Solar Energy 171 (2018) A3–A12

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Data Article

SolarData: An R package for easy access of publicly available solar datasets

Dazhi Yang

*Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), Singapore*

ARTICLE INFO

Keywords:
R
Solar data
Publicly available
Open source

ABSTRACT

Although the applications of data science and machine learning in solar engineering have increased tremendously in the past decade, most of the solar datasets come from heterogeneous and autonomous sources. For that reason, identifying spatially collocated and temporally aligned datasets has always been time consuming, or even frustrating sometimes, in data-driven solar research. In this regard, I present a new R package—`SolarData`—for easy access of some publicly available solar datasets. In version 1.0, a total of 5 datasets are included: (1) NREL physical solar model version 3; (2) NREL Oahu solar measurement grid; (3) NOAA surface radiation network; (4) SoDa Linke turbidity factor; and (5) NASA shuttle radar topography mission. This paper provides an overview of each of dataset, and gives code segments that exemplify the usage of the package. Furthermore, in the appendices, a series of self-contained R scripts are used to provide perspectives on how these datasets can be used in solar research. To promote the widespread uptake of this R package, and to facilitate future contribution and collaboration, all contents herein described are made available on GitHub. This paper is the first *Data Article*, a new submission type offered by the Solar Energy journal.

1. Introduction

Solar energy research is inter-disciplinary and has a wide spectrum. At the high end of the spectrum, resource assessment and forecasting are the two most prominent areas, which require knowledge from various domains, such as atmospheric science, climatology, meteorology, statistics, data science, or artificial intelligence. For an overview of these two areas, the reader is referred to Kleissl (2013), Vignola et al. (2016), Polo et al. (2016), and Yang et al. (2018). Whereas the inter-disciplinary knowledge advances methodological development, selecting suitable datasets is vital for empirical demonstration.

In an academic environment, methodology and data are (arguably) the two most important factors deciding whether a research work is publishable. In resource assessment and forecasting, some frequently encountered criticisms on data include:

1. low-quality data and/or lack of quality control (QC);
2. suboptimal handling of data, in terms of aggregation, disaggregation, missing data imputation, data transformation or data subsetting;
3. insufficient (spatial and/or temporal) data coverage; and
4. lack of exogenous and supporting data.

From the reviewers' point of view, expanding the data adds value to the research. However, from the authors' point of view, including more data means additional work. Based on my limited experience as an author, a reviewer, and an editor, such criticisms on data always lead to one of the two outcomes: (1) a hard-earned consensus between the authors and reviewers, and (2) a withdrawal by the authors or a rejection by the reviewers. Neither outcome is optimal. To that end, it is thought critical to establish some basic infrastructures, i.e., software packages, that can help the authors and reviewers to identify and use good-quality solar data with universal applicability. I choose R (R Core Team, 2018) for this task.

R is a free software environment for statistical computing and graphics. In 2017, it was ranked the 6th most popular programming language by IEEE Specturm. Among statisticians, it is *the* most popular software. For data science and machine learning, it is also a very popular choice, right next to Python. R is highly extensible through the use of packages. These packages are mostly on CRAN (https://cran.r-project.org/), whereas a majority of the remaining ones are on GitHub (https://github.com/). Whereas R has a command line interface, many front-end graphical user interfaces are also available, among which RStudio (https://www.rstudio.com/) is the most popular one. The installation of R and RStudio (for Windows, Mac, and Linux) is extremely simple, and can be done in less than a minute by clicking the download tabs from the respective websites.

E-mail address: yangdazhi.nus@gmail.com.

https://doi.org/10.1016/j.solener.2018.06.107
Received 26 April 2018; Received in revised form 26 June 2018; Accepted 27 June 2018
Available online 30 August 2018
0038-092X/ © 2018 Elsevier Ltd. All rights reserved.

Nomenclature

API	Application Programming Interface
CRAN	Comprehensive R Archive Network
DEM	Digital Elevation Model
DIF	DIFfuse horizontal irradiance
DNI	Direct Normal Irradiance
GHI	Global Horizontal Irradiance
LTF	Linke Turbidity Factor
NGA	National Geospatial-intelligence Agency
NOAA	National Oceanic and Atmospheric Administration
NREL	National Renewable Energy Laboratory
NSRDB	National Solar Radiation DataBase
OSMG	Oahu Solar Measurement Grid
PSM	Physical Solar Model
SoDa	Solar radiation Data
SRTM	Shuttle Radar Topography Mission
SURFRAD	SURFace RADiation budget network
TMY	Typical Meteorological Year

Symbols

μ_0	the cosine of zenith angle
E_{0n}	extraterrestrial irradiance on a normal surface
R_L	the Rayleigh limit
Z	zenith angle
Clr	Ineichen–Perez clear-sky irradiance
Prs	station pressure

Currently available solar engineering packages on CRAN include `insol` (Corripio, 2014), `sirad` (Bojanowski, 2016), `solarR` (Perpiñán, 2012). Unfortunately, none of these packages offer any function for data access. On the other hand, several meteorology packages, such as `meteoland` (De Cáceres et al., 2018), `meteoForecast` (Perpiñán and Almeida, 2018) or `stationaRy` (Iannone, 2015), contain routines for accessing, reading, and manipulating meteorological data. They are however not specific to solar data, and use only one or a few data sources.

Besides the R packages, there are other scattered code segments online that facilitate researchers to access and use solar data. For example, the climate-based optimization of renewable power allocation (COPA) model optimizes the portfolio of various renewable power production options (solar, wind, and hydro) using linear programming. Since COPA uses global reanalysis products such as MERRA or ERA-Interim, R code transforming reanalysis data into power generation is provided (https://homepage.boku.ac.at/jschmidt/TOOLS/index.html). Nevertheless, these code segments are often task-specific, and thus lack generality.

In view of the above discussions, the dataset characteristics of the new R package—`SolarData`—are designed as follows:

1. the datasets must be quality controlled, either by the data owners or in the package;
2. the datasets should have extensive spatio-temporal coverage;
3. the datasets should be suitable for a wide range of applications; and
4. the datasets need to be proven of interest to the solar community, i.e., the datasets have been used in multiple publications in the literature.

The remaining part of the paper discusses the 5 datasets (see Table 1) in version 1.0 of `SolarData`, one dataset in each section, in accordance to the above points. To run the code examples in each section, the following lines install and load the `devtools` package, which helps to install `SolarData` from GitHub.[1]

```
install.packages("devtools", repos = "http://cran.us.r-project.org")
library("devtools")
install_github("dazhiyang/SolarData")
```

[1] R version updates very frequently. At the time of submission of this paper, the newest version is 3.5.0. However, this version has some issues on the installation of one of the import packages, namely, `data.table` (see https://github.com/Rdatatable/data.table/issues/2797). For this reason, R version 3.4.4 is recommended at the moment.

2. NREL physical solar model version 3

2.1. Quick overview

The National Solar Radiation Database (NSRDB) is a collection of hourly and half-hourly values of various irradiance measurements and other meteorological data. NSRDB is managed by the National Renewable Energy Laboratory, and has provided solar resource data for 25 years. Throughout its existence, numerous updates have been made. The most recent update, namely, the physical solar model (PSM) version 3 is a satellite-derived irradiance dataset with a regular grid of 0.04° resolution in both latitude and longitude, covering most of America. Recently, this update is formally described in great details by Sengupta et al. (2018). In terms of quality control and uncertainty quantification, several previous works (Sengupta et al., 2018; Habte et al., 2017; Xie et al., 2017) have addressed the issues by contrasting the PSM data to ground-based measurements, namely, SURFRAD (see Section 4). The reader is therefore referred to those publications for a full description and uncertainty quantification of the PSM data.

2.2. Potential usages

Since PSM has an extensive temporal coverage, from 1998 to 2016, it is mostly used for resource assessment purposes. Resource assessment requires data over a long enough period, such as 20–30 years, to fully characterize the climatological and meteorological conditions at a given location and time. A commonly used way to summarize such long-term data is by converting it to a typical meteorological year (TMY) file (Wilcox and Marion, 2008; Marion and Urban, 1995). Besides PSM, other versions of TMY files are also available on NSRDB. Hence, it is of interest to compare the PSM TMY data to other versions, such as the TMY3 (Wilcox and Marion, 2008) or the SUNY data (Perez et al., 2015). Extending from resource assessment, PSM is also useful in other solar engineering applications that require long-term irradiance data over a geographical area, for instance, site adaptation (Polo et al., 2016; Martín-Pomares et al., 2017), monitoring network design (Yang, 2017; Yang and Reindl, 2015; Zagouras et al., 2013), variability quantification (Lave et al., 2017; Zagouras et al., 2014), or siting and sizing of a PV system (Rodríguez-Gallegos et al., 2018).

