Greenland Toolkit Readme

Matt Parkan
matthew.parkan@gmail.com
April 11, 2013

Software requirements

- the latest release of R, available at http://www.r-project.org/
- the latest release of RStudio (IDE for R), available at http://www.rstudio.com/
- a text editor, e.g. Notepad++ (http://notepad-plus-plus.org/), SciTE (http://www.scintilla.org/SciTE.html)
- a ftp client, e.g. Filezilla (https://filezilla-project.org/), CyberDuck (http://cyberduck.ch/)
- a desktop GIS, e.g. Quantum GIS (http://www.qgis.org/), uDig (http://udig.refractions.net/), OpenJump (http://www.openjump.org/)

Workflow overview

Part 1 - Downloading, organizing and preparing the data sets for use with R

1. Prepare the raw data folder layout.
2. Download all the data sets and place them in their specific folders.

Part 2 - Feature construction

1. Prepare the processed data folder layout.
2. Run '*extract_ephemeris.r*' to extract daily daylight duration at selected locations (detailed instructions are provided in the script).
3. Run '*extract_weather.r*' to extract daily weather features at selected locations (detailed instructions are provided in the script).
4. Run '*extract_avhrr.r*' to extract daily remote sensing features at selected locations (detailed instructions are provided in the script).
5. Run '*merge_features.r*' to merge ephemeris, weather and remote sensing features (detailed instructions are provided in the script).

Part 3 - Data inventory and pre-analysis

1. Run ‘*locate_stations.r*’ to check weather data availability and compute distance matrix.
2. Run ‘*map_stations.r*’ to plot a map of weather stations in Greenland.

Part 4 - Modelling and prediction

1. Run ‘predict_self.r’ to build training/test sets, create SVR and linear regression models and make predictions.

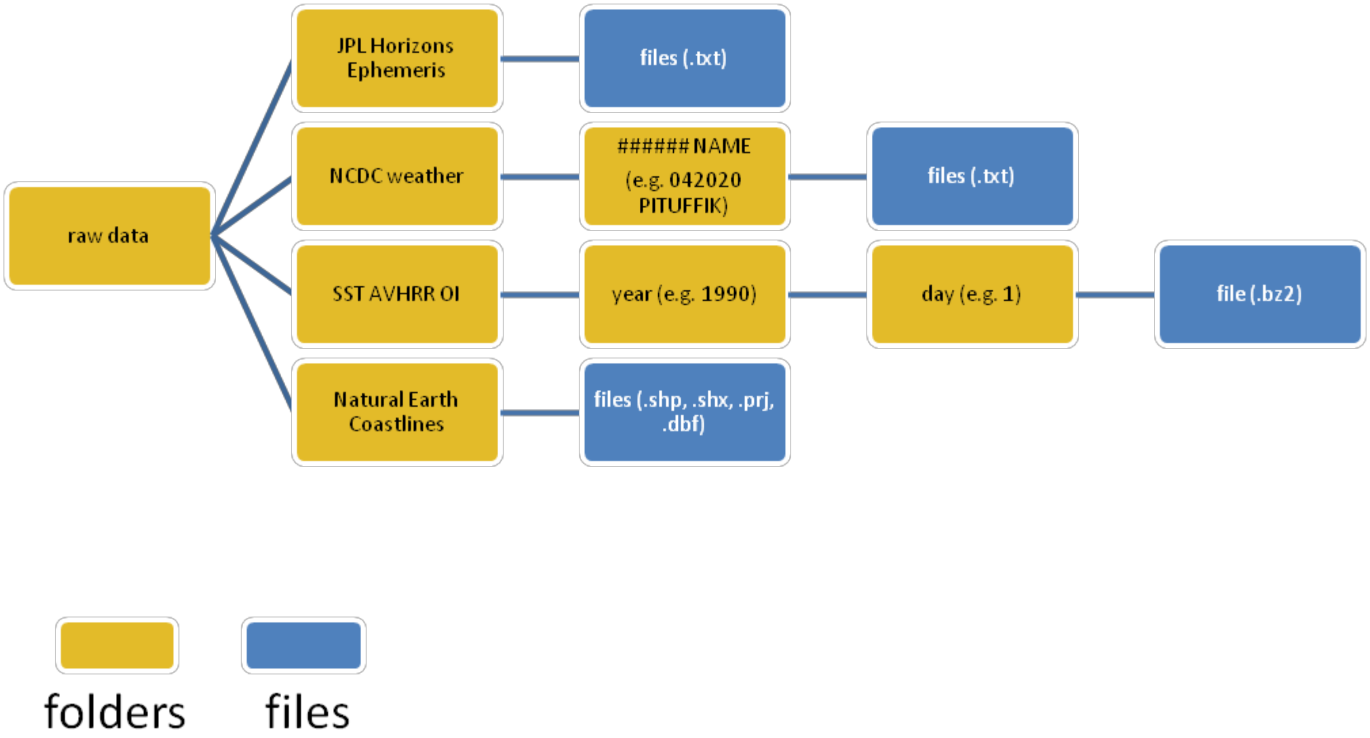
2. Run ‘predict_neighbour.r’ to build training/test sets, create SVR and linear regression models and make predictions of other stations.

