National Aeronautics and Space Administration
Goddard Earth Science Data Information and Services Center (GES DISC)

README Document for Surface PM2.5 Composition Products from the Hazardous Air Quality Ensemble System – Version 1.0

Last Revised on April 14, 2023

Goddard Earth Sciences Data and Information Services Center (GES DISC)
http://disc.gsfc.nasa.gov
NASA Goddard Space Flight Center
Code 619
Greenbelt, MD 20771 USA

Prepared By:

Suhung Shen	**Yunyao Li**
Name GES DISC GSFC Code 619	Name Center for Spatial Information Science and Systems (CSISS) George Mason University

Date April 14, 2023

Reviewed By:

Xiaohua Pan	April 14, 2023
Reviewer Name GES DISC GSFC Code 619	Date

Goddard Space Flight Center
Greenbelt, Maryland

Revision History

Revision Date	Changes	Author
April 14, 2023	version 1.0	Suhung Shen and Yunyao Li

Table of Contents

1.0 Introduction

This document provides basic information for using the Hazardous Air Quality Ensemble System (HAQES) surface PM $_{2.5}$ composition products. The dataset consists of nine products generated from HAQES with a focus on supporting research in health and air quality.

1.1 Dataset/Mission Description

HAQES is a real-time ensemble forecast of hazardous air quality events, such as wildfires, dust storms, and volcanic eruptions. Both regional and global models from multiple agencies are used to create the ensemble, including the Goddard Earth Observing System (GEOS, Randles et al., 2017) from the National Aeronautics and Space Administration (NASA), the Navy Aerosol Analysis and Prediction System (NAAPS, Lynch et al., 2016) from Naval Research Laboratory, the Global Ensemble Forecast System Aerosols (GEFS, Theurich et al., 2016), High-Resolution Rapid Refresh (HRRR, Dowell et al., 2022), and National Oceanic and Atmospheric Administration-U.S. Environmental Protection Agency (NOAA-EPA) Atmosphere-Chemistry Coupler-Community Multiscale Air Quality model (NACC-CMAQ, Campbell et al., 2022) from NOAA. The HAQES provides the forecast of surface $PM_{2.5}$ (PM25_TOT), $PM_{2.5}$ organic carbon (PM25_OC), and $PM_{2.5}$ black carbon (PM25_BC) every 3 hours. The prototypes of HAQES products were developed by the George Mason University Air Quality Laboratory as part of the NASA Health Air Quality Applied Science Team (HAQAST).

The HAQES v1.0 data provide the forecast of surface PM2.5, organic carbon, and black carbon over the contiguous United States (CONUS) domain every 3 hours for 24 hours. The forecast initial time is 12 UTC of the previous day, and the forecast time is 00 UTC to 23 UTC for the present day. The HAQES_NA_PM25_TOT data is the ensemble mean of the total PM2.5 results of the five models.

The HAQES_NA_PM25_TOT data provides an ensemble mean of PM2.5 forecasts for a total of all available species (e.g., sulfates, nitrates, organic carbon, black carbon, dust, sea salt, etc.) at a spatial resolution of 12 km x 12 km, based on outputs from the five models. The HAQES_NA_PM25_BC and HAQES_NA_PM25_OC datasets provide the same information as the HAQES_NA_PM25_TOT data but for black carbon and organic carbon, respectively.

In addition, the HAQES dataset includes spatially resolved PM2.5 measurements averaged at the county and census tract levels. The HAQES_NA_PM25_TOT_COUNTY, HAQES_NA_PM25_BC_COUNTY, and HAQES_NA_PM25_OC_COUNTY datasets provide average PM2.5 levels for each county in the United States (U.S.), while the HAQES_NA_PM25_TOT_CENSUS, HAQES_NA_PM25_BC_CENSUS, and HAQES_NA_PM25_OC_CENSUS datasets provide average PM2.5 levels for each U.S. census tract.

1.2 Data Citation and Acknowledgment

Please cite the HAQES v1.0 data in publications following the example below. The data DOI is listed in Table 1 for each product.