One of the outstanding features of PSM is that it is serially complete, i.e., gap free. Furthermore, its spatial granularity facilities studies on interactions among the irradiance data within some neighborhood. More specifically, the PSM dataset provides an excellent platform for spatio-temporal prediction research. In fact, prediction using satellite-derived data has gained much attention in the recent years (Yang et al., 2018b; Blanc et al., 2017), especially for nowcasting (Ayet and Tandeo, 2018; Lorenzo et al., 2017). Since solar irradiance is a spatio-temporal

Table 1
Summary of the five freely available datasets included in version 1.0 of `SolarData`.

Dateset	Description	Data resolution (space; time)	Data coverage (space; time)
NREL PSM	Satellite-derived irradiance and other meteorological data on a regular grid	0.04° × 0.04°; 30 min	Most of America; 1998–2016
NREL Oahu	Ground-based irradiance sensor network data	17 stations; 1 s or 3 s	1.2 km × 1.2 km; 2010 Mar–2011 Oct
NOAA SURFRAD	Ground-based irradiance and other meteorological data	7 stations; 1 min or 3 min	sparsely located in continental US; 1995–yesterday
SoDa LTF	Linke turbidity data on a regular grid	5′ × 5′; 1 mon	The globe (land and sea); 2003
NASA SRTM	Digital elevation model data on a regular grid	1″ × 1″ or 3″ × 3″; single snapshot	The globe (land only); single snapshot

process, many methods and models in the statistics literature (e.g., Cressie and Wikle, 2015) can now be applied to solar forecasting.

2.3. Getting the data

PSM data can be accessed in three ways: (1) via the NSRDB Viewer, an interactive web application, (2) via the application programming interface (API), the API key can be obtained from https://developer.nrel.gov/signup/, and (3) via Globus. Since the entire dataset is approximately 50 terabyte (Sengupta et al., 2018), it is difficult to store and manipulate the PSM dataset as a whole. In this regard, subsetting the data is almost always necessary,[2] and the API option allows one to efficiently download data for a specific location and year. The `PSM.get` function in `SolarData` does exactly that. However, before the usage of the function is exemplified, I digress and show how to define the location of interest, i.e., locations for which the PSM files will be downloaded.

There are a number of ways to generate regular and irregular lattices. For example, we can utilize the map boundary to control the lattice locations. The following code generates the boundary of California:

```
# load four libraries for this section
libs <- c("maps", "maptools", "sp", "ggplot2")
invisible(lapply(libs, library, character.only = TRUE))
# push geographical information of California in
# the object California using the R package "maps"
California <- maps::map("state", region = "california",
              fill = TRUE, plot = FALSE)
# extract the state boundaries into the object bndary using
# the R packages "maptools" and "sp"
bd <- maptools::map2SpatialPolygons(California, IDs = "california",
      proj4string=sp::CRS("+proj=longlat +datum=WGS84"))
bndary <- bd@polygons[[1]]@Polygons[[1]]@coords
# push the boundary data into a data frame and
# rename the columns to "lon" and "lat"
bndary_plot <- data.frame(bndary)
names(bndary_plot) <- c("lon", "lat")
```

It should be noted that PSM is gridded data, hence, queries on all locations within a grid box return the same file. This implies that the finest resolution one can set is 0.04°. In the following code segment, a regular lattice with a resolution of 0.2° is used.

```
# generate the regular grid with a spatial resolution of 0.2
res1 = 0.2
# generate longitude and latitude vectors and
# expand the vectors to lattice, i.e., regular grid
x1 <- seq(-124.02, -114.1, by = res1)
y1 <- seq(42.05, 32.40, by = -res1)
loc1 <- expand.grid(x1,y1)
# select those lattice locations within California
loc_reg <- loc1[which(sp::point.in.polygon(point.x = loc1[,1],
   point.y = loc1[,2], pol.x = bndary[,1], pol.y = bndary[,2])==1),]
```

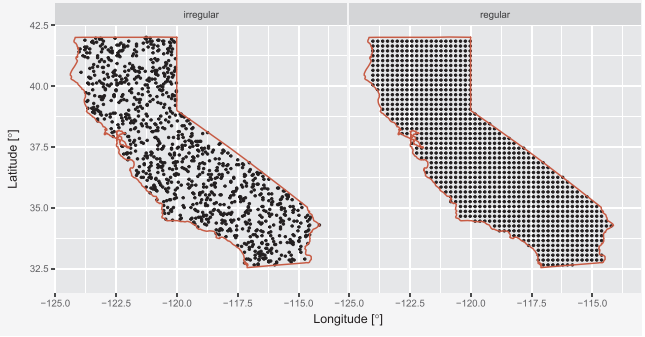
For the irregular lattice, random sampling is used to sample from the most granular lattice, i.e., 0.04°.

[2] On this point, I wrote two companion papers discussing subsetting strategy for the PSM data; the reader is referred to Yang (2018a, 2018b) for details.

```
# generate the regular grid with a spatial resolution of 0.04
# i.e., the grid of the original PSM data
res2 = 0.04
# the following code is similar to the previous segment
x2 <- seq(-124.02, -114.1, by = 0.04)
y2 <- seq(42.05, 32.40, by = -0.04)
loc2 <- expand.grid(x2,y2)
loc_irreg <- loc2[which(sp::point.in.polygon(point.x = loc2[,1],
   point.y = loc2[,2], pol.x = bndary[,1], pol.y = bndary[,2])==1),]
# function "sample" is used to sample the final irregular grid,
# following the size of the regular grid.
loc_irreg <- loc_irreg[sample(x = 1:nrow(loc_irreg), size =
             nrow(loc_reg), replace = FALSE),]
```

To visualize the above lattices, package `ggplot2` (Wickham, 2009) is used. It is noted that `ggplot2` is based on the grammar of graphics, and thus has a moderate learning curve. However, once the pieces are fit together, it becomes a very handy tool to generate research-grade plots. On this point, it is worth mentioning that the book by Oscar Perpiñán (Perpiñán, 2014) provides detailed visualization examples for time series, spatial and spatio-temporal data, which many are in fact solar data.

```
# construct a data.frame for plotting
# ggplot groups the data and plot them in different panels,
# thus a third variable "group" is added into the data.frame
data_plot <- data.frame(lon = append(loc_reg[,1], loc_irreg[,1]),
                        lat = append(loc_reg[,2], loc_irreg[,2]),
                        group = c(rep("regular", nrow(loc_reg)),
                        rep("irregular", nrow(loc_irreg))))
# ggplot code
# two types of object, "geom_point" and "geom_polygon", are used
# the plots are separated into two panels using "facet_wrap"
# set the x- and y-axis using "scale_*" and "*lab"
# lastly, set the overall theme options
p <- ggplot() +
     geom_point(data=data_plot,aes(x=lon,y=lat), size = 0.5) +
     geom_polygon(data=bndary_plot,aes(x=lon,y=lat), size = 0.5,
                  color = "red", fill = NA) +
     facet_wrap(~group) +
     coord_fixed() +
     scale_x_continuous(limits=c(-125, -113), expand = c(0, 0)) +
     scale_y_continuous(limits=c(31.5, 42.5), expand = c(0, 0)) +
     xlab(expression(paste("Longitude [", degree, "]", sep = ""))) +
     ylab(expression(paste("Latitude [", degree, "]", sep = ""))) +
     theme_gray() +
     theme(plot.margin = unit(c(0.5,0.5,0,0), "lines"),
           panel.spacing = unit(0.1, "lines"),
           text = element_text(size = 9),
           legend.position = "none")
p # print plot
```



Once the lattices, i.e., the locations where PSM data are desired, are defined, downloading can be done with a single call of PSM.get. The raw data files are saved into the working directory automatically—the function parameter directory allows one to set the working directory. For demonstration purposes, the following code segment downloads the raw files from two locations, for the year 2016. Most of the inputs for PSM.get are API request parameters. For example, attributes can be set using a vector of character stings; the options are "air_temperature", "clearsky_dhi", "clearsky_dni", "clearsky_ghi", "cloud_type", "dew_point", "dhi", "dni", "fill_flag", "ghi", "relative_humidity", "solar_zenith_angle", "surface_albedo", "surface_pressure", "total_precipitable_water", "wind_direction", and "wind_speed". Function input year takes a single character string indicating the year of data to be downloaded. Inputs interval can take either "30" or "60", corresponding to the data temporal resolution of the download. For a description on the remaining function inputs, the reader is referred to the PSM API webpage (https://developer.nrel.gov/docs/solar/nsrdb/psm3_data_download/), as well as the SolarData package documentation, which can be found in Appendix D.

```
# load the "SolarData" package
library("SolarData")
# get PSM data at locations (42.05N, 124.02W) and (44N, 120W)
loc <- matrix(c(42.05, 44, -124.02, -120), nrow = 2)
PSM.get(lat = loc[,1], lon = loc[,2],
        api.key = "YourAPIKey", # write here your API key
        attributes = "ghi,dhi,dni", # variables to be retrieved
        name = "John+Smith", # write here your name
        affiliation = "Some+Institute", # write here institute
        year = "2016", # which year to be retrieved
        leap.year = "true", # whether to include Feb 29 if leap year
        interval = "30", # data resolution, 30 or 60?
        utc = "false", # whether to use utc or local time
        reason.for.use = "research", # write here your reason
        email = "email@gmail.com", # write here your email
        mailing.list = "false", # whether to be on the mailing list
        directory = "YourDownloadDirectory") # download directory
```

Since the raw PSM data files are in csv format—one file for each location each year—they can be read using the read.csv function from the utils package. Appendix A provides an example of reading the PSM data. In that appendix, a regression-based site adaptation is performed.

