Climate Reference Metadata.

1. Introduction

Metadata was collected from a variety of sources for the Climate Reference Network (CRN) stations. The latitude and longitude datum for the stations was used to reference various GIS datasets to extract information related to the topography of the stations, the population surrounding the stations and the land cover, including data pertaining to the urbanity of the sites. The Latitude and Longitude of the stations is reported to 4 significant digits which indicates a precision of close to 10 meters at the equator. The accuracy of these datum in conjunction with the GIS data will tend to be less than this as it is dependent on the registration accuracy of the GIS data and its inherent resolution. Unless otherwise stated all projections are assumed to be WGS84, a standard Lat/lon projection. The code for processing the metadata is located at <https://github.com/stevenmosher/Ghcn_V4_Metadata> and the script for processing the data is CRNmetadata.R. Where feasible the data sources are located in github, otherwise their URL is given in the references. All data sources are open; however, some require registration to acquire. Consequently, there is no code to automate the download of the data and so the scripts do not run turnkey. The user is expected to download the data for themselves and alter the “filenames.R’ source code accordingly.

1. Filename and format.

The filename is CRN\_Metadata\_Final.csv. It is a comma separated file. It is available here: https://github.com/stevenmosher/Ghcn\_V4\_Metadata/blob/master/CRN\_Metadata\_Final.csv

1. Column names
   1. Station\_ID

Station\_ID is a unique station identifier take from the site inventory. It is an integer value that represents the WBAN of the station for CRN stations.

* 1. Name

The station name is an ascci record of the station name with spaces replaces by underlines, for example; “Murphy\_10\_W”

* 1. Longitude

The station’s longitude in degrees to 4 significant digits. The projection is WGS84

* 1. Latitude

The station’s latitude in degrees to 4 significant digits. The projection is WGS84

* 1. Elevation

The station’s elevation in meters taken from the station inventory

* 1. DEM1km

In addition to the elevation at the station’s exact latitude and longitude, the average elevation in meters surrounding the station is supplied from a Digital Elevation Model (DEM). The DEM data is supplied for cases where the elevation data may be missing, as is the case with a small percentage of GHCN stations. In addition, the DEM data may indicate that the station is in an area where the surrounding terrain is at a higher elevation and could be subject to cold air drainage. Finally, since the data includes Bathymetry the data can be used to identify stations that are located on ships at sea, oil platforms and other ‘land’ stations that may actually be located in the ocean. The DEM data source –ETOP01-- is described here: <https://www.ngdc.noaa.gov/mgg/global/relief/>. The data source has both elevation and bathymetry at a 1 arc minute resolution, or roughly 2km at the equator. Over land the elevation supplied is the elevation at the top of the ice sheet, which is significant for those stations in Greenland and Antarctica. URL <https://www.ngdc.noaa.gov/mgg/global/relief/ETOPO1/data/ice_surface/cell_registered/georeferenced_tiff/ETOPO1_Ice_c_geotiff.zip>.

* 1. DistancetoCoast

Distance to coast represents the distance of the station to a continental coast. The units is kilometers. Negative values represent a station that is inland and positive values a station that is located in the ocean. Stations can be located in the ocean for several reasons: A) slight inaccuracies in the stated location. B) stations located on small islands or atols not in the coastline database. C) stations located on ships or platforms. For CRN stations none of these conditions exist. The dataset is a WGS84 grid spanning -180 to 180 and -90 to 90. The cellsize is .01 degrees or roughly 1km at the equator. The dataset is described here: <https://oceancolor.gsfc.nasa.gov/cms/DOCS/DistFromCoast>

* 1. LCCOwnLabel

The column contains the landcover for the station lat lon as determined by a lookup into the ESA Climate Record Data Product (CRDP) built from SPOT and MERIS data. The time period covered is the 2008-2012 epoch. The data is in a WGS84 projection with a resolution of 300 meters at the equator or 10 arc minutes. The data can be obtained after registration here

<http://maps.elie.ucl.ac.be/CCI/viewer/download.php>. Overall accuracy of the classification is ca. 70%, however for the most relevant classes (Urban, Water, Croplands) User accuracies exceed 85%. Validation of the classification described here: <http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-PUG-v2.5.pdf>.

