Project no. 6524105

ATLAS

Artificial Intelligence Theoretical Foundations for Advanced Spatio-Temporal Modelling of Data and Processes

WP2: Data and Computing Facilities

Deliverables D2.1.1 and D.2.1.2

Report on Integrated Datasets and Database online

Program for Development of Projects in the field of Artificial Intelligence
https://ai.ipb.ac.rs/

Report prepared by:
Dimitrije Maletić (IPB)
Andreja Stojić (IPB)

Report reviewed internally by:
Zora Konjović (US)
Đorđe Obradović (US)

Date: 5/30/2021
Type: Public

Summary

The ATLAS platform will use all the available geo-referenced 1-hour air pollution and meteorological data from 2008 to 2018 covering Europe and the USA obtained from over a thousand measurement sites. Since the data will be used for the analysis, presentation, or fast retrieval for the creation of web pages, the integrated dataset includes two types of databases (SQLite3 and MariaDB server-based). Bash and C++ scripts, handling the raw data and creating and populating SQLite3 databases, were developed. To optimize computer resources for retrieving subsets of data and exhausting analyses, SQLite3 files for each country were created.

Keywords: Air pollution data, Meteorological data, Environmental agencies, Database, SQLite3, MariaDB, Bash

1. Introduction

To populate the ATLAS database, we have collected all the available geo-referenced 1-hour air pollution and meteorological data from 2008 to 2018 covering Europe and the USA from over a thousand measurement sites.

Two kinds of databases were implemented, since the data will be used for the analysis, presentation, or fast retrieval for the creation of web pages.

After the ATLAS project lifespan, an open data repository, including raw and the results data, will be online and accessible for verification, re-use, and mining by researchers. Data will be available through a user-friendly ATLAS platform in the form of a website, accessible from the standard web browsers.

The responsibility in terms of protection, storage, and use of data will be at IPB. The redundant copies and automatically scheduled daily backup of data will be made daily at 3 AM. Access to the data will be encrypted using a private SSL certificate.

It is planned that the dataset will be periodically refreshed and reanalyzed through regular research activities of the IPB team, being focused on environmental issues, particularly air pollution. In the phase of implementation, the costs will be covered by the Project. Once designed and hosted, the additional costs for the platform maintenance will be part of the regular expense budget for the institution's IT equipment maintenance.

2. Data

The ATLAS database contains:

- pollution data obtained from the:
    - United States Environmental Protection Agency (US EPA - https://aqs.epa.gov/aqsweb/airdata/download_files.html),
    - European Environmental Agency (EEA - https://www.eea.europa.eu/data-and-maps/data/air-base-the-european-air-quality-database-8#tab-figures-produced, https://www.eea.europa.eu/data-and-maps/data/aqereporting-2#tab-data-by-country, and http://discomap.eea.europa.eu/map/fme/AirQualityExport.htm) and

- UK Automatic Urban and Rural Network (AURN - https://uk-air.defra.gov.uk/networks/network-info?view=aurn),

- meteorological data from the National Oceanic and Atmospheric Administration (NOAA - https://www7.ncdc.noaa.gov/CDO/cdopoemain.cmd?datasetabbv=DS3505&countryabbv=&georegionabbv=&resolution=40), and

- meteorological fields implemented in the Air Resources Laboratory's Global Data Assimilation System (GDAS1 - ftp://arlftp.arlhq.noaa.gov/pub/archives/gdas1).

The database includes 112 variables in total:

- pollutant concentrations obtained from related agencies:
  - 59 volatile organic compounds (25 alkanes, 4 cycloalkanes, 8 alkenes, 1 alkyne, 2 diens, 16 aromatics, 1 aldehyde, 1 ester, 1 sulphide),
  - 1 sum of non-methane organic compounds, and
  - 8 criteria air pollutants (NO_x, NO, NO_2, SO_2, CO, O_3, PM_{10}, and $PM_{2.5}$), N_2O, NH_3, and PM_1;
- meteorological parameters:
  - 8 measured parameters obtained from NOAA (visibility, ceil height, wind speed and direction, relative humidity, dew point, atmospheric pressure, and temperature), and
  - 25 modelled GDAS1 meteorological parameters (planetary boundary layer height, downward short-wave radiation flux, sensible heat net flux at surface, etc.) calculated from the obtained meteorological fields and interpolated to 1-h resolution by using a developed C++ script, and
- 8 temporal variation variables (trend, hour, weekday, daylight, length of the day, month, season, and year) calculated by using a developed C++ script.