3. NREL Oahu solar measurement grid

The PSM dataset in Section 2 is half-hourly, and thus is not suitable for variability and forecasting studies on micro spatio-temporal scale. Within each pixel of the PSM lattice, namely, a 4 km × 4 km square, solar irradiance sampled at high frequency can decorrelate fast. In this regard, the second dataset is selected to provide opportunities for micro-scale studies.

3.1. Quick overview

The NREL Oahu solar measurement grid (OSMG) is part of the Measurement and Instrumentation Data Center. It consists of 17 horizontally installed LICOR LI-200 Pyranometer, 2 tilted LI-200, and a rotating shadowband radiometer (RSR). Whereas the LI-200 pyranometers log global (horizontal or tilted) irradiance every second, RSR log all three irradiance components (GHI, DIF, and DNI) every 3 s. All the instruments are arranged within a 1 km × 1 km area. The data is available from 2010-03 to 2011-10. The details of these instruments and their geographical arrangement can be found at https://midcdmz.nrel.gov/oahu_archive/. The raw data are stored as zip files, and can be easily accessed from the webpage.

3.2. Potential usages

The network layout of OSMG can be found at the above data website. This densely built grid was designed based on the prevailing trade wind direction over that area, namely, 60° from north. Owing to its high sampling rate, the OSMG dataset has let to a series of studies on spatio-temporal correlation of the solar irradiance random field (Munkhammar et al., 2017; Arias-Castro et al., 2014; Hinkelman, 2013), and a series of very short-term (sub-5-min horizons) forecasting studies (e Silva and Brito, 2018; Yang et al., 2015, 2017; Aryaputera et al., 2015).[3] It is believed that both areas still require much research in the future (Yang et al., 2018b), hence, this dataset is included here.

3.3. Reading the data

As mentioned earlier, the data zip files can be obtained from the archive. The individual daily files containing data from LI-200 sensors are zipped monthly, whereas RSR daily data files are zipped yearly. Since for a same day, the LI-200 file and the RSR file have the same file name, I recommend unzipping all files of interest into two folders, one for LI-200 and one for RSR, without sub-folders. In this way, the OSMG.read function will read and concatenate all files into *tidy data* (Wickham et al., 2014), which is one of the most efficient and effective approaches to manipulate data frames. In the following example, OSMG data from 2010 July 31 is read and printed.

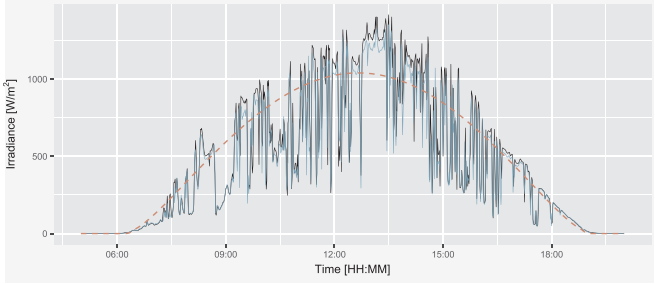
```
# define two directories, one for LI-200 files, and one for RSR files
dir_LI200 <- "YourLI200Directory"
dir_RSR <- "YourRSRDirectory"
data <- OSMG.read("20100731.txt", # file name, can be a vector
                  directory.LI200 = dir_LI200,
                  directory.RSR = dir_RSR,
                  clear.sky = TRUE, # whether to computer clear-sky
                  AP2 = TRUE, #whether to include AP2
                  agg = 60) #the aggregation interval
# filter data with zen>80 for output purpose
# this is also to introduce the pipeline operator, %>%
data %>% filter(., zen < 80)
```

```
## # A tibble: 689 x 26
##    Time                 zen   Ics   Ioh  DH1T  AP6T   DH3   DH4
##    <dttm>             <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 2010-07-31 06:55:00  80.0  101.  247.  35.6  39.8  62.1  65.5
##  2 2010-07-31 06:56:00  79.7  104.  252.  38.3  42.3  64.8  68.2
##  3 2010-07-31 06:57:00  79.5  108.  258.  40.8  43.6  66.1  69.7
##  4 2010-07-31 06:58:00  79.3  111.  263.  40.1  43.0  64.9  67.9
##  5 2010-07-31 06:59:00  79.1  115.  268.  38.2  43.4  64.2  67.0
##  6 2010-07-31 07:00:00  78.8  119.  274.  37.9  42.8  62.6  65.2
##  7 2010-07-31 07:01:00  78.6  122.  279.  36.0  39.5  56.6  59.1
##  8 2010-07-31 07:02:00  78.4  126.  285.  34.2  37.7  54.4  57.0
##  9 2010-07-31 07:03:00  78.2  129.  290.  34.8  39.3  56.4  59.3
## 10 2010-07-31 07:04:00  77.9  133.  296.  37.3  42.5  59.0  62.0
## # ... with 679 more rows, and 18 more variables: DH5 <dbl>,
## #   DH10 <dbl>, DH11 <dbl>, DH9 <dbl>, DH2 <dbl>, DH1 <dbl>,
## #   AP6 <dbl>, AP1 <dbl>, AP3 <dbl>, AP5 <dbl>, AP4 <dbl>,
## #   AP7 <dbl>, DH6 <dbl>, DH7 <dbl>, DH8 <dbl>, AP2 <dbl>,
## #   AP2.dif <dbl>, AP2.dir <dbl>
```

The clear.sky parameter is a boolean variable indicating whether the Ineichen–Perez clear-sky irradiance value (see Section 5) for each timestamp should be calculated. AP2 indicates whether the RSR 3-s data should be included. It is noted that the agg parameter must be ⩾ 3 to join the files, in the above case, the data are aggregated to 1-min intervals. To provide a simple visualization, ggplot2 is again used. Whereas the black and turquoise lines show the DH3 and AP2 time series, the red dashed line shows the clear-sky expectation.

[3] A majority of these studies consider a subset of 13 days where the sky condition is predominated by broken clouds: July 31, August 1–5, 21, 29, September 5–7, 21, and October 27. All 13 days are from the year 2010.

```
# ggplot code
# plot three lines with different colors
# function "gg_color_hue" enumerates default ggplot color
p <- ggplot(data) +
     geom_line(aes(x=Time,y=DH3), size = 0.2) +
     geom_line(aes(x=Time,y=AP2), size = 0.2,
               color = SolarData:::gg_color_hue(2)[2]) +
     geom_line(aes(x=Time,y=Ics), size = 0.3, linetype = "dashed",
               color = SolarData:::gg_color_hue(2)[1]) +
     xlab("Time [HH:MM]") +
     ylab(expression(paste("Irradiance [W/", m^2, "]", sep = ""))) +
     theme_gray() +
     theme(plot.margin = unit(c(0.5,0.5,0,0), "lines"),
           panel.spacing = unit(0.1, "lines"),
           text = element_text(size = 9),
           legend.position = "none")
p
```



Besides the OSMG.read function, two distance-calculation functions are also included in this package. These functions are useful when inter-station, along-wind, or cross-wind distances are needed, for example, in a variability study. In this regard, a full-length example that replicates the results of the anisotropic correlation models shown in Arias-Castro et al. (2014) is provided in Appendix B.

4. NOAA surface radiation network

Whereas satellite-derived products, such as PSM, provide regional and global coverage, their accuracy and reliability need to be refined and verified using high-accuracy ground-based measurements. To that end, the dataset described in this section comes from a pioneer project that aims at providing long-term, reliable, radiation measurements in differing climatic regions.

4.1. Quick overview

NOAA's surface radiation (SURFRAD) budget network was established in 1993, commenced operation in 1995. It has four initial stations, expanded to six in 1998, and a seventh station was added in 2003 (Augustine et al., 2000, 2005). At each station, all three components of solar irradiance are measured, alongside with several meteorological parameters, such as ambient temperature, station pressure, or wind speed. Moreover, SURFRAD also provides a series of longwave radiation measurements. For a full list of available parameters and instrumentation, the reader is referred to https://www.esrl.noaa.gov/gmd/grad/surfrad/, as well as the README file on the FTP server.