* 1. EF\_LF\_Desc

This column contains a description of the topology of the site. The landform data was extracted from the Global USGS dataset using the lat and lon the site. The dataset is described here:

<https://rmgsc.cr.usgs.gov/ecosystems/global.shtml> and more detail here

<https://rmgsc.cr.usgs.gov/ecosystems/docs/AAG_Global_Ecosystems_Booklet.pdf>. The dataset covers the entire world in a WGS84 format at a 250 meter resolution. The landforms are classified based on the underlying DEM data. They include:

Flat Plains, Smooth Plains, Irregular Plains, Escarpments, Low Hills, Hills , Breaks, Low Mountains, High Mountains/Deep Canyons, Surface Water. The data can be downloaded here: https://rmgsc.cr.usgs.gov/outgoing/ecosystems/Global/

* 1. WaterArea,

The water area is the total sq kilometers of water within a 10km radius of the site’s latitude and longitude. The ESA Landcover product (see 3.8) is used to determine the amount of area that is water surface within 10km of the site. Water area is provided for the following reasons. Presence of water bodies is known to diminish the UHI effect and it can be used to identify stations that are located on ships and platforms in certain cases. The product used here is the same as for item 3.8. From every location a 10km buffer is extracted and the area of the cells that are classified as water is summed. The maximum figure is roughly 314sq km.

* 1. UrbanArea10K

The Urban area is the total sq kilometers of urban landcover within a 10km radius of the site’s latitude and longitude. The ESA Landcover product ( see 3.8) is used to determine the amount of area that is urban landcover within 10km of the site. From every location a 10km buffer is interrogated and the area of the cells that are classified as urban is summed. The maximum figure is roughly 314sq km.

* 1. GPwV4\_Area

Population figures for each site are retrieved from a 30 arc minute (1km) gridded dataset. Since the population count is retrieved the population density will be a function of the area of the cell. Since area of the cell changes with latitude a record of the cell area at each site location is recorded. This provides the necessary data for calculating a population density on per sqkm basis. The population dataset is described in 3:13

* 1. GPwV4\_00

Population counts for the site location are extracted from the gridded GWP version 4 dataset. This column represents the UN adjusted population count in the year 2000. The dataset has a 30 arc minute resolution (1km at the equator) and spans from -60 lat to 85N. For locations not over land and outside the coverage area, NA ( Not Available) is recorded. At this stage we do not impute a zero population for those locations and that decision is left to the user. The methodology is described here <http://beta.sedac.ciesin.columbia.edu/data/collection/gpw-v4> The data uses areal-weighting and maintain fidelity to the input census data. So for example it does not take into consideration the landcover of the pixel or other factors such as water area or slope that other more highly modelled products use. The dataset is available here <http://beta.sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-rev10> and can be downloaded after registration. The projection is WGS84.

* 1. GPwV4\_05

Population counts for the year 2005 from GPW V4

* 1. GPwV4\_10

Population counts for the year 2010 from GPW V4

* 1. GPwV4\_15

Population counts for the year 2015 from GPW V4

* 1. GPWV4\_Area10

The column is a constant representing the approximate area of a circle with a radius of 10km from the site location. The units are sq km

* 1. GPwV4\_15\_10km

The population counts around every site are aggregated out to a radius of 10km.

* 1. Hyde\_Area

In addition to the GPW v4 data we also provide population counts from Hyde3.1 datasets. The Hyde dataset has population counts every decade back to the beginning of the recording of temperatures. The gridded data is provided in a 5 arc minute resolution ca. 9km at the equator. The approach to estimating population is covered in the paper ( see 4.3) and the briefly described here: <http://themasites.pbl.nl/tridion/en/themasites/hyde/basicdrivingfactors/population/index-2.html>. The benefit of Hyde is that it allows us to classify stations that have been discontinued before the time period covered by GPW. The datasets are available here <ftp://ftp.pbl.nl/hyde/hyde3.1/>

* 1. Hyde1970

Population count for the cell where the site is located in 1970

* 1. Hyde1980

Population count for the cell where the site is located in 1980

* 1. Hyde1990

Population count for the cell where the site is located in 1990

* 1. Hyde2000

Population count for the cell where the site is located in 2000

* 1. Hyde2005

Population count for the cell where the site is located in 2005

* 1. GpwV4\_density00

The population density of the site in people per sqkm in 2000. The area of the cell is used to create the density

* 1. Hyde\_density00

The population density of the site in people per sqkm in 2000. The area of the cell is used to create the density

* 1. GPW10km\_15\_Density

The population density for an area within 10km of the site.