2.1 US EPA data

The US EPA 1-hour data is organized as an example provided in Table 2.1.1.

Table 2.1.1. An example of US EPA data files.

Columns	Rows	
State Code	1	1
County Code	73	73
Site Num	23	23
Parameter Code	42602	42602
POC	1	1
Latitude	33.55306	33.55306
Longitude	-86.815	-86.815
Datum	WGS84	WGS84

Parameter Name	Nitrogen dioxide (NO2)	Nitrogen dioxide (NO2)
Date Local	2021-01-01	2021-01-01
Time Local	00:00	01:00
Date GMT	2021-01-01	2021-01-01
Time GMT	06:00	07:00
Sample Measure-ment	5.8	2.7
Units of Measure	Parts per billion	Parts per billion
MDL	0.1	0.1
Uncer-tainty		
Qualifier	6	6
Method Type	FEM	FEM
Method Code	200	200
Method Name	Teledyne-API Model 200EUP or T200UP - Photolytic-Chemiluminescence	Teledyne-API Model 200EUP or T200UP - Photolytic-Chemiluminescence
State Name	Alabama	Alabama
County Name	Jefferson	Jefferson

*Source: https://aqs.epa.gov/aqsweb/airdata/hourly_42602_2021.zip

More detailed documentation on measured parameters (Tables 2.1.2a and 2.1.2b), units of measure (Table 2.1.3), measurement methods (Tables 2.1.4a and 2.1.4b), qualifiers (Table 2.1.5), etc., can be found at https://aqs.epa.gov/aqsweb/airdata/FileFormats.html and https://www.epa.gov/aqs/aqs-code-list. The details on measurement sites are provided as in Table 2.1.6.

Table 2.1.2a. An example of US EPA measured parameter details.

Columns	**Rows**		
Parameter Code	43834	43837	43162
Parameter	1,1,1,2,2-Pentafluo-roethane	1,1,1,2-Tetrachloro-ethane	1,1,1-Trichloro-2,2-bis (p-chloro-phenyl) ethane
Parameter Abbrevi-ation			TRICH
Parameter Alternate Name	HFC-125		
CAS Number	354-33-6	630-20-6	50-29-3
Standard Units	Parts per billion Car-bon	Parts per billion Carbon	Nanograms/cubic meter (25 C)
Still Valid	YES	YES	YES
Round or Truncate	R	R	R

*Source: https://aqs.epa.gov/aqsweb/documents/codetables/parameters.csv

Table 2.1.2b. An example of US EPA parameter classes.

Columns	Rows		
Class Code	APP_A_PARAMETERS	APP_A_PARAMETERS	APP_A_PARAMETERS
Class Name	Parameters subject to the 40 CFR Appendix A Regulations	Parameters subject to the 40 CFR Appendix A Regulations	Parameters subject to the 40 CFR Appendix A Regulations
Parameter Code	42101	14129	85129
Parameter	Carbon monoxide	Lead (TSP) LC	Lead PM10 LC FRM/FEM

*Source: https://aqs.epa.gov/aqsweb/documents/codetables/parameter_classes.csv

Table 2.1.3. An example of US EPA units of measure.

Unit Code	Units
001	Micrograms/cubic meter (25 C)
002	Micrograms/cubic meter (0 C)
003	Nanograms/cubic meter (25 C)

*Source: https://aqs.epa.gov/aqsweb/documents/codetables/units.csv

Table 2.1.4. An example of US EPA measurement methods.

