PRACTITIONERS' GUIDE FOR THE PURPLEAIR AIR QUALITY LCS (LOW-COST SENSOR) DEPLOYMENT: Sensor installation, data management, data analysis and presentation.

Contents

INTRODUCTION

Until about 2010, measurements of particulate matter (PM) mass concentrations in the air were performed by governments or research organisations (usually in the Global North) that could afford high quality, regulatory grade instruments in the order of 10,000 euros, and a total station and educated personnel costs an order of magnitude higher. In the past decade, the methods on how air quality monitoring can be performed, who can perform it, and the temporal and spatial scales on which could be performed, has radically changed. These changes have been driven by the emergence of low-cost air quality sensors (LCSs), including low-cost PM monitors (Giordano et al, 2021). Due to their inherent characteristics (low cost, small size, portability) LCSs may be used to locate pollution hotspots, identify sources of pollution, supplement fixed-site monitoring stations, measure personal exposure to pollutants, educate, and enhance air quality awareness (EPA n.d.). There is no need for a high level of expertise to employ LCSs therefore they are used in citizen science projects all over the world. On the other hand, due to their low cost, LCSs do not meet the high-quality standards of the regulatory grade instruments, and their results must be addressed with care.

PurpleAir sensors are a popular example that can be used by citizens or communities to collect local air quality data for particulate matter (PM) and share it with the public. Data is sent via a Wi-Fi network to the cloud, it is presented on a map, and it can be inspected or downloaded for further analysis.

The inspection of real-time data on a map is very useful because it can give an instant idea of the air quality for people that perform outdoor physical activity, for vulnerable people etc. In many cases this information is enough. However, the downloading of current and historical data in the PC of the user, their management and analysis gives further insight on the concentration levels for an extended period, helps detect and correct data anomalies, find specific patterns in the data time series, and present elaborated results to a wider audience.

In 2021 a network of 7 PurpleAir sensors was installed at the bigger cities of 6 islands at the Aegean Sea to monitor the levels of $PM_{2.5}$ (figure 1). The monitors work till today (July 2023). Furthermore, one sensor was installed at the Thissio Air Monitoring Supersite, operated by the National Observatory of Athens to monitor the performance of the sensors. It was the first time the air quality was assessed in the area at such a great spatial and temporal scale. The research was conducted in the framework of the project ENIRISST+, which was co-financed by Greece and the European Union.

Co-financed by Greece and the European Union

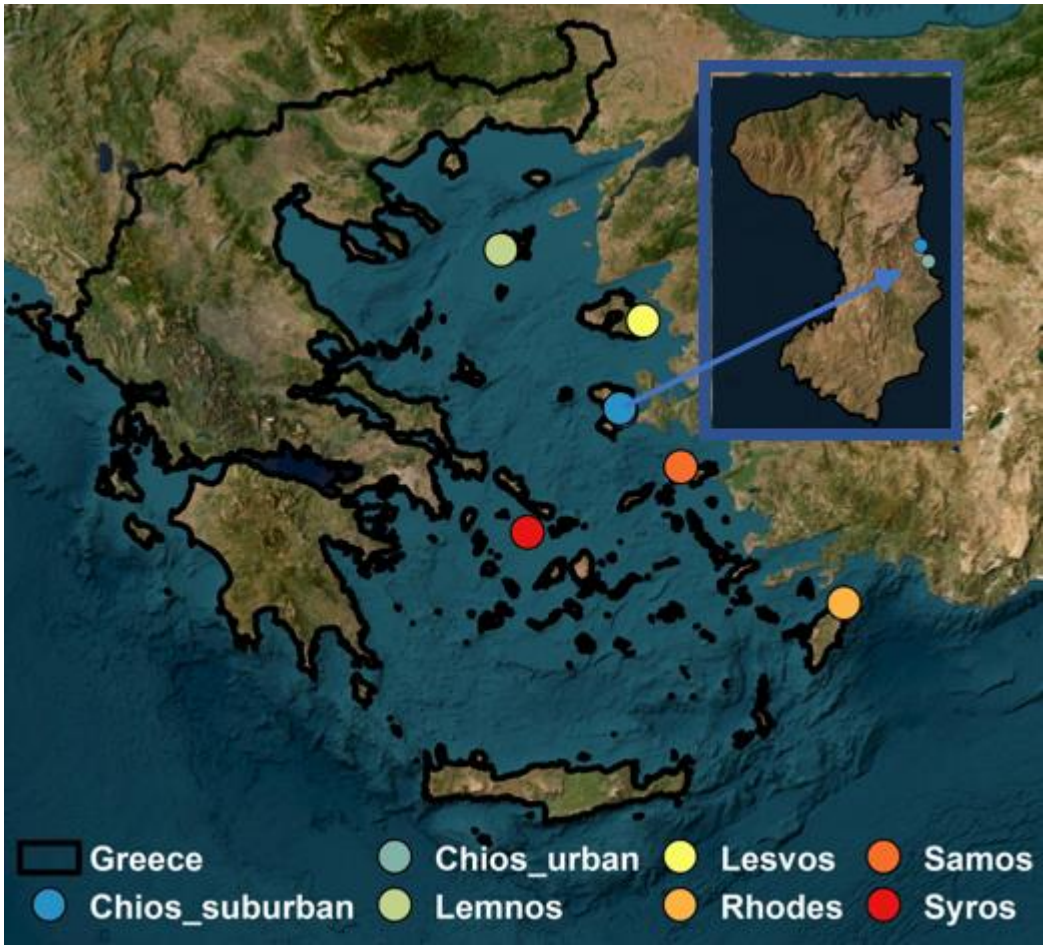


Figure 1. The network of air quality monitors at the Aegean Sea.