Example to cite the data:

> *George Mason University (2023), HAQES 3-Hourly Ensemble mean surface total PM2.5 concentration, North America, V1.0, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed:* ***[Data Access Date]****, doi:10.5067/N0BWP445491R*

1.3 Contact Information

Should you have science questions such as algorithm and data quality, please contact:

Daniel Tong, PI, qtong@gmu.edu,
or Yunyao Li, yli74@gmu.edu

2.0 Data Organization

The HAQES v1.0 data consists of nine collections (or products) for the 3-hourly mean surface PM 2.5 over CONUS.

2.1 Product list

The products and data DOIs are listed in Table 1 below.

Table 1: List of Collections in HAQES v1.0

Short Name	Format	Description	Data DOI (linked to the dataset landing page)
HAQES_NA_PM25_TOT	netCDF	HAQES 3-Hourly Ensemble mean surface total PM2.5 concentration, North America, V1.0	10.5067/N0BWP445491R
HAQES_NA_PM25_BC	netCDF	HAQES 3-Hourly Ensemble mean surface PM2.5 Black Carbon concentration,	10.5067/Q5ZFRUBCRWIW

		North America, V1.0	
HAQES_NA_PM25_OC	netCDF	HAQES 3-Hourly Ensemble mean surface PM2.5 Organic Carbon concentration, North America, V1.0	10.5067/B73NZNUN1GZ9
HAQES_NA_PM25_TOT_COUNTY	ASCII	HAQES 3-Hourly, Ensemble mean surface total PM2.5 concentration at county level, North America, V1.0	10.5067/036VLNHJM106
HAQES_NA_PM25_BC_COUNTY	ASCII	HAQES 3-Hourly Ensemble mean surface PM2.5 Black Carbon concentration at county level, North America, V1.0	10.5067/GV6VAXMJICS3
HAQES_NA_PM25_OC_COUNTY	ASCII	HAQES 3-Hourly Ensemble mean surface PM2.5 Organic Carbon concentration at county level, North America, V1.0	10.5067/OO7CW7SCN9F4
HAQES_NA_PM25_TOT_CENSUS	ASCII	HAQES 3-Hourly Ensemble mean surface total PM2.5 concentration at census level, North America, V1.0	10.5067/8V8LF6311A0J
HAQES_NA_PM25_BC_CENSUS	ASCII	HAQES 3-Hourly Ensemble mean surface PM2.5 Black Carbon concentration at census level, North America, V1.0	10.5067/YDY73WVBPISF
HAQES_NA_PM25_OC_CENSUS	ASCII	HAQES 3-Hourly Ensemble mean surface PM2.5 Organic Carbon concentration at census level, North America, V1.0	10.5067/ZDLBXTRFQ7U9

2.2 File Naming Convention

This dataset consists of files in two data formats: NetCDF-4 and ASCII with the following naming convention:

HAQES_NA_<version>_<timeFrequency>_<compositionType>.yyyymmdd.nc

HAQES_NA_<version>_<timeFrequency>_<compositionType>_<regionType>.yyyymmdd.hhz.txt

Where:

<version> = v1.0
<timeFrequency> = 3h (3-hourly mean)
<compositionType> = PM25_TOT, PM25_BC, PM25_OC
<regionType> = COUNTY, CENSUS
yyyy = 4 digits year
mm = 2 digits month
dd = 2 digits day
hh = 2 digits hour of the forecast time in UTC (Starting time of the 3 hourly interval)

File examples:

HAQES_NA_v1.0_3h_PM25_BC.20221101.nc
HAQES_NA_v1.0_3h_PM25_BC_COUNTY.20221101.00z.txt
HAQES_NA_v1.0_3h_PM25_BC_CENSUS.20221101.00z.txt

2.2 File Format and Structure

2.2.1 NetCDF-4 Files

Three collections (HAQES_NA_PM25_TOT_1, HAQES_NA_PM25_BC_1, and HAQES_NA_PM25_OC_1) files are in NetCDF-4 format. NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data that was developed by UCAR/Unidata (http://doi.org/10.5065/D6H70CW6) https://www.unidata.ucar.edu/software/netcdf/.

