Problem description: Particulate Track

**The goal of this challenge is to generate daily estimates of surface-level PM2.5 for a set of 5km x 5km geometries across three urban areas:**

1. [Los Angeles (South Coast Air Basin)](https://en.wikipedia.org/wiki/South_Coast_Air_Basin)
2. [Delhi](https://en.wikipedia.org/wiki/Delhi)
3. [Taipei](https://en.wikipedia.org/wiki/Taipei)

These locations represent areas with readily available satellite data, but varying levels of pollution and historical data. High performing models should do well in all three locations.

There will be two separate tracks for the prediction of each pollutant, PM2.5 and NO2. This is the Particulate Track (PM2.5). You can find the Trace Gas Track (NO2) [here](https://www.drivendata.org/competitions/91/competition-air-quality-no2/). More information on the dataset, performance metric, and submission specifications is provided below.

Finalists and runners-up will be determined by performance on the private test set. These participants will then have the opportunity to submit their code to be audited using an out-of-sample verification set. The top 3 eligible teams that pass this final check in each track will be awarded prizes.

* **Data**
* [Features](https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/424/#features)
* [Additional sources](https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/424/#additional)
* [Labels](https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/424/#labels)
* [Metadata](https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/424/#metadata)
* **Submission Format**
* [Format](https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/424/#sub)
* [Example](https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/424/#sub_values)
* **Performance metric**
* [R squared](https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/424/#metric)
* [Root mean square error](https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/424/#metric_2)

Data

There are two types of pre-approved data that can be used as input to your model: satellite and ancillary (meteorological and topological) data. We provide satellite data through a [public s3 bucket](https://www.drivendata.org/competitions/88/competition-air-quality-pm/data/) hosted by DrivenData. These files have already been subsetted to the correct times and geographies. Ancillary data can be accessed through public portals. You may use any approved data sources you like, but you must use **at least one** dataset from the list of approved satellite data.

Note that the data are provided in various spatiotemporal resolutions and formats. Malfunctioning instruments and other errors may lead to missing and incorrect values. The datasets may also cover different date ranges or geographical areas. It is your job to figure out how to best combine these datasets to build the best model!

Approved features (inputs)

**All data sources detailed below are pre-approved for both model training and inference**. These data have undergone a careful selection process and are freely and publicly available. Note that only the listed data products and access locations are pre-approved for use. You must go through an approved access location when using these sources for inference.

If you are interested in using additional sources that are not listed, see the process for requesting additional data sources below.

SATELLITE DATA

NASA provides many data products through the [Earthdata](https://earthdata.nasa.gov/" \t "_blank) portal. Select data sources that are likely to be of greatest use in this track are listed below. **You must use at least one of these satellite sources to train your model.**

The primary satellite data products related to estimating PM2.5 measure something called [Aerosol Optical Depth (AOD)](https://earthobservatory.nasa.gov/global-maps/MODAL2_M_AER_OD), also called [Aerosol Optical Thickness (AOT)](https://disc.gsfc.nasa.gov/information/glossary?title=Aerosol%20Optical%20Thickness). It is a unitless measure of aerosols (e.g., urban haze, smoke particles, desert dust, sea salt) distributed within a column of air from Earth's surface to the top of the atmosphere. Note that the units of PM2.5 are typically provided in μg/m3 (micrograms per cubic meter).

**MODIS/Terra and Aqua MAIAC Land Aerosol Optical Depth Daily L2G 1 km SIN Grid (**[**MCD19A2**](https://lpdaac.usgs.gov/products/mcd19a2v006/)**)**:

Multi-Angle Implementation of Atmospheric Correction ([MAIAC](https://modis-land.gsfc.nasa.gov/MAIAC.html)) is an algorithm that uses data from the two [MODIS](https://modis.gsfc.nasa.gov/about/) satellite instruments to derive high-resolution aerosol and land surface reflectance products. The MCD19A2 Version 6 gridded Level 2 product is produced daily at 1 kilometer (km) pixel resolution. These data are provided as [Hierarchical Data Format](https://en.wikipedia.org/wiki/Hierarchical_Data_Format) files.

