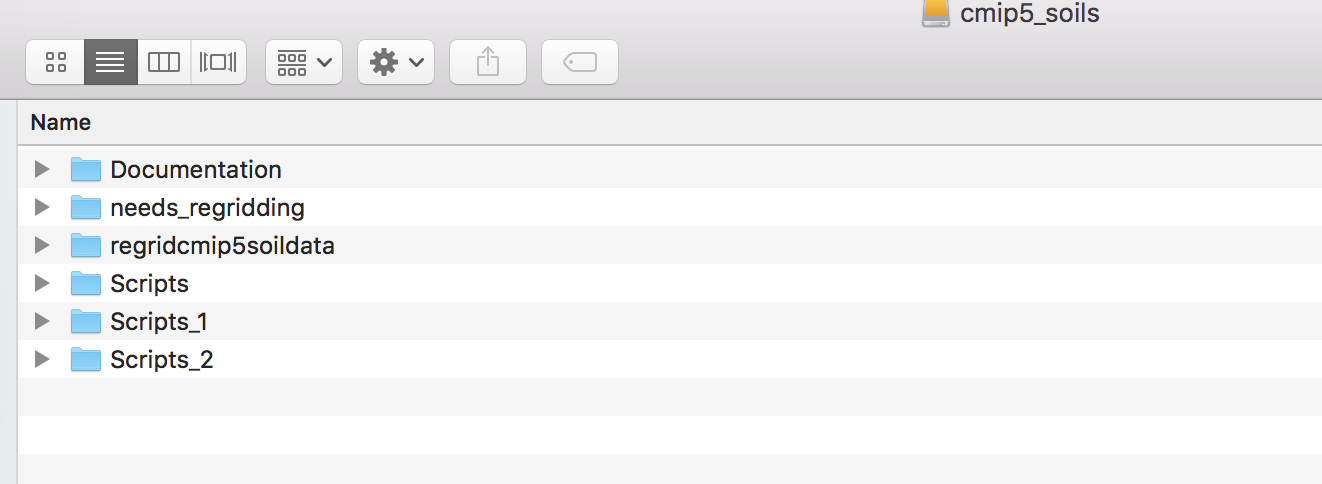
CMIP 5 Soils Processing Workflow

11/15/18

Contents

1. Data Drive Organization
2. Processing Workflow
   1. Scripts Overview
   2. NetCDF Processing Procedures
      1. Ocean Masking and Ice
      2. Soil Order Weights
3. Data Drive Organization

The hard drive is organized as follows:



The ‘Scripts’, ‘Scripts\_1’, and ‘Scripts\_2’ folders contain original scripts developed by Claire Phillips. The ‘regridcmip5soildata’ folder is where all of the regridded data and analysis are located.

Within the ‘regridcmip5soildata’ folder, there are several key folders. Regridded data is contained in the -historical, -rcp45, and -rcp85 folders for the variables tas, tsl, and mrlsl. Unused data files are located in the ‘tas’, ‘tsl’, and ‘mrlsl’ folders.

The ‘Scripts’ folder contains the updated bash and R scripts for processing regridded netcdf data, which will be described in the next section.

The ‘OUT3’ folder contains processed netcdf outputs. Finally, ‘Analysis’ is where processed CSV data and figures are located. More details on the processing workflow

1. Processing Workflow
2. Scripts Overview

Beginning with regridded files, the data processing pipeline proceeds as follows:

1. Process CMIP5 Net CDF files, by model (Bash script)

This is done by running the shell script **CMIP5Analysis\_clean\_101318.sh**, located in “regridcmip5soildata/Scripts”. This script uses the Climate Data Operators language and may be modified to run one model by setting the “index” field to a value between 0-13, which will run the corresponding model in the “MODEL” array. To run multiple models at once, leave a space between values; e.g. “index=(0 1)”. This script performs essential calculations on the regridded NetCDF files, including calculating global temperature and moisture anomalies, historic soil physics, and soil order-specific computations.

Outputs are located in the “regridcmip5soildata/OUT3” folder, but you may change the output location by changing the “OUT” variable to a different folder. Note that you will need to populate whatever output folder you choose with sub-folders corresponding to each CMIP5 model.

1. Process Net CDF files to Rdata files (R)

This step consolidates netCDF files into .Rdata files and performs additional calculations such as interpolation. It is performed by running the file **ProcessBashOutfiles\_10.30.18.r**, also located in the “regridcmip5soildata/Scripts” folder. Outputs are in the form of .Rdata intermediate files, located in “regridcmip5soildata/Analysis/RData”. The processing script is somewhat truncated from the original, which is located in “Scripts\_1/ProcessBashObservations.R” and contains Claire’s original processing code.

