Guidance on Weather Datasets for Studies

Data Sources and Use Cases for Historical Datasets

3002024375

Guidance on Weather Datasets for Studies

Data Sources and Use Cases for Historical Datasets

3002024375

Technical Update, January 2023

EPRI Project Manager

D. Larson

EPRI
3420 Hillview Avenue, Palo Alto, California 94304-1338 ▪ USA
800.313.3774 ▪ 650.855.2121 ▪ askepri@epri.com ▪ www.epri.com

EPRI PREPARED THIS REPORT.

This is an EPRI Technical Update report. A Technical Update report is intended as an informal report of continuing research, a meeting, or a topical study. It is not a final EPRI technical report.

NOTE

For further information about EPRI, call the EPRI Customer Assistance Center at 800.313.3774 or e-mail askepri@epri.com.

Together…Shaping the Future of Energy®

ACKNOWLEDGMENTS

EPRI prepared this report.

Principal Investigator
D. Larson

Project Team
S. Liu
A. Tuohy
A. Wigington
Y. Yin

This report describes research sponsored by EPRI.

This publication is a corporate document that should be cited in the literature in the following manner:

Guidance on Weather Datasets for Studies: Data Sources and Use Cases for Historical Datasets. EPRI, Palo Alto, CA: 2023. 3002024375.

ABSTRACT

Weather data is crucial to many applications in the electric power industry, including for system operations and planning processes involving variable renewable energy (VRE). However, weather datasets can vary in quality, resolution, coverage, etc., which can impact application outcomes. Therefore, it is important that consumers of weather data be informed regarding datasets and their attributes. This document provides an overview of publicly available weather datasets and practical guidance collected for staff across the industry to utilize, including those without backgrounds in meteorology.

Keywords

Weather data
Time-series
Observation
Reanalysis
Typical meteorological year
Solar irradiance

ACRONYMS AND ABBREVIATIONS

20CR: 20th Century Reanalysis

API: Application Programming Interface

ARM: Atmospheric Radiation Measurement user facility

ASOS: Automated Surface Observing Systems

AWOS: Automated Weather Observing Systems

BSRN: Baseline Surface Radiation Network

CFSR: Climate Forecast System Reanalysis

DHI: diffuse horizontal irradiance

DNI: direct normal irradiance

DOE: U.S. Department of Energy

ECMWF: European Centre for Medium-Range Weather Forecasts

ERA5: ECMWF Reanalysis v5

ESRL: Earth System Research Laboratories

FTP: File Transfer Protocol

GHI: global horizontal irradiance

HTTP: Hypertext Transfer Protocol

LiDAR: Light Detection and Ranging

MassCEC: Massachusetts Clean Energy Center

MERRA2: Modern-Era Retrospective Analysis for Research and Applications v2

MIDC: Measurement and Instrumentation Data Center

NARR: North American Regional Reanalysis

NCAR: National Center for Atmospheric Research

NCEP: National Centers for Environmental Protection

NDBC: National Data Buoy Center

NOAA: National Oceanic and Atmospheric Administration

NREL: National Renewable Energy Laboratory

NSRDB: National Solar Radiation Database

PV: photovoltaic

RH: relative humidity

SOLRAD: Solar Radiation Network

SUMR: Segmented Ultralight Morphing Rotor

SURFRAD: Surface Radiation Budget Network

TMY: typical meteorological year

UO SRML: University of Oregon Solar Radiation Monitoring Laboratory

USCRN: U.S. Climate Reference Network

WFIP1: Wind Forecast Improvement Project 1

WFIP2: Wind Forecast Improvement Project 2

WRMC: World Radiation Monitoring enter

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Many applications across the electric power industry require weather data, including historical datasets describing weather conditions observed in the past, live forecasts of conditions in the coming days, or simulated future weather based on climate trends. For example, power system operators may require solar irradiance data (transformed appropriately to power output) to estimate behind-the-meter photovoltaic (PV) power generation when direct metering is unavailable, while a planning study may require wind speed data to predict the power output and subsequent impacts of proposed wind farms which have not yet been built. Regardless of the specific use case, the choice of data source and therefore the quality, resolution, etc. of the weather dataset can impact outcomes. Therefore, consumers of weather data should be aware of available datasets and their attributes so they can then make informed decisions. This data can be publicly available, the focus of this report, or available from commercial firms.

Types of Publicly Available Weather Datasets Examined Here

There are three categories of weather datasets covered by this document:

1. **Observation Datasets:** measurements of weather variables from sensors at specific locations, e.g., ambient temperature and humidity from a weather station at an airport
2. **Reanalysis Datasets:** estimates of weather variables from gridded numerical models, which have been postprocessed to better match observations
3. **Typical Meteorological Year (TMY) Datasets:** single year datasets, e.g., an 8760 hourly timeseries, that were created from multiple years of observations such that the single year is representative of "typical" weather conditions

These three categories cover historical data and do not include any forward-looking datasets, e.g., day-ahead weather forecasts or climate models. While forward-looking weather datasets are increasingly important for several applications, such datasets are outside the scope of this document. Additionally, commercially available datasets, i.e., datasets purchased from a data provider, are also outside the scope of this document.

Types of Data Sources

Weather datasets can be accessed in multiple ways, which can be generally grouped into three categories of data sources:

1. **File Server:** A file server is a computer connected to a network that provides access to data stored as files, e.g., one CSV file per day of data. Users access a file server using a communication protocol, which may or may not require authentication. Two of the most common protocols are File Transfer Protocol (FTP) and Hypertext Transfer Protocol (HTTP), or their encrypted variants Secure File Transfer Protocol (SFTP) and Hypertext Transfer Protocol Secure (HTTPS). Transferring data from a file server may be as simple as opening a URL in a web browser or may require specialized software.

2. **Application Programming Interface (API):** An API can be thought of as a communication protocol between two computers, e.g., a server storing weather data and a user requesting access to the data. The main difference from a file server is that interacting with the data source is done programmatically, e.g., a Python script. An analogy is an API is like a waiter in a restaurant: the customer (client) tells the waiter (API) what food (data) they want, the waiter (API) gives the order to the chef (server), and when the food (data) is ready, the waiter (API) brings it to the customer (client). In other words, an API allows a user to request what they want and then the API returns the requested data, all without the user needing to know or worry about how the data was prepared.
3. **Cloud Storage:** Cloud storage can be thought of as a file server hosted using a cloud computing platform such as Amazon Web Service (AWS), Google Cloud Platform (GCP) or Microsoft Azure. As with a file server, accessing data may be as simple as opening a URL in a web browser or may require specialized software. Additionally, some cloud storage providers have APIs.