* Approved data access location: DrivenData Public S3 Bucket. Download instructions on the [data download page](https://www.drivendata.org/competitions/88/competition-air-quality-pm/data/).
* [Data product details](https://cmr.earthdata.nasa.gov/search/concepts/C1000000505-LPDAAC_ECS.html)

**MISR Level 2 FIRSTLOOK Aerosol Product (**[**MIL2ASAF**](https://asdc.larc.nasa.gov/project/MISR)**)**:

MIL2ASAF is a near real-time version of the Multi-angle Imaging SpectroRadiometer ([MISR](https://misr.jpl.nasa.gov/)) Level 2 Aerosol product. It contains information on retrieved aerosol column amount, aerosol particle properties, and ancillary information based on Level 1B2 geolocated radiances observed by MISR at 4.4 km spatial resolution. Only the [FIRSTLOOK](https://doi.org/10.5194/amt-13-593-2020) version of this product is allowed for use in this competition. These data are provided as [NetCDF](https://en.wikipedia.org/wiki/NetCDF" \t "_blank) files.

* Approved data access location: DrivenData Public S3 Bucket. Download instructions on the [data download page](https://www.drivendata.org/competitions/88/competition-air-quality-pm/data/).
* [Data product details](https://cmr.earthdata.nasa.gov/search/concepts/C1542382556-LARC.html)

Approved satellite data is hosted in a public s3 bucket in the pm25 folder. [AWS CLI](https://aws.amazon.com/cli/) will be useful for downloading data (you will probably need the --no-sign-request argument to the CLI). See pm25\_satellite\_metadata.csv for filepaths by bucket region and file format. Further instructions are detailed in airathon\_download\_instructions\_pm25.txt. CSV and text files are available on the [data download page](https://www.drivendata.org/competitions/88/competition-air-quality-pm/data/). The data is replicated to buckets hosted in the US, the EU (Germany), and Asia (Singapore). Pick the bucket closest to your machine geographically to maximize transfer speeds.

ANCILLARY DATA

**NASA Digital Elevation Model (NASADEM)**:

[NASADEM](https://earthdata.nasa.gov/esds/competitive-programs/measures/nasadem) data products were derived from original telemetry data from the Shuttle Radar Topography Mission (SRTM). In addition to Terra Advanced Spaceborne Thermal and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) Version 2 data, NASADEM also relied on Ice, Cloud, and Land Elevation Satellite (ICESat) Geoscience Laser Altimeter System (GLAS) ground control points of its lidar shots to improve surface elevation measurements that led to improved geolocation accuracy.

NASADEM can be accessed through Microsoft's Planetary Computer through the nasadem collection via the [STAC API](https://github.com/radiantearth/stac-api-spec). Using the [PySTAC library](https://pystac.readthedocs.io/en/latest/" \t "_blank) created by [Azavea](https://www.azavea.com/" \t "_blank), you can load, traverse, and access data within these STACs programmatically. This [quickstart](https://planetarycomputer.microsoft.com/docs/quickstarts/reading-stac/" \t "_blank) guide demonstrates how to search for data using the STAC API with PySTAC.

To get a SAS Token to enable access to the STAC API, use the Planetary Computer’s [Data Authentication API](https://planetarycomputer.microsoft.com/docs/reference/sas/). Alternatively, you can use the [planetary-computer](https://github.com/microsoft/planetary-computer-sdk-for-python) package to generate tokens and sign asset HREFs for access.

* Approved data access location: https://planetarycomputer.microsoft.com/api/stac/v1/collections/nasadem
* [Data access details](https://planetarycomputer.microsoft.com/dataset/nasadem#overview)
* [Data product details](https://lpdaac.usgs.gov/products/nasadem_hgtv001/)
* [Example notebook](https://github.com/microsoft/PlanetaryComputerExamples/blob/main/datasets/nasadem/nasadem-example.ipynb)

**Global Forecast System**:

The [Global Forecast System](https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php) (GFS) is a [National Centers for Environmental Prediction](https://www.ncep.noaa.gov/) (NCEP) weather forecast model that generates data for dozens of atmospheric and land-soil variables, including temperatures, winds, precipitation, soil moisture, and atmospheric ozone concentration. The system couples four separate models (atmosphere, ocean model, land/soil model, and sea ice) that work together to accurately depict weather conditions.

Only the **GFS Forecasts** dataset is pre-approved for this competition (as opposed to GFS Analysis/GFS-ANL). You may use use the 0.25°, 0.5°, or 1° grids. Note that the 0.25° dataset is only accessible via the [NCAR](https://rda.ucar.edu/) servers, which requires registration.