This guidance aims at providing information for the public (e.g., citizens and communities, non-governmental organisations, local-authorities, educational organisations) on deploying PurpleAir LCSs. First the operation principle and a description of the PurpleAir LCSs sensor is given. Then a step-by-step process for the installation of a sensor is described and troubleshooting for the most common errors is presented. Then a simple procedure for downloading, managing, analysing, and presenting data of the Purpleair air quality LCSs is proposed using the basic principles of data management.

The requirements to use this guide are:

- ➔ Access to a Purpleair air quality sensor,
- ➔ Basic knowledge of descriptive statistics and a spreadsheet software (such as Excel).

Note 1:

- ➔ In this document, guidance for sensor deployment of a particular brand of sensors (PurpleAir) are discussed. The authors do not recommend the particular or any other brand of sensors. Also, references to any specific software are made to illustrate the cases and do not constitute endorsement of the software by the authors.
- ➔ LCSs use is intended to be educational, and non-regulatory. This means that LCS cannot be used for permitting, compliance, policy, or interpretation of health effects.

Note 2:

The discussion in this guide is limited to $PM_{2.5}$ (and not to PM_{10}) because of their important effect on human health and because the measurements of $PM_{2.5}$ by PurpleAir LCSs are more accurate compared to the measurements of PM_{10}. If the PM_{10} includes a significant coarse fraction (as it happens in Sahara Desert dust episodes) the reported PM_{10} concentrations by PurpleAir LCSs appear to be questionable (Kosmopoulos et al, 2020).

SENSOR DESCRIPTION

The operation of PurpleAir sensors is based on light scattering principles (figure 2). Each unit contains two Plantower PMS5003 sensors, labelled as channel A and B. A built-in fan draws air and particles into the measurement chamber in each sensor (figure 3). The particles pass through a laser beam of 680 ± 10 nanometers (nm) wavelength light, the light of the beam is scattered by the particles and a detector detects the scattered light. The output signal is used to calculate PM_1, $PM_{2.5}$, and PM_{10} mass concentrations in the units of µg per cubic metre ($\mu g/m^3$) by a proprietary algorithm. The PurpleAir software provides uncorrected values (named correction factor CF=1 in the corresponding output file) and corrected values (named CF = atm in the file). The uncorrected values (CF = 1) are used in the rest of this work.

When a PurpleAir sensor is connected to the internet, data is sent to PurpleAir's data repository. Users can choose to make their data publicly viewable (public) or private. All PurpleAir sensors also report RH and T levels.

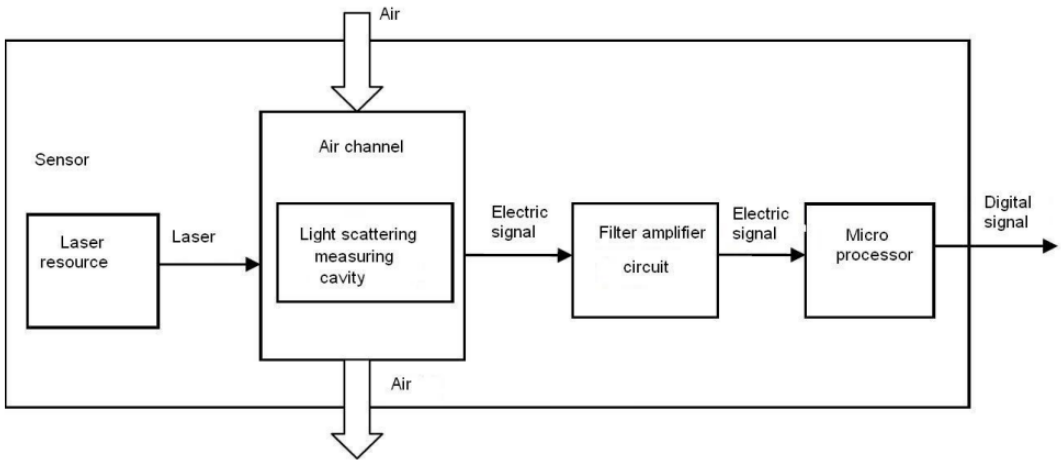


Figure 2. Functional block diagram of the Plantower PMS5003 sensor (Source: 2016 product data manual of PLANTOWER)