Columns	Rows		
Parameter	1,1,1,2,2-Pentafluoroethane	1,1,1,2-Tetrachloroethane	1,1,1,2-Tetrachloroethane
Parameter Code	43834	43837	43837
Method Code	171	105	106
Recording Mode	Intermittent	Intermittent	Intermittent
Collection Description	6L Pressurized Canister	CHARCOAL-TUBE-PERSONAL-PUMP	CHARCOAL-TUBE-PERSONAL-PUMP
Analysis Description	Precon Saturn GC/MS	CARBON DISULFIDE DESORPTION GC-ECD	CARBON DISULFIDE DESORPTION GC-ECD
Method Type			
Reference Method ID			
Equivalent Method			
Federal MDL	0.2	0.4	0.4
Min Value	-0.2	-0.4	-0.4
Max Value			
Digits	2	1	1
Round Truncate Indicator	R	R	R
Units	Parts per billion Carbon	Parts per billion Carbon	Parts per billion Carbon

*Source: https://aqs.epa.gov/aqsweb/documents/codetables/methods_all.csv

Table 2.1.4b. An example of US EPA meteorological parameter measurement methods.

Columns	Rows		
Parameter	Ammonia (precip)	Ammonia (precip)	Atmospheric Stability
Parameter Code	62604	62604	61120
Method Code	017	018	101
Recording Mode	Intermittent	Intermittent	Continuous
Collection Description	ANDERSON RAINFALL BUCKET	AEROCHEM RAINFALL BUCKET (MOD)	INSTRUMENTAL
Analysis Description	AUTOMATED PHENATE VIA TRAACS 800	FLOW INJECTION (COLORIMETRIC)	ELECTRONIC OR MACHINE AVERAGE
Method Type			
Reference Method ID			
Equivalent Method			
Federal MDL	0.01	0.01	1.0
Min Value	0.0	0.0	0.0
Max Value			
Digits	2	2	0
Round Truncate Indicator	R	R	R
Units	Milligrams/liter	Milligrams/liter	Pasquill-Gifford stability class

*Source: https://aqs.epa.gov/aqsweb/documents/codetables/methods_met.csv

Table 2.1.5. An example of US EPA qualifiers.

Qualifier Code	Qualifier Description	Qualifier Type	Qaulifier Type Code	Still Active	Legacy Code
AA	Sample Pressure out of Limits.	Null Data Qualifier	NULL	YES	9967
AB	Technician Unavaliable.	Null Data Qualifier	NULL	YES	9968
AC	Construction/Repairs in Area.	Null Data Qualifier	NULL	YES	9969

*Source: https://aqs.epa.gov/aqsweb/documents/codetables/qualifiers.csv

Table 2.1.6a. An example of US EPA measurement site details.

Columns	Rows		
State Code	1	1	1
County Code	1	1	1
Site Number	1	2	3
Latitude	32.43746	32.42847	32.33266
Longitude	-86.4729	-86.4436	-86.7915
Datum	WGS84	WGS84	WGS84
Elevation	64	0	41
Land Use	RESIDENTIAL	AGRICULTURAL	FOREST
Location Setting	SUBURBAN	RURAL	RURAL

Site Established Date	5/1/1974	1/1/1980	8/31/1989
Site Closed Date	12/31/1976	12/31/1982	11/30/1990
Met Site State Code			
Met Site County Code			
Met Site Number			
Met Site Type			
Met Site Distance			
Met Site Direction			
GMT Offset	-6	-6	-6
Owning Agency	Al Dept Of Env Mgt	Al Dept Of Env Mgt	Al Dept Of Env Mgt
Local Site Name			
Address	KING ARTHUR TRAILER COURT, PRATTVILLE, AL	COUNTY RD 4 PRATTVILLE EXPERIMENT ST	1170 COUNTY RD.15 SO., SELMA, AL. 36701
Zip Code	36067		36003
State Name	Alabama	Alabama	Alabama
County Name	Autauga	Autauga	Autauga
City Name	Prattville	Prattville	Not in a City
CBSA Name	Montgomery, AL	Montgomery, AL	Montgomery, AL
Tribe Name			
Extraction Date	5/18/2021	5/18/2021	5/18/2021

*Source: https://aqs.epa.gov/aqsweb/airdata/aqs_sites.zip

Table 2.1.6b. An example of US EPA country and measurement method details.