The SURFRAD dataset is in the form of daily ASCII text files that have an extension of.dat. These files can be retrieved via FTP. Due to the different commencing date, the length of data at each station varies. Furthermore, data from the station commencing date to 2009 January 1 are reported as 3-min averages (480 rows per file,[4] excluding headers); on and after 2009 January 1, the data are reported as 1-min averages (1440 rows per file). A quality-control sequence has been applied to all raw data by the data owners, hence, the publicly available files also include the QC flags. The details of the QC procedure and flag interpretation can be found in the README file on the FTP server.

[4] Due to missing data, the actual number of rows may be smaller than 480, or 1440 for the 1-min averaged files.

4.2. Potential usages

Owing to its good quality and long history, the SURFRAD dataset is profusely utilized, and has led to a large number of publications. A quoted Google Scholar search—"SURFRAD"—returns approximately 1500 results. By browsing through these results, it is found a large percentage of these publications are in the areas of remote sensing and atmospheric science (Franch et al., 2014; Heidinger et al., 2013; Yu et al., 2012; Wang and Liang, 2009). When the term "solar energy" is added to the search phrase, the returned results become more relevant to the present discussion, i.e., resource assessment and forecasting. In many publications, the SURFRAD dataset is used as the "ground truth", to validate numerical weather prediction (NWP) and satellite-derived irradiance models (Mathiesen and Kleissl, 2011; Perez et al., 2010, 2004). Since the NWP output and the satellite models are often biased, comparing them to ground-based data can directly incorporate these observed bias into modeling, through model output statistics or other adjustment methods. Besides being the reference measurements, the high temporal resolution of the SURFRAD data also allows the development of solar variability models (Lauret et al., 2016; Perez et al., 2011, 2012).

4.3. Getting the data

Since the SURFRAD dataset can be accessed via FTP, most free FTP clients, such as FileZilla, CoreFTP, or FireFTP, can be used to download the data. Alternatively, the SURFRAD.get function in the SolarData package can be used to download files from a particular station over a particular year. For example, to obtain the files for the first three days in 2015 from station Bondville, Illinois, the below code can be used.

```
# the function can only download one station one year at a time
SURFRAD.get(station = "bon", # station abbrv. or full name
            year = "2015", # which year to be downloaded
            day.of.year = c(1:3), # which day(s) of year
            directory = "YourDownloadDirectory")
```

4.4. Reading the data

The downloaded SURFRAD data files with .dat extension can be read using the read.table function from the utils package. Although reading the data files may sound straightforward, converting them into a usable data frame in fact requires much attention, especially for applications that desire serially complete (gap free) observations, such as forecasting. After careful consideration and many iterations, in SolarData version 1.0, the SURFRAD.read function is designed as follows:

1. The function reads five parameters, namely, zenith angle, GHI, DIF, DNI, and pressure (Prs). If other parameters are needed, they can only be specified from the source code, i.e., manually customizing the code. This design is to maximize the number of good data points after listwise deletion of missing and spurious data (see points 3 and 6 below).
2. Two boolean input variables, namely, use.original.qc and use.qc, specify whether the default SURFRAD QC and/or additional QC (see below) should be used. To skip these QC procedures, both parameters should be set to FALSE.
3. If any parameter fails the SURFRAD QC, all parameters from that timestamp are set as NA. This is because many QC tests are non-definitive, i.e., the particular parameter that causes the test to fail is unknown after the test.
4. The SURFRAD data has missing timestamps. Therefore, these missing timestamps are completed before running the solar positioning algorithm.
5. After running the solar positioning algorithm, Ineichen–Perez clear-sky irradiance (Clr) and extraterrestrial irradiance on a normal surface (E_{0n}) are calculated. In addition, the cosine of zenith angle (μ_0), the Rayleigh limit (R_L), and the GHI estimated through the closure equation, are also computed. The Rayleigh limit is defined as

$R_L = 209.3\mu_0 - 708.3\mu_0^2 + 1128.7\mu_0^3 - 911.2\mu_0^4 + 287.85\mu_0^5 + 0.046725\mu_0 \text{Prs}$.

6. To verify and improve SURFRAD QC, five QC tests are included, following the framework provided in Long and Shi (2008) and Long and Dutton (2002).
   (a) Physically-possible limits or "phy"
      • $-4 < \text{GHI} < 1.5E_{0n}\cos^{1.2}Z + 100$
      • $-4 < \text{DIF} < 0.95E_{0n}\cos^{1.2}Z + 50$
      • $-4 < \text{DNI} < E_{0n}$
   (b) Extremely-rare limits or "ext"
      • $-2 < \text{GHI} < 1.2E_{0n}\cos^{1.2}Z + 50$
      • $-2 < \text{DIF} < 0.75E_{0n}\cos^{1.2}Z + 30$
      • $-2 < \text{DNI} < 0.95E_{0n}\cos^{0.2}Z + 10$
   (c) Closure equation or "closr"
      • abs(closr) <8% for $Z < 75°$ and GHI >50
      • abs(closr) <15% for $93° > Z > 75°$ and GHI >50
   (d) Diffuse ratio test or "dr"
      • DIF/GHI <1.05 for $Z < 75°$ and GHI >50
      • DIF/GHI <1.10 for $Z > 75°$ and GHI >50
   (e) Climatological comparisons or "clim"
      • DIF/GHI <0.85 for GHI/Clr >0.85 and DIF >50
      • DIF $> R_L - 1.0$ for DIF/GHI <0.8 and GHI >50

   By specifying the function input test, these QC tests can be used individually or collectively. For the same reason mentioned in point 3 above, if any parameter fails one of the specified QC tests, all parameters from that timestamp are set as NA.
7. After the QC, negative irradiance measurements and those with $Z > 90°$ are set to zero.
8. The function input agg allows the user to aggregate the SURFRAD data into desired resolution.[5] However, due to missing data, simple aggregation is insufficient. It is well known that high-frequency irradiance can reach values higher than the clear-sky expectation, or even extraterrestrial irradiance, owing to cloud-enhancement events (Killinger et al., 2017; Schade et al., 2007). Therefore, if there are NA values within an aggregation interval, ignoring them and aggregating the rest can distort the aggregation accuracy. In this regard, an interval is only aggregated when there are more than 50% valid data points. For example, each hourly aggregated value requires at least 30 1-min data points. The final timestamps denote the end times of the aggregation intervals.
9. Since the directory folder may contain a large number of files, such as one year, a progress bar is used to track the status. It can be turned off using parameter progress.bar.
10. The final cleaned and concatenated data frame is outputted as tidy data.

The following code segment demonstrates the usage of SURFRAD.read. It assumes that the raw files for the year 2015 from station Bondville, Illinois, have already been downloaded into some folder. The aggregation interval is set to 10 min, and three QC tests are included.

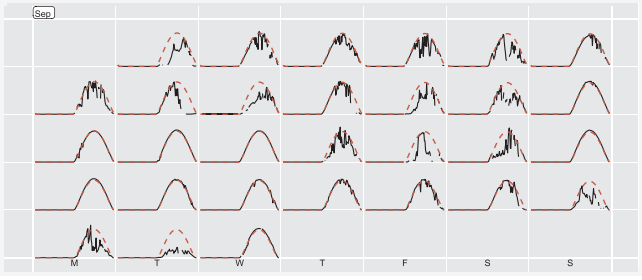
```
# set the working directory to the previous download directory
directory = "YourDataDirectory/bon/2015"
setwd(directory)
# get all the file names from the directory
# alternatively, define your own file names
files <- dir()
data <- SURFRAD.read(files, # char string or char string vector
                     use.original.qc = FALSE,
                     use.qc = TRUE, # whether to use QC
                     test = c("ext", "dr", "clim"), # type of QC
                     directory = directory,
                     agg = 10) # aggregation interval in second
data # print
```

```
## # A tibble: 52,561 x 8
##    Time                 zen dw_solar direct_n diffuse pressure
##    <dttm>             <dbl>    <dbl>    <dbl>   <dbl>    <dbl>
##  1 2015-01-01 00:00:00  105.       0.       0.      0.    1001.
##  2 2015-01-01 00:10:00  106.       0.       0.      0.    1001.
##  3 2015-01-01 00:20:00  108.       0.       0.      0.    1001.
##  4 2015-01-01 00:30:00  110.       0.       0.      0.    1001.
##  5 2015-01-01 00:40:00  112.       0.       0.      0.    1001.
##  6 2015-01-01 00:50:00  114.       0.       0.      0.    1001.
##  7 2015-01-01 01:00:00  115.       0.       0.      0.    1001.
##  8 2015-01-01 01:10:00  117.       0.       0.      0.    1001.
##  9 2015-01-01 01:20:00  119.       0.       0.      0.    1000.
## 10 2015-01-01 01:30:00  121.       0.       0.      0.    1000.
## # ... with 52,551 more rows, and 2 more variables: Ics <dbl>,
## #   Ioh <dbl>
```