The data in NetCDF-4 format are mapped on **Lambert** projection, covering the North America region:

:SouthernmostLatitude = 21.1546f ;
:NorthernmostLatitude = 53.03837f ;
:WesternmostLongitude = -131.6533f ;
:EasternmostLongitude = -58.79196f ;

2.2.2 ASCII Files

Files with extension '.txt' are in ASCII format. They are areas mean of a composition over a county or census region. The shapefiles used to derive the data are from United States Census Bureau

For the county level, it is from this website:

https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

This shape file is used:

cb_2018_us_county_500k.zip

For the census tract, it is from this website:

https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.2020.html#list-tab-5W72O1MV3NBVXX7QR8

We use the data:

Census Tracts 1:500000 (national), the newly defined 2020 ZIP Code as of February 8, 2022.

3.0 Data Contents

3.1 Data Set Attributes (File Metadata)

In addition to SDS arrays containing variables and dimension scales, global metadata in the netCDF-4 is also stored in the files. Some metadata are required by CF standard conventions, some are present to meet data provenance requirements and others as a convenience to users of the products. An example of global attributes present in all NetCDF-4 files is shown as below:

// global attributes:
:GranuleID = "HAQES_NA_v1.0_3h_PM25_BC.20221101.nc" ;
:ShortName = "HAQES_NA_PM25_BC" ;
:LongName = "HAQES 3-Hourly Ensemble mean surface PM2.5 Black Carbon

concentration, North America, V1.0" ;
:VersionID = "1.0" ;
:Format = "NetCDF-4" ;
:RangeBeginningDate = "2022-11-01" ;
:RangeBeginningTime = "00:00:00.000000" ;
:RangeEndingDate = "2022-11-01" ;
:RangeEndingTime = "23:59:59.000000" ;
:IdentifierProductDOIAuthority = "https://dx.doi.org/" ;
:IdentifierProductDOI = "10.5067/Q5ZFRUBCRWIW" ;
:ProductionDateTime = "2023-02-07T12:13:26" ;
:ProcessingLevel = "Level 4" ;
:Conventions = "CF-1.8" ;
:title = "HAQES 3-Hourly Ensemble mean surface total PM2.5 concentration, North America, V1.0" ;
:source = "HAQES version 1.0" ;
:SouthernmostLatitude = 21.1546f ;
:NorthernmostLatitude = 53.03837f ;
:WesternmostLongitude = -131.6533f ;
:EasternmostLongitude = -58.79196f ;
:MapProjection = "Lambert" ;

The ASCII data files include the metadata in the header lines. An example metadata present in ASCII files is shown as below:

GranuleID = HAQES_NA_v1.0_3h_PM25_BC_COUNTY.20221101.f00.txt
ShortName = HAQES_NA_PM25_BC_COUNTY
LongName = HAQES 3-Hourly Ensemble mean surface PM2.5 Black Carbon concentration at county level, North America, V1.0
VersionID = 1.0
Format = ASCII
RangeBeginningDate = 2022-11-01
RangeBeginningTime = 00:00:00.000000
RangeEndingDate = 2022-11-01
RangeEndingTime = 02:59:59.000000
IdentifierProductDOIAuthority = https://dx.doi.org/
IdentifierProductDOI = 10.5067/GV6VAXMJICS3
ProductionDateTime = 2023-02-07T12:13:26
ProcessingLevel = Level 4