Workflow details

Part 1 - Downloading, organizing and preparing the data sets for use with R

Preparing folders to receive the raw data

The folder layout shown below is used to archive the raw data. Before starting the download you should create the folder layout up to the first level of arborescence (i.e. JPL Horizons Ephemeris, Natural Earth Coastlines, NCDC weather, SST AVHRR OI). Each time you download NCDC weather data, you should place the files in a subfolder named with the 6 digit USAF number of the station and its name (e.g. 042020 PITUFFIK).
SST AVHRR OI data is organized by year and day of year subfolders on the source ftp server, so you just need to transfer them to the SST AVHRR OI folder on your computer.



Downloading NCDC weather station data

1. Go to http://www.ncdc.noaa.gov/cdo-web/

2. There are several ways to access the sub-daily weather data:

-*map search*, allows you to search for data via an interactive map.

-*data search*, allows you to search for data by station name and product.

-*ftp & Legacy access*, allows you to search for data by geographic region, country, station ID name. It also allows direct ftp access.

Only the *ftp & Legacy access* is covered here (n.b. most of the download steps are similar when using the map search).

3. In the *ftp & Legacy access* tab select *Surface Data, Global Hourly*.

4. Read the *WMO Resolution 40 / NOAA Policy* about data usage and click on *continue*.

5. Use the Advanced options, click on *Continue with ADVANCED OPTIONS*.

6. Specify the following options:

-Retrieve data for Country, *Greenland*

-Output format, *Text file (user selects elements of choice)*

Then, click on *Continue*.

7. Specify the following option:

-Retrieve data for *Selected Greenland Stations*.

Then, click on *Continue*.

8. Select the station you want (to limit file size **SELECT A SINGLE STATION**) and click *Continue*.

9. Select **ONLY** the following data elements:

-Air temperature observation
-Air temperature observation dewpoint
-Atmospheric Pressure observation
-Relative humidity Calculation
-Sky Condition observation
-Visibility observation
-Wind direction
-Wind observation

Then, click on *Continue*.

10. Specify the date range and the output format, use the following settings:
date range: 1978-01-01 to most current date
output format: *Delimited, without station name*
output format delimiter: *Comma*
Then, click on *Continue*.

11. Check the *Inventory review* tick box and specify your email address. Click on *Submit Request*.

12. After a while, you should receive an email with links to download the data, inventory, station information and format documentation files. Download all four files (**WITHOUT RENAMING THEM**) to a subfolder in the *NCDC weather* folder on your computer. The subfolder name must start with the USAF station number (e.g. 042020 PITUFFIK).

Downloading JPL OISST data

1. Go to http://podaac.jpl.nasa.gov/dataset/NCDC-L4LRblend-GLOB-AVHRR_OI
2. The data can be accessed by remote ftp, OPENDAP and ftp. Only the ftp access is covered here.
3. Navigate to the data access tab and copy the ftp server address.
4. You will need a ftp client to bulk download the data (e.g. *Filezilla*).
5. In your ftp client, paste the ftp server address in the host field.
6. Transfer all the desired data to the *SST AVHRR OI* folder on your computer.

Downloading Natural Earth Coastlines

1. Go to http://www.naturalearthdata.com/downloads/
2. Download the following files:
   1:10m Physical Vectors - Coastline
   1:10m Physical Vectors - Land
   1:50m Physical Vectors - Coastline
   1:50m Physical Vectors - Land
3. Place them in the '*Natural Earth Coastlines*' folder.