By using tidy data instead of the conventional data frame in R, the read results can be further manipulated with ease.[6] Furthermore, it is worth mentioning that the sugrrants package (Wang et al., 2017) introduces the calendar plots, which are suitable for visualizing time series during the data explorations stage. An example is given in the following plot.

```
# make the long table wider, i.e., stack the variables
# using the "tidyr" package
data2 <- data %>%
  tidyr::gather(group, irradiance, dw_solar:Ioh, -zen, -Time)
# prepare data for calendar plot
# note: calendar plot is not part of the package
# note: the below code is only for visualization in the paper
# note: and we need to install and load the "suggrants" package
devtools::install_github("earowang/sugrrants")
library(sugrrants)
#prepare data for calendar plot
data.plot <- data2 %>%
  mutate(Date = as.Date(data2$Time)) %>% # convert time to date
  mutate(Tm=lubridate::hour(data2$Time) # convert time to decimal
         + lubridate::minute(data2$Time)/60) %>%
  filter(lubridate::month(data2$Time)==9) %>% # select September
  sugrrants::frame_calendar(x = Tm, y = irradiance, date = Date)
  # function "frame_calendar" prepares the data for plotting
```

```
# ggplot code
# draw two lines
p <- ggplot(data.plot) +
     geom_line(data = filter(data.plot, group == "dw_solar"),
               aes(x = .Tm, y = .irradiance, group = Date),
               size = 0.4, color = "black") +
     geom_line(data = filter(data.plot, group == "Ics"),
               aes(x = .Tm, y = .irradiance, group = Date),
               size = 0.5, linetype = 2, color = "red")
# use the prettify function to finalize the plot
sugrrants::prettify(p, size=2.5, label.padding=unit(0.15, "lines")) +
  theme(plot.margin = unit(c(0.5,0.6,0,0), "lines"),
        legend.position = "none")
```



5. SoDa Linke turbidity factor

5.1. Quick overview

The Solar Radiation Data (SoDa) service offers a collection of both paid and free solar radiation and solar-related data, see http://www.soda-pro.com/web-services for the list of web services. The gridded monthly Linke turbidity factor (LTF), as tiff maps, is one of the free

[5] The numbers assigned to agg in SURFRAD.read is in minutes, whereas the agg in OSMG.read is in second.

[6] Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning) (Wickham et al., 2014).

services, derived from daily global radiation data supplied by NASA (Remund et al., 2003). The SoDa version of the LTF maps have a resolution of 5 arc minute, or equivalently, 4320 and 2160 pixels in longitude and latitude directions, respectively. These 12 maps are compressed into 4 files, one for each trimester, and can be downloaded in zip format at http://www.soda-pro.com/help/general-knowledge/linke-turbidity-factor.

5.2. Potential usages

The primary usage for LTF is to calculate Ineichen–Perez clear-sky irradiance (Ineichen and Perez, 2002) for any location in the world. Although there are many more elaborate models, such as McClear (Lefèvre et al., 2013) or REST2 (Gueymard, 2008), the Ineichen–Perez model remains to be one of the most popular models due to its simplicity—it only requires the site's altitude (see Section 6) and Linke turbidity factor as model inputs. Furthermore, in several comparison studies, it is found to be among the best performing models (Zhandire, 2017; Reno et al., 2012; Ineichen, 2006). This dataset is also used in the popular Python library `pvlib` https://github.com/pvlib/pvlib-python.

5.3. Getting the data

Once the raw tiff images are downloaded and unzipped, the `LTF.get` function can be used to retrieve the monthly LTF values at those locations, by specifying the directory that stores the images. The following code segment exemplifies the usage by getting the values at three locations, and the output is displayed.

```
# retrieve the Linke turbidity for three points
# (10N, 100E), (0N, 100W), and (90N, and 0E)
lon <- c(100, -100, 0)
lat <- c(10, 0, 90)
LTF.get(lon, lat, directory = "YourImageDirectory")
```

```
##     (100, 10) (-100, 0) (0, 90)
## Jan      3.50      3.55    1.90
## Feb      3.60      3.95    1.90
## Mar      4.15      4.00    1.90
## Apr      4.45      3.80    2.00
## May      4.50      3.50    2.00
## Jun      4.55      3.65    2.05
## Jul      4.60      3.35    2.10
## Aug      4.70      3.30    2.10
## Sep      4.65      3.80    2.00
## Oct      4.40      3.75    1.95
## Nov      3.95      3.60    1.90
## Dec      3.90      3.55    1.90
```

6. NASA shuttle radar topography mission

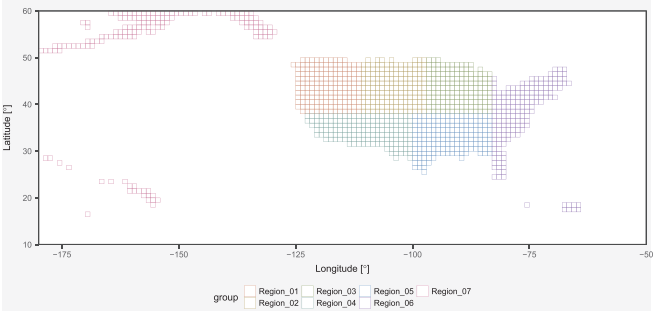
Although the Linke turbidity factor images described above are not irradiance data, they are essential in producing the Ineichen–Perez clear-sky irradiance expectations. Since the clear-sky model also requires altitude as input, in this section, NASA's shuttle radar topography mission (SRTM) dataset that provides worldwide altitude measurements is considered.

6.1. Quick overview

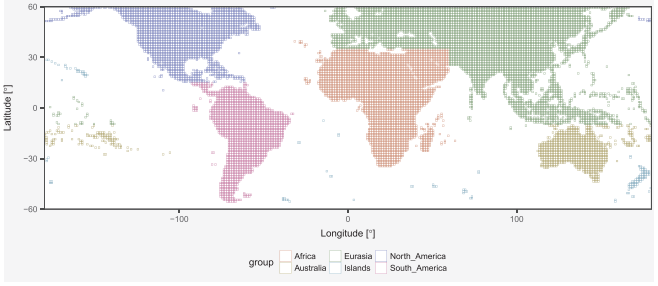
The SRTM dataset is a collaborative effort by NASA and the National Geospatial-Intelligence Agency (NGA), as well as the German and Italian space agencies. It is a near-global digital elevation model (DEM) of the Earth using radar interferometry (Farr and Kobrick, 2000). Since the raw data, i.e., version 1, produced in the year 2000 contain spurious data, the currently available version 2.1 is the results of a substantial editing effort by NGA. SRTM version 2.1 has two distribution levels: (1) SRTM1, with a resolution of 1 arcsec (30 m) in both longitude and latitude, covering the United States; and (2) SRTM3, with a resolution of 3 arcsec (90 m) covering the world. The reader is referred to the online documentation https://dds.cr.usgs.gov/srtm/ for more details.

SRTM files has an hgt extension, i.e., height files. The data is compressed and stored as "tiles". For example, file "N10W011.hgt.zip" stores the tile with a lower left corner at 10°N and 11°W. Since SRTM1 and SRTM3 have different resolutions, SRTM1 tiles are 3601 by 3601, whereas SRTM3 tiles are 1201 by 1201. Furthermore, the adjacent tiles have overlapping rows and columns at the joining edges. The directory, or the spatial stratification, of the tiles is somewhat confusing. Therefore, the `SRTM.list` function is written to help the user to visualize the tiles and manipulate the files names in R.

```
# Retrieve all available 1 arcsec tiles and plot
SRTM1 <- SRTM.list(resolution = 1, want.plot = TRUE)
```




```
# Retrieve all available 3 arcsec tiles and plot
SRTM3 <- SRTM.list(resolution = 3, want.plot = TRUE)
# zoom on the below plot to see the tiles
```



6.2. Potential usages

As an elevation dataset, SRTM is suitable for all geospatial applications. This dataset has been used for many solar engineering applications, such as clear-sky model assessment (Nemes, 2013), solar radiation estimation (Pons and Ninyerola, 2008), or PV performance estimation (Huld, 2017; Sabo et al., 2016; Huld et al., 2012). Since the incoming solar radiation depends on local terrain effects (Bosch et al., 2010; Ruiz-Arias et al., 2010), the SRTM dataset provides a third dimension in geospatial modeling, in addition to longitude and latitude.