There are similarities between the three categories, but consumers of weather data likely only need to recognize that there are differences in how data is accessed. For example, a user without programming experience should recognize that a data source available via a HTTPS file server could be accessed using their web browser, whereas another data source available via an API would require the involvement of other staff with programming experience. However, there are a few additional key points to keep in mind:

- APIs and cloud storage are relatively new, and therefore many weather datasets are still primarily (or only) accessed via file servers.
- Many organizations have security restrictions that block certain communication protocols used by file servers. For example, FTP is commonly blocked, but many file servers still use FTP. However, an increasing number of file servers provide HTTPS access, which is more secure, either in addition to or in place of FTP.
- APIs can enable more flexible data access, depending on how the API is implemented. For example, a file server may require downloading a data file that contains an entire year of data, whereas an API may allow requesting data for a specific time range, e.g., temperature data for September 6–12, 2022.
- APIs are well suited for applications requiring automated data transfer but accessing data on some file servers can also be automated.
- APIs commonly use access tokens for authentication, where a user registers for access and is given an alphanumeric string (the token) that gets included with any requests to the API, similar to a password.
- APIs can have rate limits, e.g., you can only submit X requests every hour.
- API access may be free or require a fee, e.g., a monthly subscription or a fee that scales with the number of requests made.
- While file servers and cloud storage can provide the same user experience, cloud storage can enable faster data downloads and provide data access redundancy.
- Cloud storage tends to be the preferred option for large datasets and there are cloud-specific technologies that enable efficient data transfer. For example, a reanalysis dataset is typically stored in binary formats such as GRIB (.grib) or netCDF (.nc), with one file containing data

over an entire spatial grid, e.g., a 10 km x 10 km grid over the entire globe. Therefore, a user must download an entire netCDF file even if they only want a subset of the grid, e.g., all grid points over California. But if the same data is stored in the Zarr format, then a user can download only their desired subset of grid points, resulting faster downloads and less data stored.

How to Use This Document

The purpose of this document is to provide an overview of publicly available weather datasets relevant to the electric power industry. Chapters 2–4 cover observation, reanalysis and TMY datasets, respectively, with chapters broken down into sections based on topics of interest. Chapter 5 then provides guidance on weather datasets for common applications, including notes on practical aspects. Readers can review the entire document to improve their general knowledge of weather datasets. Alternatively, readers can use the document as a reference when needing to answer specific questions, e.g., what public datasets for offshore wind speed data near the northeast Atlantic coast.

2
OBSERVATIONS

Observation datasets provide measurements of weather variables from sensors at specific locations. For ease of reference, this chapter groups observation datasets into four sections based on weather variables: (1) solar irradiance, (2) onshore wind speed and direction, (3) offshore wind speed and direction, and (4) temperature and humidity. Datasets are identified by relevant acronyms, with key attributes reported in each section and then details regarding accessing each dataset provided at the end of the chapter. Note that some datasets are listed in multiple sections as some (but not all) datasets provide observations of multiple variables.

Solar Irradiance Datasets

Solar irradiance is the power per unit area received from the Sun at the Earth's surface, typically in units of watts per square meter [W/m^2]. Solar irradiance can be measured using a sensor, e.g., a pyrometer at surface level or by the satellites. There are multiple types of solar irradiance measurements, however the main types relevant to the electric power industry are Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI).

Table 2-1 is a summary of common solar irradiance observation datasets, considering several key factors, including region, time resolution, spatial resolution, and time range. The availability of each dataset in Table 2-1 depends on the site location, the year, the time resolution, or the data level. GHI, DNI, and DHI are available in all the datasets except for NOAA USCRN (GHI only).

Table 2-1
Summary of the commonly used solar irradiance observation datasets.

Dataset	Region	Time Resolution	Spatial Resolution	Time Range
DOE ARM[1]	Global	1 min – 1 hour	Point (~150 sites)	1993–present
NOAA SOLRAD[2]	Regional - US (WA, CA, UT, NM, ND, WI, VA, TN, FL)	1–3 min	Point (9 sites)	1995–present
NOAA SURFRAD[3]	Regional - US (MT, CO, IL, MS, PA, NV, SD, OK, KS)	1–3 min	Point (8 sites)	1993–present
NOAA USCRN[4]	Regional - North America, HI	5 min – 1 month	Point (139 sites)	2000–present
NOAA WFIP2	Regional - US (OR, WA)	1–5 min	Point (10 sites)	2015–2017
NREL MIDC	Regional - US (CONUS, HI, PR)	1–5 min	Point (35 sites)	1981–present
NREL NSRDB	Global	5 min – 1 hour	Grid (2–10 km)	1998–2021
UO SRML[5]	Regional - US (ID, MT, OR, UT, WA, WY)	1 min – 1 hour	Point (38 sites)	1979–present
WRMC BSRN[6]	Global	1 min	Point (76 sites)	1992–present

Onshore Wind Speed and Direction Datasets

Table 2-2 summarizes observation datasets of onshore wind speed and direction, including the above ground level (AGL) height, i.e., the vertical distance between the surface and where the wind speed is measured. AGL height is influenced by the equipment or sensor used to measure wind speed, e.g., pitot tube, laser, cup and propellor. A LiDAR sensor, for instance, may detect wind speed up to a few hundreds of meters, whereas a cup anemometer mounted on a tower in a surface meteorological station typically monitors wind speed from a few meters to a few tens of meters. The data source has an impact on the height at which wind speed is measured as well. For instance, the way data is measured varies especially when it is gathered from multiple sources, e.g., a shared network or temporary mobile facilities.

In Table 2-2, wind speed and direction in most of the datasets are measured in meters per second [m/s] and degrees, while the unit can vary by the data source. For instance, ASOS/AWOS

[1] Most of the sites are in the US (especially the Southern Great Plains). Sites include fixed sites and mobile facilities.

[2] Time resolution depends on the year: 3 min before 2015, and 1 min on and after 2015.

[3] Sites are overlapping with DOE ARM (73/80 sites from DOE ARM). Time resolution depends on the year: 3 min before 2009, and 1 min on and after 2009.

[4] Most of the states are covered. Only 1 site in Canada. GHI only.

[5] Summary data (monthly average, daily total, and cumulative summary) is available.

[6] 10 sites in the US.

datasets accessed from the archive of Iowa State University has the option to report wind speed in meters per second, knots, or miles per hour (mph). All the datasets include wind speed and direction measurements, except for NOAA USCRN (wind speed only).

Table 2-2
Summary of the commonly used onshore wind speed and direction observation datasets.