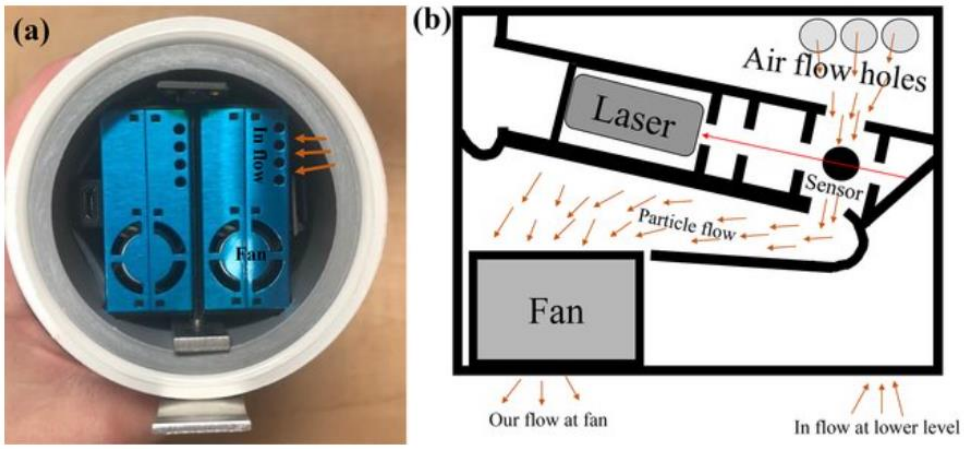


Figure 3. (a) The bottom of the PA-II unit contains two PMS5003 sensors (in blue). (b) The principle of operation of a single PMS5003 sensor. A fan draws the particles through the inflow (rounded holes) at the lower level of the sensor. The particles travel to the upper part of the sensor where they come out through the air flow holes and then pass through the laser path, causing the beam to scatter. Finally, the particles exit from the fan (Source: Ardon-Dryer et al, 2020).

SENSOR LOCATION

The location site of the sensor depends on the purpose of the study. For population exposure studies, sensor systems should be deployed at locations with different population characteristics, e.g., busy streets, indoor and outdoor, background and rural sites etc. Duration of data collection should be selected to cover temporal variations of pollutants, e.g., 4-weeks cycles in four seasons. For source apportionment studies, e.g., around airports, field study should be designed to gain information on the contribution of local source(s) in vicinity, regional background etc (Yatkin et al, 2022).

Poorly sited sensors can be influenced by local sources (e.g., backyard barbeque grill, vehicle exhaust, etc.), which can influence readings that are not truly representative of ambient air (Yatkin et al, 2022). In figure 4 the logistical considerations for finding a place to locate an air sensor are presented.

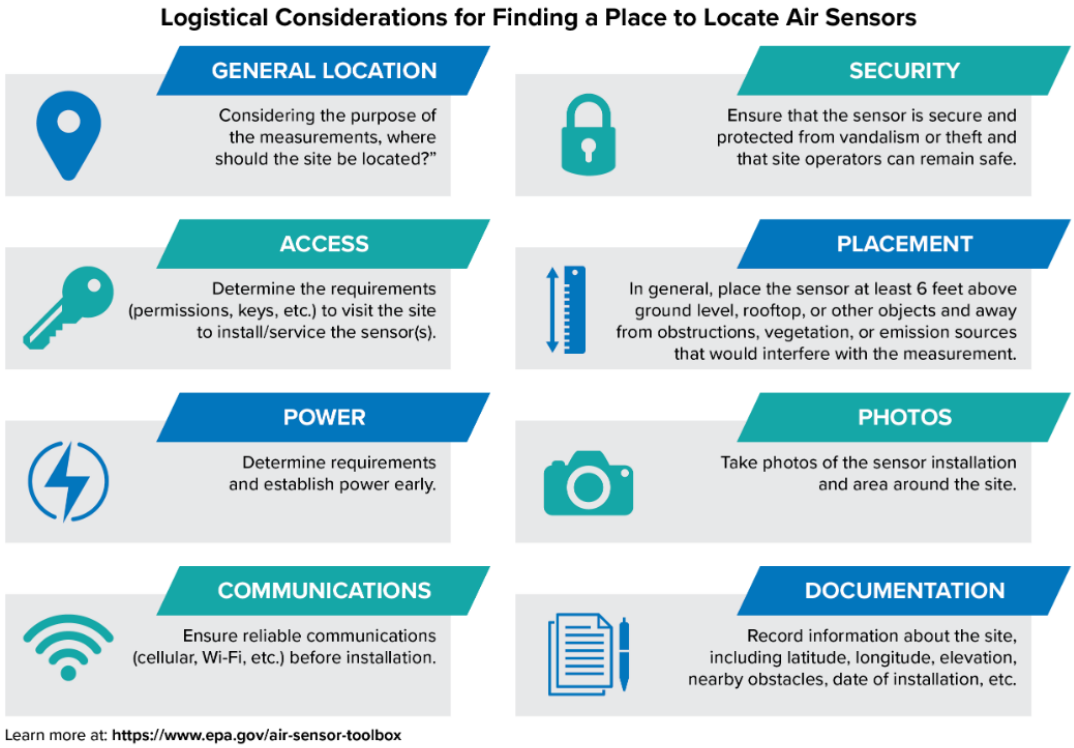


Figure 4. The logistical considerations for finding a place to locate an air sensor **(Source:** EPA, 2022)

SENSOR INSTALLATION

Before setting up your sensor you will need

- the sensor's device ID,
- an email associated with the sensor,
- a local WiFi network, and
- a WiFi-enabled device like a phone or computer.

The installation process is described in detailed by the following steps.

1st step: Plug in the cable to the power supply. Check if you need an EU power adapter because the provided cable has American plug output. If everything is ok a light will turn on at the bottom part of the sensor (figure 5).

Figure 5. The bottom part of the PurpleAir sensor. The light that turns on when the sensor is in operation is marked with red circle.