6.3. Getting the data

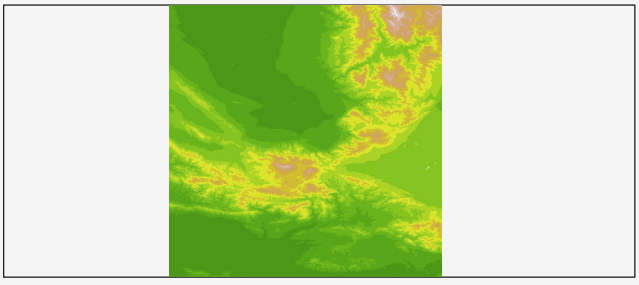
By using the `SRTM.list` function, the available tile names as well as their parent directory names can be retrieved. These names can then be used as input to the `SRTM.get` function to download and unzip the data. In the following example, four SRTM3 tiles from California are downloaded.

```
directory <- "YourDownloadDirectory"
files <- c("North_America/N34W119.hgt.zip",
           "North_America/N34W120.hgt.zip",
           "North_America/N35W119.hgt.zip",
           "North_America/N35W120.hgt.zip")
SRTM.get(resolution = 3, files, directory = directory)
```

6.4. Reading the data

Since the SRTM is raster data, the SRTM.read function is essentially a wrapper for several functions in the R package raster. This function reads and joins the tiles. The missing values are set as NA. The function parameter as.data.frame is a boolean variable indicating whether the elevation data should be outputted as a data frame. In default, it is set to FALSE, so that a raster object is generated. Assuming the above-mentioned four tiles have been downloaded, they can be quickly visualized using the following code segment. In the plot, green corresponds to lower elevation.

```
directory <- "YourDownloadDirectory"
setwd(directory)
files <- dir(pattern = ".hgt")
tiles <- SRTM.read(files, as.data.frame = FALSE)
# disable axis and legend for aesthetics
raster::plot(tiles, col = terrain.colors(12), axes=F, legend=F)
```



7. What's next?

The SolarData package, as well as the examples in the appendices, can be accessed from the GitHub page https://github.com/dazhiyang. However, since the list of good-quality freely available solar datasets goes far beyond the content of this version of the package, improving the package is an on-going effort. In this regard, I plan to update the package version whenever five new datasets are prepared. Interested individuals are cordially invited to contribute.[7] During each update, the discovered issues with the current version will also be addressed.

The naming convention of the functions in SolarData follows a "data.verb" format. The verbs such as list, get, or read are designed to help the user to obtain the data. Although some data cleaning routines have been considered, the missing data points are not handled. Therefore, other verbs such as fill or complete can be considered in future versions.

This paper is the first *Data Article*, a new submission type offered by the Solar Energy journal. Owing to the exploratory nature of this paper, future Data Article submissions can vary from this one. For detailed expectation and instruction, the reader is referred to the editorial of the current issue (Yang et al., 2018a).

Conflict of interest

None.

Appendix A. PSM example use case: Regression-based site adaptation

Satellite-based irradiance data are often biased, and site-adaptation methods can be used to correct such bias (Polo et al., 2016). In this appendix, a linear adaptation technique is considered in attempting to remove the systematic bias in the PSM data. The ground reference data comes from the SURFRAD station, Desert Rock, Nevada. The R script for this use case can be found in the "examples" folder at https://github.com/dazhiyang/SolarData.

[7] Major contributors will be invited to be the lead, corresponding, or co-author, depending on the contribution.

Appendix B. OSMG example use case: Spatio-temporal correlation analysis

Studying the spatio-temporal correlation in a solar irradiance random field has two main purposes, namely, synthetic data generation and prediction (Yang et al., 2013). For example, in spatial statistics, the weights of the optimal interpolation, i.e., kriging, can be calculated using correlations. Since the correlation between two arbitrary spatio-temporal indexes needs to be estimated, a correlation function is required. In the present discussion, a correlation function maps a geographical distance to a correlation coefficient. To that end, several correlation functions for solar irradiance have been proposed (Lonij et al., 2013; Lave and Kleissl, 2013; Perez et al., 2011). In a later work, the previously developed correlation models were consolidated and a new anisotropic correlation model was proposed (Arias-Castro et al., 2014). The empirical part of Arias-Castro et al. (2014) considered the OSMG dataset. Hence, this appendix reproduces some results of that paper. The R script OSMGExample.r can be found in the folder mentioned in Appendix A.

Appendix C. SURFRAD example use case: Irradiance component separation modeling

Predicting solar radiation on inclined surfaces require two classes of models, namely, transposition models and separation models. Whereas transposition models—see Yang (2016) for a review—convert horizontal irradiance components to tilted irradiance components, separation models estimate DIF and DNI from GHI. In a worldwide comparison study conducted by Gueymard and Ruiz-Arias (2016), 140 separation models were validated at 54 research-grade stations. Although there was no universal model found, the "quasi-universal" Engerer2 model (Engerer, 2015) was recommended for general use for 1-min data. Since SURFRAD has a 1-min resolution, this appendix provides an example of irradiance component separation modeling using Engerer2. A simplistic but well-utilized model, namely, the Erbs model (Erbs et al., 1982), is also implemented for benchmarking purposes. The R script is named SURFRADExample.r.

Appendix D. Supplementary material

Each R package has its own package documentation. Therefore, the first version of the package documentation is provided as supplementary material. However, it is noted that the SolarData package is an on-going effort. This documentation is thus subject to frequent minor changes and corrections. For the most recent version, see https://github.com/dazhiyang/SolarData.

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.solener.2018.06.107.

References

Arias-Castro, E., Kleissl, J., Lave, M., 2014. A Poisson model for anisotropic solar ramp rate correlations. Sol. Energy 101, 192–202. https://doi.org/10.1016/j.solener.2013.12.028. <http://www.sciencedirect.com/science/article/pii/S0038092X13005549>.

Aryaputera, A.W., Yang, D., Zhao, L., Walsh, W.M., 2015. Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. Sol. Energy 122, 1266–1278. https://doi.org/10.1016/j.solener.2015.10.023. <http://www.sciencedirect.com/science/article/pii/S0038092X15005745>.

Augustine, J.A., DeLuisi, J.J., Long, C.N., 2000. SURFRAD—A national surface radiation budget network for atmospheric research. Bull. Am. Meteorol. Soc. 81, 2341–2358. https://doi.org/10.1175/1520-0477(2000)081<2341:SANSRB>2.3.CO;2.

Augustine, J.A., Hodges, G.B., Cornwall, C.R., Michalsky, J.J., Medina, C.I., 2005. An update on SURFRAD—the GCOS surface radiation budget network for the continental united states. J. Atmos. Ocean. Technol. 22, 1460–1472. https://doi.org/10.1175/JTECH1806.1.

Ayet, A., Tandeo, P., 2018. Nowcasting solar irradiance using an analog method and geostationary satellite images. Sol. Energy 164, 301–315. https://doi.org/10.1016/j.solener.2018.02.068. <http://www.sciencedirect.com/science/article/pii/S0038092X18301993>.

Blanc, P., Remund, J., Vallance, L., 2017. Short-term solar power forecasting based on satellite images. In: Kariniotakis, G. (Ed.), Renewable Energy Forecasting. Woodhead Publishing Series in Energy. Woodhead Publishing, pp. 179–198. https://doi.org/10.1016/B978-0-08-100504-0.00006-8. <https://www.sciencedirect.com/science/article/pii/B9780081005040000068>.

Bojanowski, J.S., 2016. sirad: Functions for Calculating Daily Solar Radiation and Evapotranspiration. r package version 2.3-3. < https://CRAN.R-project.org/package=sirad > .

Bosch, J., Batlles, F., Zarzalejo, L., López, G., 2010. Solar resources estimation combining digital terrain models and satellite images techniques. Renew. Energy 35, 2853–2861. https://doi.org/10.1016/j.renene.2010.05.011. <http://www.sciencedirect.com/science/article/pii/S0960148110002296>.

Corripio, J.G., 2014. insol: Solar Radiation. r package version 1.1.1. < https://CRAN.R-project.org/package=insol > .

Cressie, N., Wikle, C.K., 2015. Statistics for Spatio-temporal Data. John Wiley & Sons.

De Cáceres, M., Martin, N., Granda, V., Cabon, A., 2018. meteoland: Landscape Meteorology Tools. r package version 0.7.1. < https://CRAN.R-project.org/package=meteoland > .

Engerer, N., 2015. Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia. Sol. Energy 116, 215–237. https://doi.org/10.1016/j.solener.2015.04.012. <http://www.sciencedirect.com/science/article/pii/S0038092X15001905>.

Erbs, D., Klein, S., Duffie, J., 1982. Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. Sol. Energy 28, 293–302. https://doi.org/10.1016/0038-092X(82)90302-4. <http://www.sciencedirect.com/science/article/pii/0038092X82903024>.

Farr, T.G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. Eos, Trans. Am. Geophys. Union 81, 583–585. https://doi.org/10.1029/EO081i048p00583. <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/EO081i048p00583>.

