HINDCASTING METHODS FOR SO_2 DISPERSION MODELS

Deliverable ID:	D3.1
Dissemination Level:	PU
Project Acronym:	ALARM
Grant:	891467
Call:	Call: H2020-SESAR-2019-2
Topic:	SESAR-ER4-05-2019 Environment and Meteorology for ATM
Consortium Coordinator:	UC3M
Edition date:	01 December 2021
Edition:	00.02.00
Template Edition:	02.00.02

Founding Members

EUROPEAN UNION EUROCONTROL

SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Ritthik Bhattacharya	WP3 leader/ Senior Scientist	04/11/2021
Tim Winter (SATAVIA)	WP3 Project Manager/Project Engineer	04/11/2021
Zainab Hakim (SATAVIA)	WP3 Environmental Scientist/Environmental Data Scientist	04/11/2021

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Conor Farrington	Communications Lead	04/11/2021
Hugues Brenot	WP2 leader	09/11/2021
Manuel Soler	Project Coordinator	12/11/2021

Approved for submission to the SJU By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
Hugues Brenot (BIRA)	WP2 leader	12/11/2021
Manuel Soler (UC3M)	Project Coordinator	01/12/2021
Ritthik Bhattacharya (SATAVIA)	WP3 leader	12/11/2021
Riccardo Biondi (UniPad)	WP4 leader	12/11/2021
Sigrun Matthes (DLR)	WP5 leader	12/11/2021
Tanja Bolic (SymOpt)	WP6 leader	12/11/2021
Javier García-Heras (UC3M)	WP7 leader	12/11/2021

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date

Document History

Edition	Date	Status	Author	Justification
00.00.01	03/11/2021	Initial Draft	Ritthik Bhattacharya	New document
00.00.02	04/11/2021	Complete Draft	Ritthik Bhattacharya	Sent for internal review
00.01.00	19/11/2021	Final Document	Ritthik Bhattacharya, Tim Winter, Zainab Hakim	Completed Review
00.02.00	01/12/2021	Reviewed Document	Ritthik Bhattacharya, Tim Winter, Zainab Hakim	Submission

ALARM

MULTI-HAZARD MONITORING AND EARLY WARNING SYSTEM

This Deliverable is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 891467 under European Union's Horizon 2020 research and innovation programme.

Abstract

The main objective of WP3 of the ALARM project is to blend observational sulphur dioxide (SO2) data with model hindcast SO_2 data in order to develop a bias correction metric which, in turn, can be used to build an accurate alarm forecast system for airports based on corrected model forecast data.

The objective of this Deliverable is to describe the database provided as part of WP3. The Deliverable includes a summary and a detailed description of the raw data inputs (including details about observational data used), the cleaning and storage methodology and the database output. The cleaning methodology includes comparison with model datasets and bias correction techniques. The methodology explores the applicability of bias correction techniques to correct model data at airports based on the observational data. These techniques will be used to correct forecasts for airports that then feed into an early warning system. In addition, the long historical observational dataset allows us to estimate historical extremes of SO_2 associated with an airport.

Table of Contents

List of Tables

No table of figures entries found.

List of Figures

1.– Introduction

1.1 Introduction to ALARM

The overall objective of ALARM is to develop a prototype global multi-hazard monitoring and Early Warning System (EWS).

A *global multi-hazard monitoring* system means near-real time (NRT) and continuous global Earth observations from satellite and ground-based networks, with the objective of *generating prompt alerts of natural hazards* affecting Air Traffic Management (ATM) and to provide information for enhancing situational awareness and providing resilience in crisis.

NRT data (with delay of delivery from 45 min to <4h) and tailored products from ground-based and satellite systems, as well as algorithms relying on meteorological forecast data, will be used to feed models capable of detecting (creation of alert products) and predicting (nowcasting/forecasting) the risk/displacement of:

- particles in suspension and gas derived from natural hazards (volcanic ash and SO_2, dust clouds from sandstorms, and smoke from forest fire)
- severe weather situations such as deep convection and extreme weather
- space weather regarding exposure to increased levels of radiation during flight
- environmental hotspots potentially contributing to global warming

In summary, NRT data and associated products and algorithms can be used to anticipate severe hazards and foster better decision-making.

1.2 Purpose of the document within ALARM project

The main objective of WP3 of the ALARM project is to blend observational sulphur dioxide (SO_2) data with model hindcast SO_2 data in order to develop a bias correction metric which, in turn, can be used to build an accurate alarm forecast system for airports based on corrected model forecast data.
The input from WP3 into the ALARM early warning system (EWS) will consist of:

- Bias corrected surface SO_2 concentration at airport locations
- Raw (uncorrected) timeseries and profiles of SO_2 concentration above airport locations
- Other SATAVIA capabilities mean we may be able to provide surface and/or vertical profiles of other chemical species (including CO, NO_x, sea salt, black carbon, and organic matter, among others).

2. Identify Observational Training Dataset

Task 3.1 (Identify Observational Training Dataset) focuses on identifying a set of observational datasets across the globe, which SATAVIA has quality-controlled and cleaned sufficiently to act as a training dataset. The observational data will be used as a baseline to compare and correct SO_2 data from the Copernicus Atmosphere Monitoring Service (CAMS) analysis/reanalysis model (Task 3.2), with the same algorithm then applied to SO_2 forecast data to ensure accuracy of forecast data (Task 3.3). This forecast can then be integrated into other ALARM WPs. SATAVIA has downloaded daily and hourly SO_2 data readings from multiple locations across the globe, conducting the tasks detailed below.

2.1. Download data from EPA

The US Environmental Protection Agency (EPA) is a United States government database containing SO_2 readings from more than 500 locations in the US, maintaining data standards across their datasets to provide accurate and reliable public data. The EPA works to US government data standard CIO 2133.0.

2.2. Download data from EBAS

The European Monitoring and Evaluation Programme EBAS is a database hosting observational data from atmospheric chemical composition and physical properties from observation locations around Europe. EBAS hosts data submitted by data originators in support of several national and international programs, and as such is a reliable source of observation data.

2.3. Download data from CPCB

The Central Pollution Control Board (CPCB) is part of the Ministry of Environment and Climate Change for the Government of India and has a national database for air quality management. The CPCB hosts an open database for air quality, containing observational data from over 300 locations across India – 265 of these stations provide live updates. During the assessment of the dataset, the air quality data were found to have inconsistent temporal resolution and poor correlation with the CAMS model. This made it difficult to use these data when starting to develop the algorithm in Task 3.3.

2.4. Download data from CNEMC

The China National Environmental Monitoring Centre (CNEMC) stores data for observational SO_2 data for China. However, a Chinese mobile number is required to create an account to then access and download data from the system. Therefore, this dataset will not be used for the training dataset.