3.2 Variable Data Attributes in netCDF-4 Files

The variables in the netCDF-4 files are listed as below:

```
HAQES_NA_PM25_BC_1:
        float PM25_BC(TIME, ROW, COL) ;
                                PM25_BC:long_name = "Total PM25 Black Carbon" ;
                                PM25_BC:units = "ug m-3" ;
                                PM25_BC:_FillValue = -999.f ;

HAQES_NA_PM25_OC_1:
        float PM25_OC(TIME, ROW, COL) ;
                                PM25_OC:long_name = "Total PM25 Organic Carbon" ;
                                PM25_OC:units = "ug m-3" ;
                                PM25_OC:_FillValue = -999.f ;

HAQES_NA_PM25_TOT_1:
        float PM25_TOT(TIME, ROW, COL) ;
                                PM25_TOT:long_name = "Total PM25" ;
                                PM25_TOT:units = "ug m-3" ;
                                PM25_TOT:_FillValue = -999.f ;
```

3.3 Dimensions in NetCDF-4 Files

The following Table provides a description of the variable dimensions associated with the data in netCDF-4 files.

Dimension Variable	Description	Dimensions
TIME	hours since the beginning of a day, e.g., "hours since 2022-11-01 00:00:00" (e.g., the first hour represents the three hourly mean of 00 Z to 03 Z)	8
COL	Index of array element along longitude	442
ROW	Index of array element along latitude	265

3.4 Data Fields

3.4.1 variables in NetCDF-4 Files

The following Table provides a description of the variables associated with the data in netCDF-4 files.

Variable Name	Long_Name/Description	Type	Dimensions	Undefined Value	Units
PM25_TOT	Total PM25	float	8x265x442	-999.0	ug m-3
PM25_BC	PM25 Black Carbon	float	8x265x442	-999.0	ug m-3
PM25_OC	PM25 Organic Carbon	float	8x265x442	-999.0	ug m-3
LAT	Latitude	float	265x442	-999.0	degrees_north
LON	Longitude	float	265x442	-999.0	degrees_east

3.4.2 Variables in ASCII Files

The following Table provides a description of the variables associated with the data in ASCII files.

Variable Name	Long_Name/Description	Type	Undefined Value	Units
pm25	Total PM25	float	-999.0	ug m-3
pm25bc	PM25 Black Carbon	float	-999.0	ug m-3
pm25oc	PM25 Organic Carbon	float	-999.0	ug m-3
TRACT	Tract ID	int	-	-
FIPS	County ID	int	-	-

Note that the county and tract IDs are extracted from the shapefiles and listed in UScounty_FIPS.txt and

UScensus.txt, respectively, for users' reference. And these two files are available in the same directory as the corresponding data.