Downloading JPL ephemeris data

1. Go to http://ssd.jpl.nasa.gov/?horizons

2. There are several ways to access the data. We will only cover the *horizon* email system.

3. Create the automatic request form using the template provided below.

IMPORTANT: modify the values of EMAIL_ADDR, SITE_COORD, START_TIME, STOP_TIME according to your needs

```
!$$SOF
EMAIL_ADDR = 'your@email.com'
COMMAND= '10'
CENTER= 'coord@399'
COORD_TYPE= 'GEODETIC'
SITE_COORD= '-37.633,65.6,0'
MAKE_EPHEM= 'YES'
TABLE_TYPE= 'OBSERVER'
START_TIME= '1982-01-01'
STOP_TIME= '2013-04-18'
STEP_SIZE= '1 m'
CAL_FORMAT= 'CAL'
TIME_DIGITS= 'MINUTES'
ANG_FORMAT= 'HMS'
OUT_UNITS= 'KM-S'
RANGE_UNITS= 'AU'
APPARENT= 'AIRLESS'
SOLAR_ELONG= '0,180'
SUPPRESS_RANGE_RATE= 'NO'
SKIP_DAYLT= 'NO'
EXTRA_PREC= 'NO'
R_T_S_ONLY= 'TVH'
REF_SYSTEM= 'J2000'
CSV_FORMAT= 'YES'
OBJ_DATA= 'NO'
QUANTITIES= '1,7'
!$$EOF
```

4.Send the request by email to horizons@ssd.jpl.nasa.gov

IMPORTANT: the subject MUST contain the word *JOB* (e.g. *Greenland JOB*)

5.You should receive a response email containing the requested data shortly after.

6.Save the entire email as a text (.txt) file. The file name must start with the USAF station number (e.g. 042020_ephemeris_pituffik.txt)

7.Copy only the data contents of the text file to a comma separated file (.csv). The file name must start with the USAF station number (e.g. 042020_ephemeris_pituffik.csv)

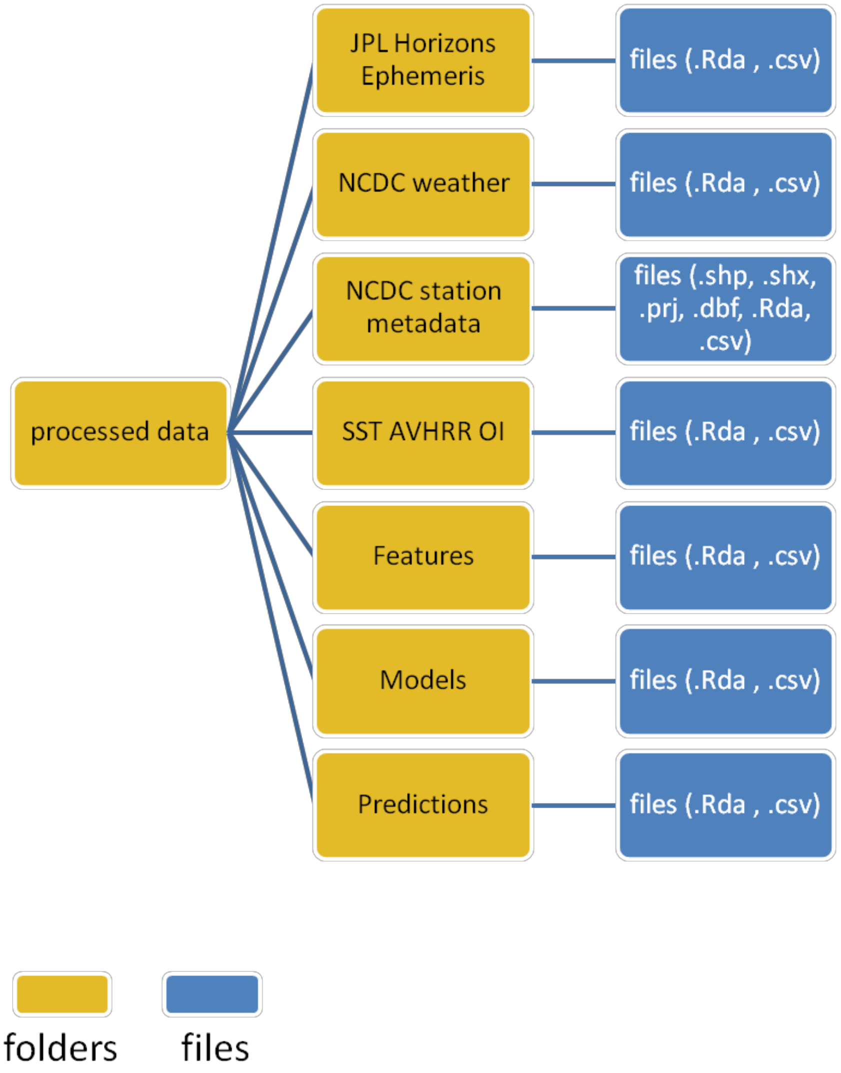
8.The content of the .csv file should have the following form:

```
1982-Jan-01 12:50,*,r,18 48 03.63,-22 59 04.4, 17 02 50.9817,
1982-Jan-01 14:35,*,t,18 48 22.82,-22 58 42.3, 18 48 08.2302,
1982-Jan-01 16:20,C,s,18 48 42.00,-22 58 19.8, 20 33 25.4786,
.........etc
```

Part 2 - Feature construction

Preparing folders to receive the processed data

Before processing the raw data you should setup the folder layout below. The *JPL Horizons Ephemeris*, *NCDC Weather* and *SST AVHRR OI* subfolders will contain the processed data. *NCDC station metadata* will contain data about weather station locations. *Features* contains the merged predictor features (i.e. ephemeris, weather, sea surface temperature and sea ice statistics). Models will contain linear regression and support vector regression predictive models. Predictions contains the model performance files.