Columns	**Rows**		
State Code	1	1	1
County Code	1	1	1
Site Number	1	1	2
Parameter Code	11103	42401	42401
Parameter Name	Benzene soluble organics (TSP)	Sulfur dioxide	Sulfur dioxide
POC	1	1	1
Latitude	32.43746	32.43746	32.42847
Longitude	-86.4729	-86.4729	-86.4436
Datum	WGS84	WGS84	WGS84
First Year of Data	1974	1974	1980
Last Sample Date	6/10/1974	8/16/1976	7/31/1982
Monitor Type	OTHER	OTHER	SLAMS

Networks			
Reporting Agency		Al Dept Of Env Mgt	Al Dept Of Env Mgt
PQAO		Al Dept Of Env Mgt	Al Dept Of Env Mgt
Collecting Agency		Al Dept Of Env Mgt	Al Dept Of Env Mgt
Exclusions			
Monitoring Objective	UNKNOWN	HIGHEST CONCENTRATION	UNKNOWN
Last Method Code	91	91	20
Last Method	HI-VOL - BENZENE EXTRACTION-SOXHLET	GAS-BUBBLER - PARAROSANILINE-SULFAMIC ACID	INSTRUMENTAL - PULSED FLUORESCENT
Measurement Scale		NEIGHBORHOOD	
Measurement Scale Definition		500 M TO 4KM	
NAAQS Primary Monitor			
QA Primary Monitor			
Local Site Name			
Address	KING ARTHUR TRAILER COURT, PRATTVILLE, AL	KING ARTHUR TRAILER COURT, PRATTVILLE, AL	COUNTY RD 4 PRATTVILLE EXPERIMENT ST
State Name	Alabama	Alabama	Alabama
County Name	Autauga	Autauga	Autauga
City Name	Prattville	Prattville	Prattville
CBSA Name	Montgomery, AL	Montgomery, AL	Montgomery, AL
Tribe Name			
Extraction Date	5/18/2021	5/18/2021	5/18/2021

*Source: https://aqs.epa.gov/aqsweb/airdata/aqs_monitors.zip

2.2 EEA data

The EEA 1-hour data is organized as an example provided in Table 2.2.1.

Table 2.2.1. An example of EEA data.

Columns	**Rows**		
Countrycode	DE	DE	DE
Namespace	http://gdi.uba.de/arcgis/rest/services/inspire/DE.UBA.AQD	http://gdi.uba.de/arcgis/rest/services/inspire/DE.UBA.AQD	http://gdi.uba.de/arcgis/rest/services/inspire/DE.UBA.AQD
AirQualityNetwork	NET.DE_BE	NET.DE_BE	NET.DE_BE
AirQualityStation	STA.DE_DEBE065	STA.DE_DEBE065	STA.DE_DEBE065
AirQualityStationEoICode	DEBE065	DEBE065	DEBE065

SamplingPoint	SPO.DE_DEBE065_CHB_dataGroup1	SPO.DE_DEBE065_CHB_dataGroup1	SPO.DE_DEBE065_CHB_dataGroup1
SamplingProcess	SPP.DE_DEBE065_CHB_automatic_GC_Duration-30minute	SPP.DE_DEBE065_CHB_automatic_GC_Duration-30minute	SPP.DE_DEBE065_CHB_automatic_GC_Duration-30minute
Sample	SAM.DE_DEBE065_1	SAM.DE_DEBE065_1	SAM.DE_DEBE065_1
AirPollutant	C6H6	C6H6	C6H6
AirPollutantCode	http://dd.eionet.europa.eu/vocabulary/aq/pollutant/20	http://dd.eionet.europa.eu/vocabulary/aq/pollutant/20	http://dd.eionet.europa.eu/vocabulary/aq/pollutant/20
AveragingTime	hour	hour	hour
Concentration	0.7000000000	0.5900000000	0.5800000000
UnitOfMeasurement	µg/m3	µg/m3	µg/m3
DatetimeBegin	2020-01-01 11:00:00 +01:00	2020-01-01 12:00:00 +01:00	2020-01-01 13:00:00 +01:00
DatetimeEnd	2020-01-01 12:00:00 +01:00	2020-01-01 13:00:00 +01:00	2020-01-01 14:00:00 +01:00
Validity	1	1	1
Verification	2	2	2

*Source: https://ereporting.blob.core.windows.net/downloadservice/DE/DE_20_7959_2021_timeseries.csv

Meta data is organized as provided in Table 2.2.2.