1. Create figures and CSV files

Current reproduced figures may be created using “ProduceEnsembleFigures.Rmd”, located in the “regridcmip5soildata/Scripts” folder. Output figures are produced in “regridcmip5soildata/Analysis/Figures”.

These figures are also somewhat truncated compared to the final paper figures used in Claire’s original analysis, and updating this script with additional figures is a priority going forward. CSV files are also produced in this step and are saved to the “regridcmip5soils/Analysis/csvs” folder. Figures that we are currently able to replicate include:

Figure 3A-D: Ensemble predictions for soil temperature and moisture changes for 12 soil orders

Figure 4A-B: Soil temperature and moisture changes through time, RCP 8.5

1. NetCDF Processing Procedures

This section describes operations performed in the “regridcmip5soildata/Scripts/CMIP5Analysis\_clean\_101318.sh” script, as well as additional operations performed in the “regridcmip5soildata/soil-wts” folder.

* 1. Ocean and Ice Masking Procedures

Objective: To remove ocean and permanently ice-covered land area from global average estimates of surface temperature, soil temperature, and soil moisture anomalies, among other metrics.

*Original Method (Developed by Claire Phillips)*

The original method uses a 'presence-absence' masking method, meaning that if the grid cell contains at least some land it has a value of 1, and a value of 0 otherwise (indicating ocean). All land-containing grid cells are then used in subsequent calculations.

This method results in an overrepresentation of grid cells along the coastline that contain a percentage of land and ocean and should have a smaller contribution to land area weighted averages.

*Updated Method (Developed by Catherine Ledna)*

Use a masking method that assigns a percentage (varying from 0 to 1) of land area to each grid cell, allowing land area per grid cell to be calculated. In subsequent calculations, each grid cell will be weighted according to its land area. This method may be extended to exclude ice-covered land regions. Relevant files are located in “regridcmip5data/area-wts” and are described below:

**Land mask file:**

Model: CCSM4

Variable: sftlf (definition here: <http://nomads.gfdl.noaa.gov:8080/DataPortal/ipcc5_vars.jsp?bundle_id=fx>)

Grid: Atmospheric

Ensemble: r0i0p0;

Period: Historical (fx)

Source: <https://www.earthsystemgrid.org/dataset/cmip5.output1.NCAR.CCSM4.historical.fx.atmos.fx.r0i0p0/file.html>

**Land-Ice Mask File:**

Model: CCSM4

Variable: sftgif (definition here: <http://nomads.gfdl.noaa.gov:8080/DataPortal/ipcc5_vars.jsp?bundle_id=fx>)

Grid: Atmospheric

Ensemble: r0i0p0;

Period: Last millennia (fx)

Source: <https://www.earthsystemgrid.org/dataset/cmip5.output1.NCAR.CCSM4.past1000.fx.land.fx.r0i0p0/file.html>

**Cell Area File:**

Model: CCSM4

Variable: areacella (Atmosphere Grid-Cell Area)

Grid: Atmospheric

Ensemble: r0i0p0

Period: historical (fx)

Source: <https://www.earthsystemgrid.org/dataset/cmip5.output1.NCAR.CCSM4.historical.fx.atmos.fx.r0i0p0/file.html>

Calculation of land area for grid cell *i*:

Where landfrac refers to the sftlf variable taken from the Land Mask file, and cellarea refers to the areacella variable taken from the Cell Area File.