Dataset	Height (AGL)	Region	Time Resolution	Spatial Resolution	Time Range
DOE ARM	Depends on location	Global	1 min – 1 hour	Point (~150 sites)	1993–present
NOAA ASOS/AWOS[7]	Depends on location, but usually ~10 m	Global	1 min – 1 hour	Point (~7,000 sites)	1901–present
NOAA NDBC[8]	Depends on location	Global	2–10 min	Point (~1,400 sites)	1988–present
NOAA SURFRAD	10 m	Regional - US (MT, CO, IL, MS, PA, NV, SD, OK, KS)	1–3 min	Point (8 sites)	1993–present
NOAA USCRN	1.5 m	Regional - North America, HI	5 min – 1 month	Point (139 sites)	2000–present
NOAA WFIP1	10–300 m	Regional - US (Northern and Southern Great Plains)	1–10 min	Point (22 sites)	2011–2012
NOAA WFIP2	From a few meters to a few hundreds of meters	Regional - US (OR, WA)	1 min – 1 hour	Point (~70 sites)	2015–2017
NOAA/ESRL PSL Wind Profiler[9]	A few meters	Global	30 min – 1 hour	Point (~140 sites)	1990–present
NREL MIDC	Depends on the location	Regional - US (CONUS, HI, PR)	1–5 min	Point (35 sites)	1981–present
NREL NSRDB[10]	N/A	Global	5 min – 1 hour	Grid (2–10 km)	1998–2021

[7] High resolution data is available in some sites. Hourly data is available globally as part of the NCEI Integrated Surface Dataset.

[8] Sites include both land stations (onshore wind) and buoy (offshore wind).

[9] Sites distributed across ~20 US States, with some additional sites in tropical areas outside of the US.

[10] Observations measured/modeled by satellites.

Table 2-2 (continued)
Summary of the commonly used onshore wind speed and direction observation datasets.

Dataset	Height (AGL)	Region	Time Resolution	Spatial Resolution	Time Range
NREL SUMR[11]	3–60 m	Regional - US (Jefferson, CO)	5 min	Point (1 site)	2019–2020
UO SRML	N/A	Regional - US (ID, MT, OR, UT, WA, WY)	1 min – 1 hour	Point (38 sites)	1979–present
WRMC BSRN	2–10 m	Global	Depends on location	Point (76 sites)	1992–present

Offshore Wind Speed and Direction Datasets

Table 2-3 lists the available offshore wind observation datasets, including the above surface level (ASL) height. There are substantially less offshore wind observation datasets compared to onshore, partially due to offshore wind being relatively new in the US. For example, the first offshore wind farm in the US opened off the coast of Block Island (Rhode Island) in 2016.

Figure 2-1 provides a map of the DOE Buoy LiDAR locations, along with annotations of the Bureau of Ocean Energy Management (BOEM) wind lease and planning areas.

Table 2-3
Summary of the available offshore wind speed and direction observation datasets.

Dataset	Height (ASL)	Region	Time Resolution	Spatial Resolution	Time Range
DOE Buoy LiDAR[12]	4–250 m	Regional - US (CA, NJ, VA, MA)	1 sec – 10 min	Point (6 sites)	2014–present
MassCEC Metocean	53–200 m	Regional - US (MA)	10 min	Point (1 site)	2016–2020
NOAA NDBC	A few meters	Global	2–10 min	Point (~1,400 sites)	1988–present

[11] Power data is available.

[12] Two LiDAR-equipped buoys sensors are deployed ~1–2 years each at four different locations. At the time of this report, two buoys are deployed offshore of California.

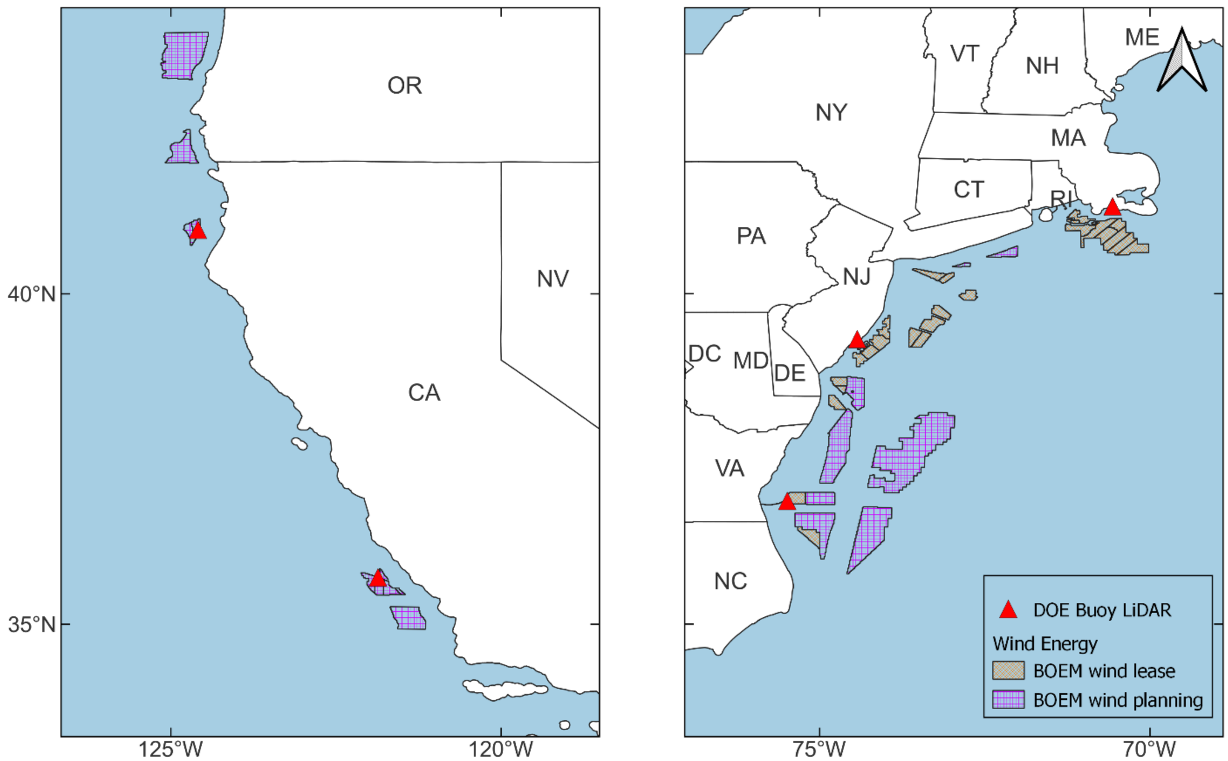

Figure 2-1
Map of DOE Buoy LiDAR sites on the west and east coast of the US.

Temperature and Humidity Datasets

Table 2-4 lists common datasets for ambient air temperature and humidity. Temperature and humidity are usually measured in degrees Celsius [ºC] and percentage [%], respectively, while some datasets also record measurements in degrees Fahrenheit [ºF] or degrees Kelvin [K]. Note that DOE Buoy LiDAR, NREL SUMR, and NOAA NDBC have temperature but not humidity data.

Table 2-4
Summary of the available temperature and humidity observation datasets.