2nd step: After a few minutes a sensor's hotspot will be created with the name PurpleAir-**** (the asterisks will be a 2–4-character code determined by the last few characters of the sensor's device ID). Open the network settings on your Wi-Fi-enabled device (phone, computer, etc.) and connect to the sensor's hotspot.

3rd step: Usually a popup window appears. If not, open a web browser and enter "http://192.168.4.1/config " into the address bar. If you still cannot reach the configuring page, try temporarily pausing or disabling data on your Wi-Fi-enabled device and re-entering the URL above.

4th step: Select the "Wi-Fi settings" tab. Multiple Wi-Fi networks appear of which you should select the one you want the sensor to be connected to. Write the password on the specific space at the bottom and click "Save".

5th step: Your sensor will begin trying to connect to the selected Wi-Fi. This process may take a couple of minutes. Once the connection is successful, the message at the top of the page will say, "Looking Good" and the signal will be green. From now on you will no longer see the PurpleAir-**** network on your network settings.

SENSOR TROUBLESHOOTING

In this section useful information are presented in the form of questions & answers to assist users with troubleshooting problems associated with setting up, maintaining and using PurpleAir sensors.

Q1.I cannot connect the sensor to my Wi-Fi network.

In the first place, you have to check that your Wi-Fi network is 2.4GHz (usually home and personal networks). Secondly, you have to provide the password of your network at the "Wi-Fi settings" mentioned in section **"SENSOR INSTALLATION"**. If nothing of the above seems to be the problem, probably you have lost the recent firmware update. In this case, you have to check the version of your sensor (in figure 6, the sensor's version is 7.02). Right - click the point of another sensor in the PurpleAir map (Real-Time Air Quality Map | PurpleAir) and write down the version shown in the sensor's info box. In case the version of your sensor is not the same you have to update it, following the below steps:

- Connect your sensor to the Wi-Fi network of your mobile phone (you have to use the password of this network).
- Keep the connection for a couple of hours (approx. 1:30 h) so as the relevant update to be completed. Please check that the version shown at the info box of your sensor is the same with the rest sensors shown in the Real-Time Air Quality Map | PurpleAir.

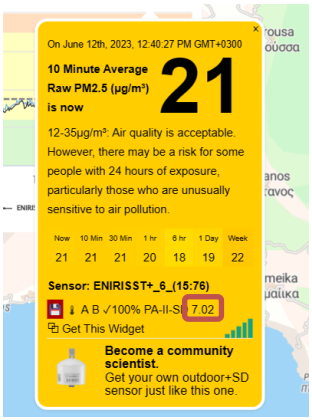


Figure 6. The information box that appears by clicking on a sensor's circle at the Real-Time Air Quality Map | PurpleAir (the sensor's version is marked with red square).

In case you want to connect your sensor at captive portal networks like coffee shops, universities and other locations that use captive portals you will need special authorization from the network administrator to allow the PurpleAir to communicate on the captive portal networks (source: Sensor Won't Connect to your WiFi Network - Sensors / Troubleshooting - PurpleAir Community).

Q2.I cannot see a PurpleAir-xxx Network.

After plugging your sensor, please check that the red light at the bottom part of your sensor is on and you can hear the mechanical activity.

If not, please check that the power cable works properly by plugging it into another device.

In any other case, please contact PurpleAir (contact@purpleair.com).

Q3. There is a great difference in the values that the two channels count or one of the channels counts a repeated value or a value that is close to zero.

It is expectable that the particles measured by channel A might be a bit different from the ones measured by channel B. However, if this difference is greater than 5 µg/m^3 for more than 5 hours or one of the channels counts the same values (e.g., 1.1, 1.4, 1.2, etc) for a long period, you have to check whether one of the laser counters in your sensor have become congested with pollen, debris or dust. It is necessary to clean out the laser counters with canned compressed air. For PA-II and PA-II-SD sensors, you must focus on the four collinear holes on each laser counter (please see figure 7). If the problem remains, you must change the laser counter. You can buy the correct one from the PurpleAir online store. For the replacement process, the following videos provide the necessary guidance (source: Sensor Maintenance - Sensors / Troubleshooting - PurpleAir Community):

- PA-I: PA-I-Indoor Laser Counter Replacement Video.mp4 - Google Drive
- PA-II or PA-II-SD: PA-II-SD Laser Counter Replacement.mp4 - Google Drive

- PurpleAir Flex: PA-II-FLEX Laser Counter Replacement.mp4 - Google Drive
- PurpleAir Zen: PurpleAir Zen Laser Counter Replacement.mp4 - Google Drive

