GSODR: Global Summary Daily Weather Data in R

20 January 2017

Summary

The GSODR package (Sparks, Hengl, and Nelson 2017) is an R package (R Core Team 2016) providing automated downloading, parsing, cleaning and of Global Surface Summary of the Day (GSOD) (United States National Oceanic and Atmospheric Administration National Climatic Data Center 2016) weather data for use in R or saving as local files in either a Comma Separated Values (CSV) or GeoPackage (GPKG) (Open Geospatial Consortium 2014) file. It builds on or complements several other scripts and packages. We take advantage of modern techniques in R to make more efficient use of available computing resources to complete the process, e.g., data.table (Dowle et al. 2015), plyr (Wickham 2011) and readr (Wickham, Hester, and Francois 2016), which allow the data cleaning, conversions and disk input/output processes to take advantage of modern computer hardware. The rnoaa (Chamberlain 2016) package already offers an excellent suite of tools for interacting with and downloading weather data from the United States National Oceanic and Atmospheric Administration, but lacks options for GSOD data retrieval. Several other APIs and R packages exist to access weather data, but most are region or continent specific, whereas GSOD is global. This package was developed to provide:

- two functions that simplify downloading GSOD data and formatting it to easily be used in research; and
- a function to help identify stations within a given radius of a point of interest.

Additional data, alternative elevation data based on a 200 meter buffer of elevation values derived from the CGIAR-CSI SRTM 90m Database (Jarvis et al. 2008) are included. These data are useful to help address possible inaccuracies and in many cases, missing elevation values, in the reported station elevations.

When using this package, GSOD stations are checked for inaccurate longitude and latitude values and any stations that are missing or having incorrect are omitted from the final data set. Station files with too many missing observations as defined by the user may be omitted from the final output to help ensure data quality. All units are converted from the United States Customary System (USCS) to the International System of Units (SI), e.g., inches to millimetres and Fahrenheit to Celsius. Wind speed is also converted from knots to metres per second. Additional useful values, actual vapour pressure, saturated water vapour pressure, and relative humidity are calculated and included in the final output. Station metadata are merged with weather data for the final data set which includes the following fields:

- **STNID** - Station number (WMO/DATSAV3 number) for the location;
- **WBAN** - number where applicable–this is the historical "Weather Bureau Air Force Navy" number - with WBAN being the acronym;
- **STN_NAME** - Unique text identifier;
- **CTRY** - Country in which the station is located;
- **LAT** - Latitude. *Station omitted in cases where values are <-90 or >90 degrees or latitude = 0 and longitude = 0*;
- **LON** - Longitude. *Station omitted in cases where values are <-180 or >180 degrees or latitude = 0 and longitude = 0*;
- **ELEV_M** - Elevation in metres;
- **ELEV_M_SRTM_90m** - Elevation in metres corrected for possible errors, derived from the CGIAR-CSI SRTM 90m database (Jarvis et al. 2008);

- **YEARMODA** - Date in YYYY-mm-dd format;
- **YEAR** - The year (YYYY);
- **MONTH** - The month (mm);
- **DAY** - The day (dd);
- **YDAY** - Sequential day of year (not in original GSOD);
- **TEMP** - Mean daily temperature converted to degrees C to tenths. Missing = NA;
- **TEMP_CNT** - Number of observations used in calculating mean daily temperature;
- **DEWP** - Mean daily dew point converted to degrees C to tenths. Missing = NA;
- **DEWP_CNT** - Number of observations used in calculating mean daily dew point;
- **SLP** - Mean sea level pressure in millibars to tenths. Missing = NA;
- **SLP_CNT** - Number of observations used in calculating mean sea level pressure;
- **STP** - Mean station pressure for the day in millibars to tenths. Missing = NA;
- **STP_CNT** - Number of observations used in calculating mean station pressure;
- **VISIB** - Mean visibility for the day converted to kilometres to tenths Missing = NA;
- **VISIB_CNT** - Number of observations used in calculating mean daily visibility;
- **WDSP** - Mean daily wind speed value converted to metres/second to tenths Missing = NA;
- **WDSP_CNT** - Number of observations used in calculating mean daily wind speed;
- **MXSPD** - Maximum sustained wind speed reported for the day converted to metres/second to tenths. Missing = NA;
- **GUST** - Maximum wind gust reported for the day converted to metres/second to tenths. Missing = NA;
- **MAX** - Maximum temperature reported during the day converted to Celsius to tenths–time of max temp report varies by country and region, so this will sometimes not be the max for the calendar day. Missing = NA;
- **MAX_FLAG** - Blank indicates max temp was taken from the explicit max temp report and not from the 'hourly' data. An "*" indicates max temp was derived from the hourly data (i.e., highest hourly or synoptic-reported temperature);
- **MIN**- Minimum temperature reported during the day converted to Celsius to tenths–time of minimum temp report varies by country and region, so this will sometimes not be the max for the calendar day. Missing = NA;
- **MIN_FLAG** - Blank indicates max temp was taken from the explicit max temp report and not from the 'hourly' data. An "*" indicates minimum temp was derived from the hourly data (i.e., highest hourly or synoptic-reported temperature);
- **PRCP** - Total precipitation (rain and/or melted snow) reported during the day converted to millimetres to hundredths; will usually not end with the midnight observation, i.e., may include latter part of previous day. A value of ".00" indicates no measurable precipitation (includes a trace). Missing = NA; *Note: Many stations do not report '0' on days with no precipitation– therefore, 'NA' will often appear on these days. For example, a station may only report a 6-hour amount for the period during which rain fell.* See FLAGS_PRCP column for source of data;
- **PRCP_FLAG** -
  - A = 1 report of 6-hour precipitation amount;