Dataset	Region	Time Resolution	Spatial Resolution	Time Range
DOE ARM	Global	1 min – 1 hour	Point (~150 sites)	1993–present
DOE Buoy LiDAR	Regional - US (CA, NJ, VA, MA)	1 sec – 10 min	Point (6 sites)	2014–present
MassCEC Metocean	Regional - US (MA)	10 min	Point (1 site)	2016–2020
NOAA ASOS/AWOS	Global	1 min – 1 hour	Point (~7,000 sites)	1901–present
NOAA NDBC	Global	2–10 min	Point (~1,400 sites)	1988–present
NOAA SURFRAD	Regional - US (MT, CO, IL, MS, PA, NV, SD, OK, KS)	1–3 min	Point (8 sites)	1993–present
NOAA USCRN	Regional - North America, HI	5 min – 1 month	Point (139 sites)	2000–present
NOAA WFIP1	Regional - US (Northern and Southern Great Plains)	1–10 min	Point (22 sites)	2011–2012
NOAA WFIP2	Regional - US (OR, WA)	1 min – 1 hour	Point (~70 sites)	2015–2017
NOAA/ESRL PSL Wind Profiler	Global	30 min – 1hour	Point (~140 sites)	1990–present
NREL MIDC	Regional - US (CONUS, HI, PR)	1–5 min	Point (35 sites)	1981–present
NREL NSRDB	Global	5 min – 1 hour	Grid (2–10 km)	1998–2021
NREL SUMR	Regional - US (Jefferson, CO)	5 min	Point (1 site)	2019–2020
UO SRML	Regional - US (ID, MT, OR, UT, WA, WY)	1 min – 1 hour	Point (38 sites)	1979–present
WRMC BSRN	Global	Depends on location	Point (76 sites)	1992–present

Data Access

Table 2-5 summarizes details regarding data access for each of the datasets in this chapter, with datasets identified using the same acronyms as in the previous tables. A few notes regarding data access:

- For most of the NOAA observation datasets, data access is done via an API or a file server (FTP or HTTP). Some datasets may be available via cloud storage, but API or file server access is still the most common for NOAA datasets and is what most users should start with.
- For NOAA AWOS/ASOS datasets, users may find the Iowa State University archive as the most user friendly. However, as it is an unofficial archive, there is no guarantee of support or long-term existence.
- DOE ARM has multiple datasets, which are identified using a combination of the site/facility and data level, which ARM refers to as a "datastream". Users should first determine which datastream corresponds to the desired dataset and then use the datastream to request the data via the API.

Table 2-5
Summary of the data formats and the common download methods for the observation datasets.

Dataset	Data Format	Download Method	URL	Comments
DOE ARM	.nc, .csv	Web portal	https://www.arm.gov/data/	Registration required
		API	https://adc.arm.gov/armlive/	Registration required. API key required
DOE Buoy LiDAR	.txt, .nc	Web portal	https://a2e.energy.gov/projects/buoy	Registration required
MassCEC Metocean	.xlsx	Web portal, FTP	https://www.masscec.com/masscec-metocean-data-initiative (URL is not accessible as of Jul 22, 2022)	FTP account and password provided
NOAA ASOS/AWOS	.txt	Web portal, FTP	https://www.ncei.noaa.gov/products/land-based-station/automated-surface-weather-observing-systems	
		Web portal	https://mesonet.agron.iastate.edu/request/download.phtml	Sample Python/R scripts available
		Cloud storage (AWS)	https://registry.opendata.aws/noaa-isd/	
NOAA NDBC	.txt	Web portal	https://www.ndbc.noaa.gov/data/	
NOAA SOLRAD	.dat	Web portal, FTP	https://gml.noaa.gov/aftp/data/radiation/solrad/	
NOAA SURFRAD	.dat	Web portal, FTP	https://gml.noaa.gov/aftp/data/radiation/surfrad/	

Table 2-5 (continued)
Summary of the data formats and the common download methods for the observation datasets.

Dataset	Data Format	Download Method	URL	Comments
NOAA USCRN	.txt	Web portal, FTP	https://www.ncei.noaa.gov/access/crn/qcdatasets.html	
NOAA WFIP1	.txt, .csv	Web portal	https://a2e.energy.gov/projects/wfip1	Registration required
NOAA WFIP2	.txt, .nc		https://a2e.energy.gov/projects/wfip2	Registration required
NOAA/ESRL PSL Wind Profiler	.txt	Web portal, FTP	https://psl.noaa.gov/data/obs/data/	
NREL MIDC	.txt, .csv, .json	Web portal	https://midcdmz.nrel.gov	
		API	https://midcdmz.nrel.gov/apps/data_api_doc.pl?BMS	
NREL NSRDB	.csv	Web portal, API	https://nsrdb.nrel.gov/data-sets/how-to-access-data	Web downloads are rate limited. API key required
	.h5	Cloud storage (AWS)	https://registry.opendata.aws/nrel-pds-nsrdb/	
NREL SUMR	.dat	Web portal	https://a2e.energy.gov/projects/sumr-d	Registration required
UO SRML	.txt	Web portal	http://solardat.uoregon.edu/SolarData.html	
WRMC BSRN	.txt	Web portal, FTP, API	https://bsrn.awi.de/data/data-retrieval-via-pangaea/	Registration required

3
REANALYSIS

Reanalysis datasets are based on outputs of gridded numerical models, with postprocessing to improve the accuracy of the results compared to observations at specific sites. There are two primary benefits of reanalysis datasets: (1) data is available on a uniform grid over a large area, e.g., North America or the entire globe and (2) data is available over long periods, e.g., 1979–present. Additionally, some reanalysis datasets provide data on both a horizontal and vertical grid, e.g., wind speed at multiple heights from surface level up to 1 km. However, reanalysis data is (1) generally not as accurate or precise as direct observations at a specific location, (2) not available in “real-time”, and (3) generally not available at sub-hourly time resolutions.

Reanalysis Datasets

Table 3-1 lists commonly used reanalysis datasets, Table 3-2 with providing information on data access for each dataset. Note that all listed reanalysis datasets have global coverage except for NCEP NARR (North America only) and there are some differences in which variables are available from each dataset, as well as the units of some variables. For example, ECMWF ERA5 has a surface solar radiation downwards variable, which is equivalent to GHI but with units of energy [J/m2] instead of power [W/m2]. However, for consistency, Table 3-1 identifies all variables using the same naming convention as the prior chapter. Readers should consult relevant documentation to confirm variable details in reanalysis datasets.

Table 3-1
Summary of the commonly used reanalysis datasets.