2.5. Download data from Other Locations

After some research, it became clear that other regions of the globe did not store consistent timeseries observational SO_2 data that are readily available to the public and available for download for research project use.

2.6. Clean and Store Data

SATAVIA collated and cleaned the downloaded data so that they were consistent across locations and easy to use when inputting into a correction algorithm. Each dataset is individually cleaned and stored depending on the state of the data at time of download. The SO_2 datasets are stored in csv format (or in pickle format where the sizes are too large for csv to be efficient) in the Amazon Web Services (AWS) cloud. Both these formats are easily accessible via Python or R. The data are stored by year for all observation stations and as a daily average for each observation location. We can additionally provide hourly time series for the US data.

2.7. Automate near-real time download of all datasets

SATAVIA plan to automate the download of the observational data in the future when the project infrastructure is in place. Whilst the algorithm is in development, the near real time download of observational data is not required. Upon completion of the correction algorithm (Task 3.3), SATAVIA will automate the download of the observational data to ensure the algorithm is up to date and producing accurate corrections.

2.8. Issues and Corrective Actions

Issues identified:

- There is no totally global dataset of SO_2 observations, so data are spatially limited. Therefore only a few regions are to be used as a baseline for the correction algorithm.
- Records that are available are often patchy and discontinuous, so data are also temporally limited
- Missing metadata means SATAVIA cannot be sure of the accuracy of some of the measurements – this is dependent on organisational data standards
- Observation locations are often not consistent with the locations of data provided in the CAMS dataset (ie at airport locations). The observational data geographically near an airport may still be at a different elevation compared to the airport elevation. Therefore, the observational data values may not be completely reliable for correcting the CAMS datapoints.

Corrective actions:

- Develop the algorithm using most reliable datasets with best locational coverage (i.e., US and Europe dataset)

3. Input Data

3.1 USA EPA Dataset

3.1.1 Ground Based Observational Data in the USA

Sulphur Dioxide data for the locations in the United States of America (US) have been downloaded from the U.S. Environmental Protection Agency (EPA). It is a United States government database containing SO_2 readings from more than 500 locations in the US, maintaining data standards across their datasets to provide accurate and reliable public data. All of these data come from the EPA Air Quality System (AQS). Data collection agencies report their data to the EPA via the AQS and this calculates types of summary data. All the files are available as comma separated value (CSV) file format that are compressed and accessible via the US EPA AirData website. For the purposes of the ALARM project, SATAVIA downloaded both daily and hourly summary data. The files contain data for every monitoring location in the data base for the specific duration (hourly/ daily). The files are separated by parameter, meaning that SO_2 files can be downloaded individually.

For daily data, the file contains a daily summary record that is:

- The aggregate of all sub-daily measurements taken at the monitor
- The single sample value if the monitor takes a single, daily sample (e.g., there is only one sample with a 24-hour duration). In this case, the mean and max daily sample will have the same value

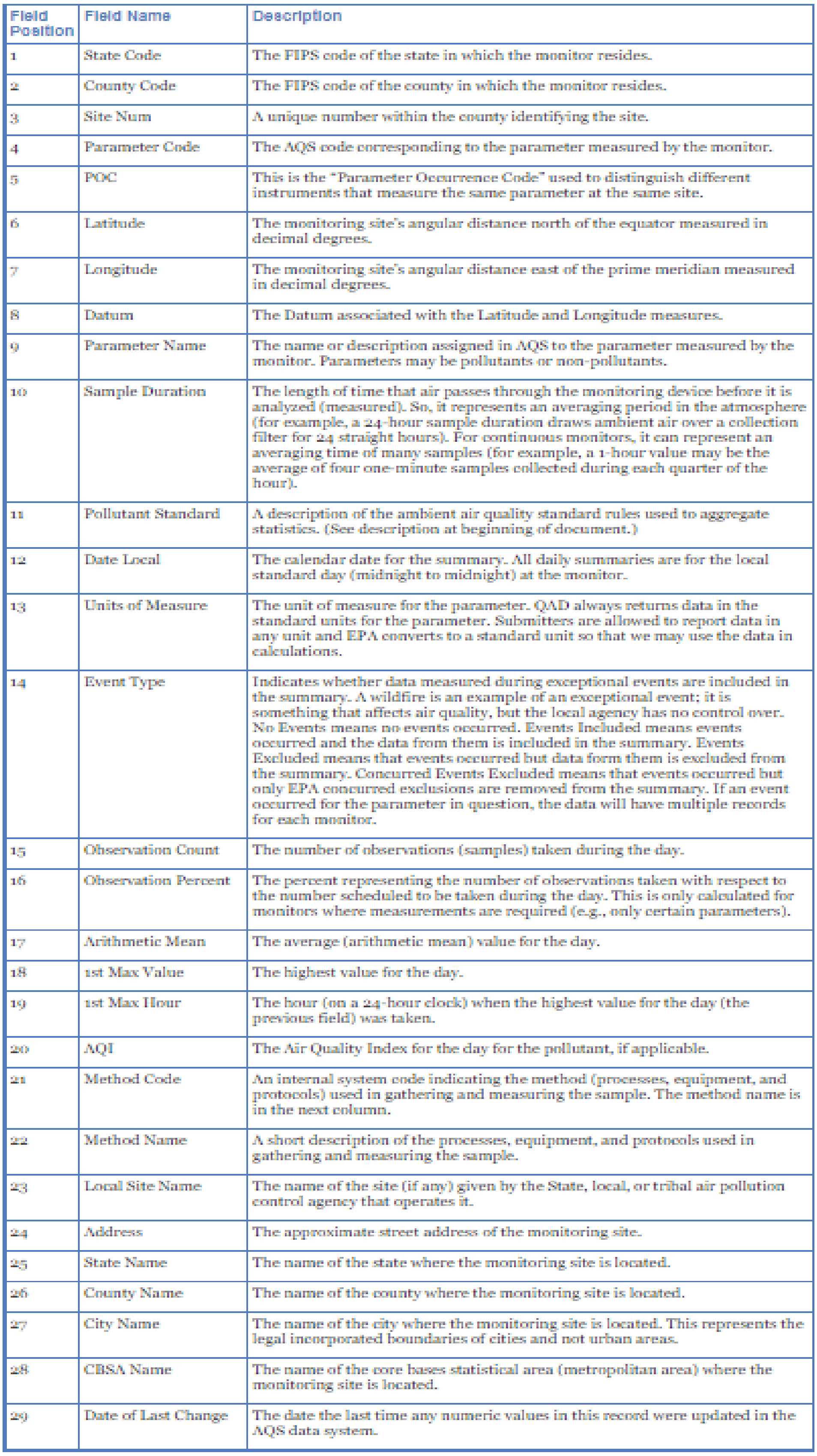
The daily summary files contain (at least) one record for each monitor that reported data for the given day. There may be multiple records for SO_2 if:

- There are calculated sample durations for the pollutant. For example, SO_2 is sometimes reported as 1-hour samples and EPA calculates 24-hour averages

EPA does receive sample data reported at durations other than hourly. For example, some SO_2 measurements are reported as 5-minute samples. These samples have not been included in these hourly files but their aggregates are included in the daily files. For the vast majority of stations, however, the daily resampling of the hourly data converges to the daily data.