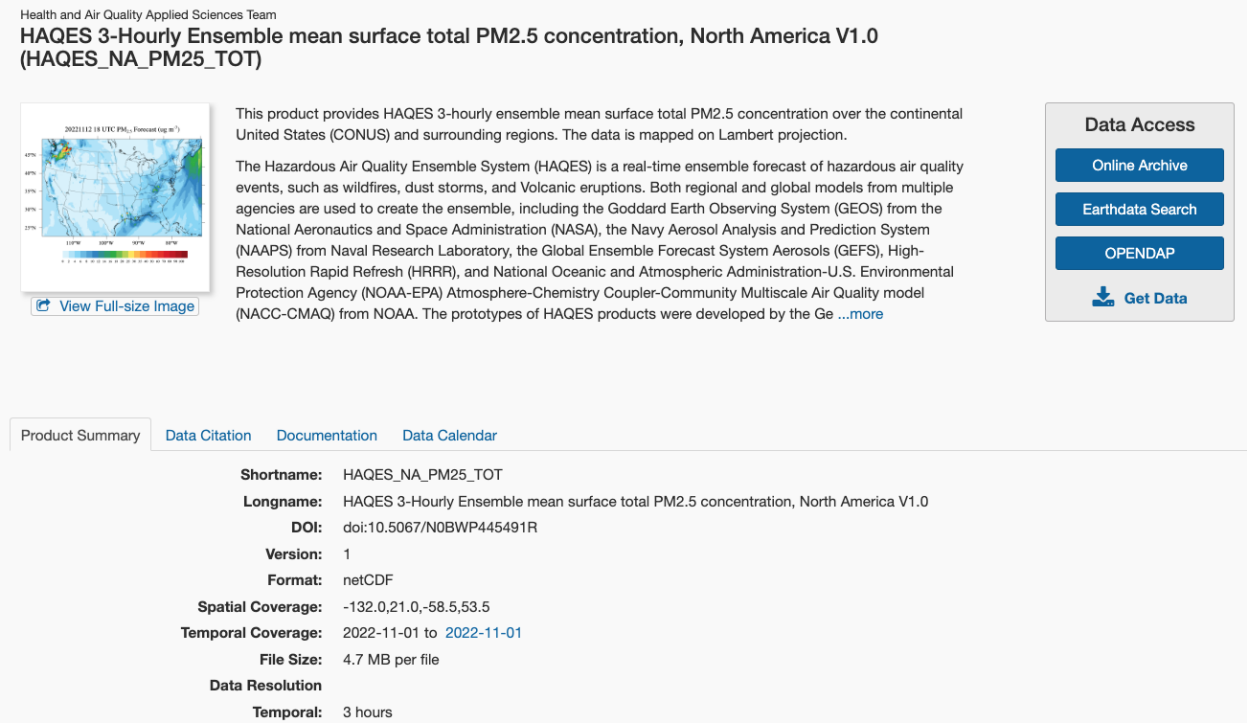
4.0 Data Access

4.1 Search products

The products of this project can be found by searching 'HAQES' in the search box after selecting 'Data Collection' on the GES DISC website: https://disc.gsfc.nasa.gov/, or on the NASA Earthdata Search interface: https://search.earthdata.nasa.gov/search.

4.2 Dataset landing page

The dataset landing page consists of the links of data access, product summary, documentation, and references, which can be found by clicking the product title from the product search result of section 4.1, for example:



4.3 Direct access

The data may be downloaded directly from the HTTPS service at:

https://acdisc.gesdisc.eosdis.nasa.gov/data/HAQAST/

5.0 Options for Reading the Data

5.1 Command Line Utilities

5.1.1 ncdump

The ncdump tool can be used as a simple browser for NetCDF-4 data files, to display the dimension names and sizes; variable names, types, and shapes; attribute names and values; and optionally, the values of data for all variables or selected variables in a netCDF file. The most common use of ncdump is with the –h option, in which only the header information is displayed.

ncdump [-c|-h] [-v ...] [[-b|-f] [c|f]] [-l len] [-n name] [-d n[,n]] filename
Options/Arguments:
[-c] Coordinate variable data and header information.
[-h] Header information only, no data
[-v var1[,...]] Data for variable(s) <var1>,... only data
[-f [c|f]] Full annotations for C or Fortran indices in data
[-l len] Line length maximum in data section (default 80)
[-n name] Name for netCDF (default derived from file name)
[-d n[,n]] Approximate floating-point values with less precision filename File name of input netCDF file
(https://www.unidata.ucar.edu/software/netcdf/workshops/2011/utilities/Ncdump.html)

5.2 Tools/Programming

The product files can be read and queried using the NetCDF4 library and tools maintained by Unidata (http://www.unidata.ucar.edu/software/netcdf/). Support for reading NetCDF is offered in many programming languages, including Python, Matlab, IDL, C/C++ and Fortran. NetCDF4 files are legal HDF5 files with additional bookkeeping information managed by the NetCDF4 library. It is therefore possible to inspect and copy data out of the NetCDF4 files by using the HDF5 utilities and libraries maintained by the HDF Group (https://www.hdfgroup.org/products/hdf5_tools/index.html) or by using the HDF5 interface in your favorite programming language. However, the two libraries should not be considered fully interchangeable.

Matlab users should note that the Matlab NetCDF4 interface is currently (as of version R2017a) not able to read attributes that are string arrays and will throw an exception if that is attempted.