* 1. EST\_POP2000

The population of the nearest populated place in the year 2000. The data source used here is the GRUMP settlement points. For every site location the distance to all populated places is created and then the closest populated place is selected. GRUMP settlement points. The settlement points consists of a database of urban locations with populations greater than 5000. It is described here <http://sedac.ciesin.columbia.edu/data/set/grump-v1-settlement-points>. The dataset is available here <http://sedac.ciesin.columbia.edu/data/set/grump-v1-settlement-points/data-download>

* 1. EST\_POP2000\_50K

The population of the closest urban location with a population count greater than or equal to 50000

* 1. EST\_POP2000\_500K
  2. EST\_POP2000\_1M
  3. EST\_POP2000\_5M
  4. DistanceToPlace

* 1. DistanceToPlace50K

Distance in kilometers to the closest settlement with a population of 50000 or greater

* 1. DistanceToPlace500K

Distance in kilometers to the closest settlement with a population of 500K or greater

* 1. DistanceToPlace1M

Distance in kilometers to the closest settlement with a population of 1M or greater

* 1. DistanceToPlace5M

Distance in kilometers to the closest settlement with a population of 5M or greater

* 1. PopulatedPlace

Name of the closest populated place

* 1. PopulatedPlace50K

Name of the closest populated place with a population of 50,000 or greater

* 1. PopulatedPlace500K

Name of the closest populated place with a population of 500,000 or greater

* 1. PopulatedPlace1M

Name of the closest populated place with a population of 1M or greater

* 1. PopulatedPlace5M

Name of the closest populated place with a population of 5M or greater

* 1. Airport\_Dist

For every station the distance to the closest airport is calculated. Airport locations are taken from the open source airport database. The data is cleaned to remove closed airports, sea air ports, heliports and balloon ports. Small, medium and large airports are included. Sample google earth views of each type are in the appendix. Data was acquired from here. http://ourairports.com/data/airports.csv

* 1. Airport\_Dist2

Distance in km to to the closest medium or large Airport is recorded

* 1. Airport\_Name

Name of the closest airport

* 1. Airport\_Name2

Name of the closest medium or large airport

* 1. Airport\_Type

Record of whether the airport is small medium or large

* 1. Airport\_Type2

Record of whether the airport is medium or large

* 1. Airport\_Lon

Longitude of the closest airport

* 1. Airport\_Lat

Latitude of the closest airport

* 1. Airport\_Lon2

Longitude of the closest medium or large airport

* 1. Airport\_Lat2

Latitude of the closest medium or large airport

* 1. Lights

Nightlights are collected for each site location by extracting the nightlights value from the gridded field. Radiance calibrated lights are used which allows for the detection of both bright lights associated with urban development and dim lights more typical of sub urban, ex urban and rural locations. The data is available here

<https://ngdc.noaa.gov/eog/> and data can be downloaded here.

<https://ngdc.noaa.gov/eog/dmsp/download_radcal.html> The file used for this metadata is the following:

<http://mapserver.ngdc.noaa.gov/viirs_data/dmsp/F16_20100111-20110731_rad_v4.geotiff.tgz>. The data is in a GWS84 projection with a 30 arc minute (1km) resolution

1. References
   1. DEM

Amante, C. and B.W. Eakins, 2009. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, NOAA. doi:10.7289/V5C8276M [Access date:12/26/2016]

* 1. Gridded Population of the World Center for International Earth Science Information Network - CIESIN - Columbia University. 2016. Gridded Population of the World, Version 4 (GPWv4): Population Count Adjusted to Match 2015 Revision of UN WPP Country Totals. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). <http://dx.doi.org/10.7927/H4SF2T42>. Accessed 22 12 2106.
  2. Hyde 3.1 Klein Goldewijk, K. , A. Beusen, M. de Vos and G. van Drecht (2011). The HYDE 3.1 spatially explicit database of human induced land use change over the past 12,000 years, Global Ecology and Biogeography**20(1)**: 73-86. [DOI: 10.1111/j.1466-8238.2010.00587.x.](http://onlinelibrary.wiley.com/doi/10.1111/j.1466-8238.2010.00587.x/abstract)
  3. GRUMP Settlement points Center for International Earth Science Information Network - CIESIN - Columbia University, International Food Policy Research Institute - IFPRI, The World Bank, and Centro Internacional de Agricultura Tropical - CIAT. 2011. Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Settlement Points. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). <http://dx.doi.org/10.7927/H4M906KR>. Accessed 22 12 2106.
  4. Balk, D.L., U. Deichmann, G. Yetman, F. Pozzi, S. I. Hay, and A. Nelson. 2006. Determining Global Population Distribution: Methods, Applications and Data. Advances in Parasitology 62:119-156. <http://dx.doi.org/10.1016/S0065-308X(05)62004-0>.
  5. Ziskin, Daniel, Kimberly Baugh, Feng Chi Hsu, Tilottama Ghosh, Chris Elvidge, 2010, "[Methods Used For the 2006 Radiance Lights](https://www.ngdc.noaa.gov/eog/pubs/APAN_30_Ziskin.pdf)", *Proceedings of the 30th Asia-Pacific Advanced Network Meeting*, 131-142

1. Supplementary Information