3.1.2 Data Format

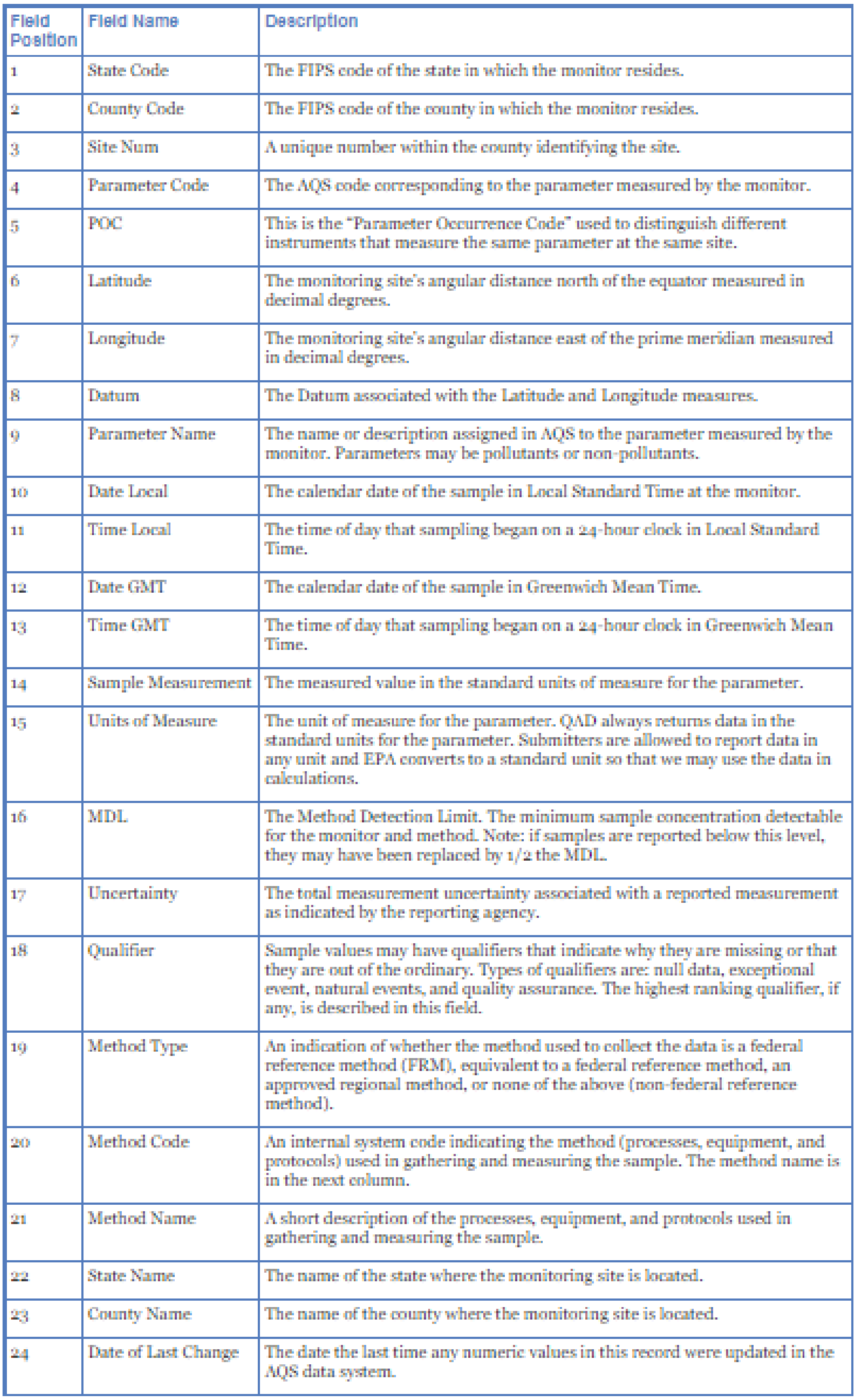
The raw daily data files for the SO_2 data are in CSV format and contain the following columns (Fig. 1):



Field Position	Field Name	Description
1	State Code	The FIPS code of the state in which the monitor resides.
2	County Code	The FIPS code of the county in which the monitor resides.
3	Site Num	A unique number within the county identifying the site.
4	Parameter Code	The AQS code corresponding to the parameter measured by the monitor.
5	POC	This is the "Parameter Occurrence Code" used to distinguish different instruments that measure the same parameter at the same site.
6	Latitude	The monitoring site's angular distance north of the equator measured in decimal degrees.
7	Longitude	The monitoring site's angular distance east of the prime meridian measured in decimal degrees.
8	Datum	The Datum associated with the Latitude and Longitude measures.
9	Parameter Name	The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants.
10	Sample Duration	The length of time that air passes through the monitoring device before it is analyzed (measured). So, it represents an averaging period in the atmosphere (for example, a 24-hour sample duration draws ambient air over a collection filter for 24 straight hours). For continuous monitors, it can represent an averaging time of many samples (for example, a 1-hour value may be the average of four one-minute samples collected during each quarter of the hour).
11	Pollutant Standard	A description of the ambient air quality standard rules used to aggregate statistics. (See description at beginning of document.)
12	Date Local	The calendar date for the summary. All daily summaries are for the local standard day (midnight to midnight) at the monitor.
13	Units of Measure	The unit of measure for the parameter. QAD always returns data in the standard units for the parameter. Submitters are allowed to report data in any unit and EPA converts to a standard unit so that we may use the data in calculations.
14	Event Type	Indicates whether data measured during exceptional events are included in the summary. A wildfire is an example of an exceptional event; it is something that affects air quality, but the local agency has no control over. No Events means no events occurred. Events Included means events occurred and the data from them is included in the summary. Events Excluded means that events occurred but data form them is excluded from the summary. Concurred Events Excluded means that events occurred but only EPA concurred exclusions are removed from the summary. If an event occurred for the parameter in question, the data will have multiple records for each monitor.
15	Observation Count	The number of observations (samples) taken during the day.
16	Observation Percent	The percent representing the number of observations taken with respect to the number scheduled to be taken during the day. This is only calculated for monitors where measurements are required (e.g., only certain parameters).
17	Arithmetic Mean	The average (arithmetic mean) value for the day.
18	1st Max Value	The highest value for the day.
19	1st Max Hour	The hour (on a 24-hour clock) when the highest value for the day (the previous field) was taken.
20	AQI	The Air Quality Index for the day for the pollutant, if applicable.
21	Method Code	An internal system code indicating the method (processes, equipment, and protocols) used in gathering and measuring the sample. The method name is in the next column.
22	Method Name	A short description of the processes, equipment, and protocols used in gathering and measuring the sample.
23	Local Site Name	The name of the site (if any) given by the State, local, or tribal air pollution control agency that operates it.
24	Address	The approximate street address of the monitoring site.
25	State Name	The name of the state where the monitoring site is located.
26	County Name	The name of the county where the monitoring site is located.
27	City Name	The name of the city where the monitoring site is located. This represents the legal incorporated boundaries of cities and not urban areas.
28	CBSA Name	The name of the core bases statistical area (metropolitan area) where the monitoring site is located.
29	Date of Last Change	The date the last time any numeric values in this record were updated in the AQS data system.