Table 2.2.2. An example of the EEA details on measurement sites.

Columns	**Rows**		
Countrycode	AD	AD	AD
Timezone	http://dd.eionet.europa.eu/vocabulary/aq/timezone/UTC+01	http://dd.eionet.europa.eu/vocabulary/aq/timezone/UTC+01	http://dd.eionet.europa.eu/vocabulary/aq/timezone/UTC+01
Namespace	AD.GovernAndorra.AQ	AD.GovernAndorra.AQ	AD.GovernAndorra.AQ
AirQualityNetwork	NET-AD001A	NET-AD001A	NET-AD001A
AirQualityStation	STA-AD0942A	STA-AD0942A	STA-AD0942A
AirQualityStationEoICode	AD0942A	AD0942A	AD0942A
AirQualityStationNatCode	942	942	942
SamplingPoint	SPO-AD0942A-0001	SPO-AD0942A-0005	SPO-AD0942A-0007
SamplingProces	SPP-AD0942A-0001-API100E	SPP-AD0942A-0005-TEOM1400A	SPP-AD0942A-0007-API400E
Sample	SAM-AD0942A-0001	SAM-AD0942A-0005	SAM-AD0942A-0007
AirPollutantCode	http://dd.eionet.europa.eu/vocabulary/aq/pollutant/1	http://dd.eionet.europa.eu/vocabulary/aq/pollutant/5	http://dd.eionet.europa.eu/vocabulary/aq/pollutant/7

ObservationDateBegin	2005-01-01T00:00:00	2005-01-01T00:00:00	2005-01-01T00:00:00
ObservationDateEnd			
Projection	EPSG:4979	EPSG:4979	EPSG:4979
Longitude	1.539138	1.539138	1.539138
Latitude	42.50969399946506	42.50969399946506	42.50969399946506
Altitude	1080	1080	1080
MeasurementType	automatic	automatic	automatic
AirQualityStationType	background	background	background
AirQualityStationArea	urban	urban	urban
EquivalenceDemonstrated	ref	no	ref
MeasurementEquipment	http://dd.eionet.europa.eu/vocabulary/aq/measurementequipment/API100E	http://dd.eionet.europa.eu/vocabulary/aq/measurementequipment/TEOM1400A	http://dd.eionet.europa.eu/vocabulary/aq/measurementequipment/API400E
InletHeight	3	2.5	3
BuildingDistance	6	6	6
KerbDistance	-999	-999	-999

*Source: https://discomap.eea.europa.eu/map/fme/metadata/PanEuropean_metadata.csv

2.3 NOAA data

The NOAA 1-hour data is organized as an example provided in Table 2.3.1 (https://www1.ncdc.noaa.gov/pub/data/noaa/isd-lite/isd-lite-technical-document.pdf and https://www1.ncdc.noaa.gov/pub/data/noaa/isd-lite/). A detailed description of the data fields is given in Table 2.3.2 (https://www1.ncdc.noaa.gov/pub/data/noaa/isd-lite/isd-lite-format.pdf), while the details on measurement station and country list were organized as provided in Tables 2.3.3 and 2.3.4.

Table 2.3.1 An example of the NOAA data.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12
2021	1	1	0	136	82	10280	30	51	-9999	-9999	-9999
2021	1	1	1	140	82	10277	40	57	-9999	-9999	-9999
2021	1	1	2	138	82	10274	20	46	-9999	-9999	-9999
2021	1	1	4	138	80	10266	30	41	-9999	-9999	0
2021	1	1	5	136	78	10261	40	72	-9999	-9999	-9999

*Source: https://www1.ncdc.noaa.gov/pub/data/noaa/isd-lite/2021/959860-99999-2021.gz

Table 2.3.2 An explanation of the NOAA data fields.