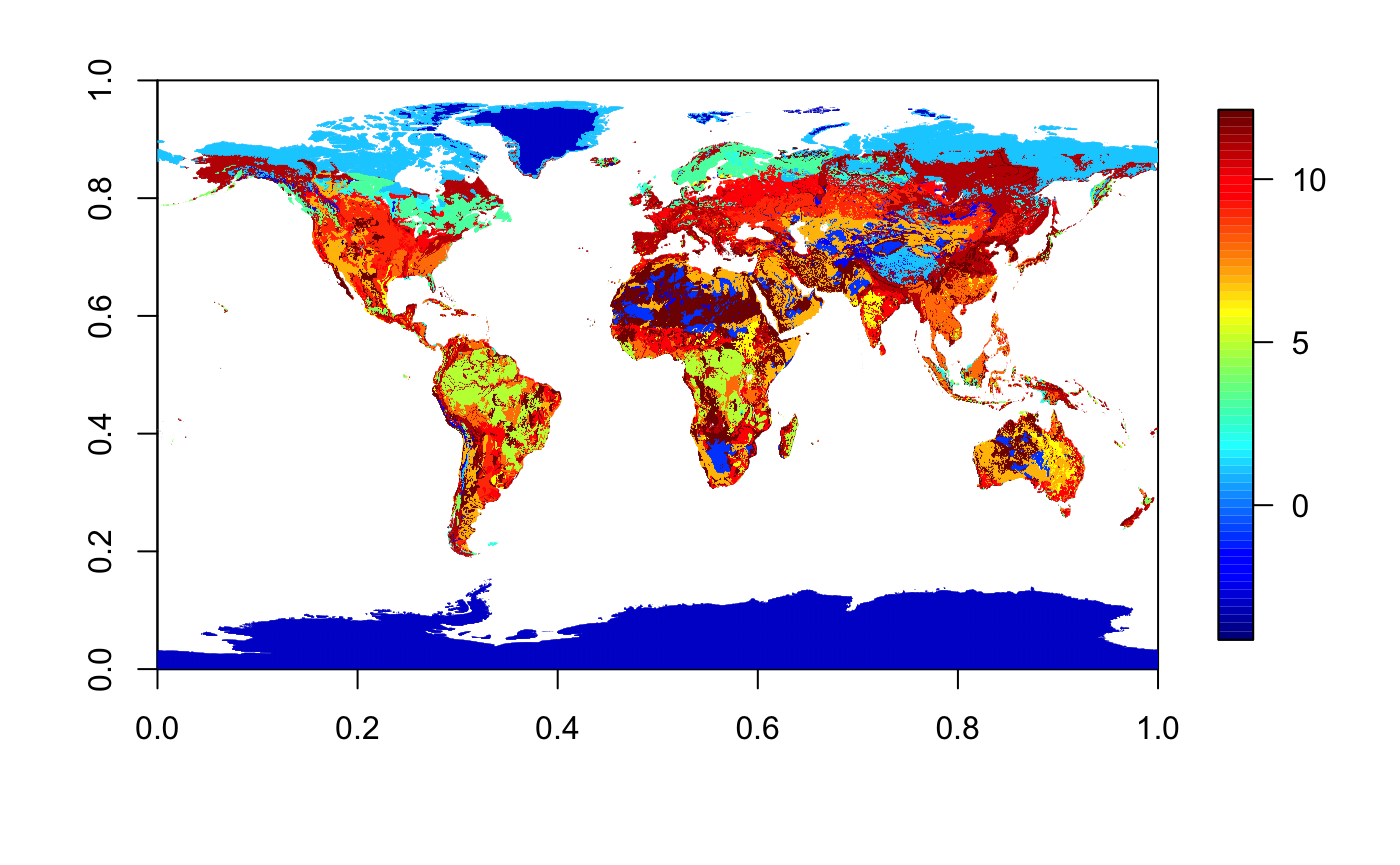
Calculation for global area-weighted value at time *t,* for grid cells 1…N: (substitute area\_i for landareai, soilareai, etc…)

* 1. Soil Order Weights

This section describes operations used to create the soil order masks (“soil-wts” folder) and the computation of weighted soil-order calculations (“CMIP5Analysis\_clean\_101318.sh).

*Creation of Soil Order Masks*

All relevant data and scripts are located in the “soil-wts” folder. Soil order masks were downloaded from the Natural Resources Conservation Service (NRCS) at 0.33 degree resolution. The NRCS map gave a range of 16 soil orders (ranging from a value of -4 to 10 on the map below).



This map was then split out into twelve soil orders, giving each map a value of 0 or 1 for presence or absence of the soil on land, using the script make\_soil\_order\_maps\_oceanmask\_cl.py (originally developed by Charlie Koven and slightly modified to fix a missing value error). This script scales the original NRCS map from 10800x5400 (0.33 degrees) to 1080x540 in order to make the files easier to regrid. This is done by taking an average of 100 neighboring grid cells (10x10) to produce a new grid value. Area weighting is not taken into account; cells in each neighborhood are assumed to have the same area. Because the grid resolution is so fine and the cells being averaged are neighbors, this is presumably a safe assumption as the total variation between neighboring grid cell areas will be quite small.

With a new resolution of 1080x540, the updated grids are regridded to the CCSM4 grid of 288x192 using the ncremap tool in nco (regrid\_soil\_order.sh). The command used is as follows:

* 1. ncremap -i ifile -r 0.1 -g gridfile -o outfile

The -r 0.1 flag requests renormalization of values with a weight threshold of above 10 percent – e.g. destination grid cells with at least 10% of their area contributed by valid source grid cells will contain valid (not missing) values that are the area-weighted mean of the valid source values. In this process, the original higher-resolution grid cells were averaged into a new value for the grid cell, creating a fraction ranging from 0 to 1.

*Soil Order Weights*

Soil order weights (files with the \_regrid.nc designation in the “soil-wts” folder) are used to calculate the fraction of area in each grid cell corresponding to each soil order. Soil order weights are computed as a fraction of total cell area, with no land-ocean distinction for coastal cells. For mixed land and ocean cells, this means that the soil area fraction will be a percentage of cell area, not a fraction of land area. Therefore, we calculate soil area for grid cell *i* as:

When performing soil order-specific calculations, the weighting procedure follows the same method as land area-weighted values.