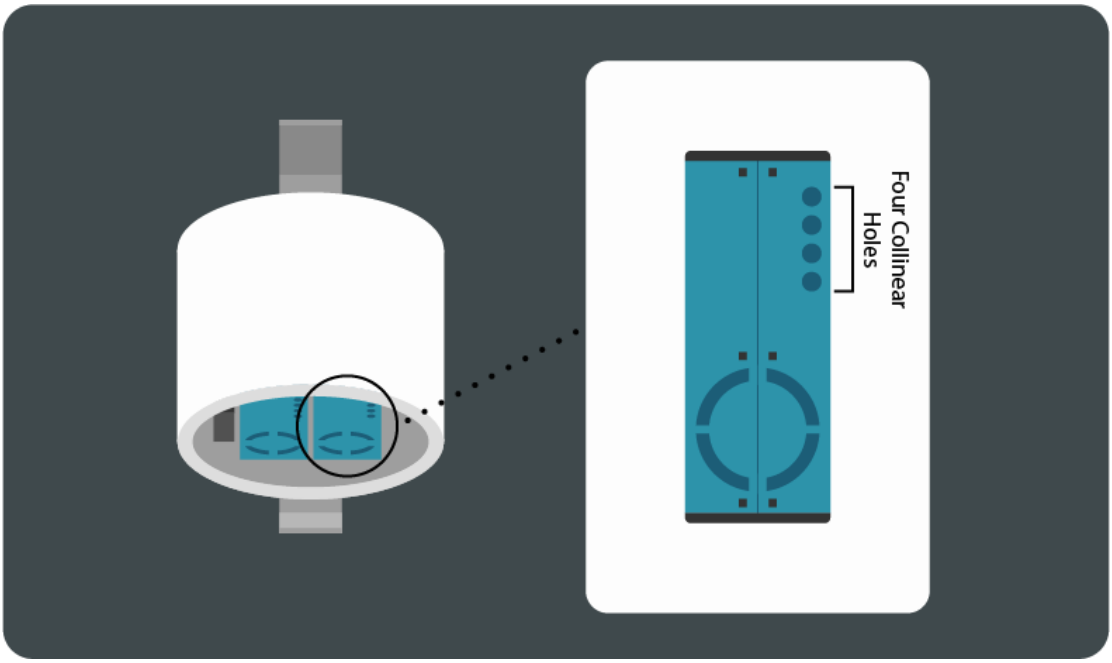


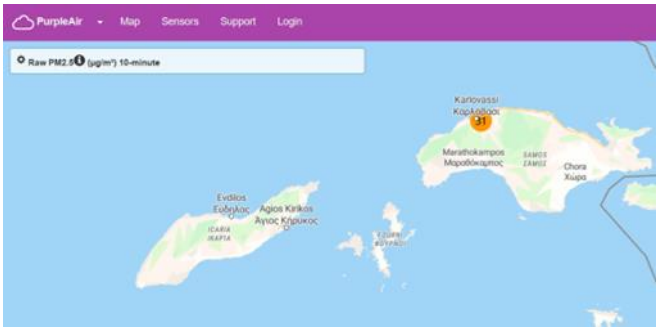
Figure 7. The four collinear holes that should be cleaned with canned compressed air in case that the laser counters have become congested with pollen, debris or dust (source: Sensor Maintenance - Sensors / Troubleshooting - PurpleAir Community)

DATA DOWNLOADING

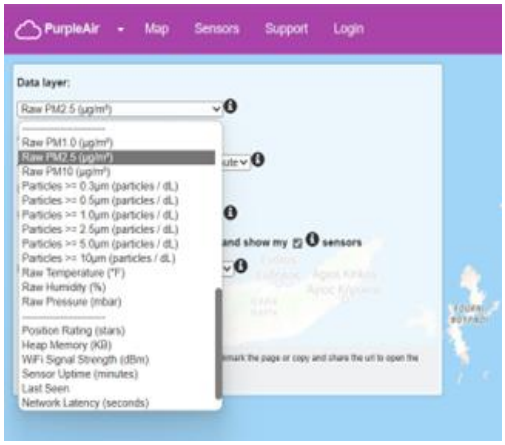
The data presented in the PurpleAir map as well as historical data can be downloaded for further use. There are three ways to download the data recorded by each sensor.

1) **Map download**

   Once zooming to the sensor, you want to download data (figure 8a), you can click on the bar (top left, figure 8b) to select the "Data layer" (parameter) to be displayed on the map (e.g. Raw $PM_{2.5}$, Raw PM_{10}, Humidity, etc.) as well as the average period (figure 8c).



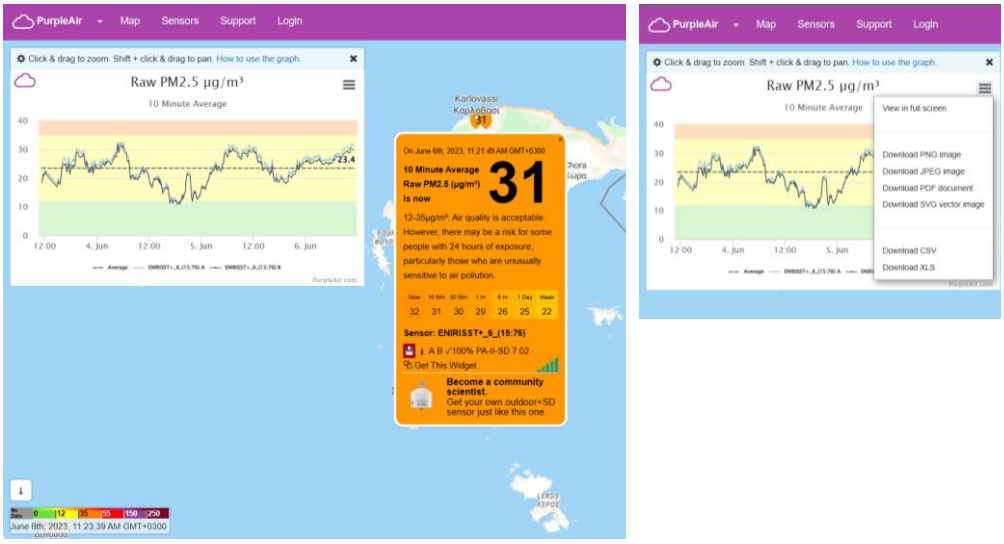
(a)



(b) **(c)**

Figure 8. Procedure for downloading data from the Real-Time Air Quality Map | PurpleAir.