Field	Field length	Position	Variable	Variable description	Units	Value	Scaling factor	Missing value	Note
1	4	1-4	Observation Year	Year of observation		Rounded to nearest whole hour			
2	2	6-7	Observation Month	Month of observation		Rounded to nearest whole hour			
3	2	9-11	Observation Day	Day of observation		Rounded to nearest whole hour			
4	2	12-13	Observation Hour	Hour of observation		Rounded to nearest whole hour			
5	6	14-19	Air Temperature	The temperature of the air	Degrees Celsius		10	-9999	
6	6	20-24	Dew Point Temperature	The temperature to which a given parcel of air must be cooled at constant pressure and water vapor content in order for saturation to occur.	Degrees Celsius		10	-9999	
7	6	26-31	Sea Level Pressure	The air pressure relative to Mean Sea Level (MSL).	Hectopascals		10	-9999	
8	6	32-37	Wind Direction	The angle between true north and the direction from which the wind is blowing.	Angular Degrees (measured in a clockwise direction)		1	-9999	Wind direction for calm winds is coded as 0.
9	6	38-43	Wind Speed Rate	The rate of horizontal travel of air past a fixed point.	Meters per second		10	-9999	
10	6	44-49	Sky Condition Total Coverage Code	The code that denotes the fraction of the total celestial dome covered by clouds or other obscuring phenomena.				-9999	
11	6	50-55	Liquid Precipitation	The depth of liquid precipitation that is measured	Millimeters		10	-9999	Trace precipitation is coded as -1

			Depth Dimension - One Hour Duration	over a one-hour accumulation period.					
12	6	56-61	Liquid Precipitation Depth Dimension - Six Hour Duration	The depth of liquid precipitation that is measured over a six-hour accumulation period.	Millimeters		10	-9999	Trace precipitation is coded as -1

*Source: https://www1.ncdc.noaa.gov/pub/data/noaa/isd-lite/isd-lite-format.txt

Table 2.3.3 An example of the NOAA measurement station details.

Columns	**Rows**		
USAF	007018	007026	007070
WBAN	99999	99999	99999
STATION NAME	WXPOD 7018	WXPOD 7026	WXPOD 7070
CTRY		AF	AF
STATE			
ICAO			
LAT	+00.000	+00.000	+00.000
LON	+000.000	+000.000	+000.000
ELEV(M)	+7018.0	+7026.0	+7070.0
BEGIN	20110309	20120713	20140923
END	20130730	20170822	20150926

*Source: https://www1.ncdc.noaa.gov/pub/data/noaa/isd-history.csv

Table 2.3.4 An example of the NOAA country list details.

Column1		**Column2**	**Column3**	**Column4**
KS	**KOREA**	SOUTH		
US	**UNITED STATES**			
YY	**ST. MARTEEN**	ST. EUSTATIUS	AND SABA	

*Source: https://www1.ncdc.noaa.gov/pub/data/noaa/country-list.txt

2.4 GDAS data

Surface meteorological data was calculated and interpolated (longitude/latitude for each EEA and US EPA measurement station) from GDAS1 meteorological field data (Table 2.4.1) by using the developed C++ script. The obtained 3-hour resolution data was interpolated to 1-hour data by using the linear interpolation method.

Table 2.4.1. GDAS1 surface meteorological parameters.

Field	**Units**	**Label**	**Data Order**
Pressure at surface	hPa	PRSS	S1
Pressure reduced to mean sea level	hPa	MSLP	S2
Accumulated precipitation (6 h accumulation)	m	TPP6	S3

u-component of momentum flux (3- or 6-h average)	N/m2	UMOF	S4
v-component of momentum flux (3- or 6-h average)	N/m2	VMOF	S5
Sensible heat net flux at surface (3- or 6-h average)	W/m2	SHTF	S6
Downward short-wave radiation flux (3- or 6-h average)	W/m2	DSWF	S7
Relative Humidity at 2m AGL	%	RH2M	S8
U-component of wind at 10 m AGL	m/s	U10M	S9
V-component of wind at 10 m AGL	m/s	V10M	S10
Temperature at 2m AGL	K	TO2M	S11
Total cloud cover (3- or 6-h average)	%	TCLD	S12
Geopotential height	gpm*	SHGT	S13
Convective available potential energy	J/Kg	CAPE	S14
Convective inhibition	J/kg	CINH	S15
Standard lifted index	K	LISD	S16
Best 4-layer lifted index	K	LIB4	S17
Planetary boundary layer height	m	PBLH	S18
Temperature at surface	K	TMPS	S19
Accumulated convective precipitation (6 h accumulation)	m	CPP6**	S20
Volumetric soil moisture content	frac.	SOLM	S21
Categorial snow (yes=1, no=0) (3- or 6-h average)		CSNO	S22
Categorial ice (yes=1, no=0) (3- or 6-h average)		CICE	S23
Categorial freezing rain (yes=1, no=0) (3- or 6-h average)		CFZR	S24
Categorial rain (yes=1, no=0) (3- or 6-h average)		CRAI	S25
Latent heat net flux at surface (3- or 6-h average)	W/m2	LHTF	S26
Low cloud cover (3- or 6-h average)	%	LCLD	S27
Middle cloud cover (3- or 6-h average)	%	MCLD	S28
High cloud cover (3- or 6-h average)	%	HCLD	S29