  - B = Summation of 2 reports of 6-hour precipitation amount;
  - C = Summation of 3 reports of 6-hour precipitation amount;
  - D = Summation of 4 reports of 6-hour precipitation amount;
  - E = 1 report of 12-hour precipitation amount;
  - F = Summation of 2 reports of 12-hour precipitation amount;
  - G = 1 report of 24-hour precipitation amount;
  - H = Station reported '0' as the amount for the day (e.g., from 6-hour reports), but also reported at least one occurrence of precipitation in hourly observations–this could indicate a trace occurred, but should be considered as incomplete data for the day;
  - I = Station did not report any precip data for the day and did not report any occurrences of precipitation in its hourly observations–it's still possible that precipitation occurred but was not reported;
- **SNDP** - Snow depth in millimetres to tenths. Missing = NA;
- **I_FOG** - Indicator for fog, (1 = yes, 0 = no/not reported) for the occurrence during the day;
- **I_RAIN_DRIZZLE** - Indicator for rain or drizzle, (1 = yes, 0 = no/not reported) for the occurrence during the day;
- **I_SNOW_ICE** - Indicator for snow or ice pellets, (1 = yes, 0 = no/not reported) for the occurrence during the day;
- **I_HAIL** - Indicator for hail, (1 = yes, 0 = no/not reported) for the occurrence during the day;
- **I_THUNDER** - Indicator for thunder, (1 = yes, 0 = no/not reported) for the occurrence during the day;
- **I_TORNADO_FUNNEL** - Indicator for tornado or funnel cloud, (1 = yes, 0 = no/not reported) for the occurrence during the day;
- **ea** - Mean daily actual vapour pressure;
- **es** - Mean daily saturation vapour pressure;
- **RH** - Mean daily relative humidity.

References

Chamberlain, Scott. 2016. *Rnoaa: 'NOAA' Weather Data from R.* https://CRAN.R-project.org/package=rnoaa.

Dowle, M, A Srinivasan, T Short, S Lianoglou with contributions from R Saporta, and E Antonyan. 2015. *Data.table: Extension of Data.frame.* https://CRAN.R-project.org/package=data.table.

Jarvis, Andy, Hannes I Reuter, Andy Nelson, and Edward Guevara. 2008. "Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database." http://srtm.csi.cgiar.org.

Open Geospatial Consortium. 2014. "GeoPackage Encoding Standard." http://www.opengeospatial.org/standards/geopackage.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sparks, Adam, Tomislav Hengl, and Andrew Nelson. 2017. *GSODR: Global Summary Daily Weather Data in*

R. http://ropensci.github.io/GSODR/.

United States National Oceanic and Atmospheric Administration National Climatic Data Center. 2016. "Global Surface Summary of Day (GSOD)." https://data.noaa.gov/dataset/global-surface-summary-of-the-day-gsod.

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

Wickham, Hadley, Jim Hester, and Romain Francois. 2016. *Readr: Read Tabular Data*. https://CRAN.R-project.org/package=readr.