Franch, B., Vermote, E., Claverie, M., 2014. Intercomparison of landsat albedo retrieval techniques and evaluation against in situ measurements across the US SURFRAD network. Remote Sens. Environ. 152, 627–637. https://doi.org/10.1016/j.rse.2014.07.019. <http://www.sciencedirect.com/science/article/pii/S0034425714002685>.

Gueymard, C.A., 2008. REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation – Validation with a benchmark dataset. Sol. Energy 82, 272–285. https://doi.org/10.1016/j.solener.2007.04.008. <http://www.sciencedirect.com/science/article/pii/S0038092X07000990>.

Gueymard, C.A., Ruiz-Arias, J.A., 2016. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. Sol. Energy 128, 1–30. https://doi.org/10.1016/j.solener.2015.10.010. (special issue: Progress in Solar Energy). <http://www.sciencedirect.com/science/article/pii/S0038092X15005435>.

Habte, A., Sengupta, M., Lopez, A., 2017. Evaluation of the National Solar Radiation Database (NSRDB): 1998-2015. Technical Report NREL/TP-5D00-67722. National Renewable Energy Laboratory, Golden, CO, United States.

Heidinger, A.K., Laszlo, I., Molling, C.C., Tarpley, D., 2013. Using SURFRAD to verify the NOAA single-channel land surface temperature algorithm. J. Atmos. Ocean. Technol. 30, 2868–2884. https://doi.org/10.1175/JTECH-D-13-00051.1.

Hinkelman, L.M., 2013. Differences between along-wind and cross-wind solar irradiance variability on small spatial scales. Sol. Energy 88, 192–203. https://doi.org/10.1016/j.solener.2012.11.011. <http://www.sciencedirect.com/science/article/pii/S0038092X12004021>.

Huld, T., 2017. PVMAPS: software tools and data for the estimation of solar radiation and photovoltaic module performance over large geographical areas. Sol. Energy 142, 171–181. https://doi.org/10.1016/j.solener.2016.12.014. <http://www.sciencedirect.com/science/article/pii/S0038092X16306089>.

Huld, T., Müller, R., Gambardella, A., 2012. A new solar radiation database for estimating PV performance in Europe and Africa. Sol. Energy 86, 1803–1815. https://doi.org/10.1016/j.solener.2012.03.006. <http://www.sciencedirect.com/science/article/pii/S0038092X12001119>.

Iannone, R., 2015. stationaRy: Get Hourly Meteorological Data from Global Stations. r package version 0.4.1. < https://CRAN.R-project.org/package=stationaRy > .

Ineichen, P., 2006. Comparison of eight clear sky broadband models against 16 independent data banks. Sol. Energy 80, 468–478. https://doi.org/10.1016/j.solener.2005.04.018. (urban Ventilation). <http://www.sciencedirect.com/science/article/pii/S0038092X05001635>.

Ineichen, P., Perez, R., 2002. A new airmass independent formulation for the Linke turbidity coefficient. Sol. Energy 73, 151–157. https://doi.org/10.1016/S0038-092X(02)00045-2. <http://www.sciencedirect.com/science/article/pii/S0038092X02000452>.

Killinger, S., Engerer, N., Müller, B., 2017. QCPV: A quality control algorithm for distributed photovoltaic array power output. Sol. Energy 143, 120–131. https://doi.org/10.1016/j.solener.2016.12.053. <http://www.sciencedirect.com/science/article/pii/S0038092X16306600>.

Kleissl, J., 2013. Solar Energy Forecasting and Resource Assessment. Academic Press.

Lauret, P., Perez, R., Aguiar, L.M., Tapachès, E., Diagne, H.M., David, M., 2016. Characterization of the intraday variability regime of solar irradiation of climatically distinct locations. Sol. Energy 125, 99–110. https://doi.org/10.1016/j.solener.2015.11.032. <http://www.sciencedirect.com/science/article/pii/S0038092X15006490>.

Lave, M., Broderick, R.J., Reno, M.J., 2017. Solar variability zones: satellite-derived zones that represent high-frequency ground variability. Sol. Energy 151, 119–128. https://doi.org/10.1016/j.solener.2017.05.005. <http://www.sciencedirect.com/science/article/pii/S0038092X17303821>.

Lave, M., Kleissl, J., 2013. Cloud speed impact on solar variability scaling – Application to the wavelet variability model. Sol. Energy 91, 11–21. https://doi.org/10.1016/j.solener.2013.01.023. <http://www.sciencedirect.com/science/article/pii/S0038092X13000406>.

Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B., Gschwind, B., Qu, Z., Wald, L., Schroedter-Homscheidt, M., Hoyer-Klick, C., Arola, A., Benedetti, A., Kaiser, J.W., Morcrette, J.J., 2013. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmos. Meas. Techn. 6, 2403–2418. https://doi.org/10.5194/amt-6-2403-2013. <https://www.atmos-meas-tech.net/6/2403/2013/>.

Long, C.N., Dutton, E.G., 2002. BSRN Global Network Recommended QC Rests, V2. Technical Report. BSRN.

Long, C.N., Shi, Y., 2008. An automated quality assessment and control algorithm for surface radiation measurements. Open Atmos. Sci. J. 2, 23–37.

Lonij, V.P., Brooks, A.E., Cronin, A.D., Leuthold, M., Koch, K., 2013. Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. Sol. Energy 97, 58–66. https://doi.org/10.1016/j.solener.2013.08.002. <http://www.sciencedirect.com/science/article/pii/S0038092X13003125>.

Lorenzo, A.T., Morzfeld, M., Holmgren, W.F., Cronin, A.D., 2017. Optimal interpolation of satellite and ground data for irradiance nowcasting at city scales. Sol. Energy 144, 466–474. https://doi.org/10.1016/j.solener.2017.01.038. <http://www.sciencedirect.com/science/article/pii/S0038092X17300555>.

Marion, W., Urban, K., 1995. User's Manual for TMY2s. Technical Report. National Renewable Energy Laboratory, Golden, CO, United States. <http://rredc.nrel.gov/solar/pubs/tmy2/>.

Martín-Pomares, L., Martínez, D., Polo, J., Perez-Astudillo, D., Bachour, D., Sanfilippo, A., 2017. Analysis of the long-term solar potential for electricity generation in Qatar. Renew. Sustain. Energy Rev. 73, 1231–1246. https://doi.org/10.1016/j.rser.2017.01.125. <http://www.sciencedirect.com/science/article/pii/S136403211730134X>.

Mathiesen, P., Kleissl, J., 2011. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. Sol. Energy 85, 967–977. https://doi.org/10.1016/j.solener.2011.02.013. <http://www.sciencedirect.com/science/article/pii/S0038092X11000570>.

Munkhammar, J., Widén, J., Hinkelman, L.M., 2017. A copula method for simulating correlated instantaneous solar irradiance in spatial networks. Sol. Energy 143, 10–21. https://doi.org/10.1016/j.solener.2016.12.022. <http://www.sciencedirect.com/science/article/pii/S0038092X16306168>.

Nemes, C., 2013. A clear sky irradiation assessment using the European Solar Radiation Atlas model and Shuttle Radar Topography Mission database: A case study for Romanian territory. J. Renew. Sustain. Energy 5, 041807. https://doi.org/10.1063/1.4813001.

Perez, R., Ineichen, P., Kmiecik, M., Moore, K., Renne, D., George, R., 2004. Producing satellite-derived irradiances in complex arid terrain. Sol. Energy 77, 367–371. https://doi.org/10.1016/j.solener.2003.12.016. (the American Solar Energy Society's Solar 2003 Special Issue). <http://www.sciencedirect.com/science/article/pii/S0038092X03004687>.

Perez, R., Kivalov, S., Schlemmer, J., Hemker, K., Hoff, T., 2011. Parameterization of site-specific short-term irradiance variability. Sol. Energy 85, 1343–1353. https://doi.org/10.1016/j.solener.2011.03.016. <http://www.sciencedirect.com/science/article/pii/S0038092X11000995>.

Perez, R., Kivalov, S., Schlemmer, J., Hemker, K., Hoff, T.E., 2012. Short-term irradiance variability: preliminary estimation of station pair correlation as a function of distance. Sol. Energy 86, 2170–2176. https://doi.org/10.1016/j.solener.2012.02.027. (progress in Solar Energy 3). <http://www.sciencedirect.com/science/article/pii/S0038092X12000928>.

Perez, R., Kivalov, S., Schlemmer, J., Hemker, K., Renné, D., Hoff, T.E., 2010. Validation of short and medium term operational solar radiation forecasts in the US. Sol. Energy 84, 2161–2172. https://doi.org/10.1016/j.solener.2010.08.014. <http://www.sciencedirect.com/science/article/pii/S0038092X10002823>.

Perez, R., Schlemmer, J., Hemker, K., Kivalov, S., Kankiewicz, A., Gueymard, C., 2015. Satellite-to-irradiance modeling – A new version of the SUNY model. In: 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC), pp. 1–7. https://doi.org/10.1109/PVSC.2015.7356212.