*Source: https://www.ready.noaa.gov/gdas1.php

2.5 Temporal variation data

Temporal variation data was produced from ‘date’ variable, while the ‘length of the day’ was calculated by using the developed C++ script.

3. Database implementation

Different types of data files will be obtained from various websites and used to populate the ATLAS database. Since the data will be used for the analysis, presentation, or fast retrieval for the creation of web pages, two types of databases were implemented.

The first database was developed for general usage. The data was processed and segmented into a moderate number of portable serverless SQLite3 database files stored in compressed form. This provides many benefits such as:

- enhanced portability of the data,
- short data retrieval and processing by the user-selected applications,

- database access by Python, R, C++, Bash, and ROOT applications or docker images/notebooks.

The second, MariaDB server-based database, includes the data retrieved and stored to be processed for web page creation.

General usage of the ATLAS platform powered by PARADOX supercomputer will include a user (web) interface to initiate processing through the batch job submission to the PARADOX worker nodes (WN). Before initiating resource-demanding tasks on WN, the user will need to test their codes and ideas, which will happen more frequently and, probably, using their personal computers. When the batch job is submitted, the part of the segmented database will be copied to the WN disk space, de-compressed, and accessed for data retrieval. Subsequently, the data will be processed using various ATLAS-predefined or user-defined applications.

3.1 Database population

The most demanding task in the SQLite3 database population was joining the data of all of the obtained pollutants by measurement station and date. This task is important, since:

- joining a significant number of tables with a huge number of entries will take great computer resources each time a user needs to get data for a specific subset of measurement stations,

- measurement stations have different sets of variables - creating a table with 200 columns representing 200 pollutants obtained from agency databases is not optimal regarding the size of the database and the speed of retrieving the data from the database,

- creating a separate SQLite3 database for each measurement station with its own set of pollutants will demand a huge number of small SQLite3 files.

The compromise was to create a single SQLite3 file for each country. Tens of Bash scripts, handling the downloaded raw data from various websites and creating and populating SQLite3 databases, were developed. The bulk of data was processed initially, but the developed tools will be also used when new measurement data is acquired and added to databases through an automated process of new data retrieval.

The size of the compressed SQLite3 database files is around 68 GB (6.0 GB for EEA data provided in 92 files, 38 GB for NOAA data provided in 68 files, 7.3 GB for GDAS1 SQLite zips provided in 33 files, 14 GB for GDAS0p5 SQLite zips provided in 27 files, and 3.3 GB for US EPA data SQLite provided in 698 files), while the uncompressed size is about five times bigger.

3.2 Database access

Dimitrije

4. Conclusions

All of the available geo-referenced 1-hour air pollution and meteorological data from 2008 to 2018 covering Europe and the USA were obtained from over a thousand measurement sites. According to the data usage (analysis, presentation, or fast retrieval for the creation of web pages), the data were organized into two types of databases (SQLite3 and MariaDB server-based). Bash and C++ scripts

handling the raw data and creating and populating SQLite3 databases were developed. SQLite3 files were created for each country to optimize computer resources for retrieving subsets of data and exhausting analyses.

References

[1] European Environment Agency, https://www.eea.europa.eu/
[2] United States Environmental Protection Agency, https://www.epa.gov/
[3] National Oceanic and Atmospheric Administration, https://www.noaa.gov/
[4] Global Data Assimilation System, https://www.ready.noaa.gov/gdas1.php