By clicking on the circle of the sensor the sensor's ID box appears (the orange box in Figure 9a) coloured accordingly to the legend bar shown at the bottom left of the map (in the example it is orange because the 10-minutes average $PM_{2.5}$ concentration is 31 µg/m^3 and it belongs to the third colour class). The graph shows the data for the selected period. The dashed line presents the average value for this period. By moving the mouse over the graph or tapping on a specific point on the graph the value, date and time will be displayed. There are available options to download an image of the graph as a PNG, JPG, PDF, or SVG. This data can also be downloaded either in CSV or in XLS format by clicking on three horizontal lines menu icons (hamburger menu) (see Figure 9b). Depending on the averaging period you have chosen there is a limit on the values included in the downloaded csv or xls file. When selecting "*Reporting or modified within: All-time*", the period of downloaded values is as presented in table 1 while values for the two sensors are presented separately. Concentrations in the downloaded file are presented in local time (for Greece that is UTC +2 for wintertime and UTC+3 for summertime). More specifically, four columns are written; 1st column: local time, 2nd column the average value, 3rd column: concentration for channel A and 4th column: concentration for channel B.



(a) **(b)**

Figure 9. The info box as well as the graph that appear by selecting sensor (clicking on the circle representing the specific sensor).

Table 1. The capacity of the downloaded files according to the selected averaging period of concentrations.

Averaging period	**Number of values in the downloaded file**
Real-time	1140 values
10-minute	approx. 432 values
30-minute	approx. 336 values
1-hour	approx. 336 values
6-hour	approx. 360 values
1-day	365 values (the last year)
1-week	approx. 79 values
1-Month	68 values (practically from the installation of the sensor)
1-Year	all the years the sensor operates are included

For the rest choices the existence of API keys is necessary. The user should request his/her personal API keys from contact@purpleair.com to be able to interact with the PurpleAir API. There are two types of keys:

`READ` keys are used when reading entries.

`WRITE` keys are used when adding, updating or deleting entries.

2) Download using PurpleAir API

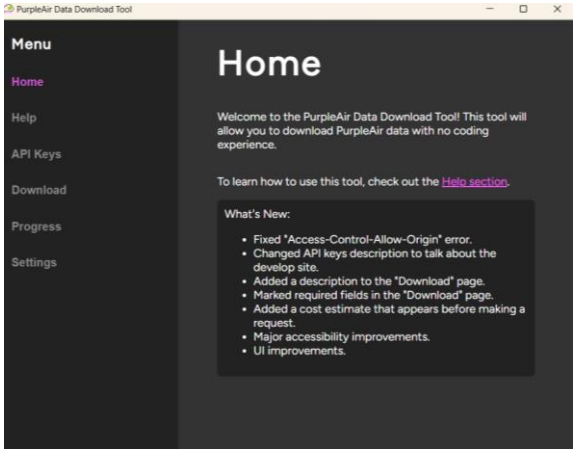
In case you want to make API calls with the PurpleAir API, please read the instructions given at the following link: Making API Calls with the PurpleAir API - Data / API - PurpleAir Community

3) PurpleAir Data Download Tool

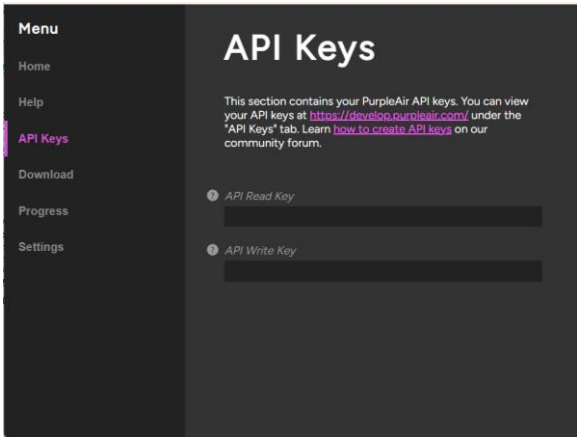
Probably the easiest way to download data, for a longer period than the one provided by the map graph, is the ***PurpleAir Data Download Tool*** which is an open-source software. Depending on the operating system of your computer you can follow the links below in order to download and install the program (PurpleAir Data Download Tool - General / Announcements - PurpleAir Community).

- MacOS
- Windows (32 bit)
- Windows (64 bit)

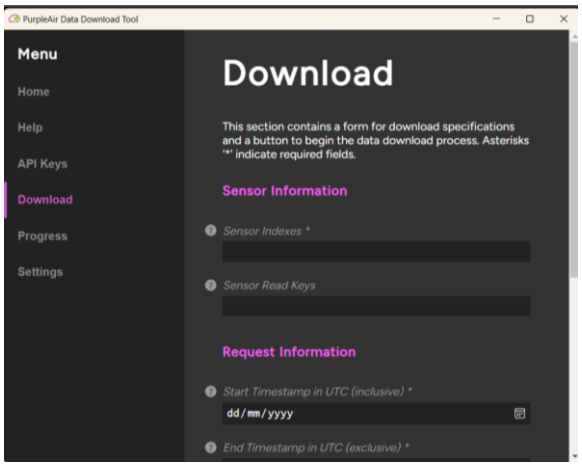
Afterwards, you have to open the folder where the file was downloaded and run the executable by clicking twice. If a pop-up message from the windows defender appears mentioning that the software is malicious, click "more info" and then "run anyway". This message appears because the software is not known to the database of the windows defender. An icon is now on your desktop. By right clicking it the ***PurpleAir Data Download Tool*** opens. The following steps will guide you through the usage of the Tool.