Figure 1 - USA EPA SO2 Raw Data Columns (www.epa.gov, 2021)

The raw hourly data files for SO_2data are in CSV format and contain the following columns (Fig. 2):



Field Position	Field Name	Description
1	State Code	The FIPS code of the state in which the monitor resides.
2	County Code	The FIPS code of the county in which the monitor resides.
3	Site Num	A unique number within the county identifying the site.
4	Parameter Code	The AQS code corresponding to the parameter measured by the monitor.
5	POC	This is the "Parameter Occurrence Code" used to distinguish different instruments that measure the same parameter at the same site.
6	Latitude	The monitoring site's angular distance north of the equator measured in decimal degrees.
7	Longitude	The monitoring site's angular distance east of the prime meridian measured in decimal degrees.
8	Datum	The Datum associated with the Latitude and Longitude measures.
9	Parameter Name	The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants.
10	Date Local	The calendar date of the sample in Local Standard Time at the monitor.
11	Time Local	The time of day that sampling began on a 24-hour clock in Local Standard Time.
12	Date GMT	The calendar date of the sample in Greenwich Mean Time.
13	Time GMT	The time of day that sampling began on a 24-hour clock in Greenwich Mean Time.
14	Sample Measurement	The measured value in the standard units of measure for the parameter.
15	Units of Measure	The unit of measure for the parameter. QAD always returns data in the standard units for the parameter. Submitters are allowed to report data in any unit and EPA converts to a standard unit so that we may use the data in calculations.
16	MDL	The Method Detection Limit. The minimum sample concentration detectable for the monitor and method. Note: if samples are reported below this level, they may have been replaced by 1/2 the MDL.
17	Uncertainty	The total measurement uncertainty associated with a reported measurement as indicated by the reporting agency.
18	Qualifier	Sample values may have qualifiers that indicate why they are missing or that they are out of the ordinary. Types of qualifiers are: null data, exceptional event, natural events, and quality assurance. The highest ranking qualifier, if any, is described in this field.
19	Method Type	An indication of whether the method used to collect the data is a federal reference method (FRM), equivalent to a federal reference method, an approved regional method, or none of the above (non-federal reference method).
20	Method Code	An internal system code indicating the method (processes, equipment, and protocols) used in gathering and measuring the sample. The method name is in the next column.
21	Method Name	A short description of the processes, equipment, and protocols used in gathering and measuring the sample.
22	State Name	The name of the state where the monitoring site is located.
23	County Name	The name of the county where the monitoring site is located.
24	Date of Last Change	The date the last time any numeric values in this record were updated in the AQS data system.

Figure 2 - USA EPA SO2 Raw Data Columns (www.epa.gov, 2021)

3.1.3 – Cleaning and Formatting

The data are cleaned using Python and within the Jupyter Notebook IDE. The purpose of cleaning the data from their raw data form is to ensure that a uniform database of observational data is available to act as a training dataset for the bias correction algorithms developed for deliverable 3.2.

In order to clean the US EPA data into a consistent format, the following steps were taken:

- Connect to shared database of raw data where the files had been downloaded to and download the files from this location onto a local machine or temporarily onto cloud compute environment
- Use Python methods to read the csv into a pandas dataframe for ease of data manipulation
- Create a Python function that cleans the data to a manageable format. For each observation location for a certain time step (daily or hourly):
  - Select only: 'State Code', 'County Code', 'Site Number', 'City Name', 'County Name', 'State Name', 'Latitude', 'Longitude', 'Date Local' and 'Arithmetic Mean' from the columns as these contain the relevant information
  - Create a singular unique station identification number by joining state, county and site code
  - Reduce columns to: 'station_number', 'station_location', 'Latitude', 'Longitude', 'Date Local' and 'Arithmetic Mean'
  - Drop any duplicate stations (using POC 1 i.e. the primary monitor, where multiple monitors are available) or days where the amount of data collected is insufficient to aggregate
  - Read and upload dataframes into cloud database

3.2 EU EBAS Dataset

3.2.1 Ground Based Observational Data in Europe

Sulphur dioxide data for the European stations are downloaded from “The co-operative programme for monitoring and evaluation of the Long-Range transmission of air pollutants in Europe” (unofficially 'European Monitoring and Evaluation Programme' = EMEP). It is a scientifically based and policy driven programme under the Convention Long-Range Transboundary Air Pollution (CLRTAP) for international co-operation to solve transboundary air pollution problems. The data from EMEP is hosted by EBAS. EBAS is a database infrastructure developed and operated by NILU – Norwegian Institute for Air Research.

The EMEP stations have the following characteristics: they must be placed in rural areas, away from building areas; their placement must be a site at least 40 km from industrial sources of pollutants; they must not be placed in valleys or peaks of mountains; and the setting of the station must not experience strong winds. The stations are equipped with systems for wet deposition collection, particulate sampling, and air sampling. At each station meteorological parameters (temperature, relative humidity, wind speed and direction, atmospheric pressure, quantity of precipitation) are measured together with the pH, conductivity, anion content (chlorides, sulphates, nitrates), and cations (sodium, potassium, magnesium, calcium, ammonium) in wet deposition and the concentrations of SO_2, NO_2, O_3, VOCs, SO_4^{2-}, $HNO_3 + NO_3^-$, NH_3^+, NH_4^+, and TSP in the air. In this study we use the concentrations of SO_2. Air composition is analysed every day in these stations.