0.5° or 1° grid:

* Approved data access location: [NCEI Servers](https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast)
* [Data access details](https://www.ncdc.noaa.gov/nomads/documentation/user-guide/advanced-data-access-methods)
* [Data product details](https://www.nco.ncep.noaa.gov/pmb/products/gfs/)

0.25° grid:

* Approved data access location: [NCAR Servers](https://rda.ucar.edu/datasets/ds084.1)
* [Data access details](https://rda.ucar.edu/datasets/ds084.1/#!access) (requires sign-in)
* [Data product details](https://rda.ucar.edu/datasets/ds084.1/#metadata/detailed.html?_do=y)
* [Python API Client](https://github.com/NCAR/rda-apps-clients)

**ECMWF IFS CY41r2 High-Resolution Operational Forecasts**:

The Integrated Forecast System (IFS) model cycle CY41r2 [High-Resolution Operational Forecast](https://rda.ucar.edu/datasets/ds113.1/) is a dataset by the [European Centre for Medium-Range Weather Forecasts (ECMWF)](https://www.ecmwf.int/). It provides highly accurate weather forecasts with a nominal grid point spacing of nine kilometers.

Note that the approved data access location is accessible via [NCAR](https://rda.ucar.edu/) servers, which requires registration.

* Approved data access location: [NCAR Servers](https://rda.ucar.edu/datasets/ds113.1/)
* [Data access details](https://rda.ucar.edu/datasets/ds113.1/#!access) (requires sign-in)
* [Data product details](https://confluence.ecmwf.int/display/FUG/HRES+-+High-Resolution+Forecast#HRESHighResolutionForecast-AdditionalSourcesofInformation)
* [Python API Client](https://github.com/NCAR/rda-apps-clients)

**CALIPSO Lidar (**[**CAL\_LID\_L15-Standard-V1-00**](https://cmr.earthdata.nasa.gov/search/concepts/C1605962520-LARC_ASDC.html)**)**:

Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) Lidar Level 1.5 Profile, Version 1-00 data product is a continuous segment of calibrated, geolocated, cloud-cleared, and spatially averaged profiles of lidar attenuated backscatter. It provides information about the vertical distribution of aerosol. Data for this product was collected using the CALIPSO Cloud-Aerosol Lidar with Orthogonal Polarization ([CALIOP](https://www.nasa.gov/mission_pages/calipso/spacecraft/index.html#:~:text=CALIOP%20is%20a%20two%2Dwavelength,the%20532%20nm%20backscattered%20signal.)) instrument. CALIPSO was launched on April 28, 2006 to study the impact of clouds and aerosols on the Earth's radiation budget and climate.

CALIPSO files are in [HDF4](https://hdfeos.org/software/pyhdf.php) format and available through the [OPeNDAP Hyrax Server](https://earthdata.nasa.gov/collaborate/open-data-services-and-software/api/opendap/opendap-user-guide" \t "_blank). You can use the [Pydap](https://www.pydap.org/en/latest/" \t "_blank) library to access data.

* Approved data access location: [OPeNDAP Hyrax Server](https://opendap.larc.nasa.gov/opendap/CALIPSO/LID_L15-Standard-V1-00/contents.html" \t "_blank)
* [Data access details](https://cmr.earthdata.nasa.gov/search/concepts/C1605962520-LARC_ASDC.html)
* [Data product details](https://www-calipso.larc.nasa.gov/resources/calipso_users_guide/)

**Goddard Earth Observing System – Composition Forecasting**:

The NASA [GEOS Composition Forecasting (GEOS-CF)](https://gmao.gsfc.nasa.gov/weather_prediction/GEOS-CF/) system produces global, three-dimensional distributions of atmospheric composition. The horizontal spatial resolution is 0.25 degrees, and the temporal resolution is hourly for most outputs. Using meteorological analyses from other GEOS systems, the GEOS-CF products include a running near-time estimate of surface pollutant distributions and the composition of the troposphere and stratosphere. A single five-day forecast is generated daily, beginning at 12 UTC.

Only the **aqc\_tavg\_1hr\_g1440x721\_v1** data collection is approved for this challenge. This collection provides air-quality relevant outputs, including surface-level concentrations of NO2 and PM2.5. You may use either the forecast (fcast) or meteorological replay (assim) versions of this collection. The approved data access location is an [OPeNDAP](https://www.opendap.org/" \t "_blank) server. The data is provided in [NetCDF](https://en.wikipedia.org/wiki/NetCDF" \t "_blank) format and can be read with HDF5 software. Example code is provided below.