Dataset	Region	Time Resolution	Horizontal Spatial Resolution	Time Range	Variable
NCEP/NCAR reanalysis 1	Global	6-hourly, daily, monthly, long-term mean	Grid (2.5°; ~278 km)	1948–present	GHI, DNI, DHI, wind speed, temperature, humidity
NCEP/DOE reanalysis 2	Global	6-hourly, daily, monthly, long-term mean	Grid (2.5°; ~278 km)	1979 – present	GHI, wind speed, temperature, humidity
NCEP NARR	Regional - North America	3-hourly, daily, monthly, long-term mean	Grid (0.3°; ~33 km)	1979 – present	GHI, wind speed, temperature, humidity
NOAA 20CR[13]	Global	3-hourly, 6-hourly, daily, monthly, long-term mean	Grid (1–2°; ~111–222 km)	1836–2015	GHI, DNI, DHI, wind speed, temperature, humidity
NOAA CFSR	Global	Hourly, 6-hourly, monthly	Grid (0.5–2.5°; ~55–278 km)	1979–2011	GHI, wind speed, temperature, humidity
ECMWF ERA5	Global	Hourly, monthly	Grid (0.1–0.25°; ~11–28 km)	1950–present	GHI, DNI, wind speed, temperature, humidity
NASA MERRA2	Global	Hourly, monthly	Grid (0.5° x 0.625°; ~55 km x 69 km)	1980–present	GHI, wind speed, temperature, humidity

Data Access

Table 3-2 provides details on accessing the reanalysis datasets in this chapter. A few additional notes:

1. For NOAA reanalysis datasets and NASA MERRA2, users should first try accessing data via a file server (FTP or HTTP).
2. For ERA5 reanalysis datasets, users should first try the Climate Data Store (CDS) web interface, which allow users to submit requests into a queue and then be emailed when the

[13] Four model versions (V1, V2, V2c, and V3). Time range refers to V3 (longest). Time and spatial resolution depend on the model version.

data is ready for download. Note however that large file sizes and high usage of CDS by users across the globe can mean data downloads take multiple hours (from request to final download).

3. There is an archive of ERA5 reanalysis data available on AWS, which contains data for the 18 essential weather variables, e.g., air temperature, wind speed and GHI. For those based in the US who are comfortable accessing data on cloud storage, this AWS archive can provide faster and easier access to ERA5 data. Also, the AWS archive allows accessing the ERA5 data in Zarr format, which enables more efficient querying of data, e.g., only download data for specific grid points rather than downloading all grid points and then extracting the desired subset of points.
4. NCAR/UCAR's RDA provides organized tables for variables and units by data products (data levels), which can be a good reference if users want to check which variables to use before downloading the dataset. Additionally, RDA provides sample download scripts for users.

Table 3-2
Summary of the data formats and the common download methods for the reanalysis datasets.

Dataset	Data Format	Download Method	URL	Comments
NCEP/NCAR reanalysis I	.nc	Web portal, FTP	https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html	
	.grib	Web portal	CISL RDA: NCEP/NCAR Global Reanalysis Products, 1948-continuing (ucar.edu)	Registration required
NCEP/DOE reanalysis II	.nc	Web portal, FTP	https://psl.noaa.gov/data/gridded/data.ncep.reanalysis2.html	
	.grib	Web portal	https://rda.ucar.edu/datasets/ds091.0/	Registration required
NCEP NARR	.nc	Web portal, FTP	https://psl.noaa.gov/data/gridded/data.narr.html	
	.grib	Web portal	https://rda.ucar.edu/datasets/ds608.0/	Registration required
NOAA 20CR	.nc	Web portal, FTP	https://psl.noaa.gov/data/20thC_Rean/	V2, V2c, V3 data available. URL is directed to V3
	.nc, .grib	Web portal	https://rda.ucar.edu/datasets/ds131.3/	Registration required. V1, V2, V2c, V3 data available. URL is directing to V3
	.nc, .grib	Web portal	https://portal.nersc.gov/project/20C_Reanalysis/	V2, V2c, V3 data available

Table 3-2 (continued)
Summary of the data formats and the common download methods for the reanalysis datasets.

Dataset	Data Format	Download Method	URL	Comments
NOAA CSFR	.grib, .bufr	Web portal	https://www.ncei.noaa.gov/products/weather-climate-models/climate-forecast-system	
	.grib	Web portal	https://rda.ucar.edu/#!lfd?nb=y&b=proj&v=NCEP%20Climate%20Forecast%20System%20Reanalysis	Registration required
	.grib	Web portal	https://nomads.ncep.noaa.gov/	
ECMWF ERA5	.grib	Web portal, API	https://cds.climate.copernicus.eu/#!/search?text=ERA5&type=dataset&keywords=((%20%22Product%20type:%20Reanalysis%22%20))	Registration required. API key required
	.nc, .grib	Web portal	https://rda.ucar.edu/index.html?hash=cgi-bin/dssearch?words=era5	Registration required
	.nc, .grib	API	https://www.ecmwf.int/en/computing/software/ecmwf-web-api	API key required. Some open-source packages available: cdsapi[14], ecmwf-api-client[15], ecmwf-opendata[16], herbie[17].
	.nc, .zarr	Cloud storage (AWS)	https://registry.opendata.aws/ecmwf-era5/	Only 18 variables available (compared to 185 variables in the full ERA5 dataset)
NASA MERRA2	.nc	Web portal, FTP	https://disc.gsfc.nasa.gov/datasets?project=MERRA-2	
		Web portal	https://rda.ucar.edu/index.html?hash=cgi-bin/dssearch?words=merra2	Registration required

[14] cdsapi Python package: https://github.com/ecmwf/cdsapi
[15] ecmwf-api-client Python package: https://github.com/ecmwf/ecmwf-api-client
[16] ecmwf-opendata: https://github.com/ecmwf/ecmwf-opendata
[17] Herbie Python package: https://github.com/blaylockbk/Herbie

4

TYPICAL METEOROLOGICAL YEAR

Typical meteorological year (TMY) datasets refer to single year datasets that were created from multiple years of data, such that the resulting year-long dataset represents “typical” weather conditions. TMY datasets are commonly used in energy modeling and building design analyses, or in computationally intensive applications where evaluating multiple years of weather data is prohibitive. Due to the nature of being “typical”, TMY data cannot be used for weather prediction or studying extreme weather events but should still bear the climatic characteristics of diurnal and seasonal cycles within a year. As first defined by Hall et al. (1978)[18], a TMY dataset at each location consists of 12 months of hourly values, with the data being representative of “typical” meteorological conditions over many years. The production of a reliable TMY dataset usually requires an original dataset of at least 10 consecutive years, with a desirable length of 30 years to fully characterize the statistical impact short-term weather variations have on the solar irradiance, such as sunspot cycle and El Nino events (Vignola et al., 2012)[19].

Common meteorological variables incorporated in a TMY dataset include solar irradiance, temperature, wind speed, and relative humidity. TMY datasets used in the US are commonly created using the methodologies described in Hall et al. (1978)[18], while TMY datasets used in Europe tend to use a similar methodology based on the ISO 15927-4:2005 standard[20]. However, the key differentiators between TMY datasets are (1) the multi-year weather datasets used as input and (2) the weighting used for selecting the “typical” data periods based on the selected weather variables. More details on TMY methodologies are given at the end of the chapter.