3.2.2 Data Format

Data downloaded from the EBAS database through its web-portal are provided in a format based on the NASA Ames 1001 format. The EBAS-Nasa Ames format consists of a tabular data section preceded by a header containing the metadata. The first value in each data line, the independent variable according to NASA-Ames terminology, is the start time of the averaging period reported in the line, stated as day of year beginning at 0 on 1 January 00 UTC of the year. It is followed by the end time of the averaging period, which is the first dependent variable according to NASA-Ames terminology. The actual data follow as further columns of dependent variables. The extent of the available timeseries depends on the individual station. We have collected data from 2012 to 2020.

3.2.3 Cleaning and Filtering

These files have been accessed and formatted using Python. The data are then carefully filtered using the following steps:

- Data are first filtered out on the bases of their analytical method used to measure SO_2. In this case we have taken the observations measured using uv_fluoresc, filter_2pack, filter_3pack, online_IC analytical methods
- These data have a level 2 filtering and is flagged. The data are filtered only if they are flagged as 0 (with 0 being the most reliable values and 0.999 being highly errored values)
- Hourly data is aggregated daily, if any day records less than 20 hours, the day is disregarded.

Details of cleaning methodology can be accessed in: https://ebas-submit.nilu.no/templates/Inorganic-air-aerosol-chemistry-filter-based/lev2

The data cleaning methodology is derived from the data reporting standards based upon:

1. experience with the data from earlier measurements
2. relations between chemical components in air and precipitation
3. knowledge about spatial variation
4. knowledge about temporal variation
5. comparisons between measurements and estimates from theory or models.

Individual station data are then combined to create a data frame consisting of all the stations and their corresponding measurements.

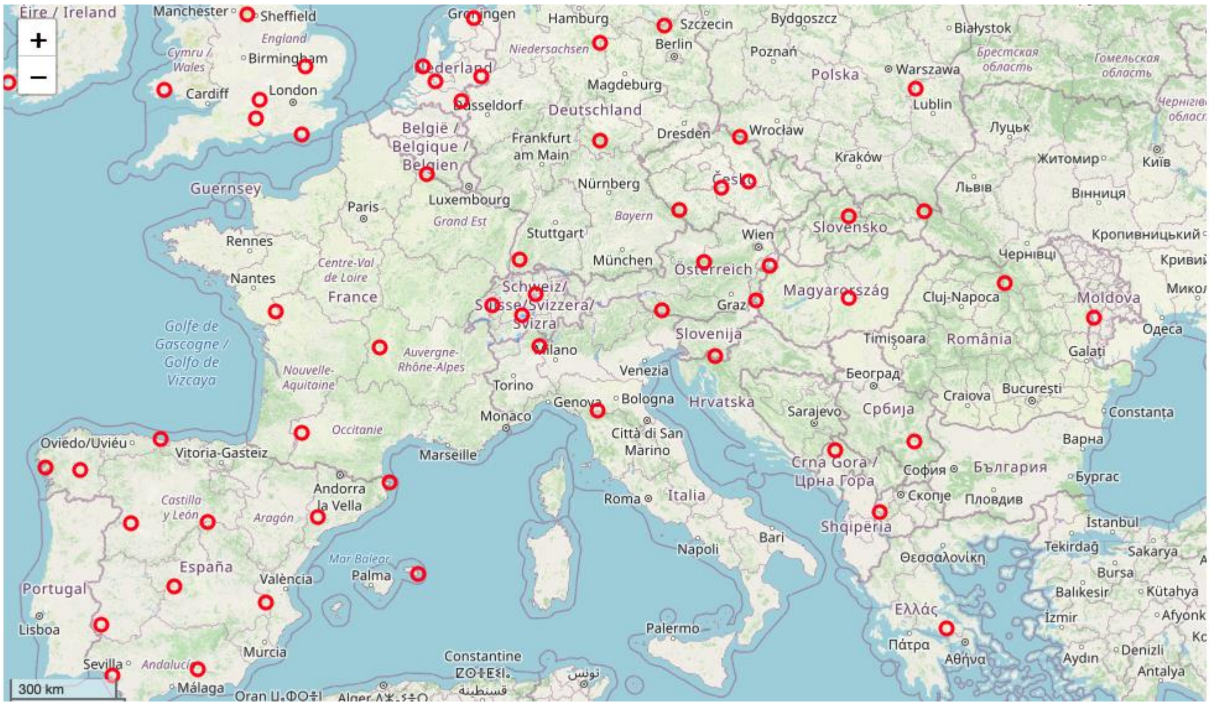


Figure 3: Observational locations (EBAS) over a part of Europe where long term daily timeseries are available. The relative sparsity of the observational locations over Europe make it challenging to leverage this data for bias correction at airport locations.

3.3 CAMS SO_2 Dataset

3.3.1 Data Source

We identified the ECMWF implementation of the Copernicus Atmosphere Monitoring Service (CAMS) analysis, reanalysis, and forecast datasets as the most appropriate source of model SO_2 data. This model data and the bias correction approaches (explored in section 3.3.3) have allowed us to explore the usability of the observational datasets to correct model forecasts. The correction method involves two steps. First it is calibrated using “training” data of past observations and past modelled data. Secondly it is applied to the future model forecasts, such that the corrected forecasts can then act as a proxy for actual observations into the future. Here we have explored the observational data, the model data and the bias correction algorithms by dividing the historical period into “train” and “test” sets and tested the quality and applicability of all these (three) components.

In practice, these methods will be applied to model forecasts for near real-time corrected forecasts. These corrected forecasts will help us create accurate time series of SO_2 at airports in the near-future and feed into creating alarms once the SO_2has crossed a certain threshold (or other exceedance metrics). The threshold will be based on historical observations at any given location.

We have downloaded data from 2012 to 2020 and stored these in blob storage as netCDF files. In addition, we have extracted data at the airport locations and stored these as parquet files in Microsoft blob stores. The extracted airports data are about 50 GB in size and can be accessed on request via the SATAVIA API (Deliverable 3.1).

We use the CAMS SO_2 dataset in order to explore bias correction techniques at airport locations. CAMS produces twice daily global forecasts for atmospheric composition. The forecasts consist of more than 50 chemical species and seven different types of aerosol, including SO_2. Once the global SO_2 analysis/ reanalysis data is obtained it is mapped on to airport locations as described in the section 3.3.2.

The initial conditions of each forecast are obtained by combining a previous forecast with current satellite observations using data assimilation. This best estimate of the state of the atmosphere at the initial forecast time step, called the analysis, provides a globally complete and consistent dataset allowing for estimates at locations where observation data coverage is low or for atmospheric pollutants for which no direct observations are available.