* Approved data access location: [GrADS OpeNDAP Data Server](https://opendap.nccs.nasa.gov/dods/gmao/geos-cf" \t "_blank)
* [Data access details](https://gmao.gsfc.nasa.gov/pubs/docs/Knowland1204.pdf)
* [Data product details](https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020MS002413)

The following sample code demonstrates how to access data in Python using the packages [Pydap](https://www.pydap.org/en/latest/" \t "_blank), [numpy](https://numpy.org/" \t "_blank), and [xarray](https://docs.xarray.dev/en/stable/" \t "_blank). Specifically, it retrieves the historical estimate of surface-level NO2 for the month of September 2019 for a 2-degree latitude by 2-degree-longitude region over San Francisco.

import pydap

import xarray as xr

import numpy as np

url = 'https://opendap.nccs.nasa.gov/dods/gmao/geos-cf/assim/aqc\_tavg\_1hr\_g1440x721\_v1'

DATASET = xr.open\_dataset(url)

s\_type = 'no2'

start\_time = np.datetime64('2019-09-01 00:00:00')

end\_time = start\_time + np.timedelta64(30,'D')

minimum\_latitude = 37

minumum\_longitude = -123

maximum\_latitude = 39

maximum\_longitude = -121

DATA\_SUBSET = DATASET[s\_type].loc[{'time':slice(start\_time,end\_time),'lat':slice(minimum\_latitude,maximum\_latitude),'lon':slice(minumum\_longitude,maximum\_longitude)}]

A note on time: Keep in mind, estimates should only factor in data available at the time of inference (training data excluded). You may **only use data that would be available up through the day of estimation** when generating predictions. Use of any future data after the day of estimation, excluding training data, is prohibited.

Additional data sources

Only the approved sources listed above may be used in this challenge track. Any additional data must be reviewed and approved in order to be eligible for use.

If you would like for any additional sources to be approved, you are welcome to submit an [official request form](https://forms.gle/4b2vwg5XQW9HuN4A6) and the challenge organizers will review the request. Only select sources that demonstrate a strong case for use will be considered.

To qualify for approval, data sources must meet the following minimum requirements:

* Freely and publicly available to all participants
* Produced reliably by an operational data product
* Does not incorporate reference-grade surface monitor data in any way
* Provides clear value beyond existing approved sources

Keep in mind that data sources used in this challenge cannot be derived from models that use reference-grade monitoring data as input, or include data collected from reference-grade monitors of NO2 or PM2.5 pollutants. [Low-cost sensor data](https://openaq.org/#/locations?page=1&grade=Low-Cost_Sensor) may be considered as a separate category, as they explicitly include sensors that are less expensive to manufacture and are typically less accurate and robust compared with reference-grade monitors. Any data in this category must be approved before use, based on clear documentation of how the data is incorporated and a demonstrated case for added value.

**Any requests to add approved data sources must be received by February 21 to be considered.** An announcement will be made to all challenge participants if your data source has been approved for use.

Labels (outputs)

Reference grade ground monitor data will be provided for all training times and geographies as target measures for model output. PM2.5 is reported in μg/m3, or micrograms per cubic meter.

The training period for PM2.5 spans Feb 2018 - Dec 2020. The test period spans two disjoint periods: Jan 2017 - Jan 2018 and Jan 2021 - Aug 2021.

train\_labels.csv contains the following columns:

* datetime (string): The UTC datetime of the measurement in the format YYYY-MM-DDTHH:mm:ssZ. A value represents the average between 12:00am to 11:59pm local time. The datetime provided represents the start of that 24 hour period in UTC time. Remember that for each observation, you may only use input values that are available *before* this time.
* grid\_id (string): A 5-character alphanumeric ID that uniquely identifies a 5x5cm grid cell
* value (float): A float indicating the average daily reference-grade monitor measurement for PM2.5

A unique row is identified by the combination of datetime and grid\_id.

After entering the challenge, you can download the training data on the [data download](https://www.drivendata.org/competitions/88/competition-air-quality-pm/data/) page.

Metadata

Additionally, we provide two metadata csv files.

grid\_metadata.csv contains metadata about each grid cell and contains the following columns:

* grid\_id (string): A 5-character alphanumeric ID that uniquely identifies a 5x5 km grid cell
* location (string): The location associated with a grid cell (Delhi, Los Angeles(SoCAB), or Taipei)
* tz: The timezone used to localize the dates. Note that the dates for Los Angeles (SoCAB) ignore daylight savings time and use the Etc/GMT+8 timezone.
* wkt ([WKT](https://en.wikipedia.org/wiki/Well-known_text_representation_of_geometry)): The geometry / polygonal coordinates of the grid cell

pm25\_satellite\_metadata.csv contains metadata about hosted satellite data. Each file for a particular dataset is referred to as a granule:

* granule\_id (str): The filename for each granule
* time\_start (datetime): The start time of the granule in YYYY-DD-MMTHH:mm:ss.sssZ
* time\_end (datetime): The end time of the granule in YYYY-DD-MMTHH:mm:ss.sssZ
* product (str): The concise name for the satellite data source
* location (str): One of the three locations for this challenge
* us\_url (str): The file location of the granule in the public s3 bucket in the US East (N. Virginia) region
* eu\_url (str): The file location of the granule in the public s3 bucket in the Europe (Frankfurt) region
* as\_url (str): The file location of the granule in the public s3 bucket in the Asia Pacific (Singapore) region
* cksum (int): The result of running the unix cksum command on the granule
* granule\_size (int): The filesize in bytes

The cksum and granule\_size columns are especially useful for confirming that downloaded files are not corrupted.