TMY Datasets

Table 4-1 lists commonly used and publicly available TMY datasets. Most TMY datasets are developed with a core solar irradiance dataset, which may then be combined with various sources of other meteorological variables. NSRDB is commonly used as the core solar irradiance data in most US-based TMY datasets, whereas Europe-based TMY datasets tend to use ECMWF reanalysis. Note that there are many local datasets not listed in these tables but are included in some global datasets. For example, the Chinese Standard Weather Data (CSWD), the Solar and Wind Energy Resource Assessment (SWERA), and the International Weather for Energy Calculations (IWEC) datasets are incorporated into the Climate Onebuilding and EnergyPlus[21], but are otherwise not publicly available.

---

[18] Hall, I. J., Prairie, R. R., Anderson, H. E., & Boes, E. C. (1978). ”Generation of a typical meteorological year” (No. SAND-78-1096C; CONF-780639-1). Sandia Labs, Albuquerque, NM (USA).

[19] Vignola, F., Grover, C., Lemon, N., & McMahan, A. (2012). “Building a bankable solar radiation dataset.” *Solar Energy*, *86*(8), 2218-2229. DOI: https://doi.org/10.1016/j.solener.2012.05.013

[20] EN ISO 15927-4:2005, Hygrothermal performance of buildings – calculation and presentation of climatic data. Part 4: data for assessing the annual energy for heating and cooling, 2005. https://www.iso.org/standard/41371.html

[21] EnergyPlus documentation for source weather data formats: https://bigladdersoftware.com/epx/docs/9-4/auxiliary-programs/source-weather-data-formats.html

Table 4-1
Summary of the main TMY datasets and their sources. Note all TMY datasets have hourly time resolution.

Dataset	Input Data Source	Region	Spatial Resolution	Variable
California Typical Weather Files	NOAA Integrated Surface Database and NREL NSRDB	California, US	Point (~100 sites)	GHI, DNI, temperature, wind speed
Canadian Weather Year for Energy Calculation (CWEC)	Canadian Weather Energy and Engineering Datasets (CWEEDS)	Canada	Point (564 sites)	GHI, wind speed, temperature, dewpoint temperature
Climate Onebuilding	NOAA Integrated Surface Database and ERA5 (solar)	Global	Point (~16,100 sites)	GHI, DHI, DNI, temperature, dewpoint temperature, wind speed, humidity
DOE2 Weather Dataset [22]	NREL NSRDB	US	Point (238 or 1020 sites)	GHI, DNI, temperature, dewpoint temperature, wind speed
EnergyPlus Weather Data	Various sources [23]	Global	Point (~3,000 sites)	GHI, DHI, DNI, temperature, dewpoint temperature, wind speed, humidity
European Commission TMY Generator	Satellites (solar radiation), ERA-I (all other)	Global [24]	Grid (0.05°; ~5 km)	GHI, wind speed, temperature, humidity
NREL TMY	NREL NSRDB	US	Point (238 sites)	GHI, DHI, DNI, temperature, dewpoint temperature, wind speed, humidity
NREL TMY2			Point (1,020 sites)	
NREL TMY3				

Data Access

Table 4-2 summarizes access methods for the TMY datasets in this chapter, with most datasets available via a web portal or API. However, reading and processing of a TMY dataset depends on the file format. For example, many TMY datasets are available in the EnergyPlus Weather

[22] Weather data provided by this source can only be interpreted by eQuest software.

[23] https://bigladdersoftware.com/epx/docs/8-3/auxiliary-programs/data-sources-uncertainty.html

[24] Though the interactive system (PVGIS) provides a location selection worldwide, the European Commission TMY Generator mainly covers Europe.

Format .epw, which is the default format from the open-source building energy simulation software EnergyPlus[25]. But the .epw format is also compatible with pvlib[26], an open-source Python package for modeling solar photovoltaic processes. Other formats listed in Table 4-2 can either be read by specific software or by programming languages. Additional information about file formats is provided in footnote[27].

Table 4-2
Summary of the data access details for the TMY datasets in Table 4-1.

Dataset	Data Format[27]	Download Method	URL
California Typical Weather Files	.epw, .fin4, .ddy2, .binm	Web portal	https://www.calmac.org/weather.asp
		API	https://www.calmac.org/apiguide.asp
Canadian Weather Year for Energy Calculation (CWEC)	.epw, .csv, .wy3	Cloud storage (Google Drive)	https://climate.weather.gc.ca/prods_servs/engineering_e.html
Climate Onebuilding	.epw, .clm, .wea, .pvsyst, .ddy, .rain, .stat	Web portal	https://climate.onebuilding.org/sources/default.html
DOE2 Weather Dataset	.bin	Web portal	https://doe2.com/index_Wth.html
EnergyPlus	.epw, .ddy, .stat	Web portal	https://energyplus.net/weather
EU Commission TMY	.epw, .csv, .json	Web portal	https://re.jrc.ec.europa.eu/pvg_tools/en/#TMY
		API	https://joint-research-centre.ec.europa.eu/pvgis-photovoltaic-geographical-information-system/getting-started-pvgis/api-non-interactive-service_en
NREL TMY, TMY2, and TMY3	.csv, .json	API	https://developer.nrel.gov/docs/solar/nsrdb/psm3-tmy-download/

[25] https://energyplus.net/

[26] William F. Holmgren, Clifford W. Hansen, and Mark A. Mikofski. (2018) "pvlib python: a python package for modeling solar energy systems." Journal of Open Source Software, 3(29), 884. https://doi.org/10.21105/joss.00884

[27] Data formats in this table include: .epw (EnergyPlus Weather Format), .clm (ESP-r weather format), .wea (Daysim weather format), .pvsyst (PV Solar weather design format), .ddy (ASHRAE Design Conditions or "file" design conditions in EnergyPlus format), .rain (hourly precipitation in meters per hour, where available), .stat (expanded EnergyPlus weather statistics) (source: https://climate.onebuilding.org/), .bin (binary files only interpretable by eQuesti software), .binm, and .wy3 (binary files) files.

Notes on TMY Methodologies

The essence of a TMY dataset is the representativeness of each month's data. The most used metric is the Finkelstein-Schafer (FS) statistic[28], which measures how different a variable is in a month from its long-term value based on its cumulative frequency distribution (CFD). The FS is calculated for each month over a multi-year dataset, then the month with the smallest cumulative FS is deemed as the "typical" month and is included in the final TMY dataset. For example, given 10 years of historical data, the FS is calculated for each January across the 10 years and the January with the smallest FS—out of the 10 Januarys—is selected for inclusion in the TMY dataset. The process is repeated for each February to select the "typical" February, each March to select the "typical" March, etc. Figure 4-1 provides a visualization of this concept, where the grey lines show potential selections, the bold blue line shows the long-term mean of all potential selections, and the bold black line is the selection with the smallest CFD of the FS statistic and therefore the most "typical" data.