5.2.1 Python

1) The following code snippet shows how to read variable row, col, and PM25_TOT from a netCDF-4 file, HAQES_NA_v1.0_3h_PM25_TOT.20230201.nc, and shows some basic information about the variable's size.

```
import netCDF4

#Read netcdf data file
Data_file = 'HAQES_NA_v1.0_3h_PM25_TOT.20230201.nc'
nc_fid = netCDF4.Dataset( Data_file ,mode='r',format='NETCDF4')

#Read in the variables
row = nc_fid.variables['ROW'][:]
col = nc_fid.variables['COL'][:]
pm25 = nc_fid.variables['PM25_TOT'][:]

#Check the shape of each variable
pm25.shape
row.shape
```

2) The following code snippet shows how to read PM25_TOT from a COUNTY text file, for example, HAQES_NA_v1.0_3h_PM25_TOT_COUNTY.20220102.06z.txt, and how to read the county shapefile. Note that you must install Python shapefile Library (PyShp) to read Shapefiles (conda install -c conda-forge pyshp or pip install pyshp)

```
import numpy as np
import csv
import os
import pandas as pd
import shapefile

#Read data file
Data_file= 'HAQES_NA_v1.0_3h_PM25_TOT_COUNTY.20220102.06z.txt'
df_county = pd.read_csv(Data_file,low_memory=False, dtype={"FIPS": "str"},
skiprows=20)
df_county = df_county.rename(columns={'FIPS':'GEOID'})
```

```
#Read county shapefile (first downloaded the Shape_file to the current working
directory, following instruction in Section 2.2.2)
Shape_file = './cb_2018_us_county_500k/cb_2018_us_county_500k.shp'
sf = shapefile.Reader(Shape_file)
fields = [x[0] for x in sf.fields][1:]
records = sf.records()
shps = [s.points for s in sf.shapes()]

#Write into a dataframe
df_shp = pd.DataFrame(columns=fields, data=records)
df_shp = df_shp.assign(coords=shps)
print(df_shp.columns.tolist())

# merge
dfmrg = pd.merge(df_shp,df_county,on=['GEOID'], how='outer')
print(dfmrg)
```

3) The following code snippet shows how to read PM25_TOT from the CENSUS text file, such as HAQES_NA_v1.0_3h_PM25_TOT_CENSUS.20220102.03z.txt, and how to read the census tract shapefile.

```
import numpy as np
import csv
import os
import pandas as pd
import shapefile

#Read data file
Data_file= 'HAQES_NA_v1.0_3h_PM25_TOT_CENSUS.20220102.03z.txt'
df_census = pd.read_csv(Data_file,low_memory=False, dtype={"TRACT": "str"},
skiprows=20)
df_census = df_census.rename(columns={'TRACT':'GEOID'})
print(df_census.columns.tolist())

#Read census tract shapefile (first downloaded the Shape_file to the current directory,
following instruction in Section 2.2.2)
Shape_file = './cb_2020_us_tract_500k/cb_2020_us_tract_500k.shp'
sf = shapefile.Reader(Shape_file)
fields = [x[0] for x in sf.fields][1:]
records = sf.records()
shps = [s.points for s in sf.shapes()]

#write into a dataframe
df_shp = pd.DataFrame(columns=fields, data=records)
```

```
df_shp = df_shp.assign(coords=shps)
print(df_shp.columns.tolist())

# merge
dfmrg = pd.merge(df_shp,df_census,on=['GEOID'], how='outer')
```

5.2.2 Panoply

The netCDF-4 files could be visualized with the NASA free data tool Panoply. Example is available as below:

Quick View Data with Panoply

6.0 GES DISC Data Services

If you need assistance or wish to report a problem:
Email: gsfc-dl-help-disc@mail.nasa.gov
Voice: 301-614-5224
Fax: 301-614-5268
Address:
Goddard Earth Sciences Data and Information Services Center NASA Goddard Space Flight Center Code 619 Greenbelt, MD 20771 USA

6.1 "How To" Articles

The GESDISC website provides many useful data recipes and articles under the sections of "How To" , "FAQ" (frequently asked questions), "News", "Glossary", and "Help" . Some samples of these articles include:

- Earthdata Login for Data Access
- How to Download Data Files from HTTPS Service with wget
- How to Obtain Data in NetCDF Format via OpeNDAP
- Quick View Data with Panoply
- How to read and plot NetCDF MERRA-2 data in Python

7.0 Acknowledgments

This study is financially supported by NASA Health and Air Quality Program. Model prediction datasets used for constructing the ensemble forecast are provided by NASA Goddard Space Flight Center (Anton Darmenov), NOAA Air Resources Laboratory (Barry Baker, Patrick Campbell and Youhua Tang), and Naval Research Laboratory (Peng Xian and Edward Hyer). Ground measurements collected by the EPA and satellite AOD data by NOAA and NASA are gratefully acknowledged.

8.0 References

If you are using the HAQES v1.0 dataset in your research, we kindly request that you cite the following papers:

Li, Y., Tong, D., Ma, S., Zhang, X., Kondragunta, S., Li, F., Saylor, R. (2021). Dominance of wildfires impact on air quality exceedances during the 2020 record-breaking wildfire season in the United States. *Geophysical Research Letters*, 48, e2021GL094908. https://doi.org/10.1029/2021GL094908

Makkaroon, P. D.Q. Tong, Y. Li, E. J. Hyer, P. Xian, S. Kondragunta, P. C. Campbell, Y. Tang, B. D., Baker, M. D. Cohen, A. Darmenov, A. Lyapustin, R. D. Saylor, Y. Wang, I. Stajner, Development and Evaluation of a North America Ensemble Wildfire Forecast: Initial Application to the 2020 Western United States "Gigafire". Manuscript is Under Review.

Other reference mentioned in this README:

Campbell, P. C., Tang, Y., Lee, P., Baker, B., Tong, D., Saylor, R., Stein, A., Huang, J., Huang, H.-C., Strobach, E., McQueen, J., Pan, L., Stajner, I., Sims, J., Tirado-Delgado, J., Jung, Y., Yang, F., Spero, T. L., & Gilliam, R. C. (2022). Development and evaluation of an advanced National Air Quality Forecasting Capability using the NOAA Global Forecast System version 16. Geoscientific Model Development, 15(8), 3281–3313. https://doi.org/10.5194/gmd-15-3281-2022

Dowell, D. C., Alexander, C.R., James, E.P., Weygande, S.S., Benjamin, S.G., Manikin, G.S., Blake, B.T., Brown, J.M. et al. (2022). The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description. Wea. Forecasting, 37, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Lynch, P., Reid, J. S., Westphal, D. L., Zhang, J., Hogan, T. F., Hyer, E. J., Curtis, C. A., Hegg, D. A., Shi, Y., Campbell, J. R., Rubin, J. I., Sessions, W. R., Turk, F. J., & Walker, A. L. (2016). An 11-year global gridded aerosol optical thickness reanalysis (v1.0) for atmospheric and climate sciences. Geoscientific Model Development, 9(4), 1489–1522. https://doi.org/10.5194/gmd-9-1489-2016

Randles, C. A., Silva, A. M. da, Buchard, V., Colarco, P. R., Darmenov, A., Govindaraju, R., Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., & Flynn, C. J. (2017). The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part I: System Description and Data Assimilation Evaluation. Journal of Climate, 30(17), 6823–6850. https://doi.org/10.1175/JCLI-D-16-0609.1

Theurich, G., DeLuca, C., Campbell, T., Liu, F., Saint, K., Vertenstein, M., Chen, J., Oehmke, R., Doyle, J., Whitcomb, T., Wallcraft, A., Iredell, M., Black, T., Silva, A. M. D., Clune, T., Ferraro, R., Li, P., Kelley, M., Aleinov, I., … Dunlap, R. (2016). The Earth System Prediction Suite: Toward a Coordinated U.S. Modeling Capability. Bulletin of the American Meteorological Society, 97(7), 1229–1247. https://doi.org/10.1175/BAMS-D-14-00164.1