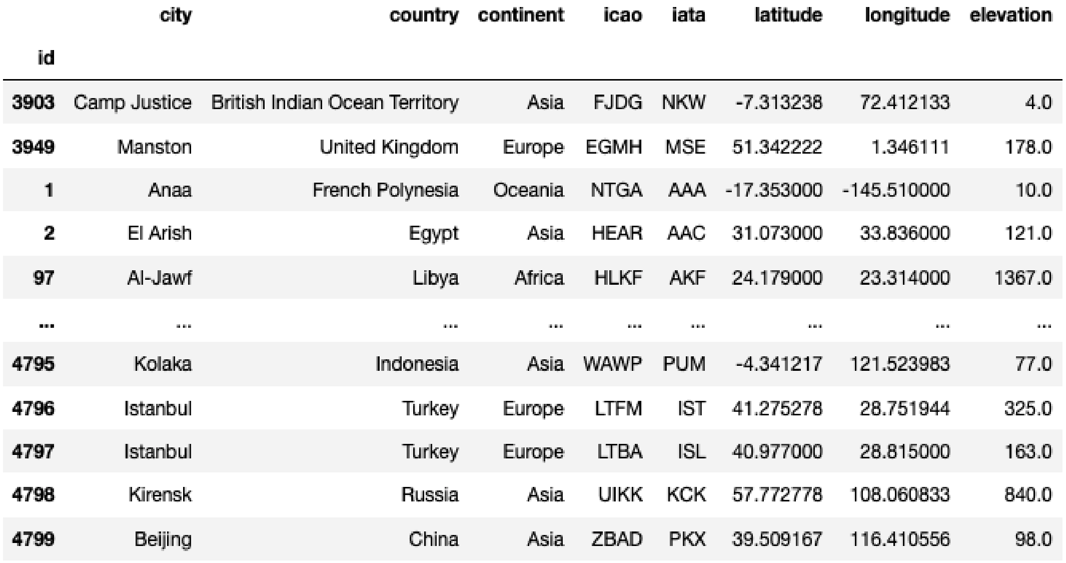
The forecast comes from a dynamical model of the atmosphere to determine the evolution of the concentrations of all chemical species over time for the next five days. Apart from the required initial state obtained above, it also uses inventory-based or observation-based emission estimates as a boundary condition at the surface. The model data is a gridded product with 60 model levels before July 2019 and 137 levels after then, downloaded in the netCDF format. The horizontal resolution is 0.4 X 0.4 degrees.

The CAMS global forecasting system is upgraded about once a year resulting in technical and scientific changes. These changes include changes to horizontal or vertical resolution, addition of new species, inventories, etc. These changes make it necessary to consider systematic changes when considering long dataset. SO_2 is forecast as mixing ratio (kg/Kg). To convert the mass mixing ratio to ppbv (used in observations) we use:

Concentration in ppbv = (Mol. mass of air / Mol. mass of so2) * 1e9 * concentration (in Kg/Kg)

3.3.2 Data interpolation to airports

The CAMS dataset is a gridded dataset that needs to be interpolated to the airport locations. Airport locations (unique ids) are defined in 3D by latitude, longitude, and elevation (Figure 4). Horizontally, the nearest grid cell is chosen to the airport location. This is found by choosing the nearest latitude and longitude. Vertically, if the airport is *above* model surface height no adjustment is made to the model levels or temperatures. Contaminant concentrations and temperatures etc have been interpolated between the heights of the model levels (corresponding to the airport height). If the airport is *below* the model surface height, the model levels and temperatures have been adjusted using the surface height and 2m temperature. For contaminants, the lowest model level is shifted down to the airport surface height (i.e the airport surface SO_2 value is assumed to be the value at the lowest model level). All model levels above have been shifted down by half this height difference. This gives us the vertical profile of SO_2 at an airport.



id	city	country	continent	icao	iata	latitude	longitude	elevation
3903	Camp Justice	British Indian Ocean Territory	Asia	FJDG	NKW	-7.313238	72.412133	4.0
3949	Manston	United Kingdom	Europe	EGMH	MSE	51.342222	1.346111	178.0
1	Anaa	French Polynesia	Oceania	NTGA	AAA	-17.353000	-145.510000	10.0
2	El Arish	Egypt	Asia	HEAR	AAC	31.073000	33.836000	121.0
97	Al-Jawf	Libya	Africa	HLKF	AKF	24.179000	23.314000	1367.0
...	...	...	...	...	...	...	...	...
4795	Kolaka	Indonesia	Asia	WAWP	PUM	-4.341217	121.523983	77.0
4796	Istanbul	Turkey	Europe	LTFM	IST	41.275278	28.751944	325.0
4797	Istanbul	Turkey	Europe	LTBA	ISL	40.977000	28.815000	163.0
4798	Kirensk	Russia	Asia	UIKK	KCK	57.772778	108.060833	840.0
4799	Beijing	China	Asia	ZBAD	PKX	39.509167	116.410556	98.0

Figure 4 – Mapping of airport id to location: latitude, longitude, elevation (in m above sea level)

The airports where the data are extracted are visualized in Figure 5.

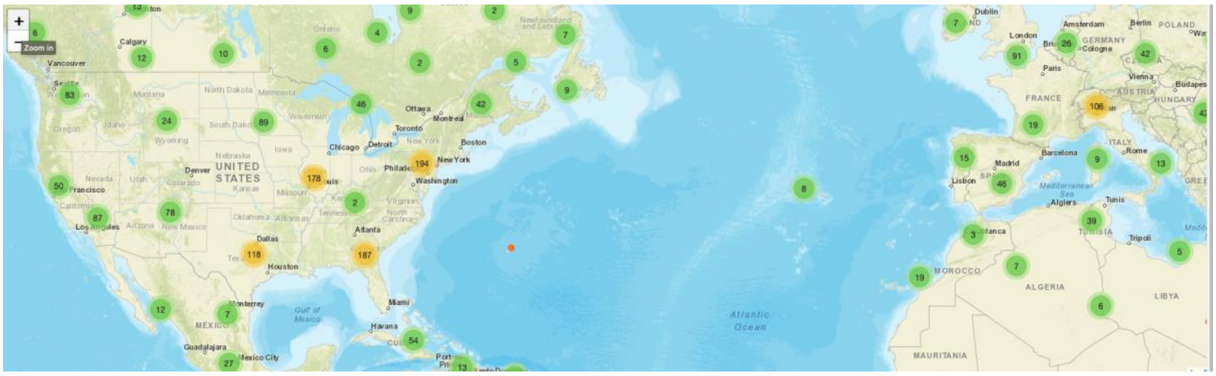


Figure 5 : A snapshots of airports (number of airports per region) where modelled (CAMS) so2 data has been archived. This data is stored globally and our present work encompasses US and Europe due to longer term observational datasets available there.