First developed by Hall et al. (1978)[18], this approach using the FS statistic is sometimes referred to as the Sandia method and is the basis for many TMY datasets. The datasets listed in Table 4-1 are all produced with the Sandia method, except the EU Commission TMY and Climate Onebuilding, which are produced with the ISO 15927-4:2005[20] method. However, a key difference in the datasets is in the weighting of the meteorological variables when being selected to calculate the FS statistic, with most datasets placing the largest weighting on the solar radiation. For example, the NREL TMY dataset adopted a ratio of 4:4:4:12 between dry-bulb, dew point, wind speed, and solar radiation, while TMY2 adopted a ratio of 4:4:2:10 (Huang, 2020[29]). Research conducted by Su et al., (2009)[30], however, indicated the selection of weights for different meteorological variables did not make significant difference in the results, with a reported variation of less than 10%. On the other hand, Pernigotto et al., (2014)[31] argued that 10% of uncertainty of weather data in building energy simulation could lead to poor result when combined with all other uncertainties.

---

[28] Finkelstein, J.M. and Schafer, R.E., (1971). "Improved goodness-of-fit tests". *Biometrika*, *58*(3), pp.641-645.

[29] Huang, J., (2020) "Update of California Weather Files for Use in Utility Energy Efficiency Programs and Building Energy Standard Compliance Calculations", California Measurement Advisory Council publications. https://www.calmac.org/publications/Update_of_CA_Weather_Files_for_Energy_Efficiency_and_Building_Energy_Compliance_CALMAC_ID_PGE0450.pdf

[30] Su, F., Huang, J., Xu, T. and Zhang, C., (2009). "An evaluation of the effects of various parameter weights on typical meteorological years used for building energy simulation". In *Building Simulation* (Vol. 2, No. 1, pp. 19-28). Springer Berlin Heidelberg.

[31] Pernigotto, G., Prada, A., Gasparella, A. and Hensen, J.L., (2014). Analysis and improvement of the representativeness of EN ISO 15927-4 reference years for building energy simulation. *Journal of Building Performance Simulation*, *7*(6), pp.391-410.

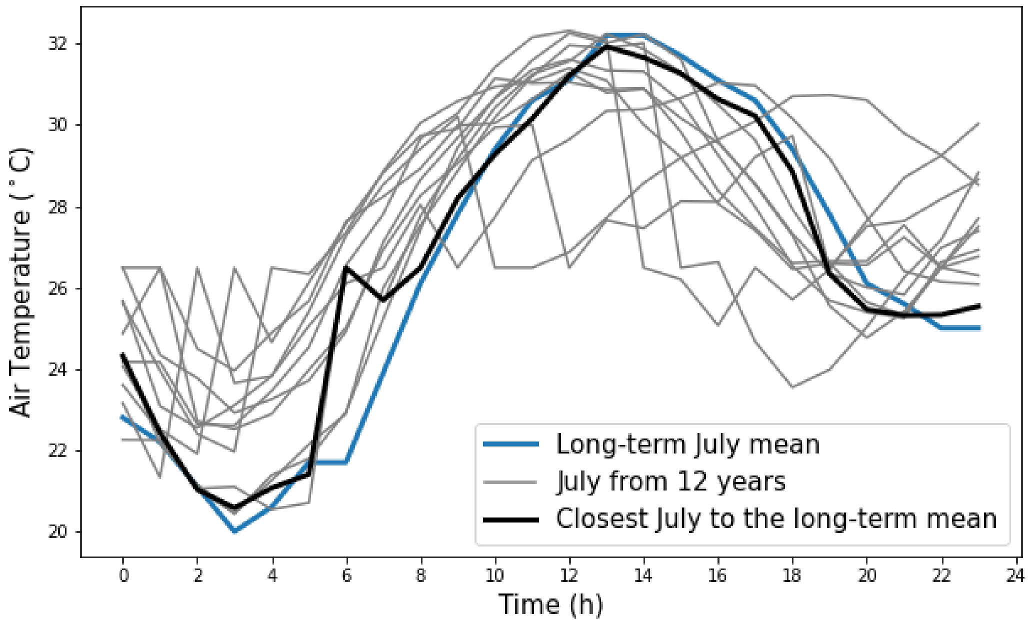

Figure 4-1
Illustration of potential choices versus the most "typical" July based on the FS metric.

5
GUIDANCE ON DATASET SELECTION

Selecting a weather dataset can be a complicated process, involving a balance of both technical and practical considerations. Here we provide guidance on common scenarios based on EPRI subject matter expertise. While not an exhaustive list, the commentary in this chapter should provide readers with a good starting point when selecting a dataset. Additionally, this chapter should help readers understand some of the key factors that can help narrow down choices.

Q: What dataset to use if I have no specific needs regarding data source, resolution, etc.?
A: Most users should start with an observation dataset such as the NOAA ASOS/AWOS, which provides standard weather variables (temperature, humidity, etc.) for ~7,000 sites and can be readily accessed through web interfaces such as the Iowa State University archive. Additionally, many utilities already use data from weather stations that are part of the ASOS/AWOS networks. However, if solar irradiance is required, then the NREL NSRDB is a good option since it provides data on a grid.

Q: What dataset to use if I need gridded data?
A: The NREL NSRDB or ECMWF ERA5 datasets are good starting points. NSRDB has the advantage of offering sub-hourly data (5-minutes as of 2018) and on a grid with finer spatial resolution than ERA5. But ERA5 covers a longer time span (1978–present), includes more variables and updates more frequently (5 days behind real-time) than NSRDB.

Q: What dataset to use if I need sub-hourly, e.g., 5-minute data?
A: There are several observation datasets with sub-hourly data. The NOAA ASOS/AWOS datasets are a good starting point, with NREL NSRDB as a good option if needing solar irradiance.

Q: What dataset to use if I need 20+ years of data?
A: Some observation datasets span 20+ years, but reanalysis datasets are generally a better option due to data availability. In particularly, ECMWF ERA5 is commonly used for long-term analyses.

Q: What dataset to use if I need reanalysis data?
A: ECMWF ERA5 is a good starting point. ERA5 is also one of the most used reanalysis datasets and therefore there are many resources available on accessing and using ERA5 data, including comparisons to other datasets.

Q: What dataset to use if I need TMY data?
A: Climate Onebuilding is a good starting point given it covers ~16,100 sites across the globe. EnergyPlus Weather Data and the European Commission TMY Generator are also reasonable choices for sites across the globe. However, region-specific TMY datasets may be preferable when available since such datasets can be more representative of specific regions, e.g., the California Typical Weather Files for studies focused on California.