Reading the JPL ephemeris with ‘*extract_ephemeris.r*’

Description:
The script computes daily daylight duration from JPL HORIZON ephemeris and saves the result as .csv and .Rda files in the output directory.

Usage:

1. Install all required libraries
2. Define working directories (input and output paths)
3. Define USAF number of desired stations (check station inventory)
4. Run

Reading the NCDC weather data with ‘*extract_weather.r*’

Description:
The script reads NCDC weather data, creates and exports weather features to .csv and .Rda files.

Usage:
1. Install all required libraries
2. Define working directories (input and output paths)
3. Define USAF number of desired stations (check station inventory), you may also process all stations in the input directory.
4. Specify the minimum number of observations threshold (features will not be created if the number of observations is below this threshold)
5. Specify the date range
6. Specify number of lag days for each variable
7. Run

Reading the OISST data with ‘*extract_avhrr.r*’

Description:
The script reads OISST netcdf files and extracts area statistics in circular buffers around specific locations (i.e. weather stations). It then saves the result as .csv and .Rda files in the output directory.

Usage:
1. Install all required libraries
2. Define working directories (input and output paths)
3. Define path to station metadata file (.Rda)
4. Define path to quadrangle area file (.Rda)
5. Define global output file name (without extension)
6. Define USAF number of desired stations (check station inventory)
7. Define circular buffer ranges (in km) around stations used to compute area statistics
8. Specify number of lag days for each variable-range couple
9. Indicate if buffer polygons should be (re)computed (TRUE=yes, FALSE=no)
10. Indicate if approximate area quadrangles should be (re)computed (TRUE=yes, FALSE=no)
11. Indicate if raster values should be extracted (TRUE=yes, FALSE=no)
12. Run

Merging features with ‘*merge_features.r*’

Description:
The script merges ephemeris, weather and remote sensing features into a single data frame and saves the result as .csv and .Rda files in the output directory.

Usage:
1. Define input directories
2. Define output directory
3. Define USAF number of desired stations (check station inventory)
4. Run

Part 3 - Data inventory and pre-analysis

Checking weather data availability and computing distance matrix with ‘locate_stations.r’

Description:
The script computes inter-station distance matrix and a creates a metadata file. It then saves the result as .csv and .Rda files in the output directory.

Usage:
1. Install all required libraries
2. Define working directories
3. Define inventory file name
4. Define output file names
5. Run

Plotting a map of the stations and circular buffers with ‘plot_station_map.r’

Description:
The script plots a map of weather stations and buffers in Polar Stereographic projection. It then saves the result as .eps, .pdf and .png files in the figure directory.

Usage:
1. Install all required libraries
2. Define path to Natural Earth data
3. Define path to station metadata file (.Rda)
4. Define path to folder for saving figures
5. Define USAF number of desired stations (check station inventory)
6. Define circular buffer ranges (in km) around stations
7. Run

Part 4 - Modelling and prediction

Predicting temperature at station with ‘predict_self.r’

Description: The script builds training/test sets, create SVR and linear regression models and make predictions.

Usage:
1. Install all required libraries
2. Define input directory
3. Define output directory
4. Define USAF number of desired stations (check feature availability before)
5. Define which month(s) should be predicted
6. Indicate how many runs should be performed
7. Indicate if support vector regression (epsilon) should be performed? (TRUE= yes, FALSE=no)
8. Indicate if linear regression should be performed? (TRUE= yes, FALSE=no)
9. Run

Predicting temperature at another station with ‘predict_neighbour.r’

Description: The script builds training/test sets, create SVR and linear regression models and make predictions on a different station.

Usage:
1. Install all required libraries
2. Define input directories (features and weather)
3. Define output directory
4. Specify USAF number of stations for which temperatures will be predicted
5. Specify USAF number of stations used as a predictor
6. Define which month(s) should be predicted
7. Indicate how many runs should be performed
8. Indicate if support vector regression (epsilon) should be performed? (TRUE= yes, FALSE=no)
9. Indicate if linear regression should be performed? (TRUE= yes, FALSE=no)
10. Run