3.3.3 - Bias correction approaches

The CAMS forecast will be corrected to provide accurate real time forecasts (especially of higher percentiles) at the airport locations. Forecast of these extremes will feed into the alarm forecast system at the airports. The methods used to correct the forecasts have been trained on the historical CAMS analysis and surface observations (described in Sections 3.1 and 3.2). The methods we have explored include:

- Multiplicative correction (e.g., Borrego *et al*., 2011): corrects the test model dataset using the ratio of means of observed and model train datasets
- Quantile correction (e.g., Heo *et al*., 2019): often used to bias correct non-normal quantities like precipitation (or chemical contaminants)

- Power law correction (e.g., Bannister *et al*., 2019): allows mapping the modelled distribution on to the observed distribution by utilising a scale and a shape factor

While multiplicative correction corrects the mean and the variance, quantile and power law correction correct the entire distribution. The distribution involves the mean, the variance, and the shape factors (skewness and other higher moments). These methods have been explored here in order to establish the *validity* of using the surface observations to correct the model dataset. Further exploration is needed when these (and more approaches) will be used to correct the forecast data.

3.3.3.1 - Padding of training dataset

The time series of contaminants shows a seasonal variation. Hence the bias correction works best when window of days (for example 5 days) in the testing dataset is corrected using the same window of days in the training dataset, padded on both sides by a fixed number of days (Terink et al. 2010). This padding allows the increasing of training data sample which makes the bias correction more robust. The padding is robust to window and pad sizes. In addition, this approach leads to the convergence of bias corrected datasets using the Quantile and the Power law approaches.

These bias corrections allowed us to understand the validity of the observations to bias correct surface airport time series.

Two key points were observed:

1. The observation location needs to be well co-located with an airport locations. There are numerous locations in the US and a few within Europe where this criterion is satisfied
2. For the bias correction to work optimally, the observation time series must historically show at least moderate correlation with the model time series.

For locations where the above criteria are satisfied, the padded correction approach improves the bias and RMSE substantially, thus validating the usage of the observational dataset. An example is shown in Figure 1. In near-real time forecast corrections this methodology will be used to correct the forecasts.

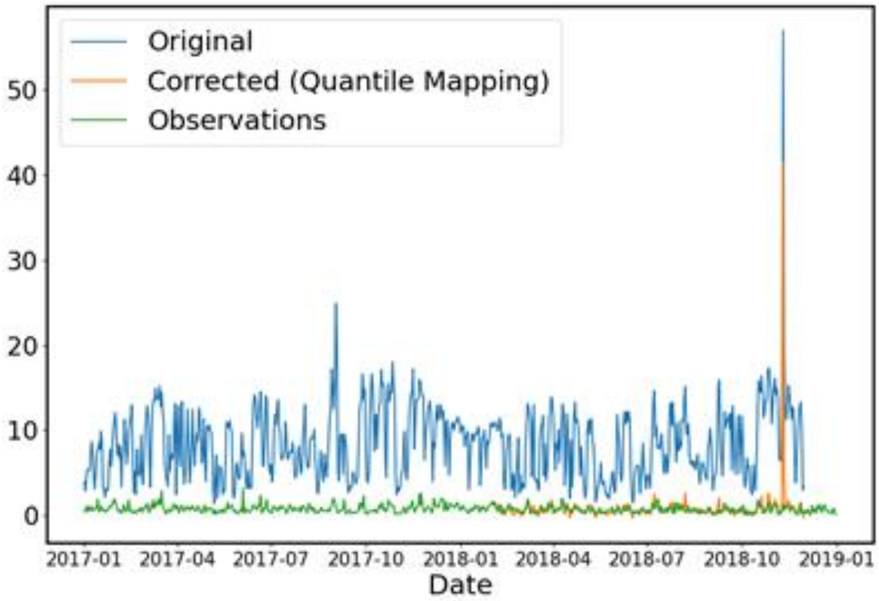


Figure 6 - Modelled data, observations, and corrected model data at Birmingham Airport, Alabama. Quantile mapping has been applied to the training data (days to correct from the previous years padded by 30 days on each side)

Figure 6 shows the correction for Birmingham airport (Alabama, U.S.A.) for which the observation location is well co-located with the airport location. Note the so2 for this location is mostly anthropogenic in source.[1]

The so2 amounts are attenuated by the meteorology, and the topography.[2] The "test period" is February 2018 onwards and the "training" data set is January 2017 to the beginning of the "testing" period. As mentioned above, each (5-day) window of the model data in the "test" period is corrected using information from the same period (padded appropriately on both sides; here 30 days). The corrected model data is seen to maintain the moderate correlation, as well as reduce the root-mean-square-error from 8.91 to 0.42 (the absolute bias is reduced from 7.57 to 0.01). In addition, the power law approach leads to very similar results.

[1] (https://adem.alabama.gov/programs/air/airquality/2017AmbientAirPlan.pdf).

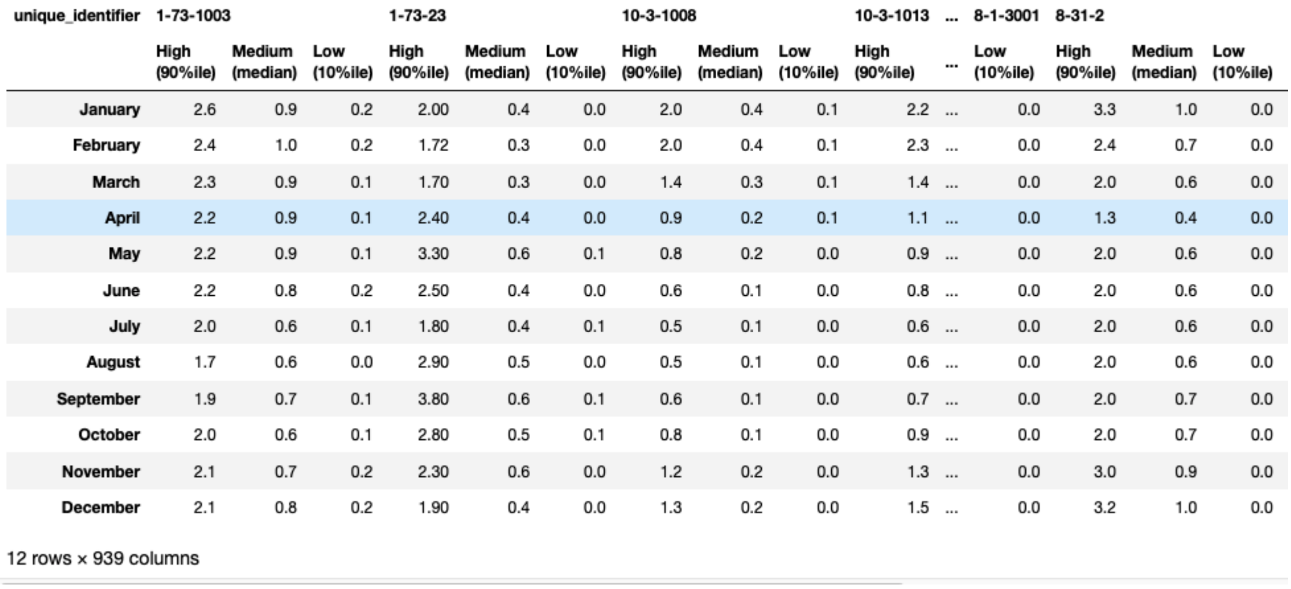
[2] (https://www.iqair.com/usa/alabama/birmingham).