Q: What dataset to use if I need to simulate utility-scale photovoltaic (PV) power production?
A: Observation datasets with solar irradiance are a good option, but data availability depends on

the specific location. If there is no observation dataset for the chosen location, then NREL NSRDB is generally a good option. Reanalysis datasets may also be a good option. However, many reanalysis datasets only include GHI, which can be challenging for applications requiring DNI and DHI, e.g., simulations of tracking PV plants that require calculating plane-of-array irradiance from GHI, DNI and DHI measurements.

Q: What dataset to use if I need to simulate distributed PV?
A: The NREL NSRDB dataset is a good starting point since data is available on a uniform grid.

Q: What dataset to use if I need to simulate onshore wind power?
A: Observation datasets are generally good starting points, but data availability depends heavily on location, including height of wind speed data. An alternative is to use reanalysis datasets, such as ECMWF ERA5, which can provide data on uniform grids, including wind speed at multiple heights, e.g., 10 m and 100 m.

Q: What dataset to use if I need to simulate offshore wind power?
A: Availability of offshore observation data is relatively limited compared to onshore, both in terms of location and time range. Therefore, it is reasonable for users to start with an observation dataset—if there is one available for the region of interest—and use reanalysis datasets as a backup.

Q: What dataset to use if I need to calculate a weather-normalized load?
A: The NOAA ASOS/AWOS datasets are good starting points since they provide the necessary weather variables (temperature and humidity), and the data is relatively easy to access. Additionally, many utilities already use data from weather stations that are part of the ASOS/AWOS networks.

Q: What dataset to use if I do not have programming experience?
A: Many observation datasets are available via web portals that do not require programming to access and in formats that can be opened in Microsoft Excel, e.g., .csv, .txt and .dat. As an example, the Iowa State University archive of the NOAA ASOS/AWOS datasets provides a user-friendly web interface, including the ability to search for locations, specify a time range and time resolution, and file format.

Q: What dataset to use if I want the most accurate data?
A: Observation datasets generally provide the most accurate (and precise) weather data. However, data quality can vary between locations—even within the same data source—and over time, and no dataset is perfect. With that said, datasets such as NOAA SURFRAD and NOAA SOLRAD tend to provide more accurate data due to professional maintenance and calibration efforts by NOAA. Unfortunately, there are only ~20 total stations across the SURFRAD and SOLRAD networks.

Q: What dataset to use if I want to automate data access?
A: Datasets with API access are generally the best for automated data access. For example, the NREL NSRDB dataset has an official API and registered users receive an access token, which can be used to authenticate to the API in scripts without any human intervention.

Q: What dataset to use if I need "real-time" data?
A: Observation datasets are usually best for "real-time" data access, though the latency between

measurement time and when the data is available to an end user can vary. The NOAA ASOS/AWOS datasets are a good starting point, but other observation datasets may also work well. However, depending on the specific application and its requirements, commercial datasets or weather forecasting information may be more suitable than publicly available observation datasets. For example, some commercial firms offer “real-time” estimates of solar irradiance based on geostationary satellite imagery, which can be used in place of direct measurements from an irradiance sensor installed at the site of interest.

6
SUMMARY

The use of weather data is an important aspect for many applications across the electric power industry. Selecting a weather dataset can involve both technical and practical considerations, and the choice of data can in turn impact application outcomes. Therefore, it is important for consumers of weather data to be informed of available weather datasets to make informed decisions. This document provided an overview of publicly available observation, reanalysis, and typical metrological year (TMY) weather datasets relevant to the electric power industry. Given their historical nature, these datasets are best used in planning studies where past weather trends can be used to identify trends and needs for planning applications. Additionally, this document provided guidance on selecting weather datasets, including practical considerations.

The following are key takeaways:

1. **Data availability**—in terms of location, time range, time resolution and weather variables—is a primary factor in selecting weather datasets. However, there are many publicly available weather datasets that are well-suited for applications across the electric power industry.

2. **Observation datasets** are generally a good option for users without programming experience, applications requiring sub-hourly data, and applications requiring measurements with the highest accuracy and precision. However, no dataset is perfect and observation data quality can vary based location, time range, weather variable, etc.

3. **Reanalysis datasets** can be used for most of the same applications as observation datasets, but with the advantage of generally being available for longer time ranges, more locations and including more weather variables. For example, the ECMWF ERA5 dataset covers 60+ years (1959–present) on a ~30 km horizontal grid across the globe. However, accessing and using reanalysis datasets requires more advanced programming expertise and processing time, and such data is generally less accurate and less precise compared to observation datasets.

4. **Typical Meteorological Year (TMY)** datasets are good options for applications built around single-year analyses, are generally easy to access and use without programming experience, and most TMY datasets are generated using the same methodology. However, TMY datasets are typically only available at hourly time resolutions and by their definition, only capture "typical" weather, rather than the full range of conditions that may occur.

Further work in this area will examine how such datasets could be leveraged together with climate models to develop future scenarios considering climate change. Additionally, work could also be carried out to show how the datasets can be used in important industry applications and whether—and how much—the characteristics of the datasets, e.g., resolution, accuracy and length of data, can impact on the value of the data for the specified end use.

Export Control Restrictions

Access to and use of this EPRI product is granted with the specific understanding and requirement that responsibility for ensuring full compliance with all applicable U.S. and foreign export laws and regulations is being undertaken by you and your company. This includes an obligation to ensure that any individual receiving access hereunder who is not a U.S. citizen or U.S. permanent resident is permitted access under applicable U.S. and foreign export laws and regulations.

In the event you are uncertain whether you or your company may lawfully obtain access to this EPRI product, you acknowledge that it is your obligation to consult with your company's legal counsel to determine whether this access is lawful. Although EPRI may make available on a case by case basis an informal assessment of the applicable U.S. export classification for specific EPRI products, you and your company acknowledge that this assessment is solely for informational purposes and not for reliance purposes.

Your obligations regarding U.S. export control requirements apply during and after you and your company's engagement with EPRI. To be clear, the obligations continue after your retirement or other departure from your company, and include any knowledge retained after gaining access to EPRI products.

You and your company understand and acknowledge your obligations to make a prompt report to EPRI and the appropriate authorities regarding any access to or use of this EPRI product hereunder that may be in violation of applicable U.S. or foreign export laws or regulations.

About EPRI

Founded in 1972, EPRI is the world's preeminent independent, non-profit energy research and development organization, with offices around the world. EPRI's trusted experts collaborate with more than 450 companies in 45 countries, driving innovation to ensure the public has clean, safe, reliable, affordable, and equitable access to electricity across the globe. Together, we are shaping the future of energy.

3002024375

EPRI
3420 Hillview Avenue, Palo Alto, California 94304-1338 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com