Perpiñán, O., 2012. solaR: solar radiation and photovoltaic systems with R. J. Stat. Softw. 50, 1–32. <http://www.jstatsoft.org/v50/i09/>.

Perpiñán, Ó., 2014. Displaying Time Series, Spatial, and Space-time Data with R. CRC Press.

Perpiñán, O., Almeida, M.P., 2018. meteoForecast. r package version 0.52. < https://github.com/oscarperpinan/meteoForecast/ > .

Polo, J., Wilbert, S., Ruiz-Arias, J., Meyer, R., Gueymard, C., Súri, M., Martín, L., Mieslinger, T., Blanc, P., Grant, I., Boland, J., Ineichen, P., Remund, J., Escobar, R., Troccoli, A., Sengupta, M., Nielsen, K., Renne, D., Geuder, N., Cebecauer, T., 2016. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Sol. Energy 132, 25–37. https://doi.org/10.1016/j.solener.2016.03.001. <http://www.sciencedirect.com/science/article/pii/S0038092X16001754>.

Pons, X., Ninyerola, M., 2008. Mapping a topographic global solar radiation model implemented in a GIS and refined with ground data. Int. J. Climatol. 28, 1821–1834. https://doi.org/10.1002/joc.1676. <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.1676>.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Remund, J., Wald, L., Lefèvre, M., Ranchin, T., Page, J., 2003. Worldwide linke turbidity information. In: ISES Solar World Congress 2003. International Solar Energy Society

(ISES).
Reno, M.J., Hansen, C.W., Stein, J.S., 2012. Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis. Technical Report SAND2012-2389. SANDIA Lab.
Rodríguez-Gallegos, C.D., Gandhi, O., Yang, D., Alvarez-Alvarado, M.S., Zhang, W., Reindl, T., Panda, S.K., 2018. A siting and sizing optimization approach for PV–battery–diesel hybrid systems. IEEE Trans. Ind. Appl. 54, 2637–2645. https://doi.org/10.1109/TIA.2017.2787680.
Ruiz-Arias, J., Cebecauer, T., Tovar-Pescador, J., Šúri, M., 2010. Spatial disaggregation of satellite-derived irradiance using a high-resolution digital elevation model. Sol. Energy 84, 1644–1657. https://doi.org/10.1016/j.solener.2010.06.002. <http://www.sciencedirect.com/science/article/pii/S0038092X10002136>.
Sabo, M.L., Mariun, N., Hizam, H., Radzi, M.A.M., Zakaria, A., 2016. Spatial energy predictions from large-scale photovoltaic power plants located in optimal sites and connected to a smart grid in Peninsular Malaysia. Renew. Sustain. Energy Rev. 66, 79–94. https://doi.org/10.1016/j.rser.2016.07.045. <http://www.sciencedirect.com/science/article/pii/S1364032116303732>.
Schade, N.H., Macke, A., Sandmann, H., Stick, C., 2007. Enhanced solar global irradiance during cloudy sky conditions. Meteorol. Z. 16, 295–303.
Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., Shelby, J., 2018. The national solar radiation data base (NSRDB). Renew. Sustain. Energy Rev. 89, 51–60. https://doi.org/10.1016/j.rser.2018.03.003. <http://www.sciencedirect.com/science/article/pii/S136403211830087X>.
e Silva, R.A., Brito, M.C., 2018. Impact of network layout and time resolution on spatio-temporal solar forecasting. Sol. Energy 163, 329–337. https://doi.org/10.1016/j.solener.2018.01.095. <http://www.sciencedirect.com/science/article/pii/S0038092X18301166>.
Vignola, F., Michalsky, J., Stoffel, T., 2016. Solar and Infrared Radiation Measurements. CRC Press.
Wang, E., Cook, D., Hyndman, R., 2017. sugrrants: Supporting Graphs for Analysing Time Series. r package version 0.1.1. < https://CRAN.R-project.org/package=sugrrants > .
Wang, K., Liang, S., 2009. Evaluation of ASTER and MODIS land surface temperature and emissivity products using long-term surface longwave radiation observations at SURFRAD sites. Remote Sens. Environ. 113, 1556–1565. https://doi.org/10.1016/j.rse.2009.03.009. (monitoring Protected Areas). <http://www.sciencedirect.com/science/article/pii/S0034425709000881>.
Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. <http://ggplot2.org>.
Wickham, H., et al., 2014. Tidy Data. J. Stat. Softw. 59, 1–23.
Wilcox, S., Marion, W., 2008. Users Manual for TMY3 Data Sets. Technical Report NREL/TP-581-43156. National Renewable Energy Laboratory, Golden, CO, United States. <https://www.nrel.gov/docs/fy08osti/43156.pdf>.
Xie, Y., Sengupta, M., Habte, A., Lopez, A., 2017. Evaluation of the National Solar Radiation Database (NSRDB) using ground-based measurements. In: AGU Fall Meeting Abstracts.
Yang, D., 2016. Solar radiation on inclined surfaces: Corrections and benchmarks. Sol. Energy 136, 288–302. https://doi.org/10.1016/j.solener.2016.06.062. <http://www.sciencedirect.com/science/article/pii/S0038092X16302432>.
Yang, D., 2017. On adding and removing sensors in a solar irradiance monitoring network for areal forecasting and PV system performance evaluation. Sol. Energy 155, 1417–1430. https://doi.org/10.1016/j.solener.2017.07.061. <http://www.sciencedirect.com/science/article/pii/S0038092X17306461>.
Yang, D., 2018a. Kriging for NSRDB PSM version 3 satellite-derived irradiance. Sol. Energy 171, 876–883. https://doi.org/10.1016/j.solener.2018.06.055 <https://www.sciencedirect.com/science/article/pii/S0038092X18306066>.
Yang, D., 2018b. Spatial prediction using kriging ensemble. Sol. Energy 171, 977–982. https://doi.org/10.1016/j.solener.2018.06.105 < http://www.sciencedirect.com/science/article/pii/S0960148113002759 > .
Yang, D., Gu, C., Dong, Z., Jirutitijaroen, P., Chen, N., Walsh, W.M., 2013. Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging. Renew. Energy 60, 235–245. https://doi.org/10.1016/j.renene.2013.05.030. <http://www.sciencedirect.com/science/article/pii/S0960148113002759>.
Yang, D., Gueymard, C.A., Kleissl, J., 2018a. Editorial: submission of Data Article is now open. Sol. Energy 1–2. https://doi.org/10.1016/j.solener.2018.07.006.
Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T.C., Coimbra, C.F.M., 2018b. History and trends in solar irradiance and PV power forecasting: a preliminary assessment and review using text mining. Sol. Energy 168, 60–101. https://doi.org/10.1016/j.solener.2017.11.023. <http://www.sciencedirect.com/science/article/pii/S0038092X17310022>.
Yang, D., Quan, H., Disfani, V.R., Liu, L., 2017. Reconciling solar forecasts: geographical hierarchy. Sol. Energy 146, 276–286. https://doi.org/10.1016/j.solener.2017.02.010. <http://www.sciencedirect.com/science/article/pii/S0038092X17301020>.
Yang, D., Reindl, T., 2015. Solar irradiance monitoring network design using the variance quadtree algorithm. Renew.: Wind, Water, Sol. 2, 1–8. https://doi.org/10.1186/s40807-014-0001-x.
Yang, D., Ye, Z., Lim, L.H.I., Dong, Z., 2015. Very short term irradiance forecasting using the lasso. Sol. Energy 114, 314–326. https://doi.org/10.1016/j.solener.2015.01.016. <http://www.sciencedirect.com/science/article/pii/S0038092X15000304>.
Yu, Y., Tarpley, D., Privette, J.L., Flynn, L.E., Xu, H., Chen, M., Vinnikov, K.Y., Sun, D., Tian, Y., 2012. Validation of GOES-R satellite land surface temperature algorithm using SURFRAD ground measurements and statistical estimates of error properties. IEEE Trans. Geosci. Remote Sens. 50, 704–713. https://doi.org/10.1109/TGRS.2011.2162338.
Zagouras, A., Inman, R.H., Coimbra, C.F., 2014. On the determination of coherent solar microclimates for utility planning and operations. Sol. Energy 102, 173–188. https://doi.org/10.1016/j.solener.2014.01.021. <http://www.sciencedirect.com/science/article/pii/S0038092X14000395>.
Zagouras, A., Kazantzidis, A., Nikitidou, E., Argiriou, A., 2013. Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. Sol. Energy 97, 1–11. https://doi.org/10.1016/j.solener.2013.08.005. <http://www.sciencedirect.com/science/article/pii/S0038092X13003150>.
Zhandire, E., 2017. Predicting clear-sky global horizontal irradiance at eight locations in South Africa using four models. J. Energy Southern Africa 28, 77–86.