4 - Outputs

Deliverable 3.1 encompasses an airport database of historical SO_2 concentrations, including timeseries and quantitative statistic descriptions.

Having downloaded, cleaned, and archived observational (surface) SO_2 data (Task 3.1) and the equivalent modelled SO_2 data for airport locations (Task 3.2), the following are currently available:

1. Raw model (historical) SO_2 concentrations at 4788 global airports (from 2012-2020). Access can be provided through SATAVIA's API
2. Raw model and observed SO_2 concentration for around 500 observational locations (surface only). US data include both daily and hourly values.



unique_identifier	1-73-1003			1-73-23			10-3-1008			10-3-1013	...	8-1-3001	8-31-2		
	High (90%ile)	Medium (median)	Low (10%ile)	High (90%ile)	Medium (median)	Low (10%ile)	High (90%ile)	Medium (median)	Low (10%ile)	High (90%ile)	...	Low (10%ile)	High (90%ile)	Medium (median)	Low (10%ile)
January	2.6	0.9	0.2	2.00	0.4	0.0	2.0	0.4	0.1	2.2	...	0.0	3.3	1.0	0.0
February	2.4	1.0	0.2	1.72	0.3	0.0	2.0	0.4	0.1	2.3	...	0.0	2.4	0.7	0.0
March	2.3	0.9	0.1	1.70	0.3	0.0	1.4	0.3	0.1	1.4	...	0.0	2.0	0.6	0.0
April	2.2	0.9	0.1	2.40	0.4	0.0	0.9	0.2	0.1	1.1	...	0.0	1.3	0.4	0.0
May	2.2	0.9	0.1	3.30	0.6	0.1	0.8	0.2	0.0	0.9	...	0.0	2.0	0.6	0.0
June	2.2	0.8	0.2	2.50	0.4	0.0	0.6	0.1	0.0	0.8	...	0.0	2.0	0.6	0.0
July	2.0	0.6	0.1	1.80	0.4	0.1	0.5	0.1	0.0	0.6	...	0.0	2.0	0.6	0.0
August	1.7	0.6	0.0	2.90	0.5	0.0	0.5	0.1	0.0	0.6	...	0.0	2.0	0.6	0.0
September	1.9	0.7	0.1	3.80	0.6	0.1	0.6	0.1	0.0	0.7	...	0.0	2.0	0.7	0.0
October	2.0	0.6	0.1	2.80	0.5	0.1	0.8	0.1	0.0	0.9	...	0.0	2.0	0.7	0.0
November	2.1	0.7	0.2	2.30	0.6	0.0	1.2	0.2	0.0	1.3	...	0.0	3.0	0.9	0.0
December	2.1	0.8	0.2	1.90	0.4	0.0	1.3	0.2	0.0	1.5	...	0.0	3.2	1.0	0.0

12 rows × 939 columns

Figure 7- Summary statistics (low, medium, and high values) for observational locations within the US for each month in the historical dataset.

Figure 7 demonstrates the extrema and the median values of SO_2 for each unique observational location within the U.S.A. As noted in section 3.1.3, a unique location corresponds to a combination of state, county, and site codes. Hourly data (available as mentioned above) have been used to create the summary statistics. This better estimates the SO_2 extremes compared to daily averages. A long time series from 2012 – 2021 has been used to generate this and seasonality can clearly be observed in the time series. Access to these data is available upon request.

5 – Conclusion

The data provided are historical observational data for the US and Europe that have been cleaned to create a long consistent daily timeseries from 2012 to 2020. In addition, for the US, consistent long (2012-2020) hourly observational data for observational locations have been provided. Finally, since exploration of these timeseries reveal seasonality**,** dataframes for the “High (90%ile)”, “Medium (50%ile)”, and “Low (10%ile)” values associated with each observational location for each month of the year have been calculated and provided. **These “High” values help us define the exceedance thresholds for the observational locations (and hence the nearby airport location) for ALARM.**

Airport data have been downloaded from 2012 to 2020 and stored in blob storage as netCDF files. In addition, we have extracted data at the airport locations as detailed above and stored these as parquet files in Microsoft blob stores. The extracted airports data total ~50 GB in size and can be accessed on request via the SATAVIA API (Deliverable 3.1). Furthermore, airport descriptions (ID, geographical information etc) have been provided as a csv file.

6 - References

Bannister, D., Orr, A., Jain, S.K., Holman, I.P., Momblanch, A., Phillips, T., Adeloye, A.J., Snapir, B., Waine, T.W., Hosking, J.S. and Allen-Sader, C., 2019. Bias correction of high-resolution regional climate model precipitation output gives the best estimates of precipitation in Himalayan catchments. *Journal of Geophysical Research: Atmospheres*, *124*(24), pp.14220-14239.

Borrego, C., Monteiro, A., Pay, M.T., Ribeiro, I., Miranda, A.I., Basart, S. and Baldasano, J.M., 2011. How bias-correction can improve air quality forecasts over Portugal. *Atmospheric Environment*, *45*(37), pp.6629-6641.

Heo, J.H., Ahn, H., Shin, J.Y., Kjeldsen, T.R. and Jeong, C., 2019. Probability distributions for a quantile mapping technique for a bias correction of precipitation data: A case study to precipitation data under climate change. *Water*, *11*(7), p.1475.

Terink, W., Hurkmans, R.T.W.L., Torfs, P.J.J.F. and Uijlenhoet, R., 2010. Evaluation of a bias correction method applied to downscaled precipitation and temperature reanalysis data for the Rhine basin. *Hydrology and earth system sciences*, *14*(4), pp.687-703.

uc3m
Universidad
Carlos III
de Madrid

BIRA•IASB

UNIVERSITAS • STUDII PADUANI •
MCCXXII

DLR

symopt

SATAVIA