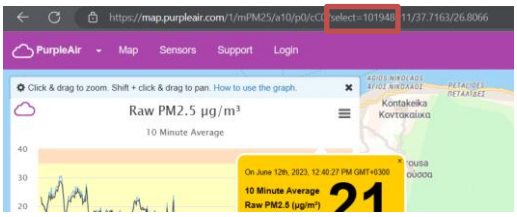


1st step: At the bar on the left, click the *API Keys* tab. In this section you have to write your personal API Read and Write Keys (the ones you received from the Purple Air contact email).



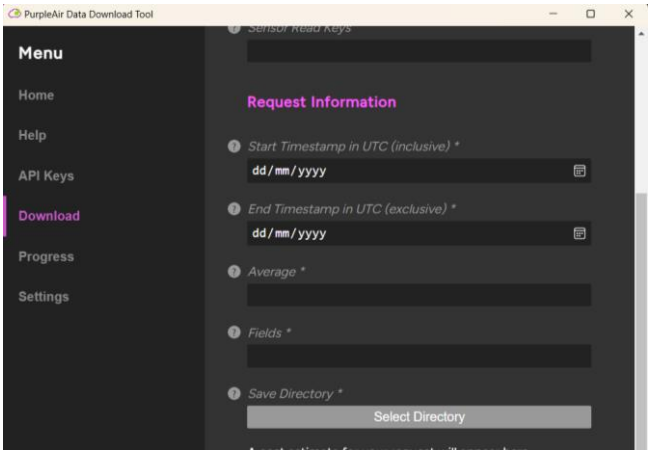
2nd step: Click the *Download* tab and fill in the sensor's indexes that you want to download data. You can find the sensor's index at the url (after the word "select=") by clicking on the sensor on the map. For multiple indexes please separate the numbers with commas (e.g., 101948,101361; be careful not to use space between the numbers!). For public sensors you don't have to fill in the "Sensor read Keys".



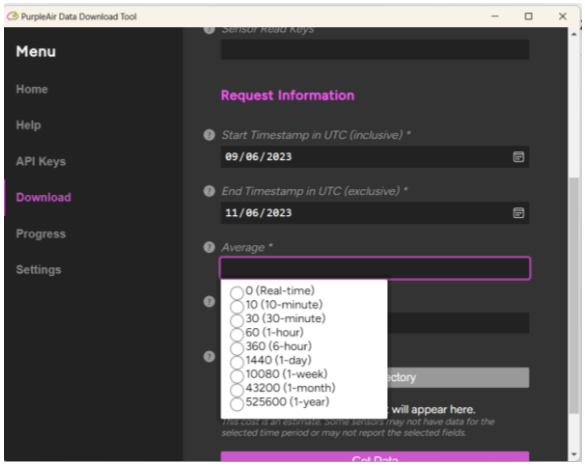


3rd step - Request information:

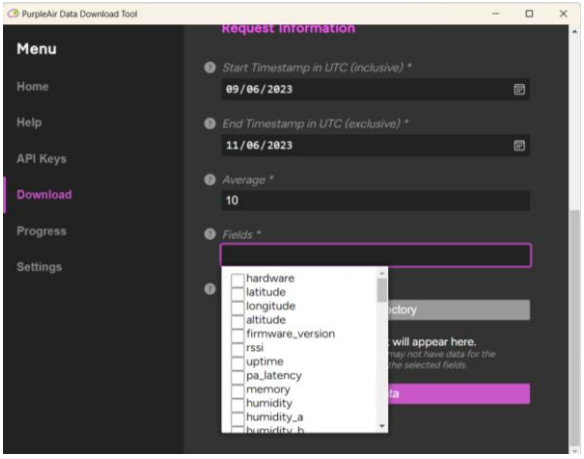
- Fill in the start and the end timestamp (be careful in UTC!!!, day/month/year). You can also click on the calendar icon on the right.



- Average stands for the period you want the export value to be calculated. By clicking on the tab multiple choices appear.



- Fields stands for the parameter you want to download data for. You can select more than one field at one request (by clicking on the field tab). Similarly, multiple fields should be separated with commas and no space between). The explanation of each field can be found here API - PurpleAir at the tab "Sensors - Get Sensors Data"' (please scroll down to find the field you are interested in).



- Finally, the folder where the downloaded file will be saved should be selected (Save directory).
- Click on the "Get Data" tab and the downloading process will begin.

Useful note1: The values downloaded through the map graph are in Local Time, however in the Tool you must fill in the preferred time in UTC.