**TIP**: To identify the corresponding granule for a given observation from train\_labels.csv, you can use the grid\_metadata.csv to first find the location. Then, using the location and datetime, you can find s3 locations of relevant granules using satellite\_metadata.csv. Remember that you are not allowed to use future data, so the time\_end of an granule must occur on the same day of a given observation.

Submission format

The format for submissions is a .csv with columns: datetime, grid\_id, value. The combined grid\_id and datetime are the row identifiers and should exactly match the provided submission format example. You will fill in the value column with your predictions.

To create a submission, download the submission format and replace the placeholder values in the prediction column with your predicted values. Prediction values must be floats or they will not be scored correctly.

EXAMPLE

For example, if you predicted a value of 1.0 for the first 5 grid cells for January 7, 2017, your predictions would look like the following:

| **datetime** | **grid\_id** | **value** |
| --- | --- | --- |
| **2017-01-07T16:00:00Z** | **1X116** | 1.0 |
| **2017-01-07T16:00:00Z** | **9Q6TA** | 1.0 |
| **2017-01-07T16:00:00Z** | **KW43U** | 1.0 |
| **2017-01-07T16:00:00Z** | **VR4WG** | 1.0 |
| **2017-01-07T16:00:00Z** | **XJF9O** | 1.0 |

And the first few rows and columns of the .csv file that you submit would look like:

datetime,grid\_id,value

2017-01-07T16:00:00Z,1X116,1.0

2017-01-07T16:00:00Z,9Q6TA,1.0

2017-01-07T16:00:00Z,KW43U,1.0

2017-01-07T16:00:00Z,VR4WG,1.0

2017-01-07T16:00:00Z,XJF9O,1.0

You can see an example of the format that your submission must conform to, including headers and row names, in submission\_format.csv.

In this challenge, predictions for a given test grid cell may use approved features and predictions from other test samples that would be available at the time of inference (i.e., it is not necessary for each test sample to be processed independently without the use of information from other cases in the test set).

REQUIREMENTS FOR WINNING SOLUTIONS

Note that while only grid cells listed in the submission format are required for evaluation during the challenge, winning solutions must be able to produce predictions for the same grid cells on a single new day.

Finalists will need to deliver a solution that includes a trained model that ingests a new date, processes the relevant features, and outputs predicted pollutant concentrations on that date without additional training. Algorithms should be able to run inference within a reasonable runtime of one hour or less to predict one day’s concentrations for all three cities on a single GPU node. Submitted models will be run on an out-of-sample time period to confirm that they are able to execute successfully with comparable performance to the test set.

A readme file accompanying these solutions shall clearly describe the data sources used in both training and inference, including a complete description of where they are incorporated and how they ensure time is treated correctly during inference, along with easy-to-follow instructions for running all parts of the solution. Additonal instructions will be provided to finalists at the end of the submission period.

Performance metric

To measure your model’s performance, we’ll use a metric called the coefficient of determination **R2** (R squared). R2 indicates the proportion of the variation in the dependent variable that is predictable from the independent variables. This is an accuracy metric, so a higher value is better.

R2 is defined as:

R2=1−residual sum of squared errorstotal sum of squared errors=1−∑ni=1(yi−yi^)2∑ni=1(yi−y¯)2R2=1−residual sum of squared errorstotal sum of squared errors=1−∑i=1n(yi−yi^)2∑i=1n(yi−y¯)2

* nn = number of values in the dataset
* yiyi = iith true value
* yi^yi^ = iith predicted value
* y¯y¯ = average of all true yy values

For each submission, a secondary metric called **Root Mean Square Error (RMSE)** will also be reported on the leaderboard. RMSE is the square root of the mean of squared differences between estimated and observed values. This is an error metric, so a lower value is better.

RMSE is defined as:

RMSE=1N∑i=0N(yi−yi^)2−−−−−−−−−−−−−⎷RMSE=1N∑i=0N(yi−yi^)2

where:

* yi^yi^ is the iith predicted value
* yiyi is the iith true value
* NN is the number of samples

While both R2 and RMSE will be reported, only R2 will be used to determine your official ranking and prize eligibility.