Useful note2: PurpleAir has implemented a points-based system for API usage. Viewing and downloading data on the PurpleAir map remains free. However, if you want to download data through the ***PurpleAir Data Download Tool*** you have to contact PurpleAir (contact@purpleair.com) in order to update your account by adding points to start with (you should provide your API Keys which represent your "organisation"). **Your API keys will only work once points are added to your "organization". You can add these points at** PurpleAir Develop**. You can find more information on adding points and using the dashboard here: https://community.purpleair.com/t/new-api-online-dashboard/3981.**

DATA MANAGEMENT

File size and data storage requirements

Either selecting to download the data in .csv or .xls format, the file size is very small (max. 220 KB/sensor). In case you choose to download the graph as an image (PNG, JPG, PDF, or SVG) the file size is less than 110 KB per sensor.

Inspecting a new datafile

Upon downloading of data from the PurpleAir application, these are the questions that must be answered:

1. What is the type of files downloaded and how the files are managed?
2. How many cases (rows) and variables (columns) are in the file?
3. Did the variable values, names, and labels transfer correctly?
4. Are there any time considerations in the data time series? (Synchronisation, averaging)
5. Are all the data values within a reasonable range?
6. How much data is missing and in what patterns?

Type of files and file management

The data can be downloaded from the PurpleAir map as a csv or as an xls file. An xls file is the rectangular data file type. The most important aspect of the layout of a rectangular data file is that each row represents a record, and each column represents a variable. This type of files can later be inserted in a more sophisticated statistical program.

An archived version of the original (unmanipulated) data file should be stored somewhere in the computer that it can't be changed in case there is a need to reverse a coding decision later, or the edited file becomes corrupted and must be re-created. It's also good practice to store versions of the file after each major round of editing in case it is decided that some changes made are not valid and a previous version of the file is needed.

Number of cases and variables

You should know how many cases are expected to be in the data file you have downloaded. For example, if the averaging period is 1 h and the measurements are taken for 15 days, then it is anticipated that the file contains 24x15 = 360 rows (or cases) plus the headings. It is also good practice to check if the anticipated number of rows are correct after each round of editing.

Assuming the number of cases is correct, you also need to confirm that the correct number of variables (columns) are included in the file.

In case that the excel file will be imported to a statistical program (such as SPS), someone needs to look up the naming conventions for each program and create variable names that will be compatible with all the programs that will be used.

Time considerations in the time series of the data

Correct date readings: You must consider that, in general, date data is stored as a number reflecting the number of units of time (days) from a particular reference date. Unfortunately, each program seems to use a different reference date, and some use different units as well, with the consequence that date values often do not transfer correctly from one program to another.

Time zone differences. Sometimes the concentrations are reported in UTC (Coordinated Universal Time). UTC is the primary time standard by which the world regulates clocks and time. All time zones are defined by their offset from UTC. For example, in Greece, the standard time is Eastern European Time (UTC+02:00). Daylight saving time, which moves one hour ahead to UTC+03:00 is observed from the last Sunday in March to the last Sunday in October. This means that when a $PM_{2.5}$ concentration is reported to occur somewhere in Greece on April 1st, 11:40 AM UTC time, the local Greek time is 14:40.

When we download data from international platforms, sometimes the option is given to download it in UTC or in local (sensor) time. But often these platforms do not consider the daylight-saving time convention for the conversion from UTC to local time of each sensor. So, it is preferable to download our data in UTC time and convert to local time ourselves. When the data is presented to the public it is useful to report in Local Time.

Data synchronisation. Data synchronisation is essential in comparisons of the LCSs readings with the reference instrument concentrations that are often reported in Local Time.

Averaging period: Each laser counter within a sensor alternates five-second readings averaged over two minutes. There is the option to download data from the map at averaging time intervals of 10 min, 30 min, 1 h, 6 h, 1 day, 1 month and 1 year. The selection of the averaging period for our analysis depends on its aim. If the aim of the analysis is to search for the impact of events such as arrivals/departures of ships at a nearby port or short-lived road traffic jams, or the effect of short events such as smoking (in indoor air pollution) then it would be more suitable to select a short averaging time (real time reporting - which means 2 min intervals - or 10 min intervals). Note that the 2 min averaging time could result in high noise levels. On the other hand, if the purpose is to compare the sensor readings with EU limits or WHO guidelines, then it would be sufficient to download the data in hourly or daily averages.

For what time interval the averaging stands for? The interval (start time – end time) that the averaged data correspond to must be known. For the Purpleair LCSs, an hourly averaged concentration reported at 14:00 is the result of the average of the real time reporting - which means concentrations every 2 min intervals - of the next hour (start time 14:00 – end time 14:58). This information is useful to synchronise data from different instruments.

Data quality

Are the sensor channels A and B readings similar?

The PA-II PurpleAir sensor has two identical PM sensor components (Plantower PMS 5003 sensors) referred to as “channel A” and “channel B” and the divergence between their measurements can be observed. This is a very useful indicator of sensor condition and data quality.

In the example below, a graph (figure 10) that shows data from a sensor in Ermoupoli, Syros, Greece is presented (Purpleair, n.d.).

One can notice blue and black lines on this graph. The blue line represents channel A and the black line, channel B. Each channel is connected to a laser counter inside the sensor. By comparing the correlation of these two channels a measure of confidence is created—called the Confidence Score—that serves as an assurance of data quality. The Confidence Score in the first graph is 28% indicating a poor similarity of readings whereas in the second graph it is 100% meaning that there is a good similarity of readings between the two channels.

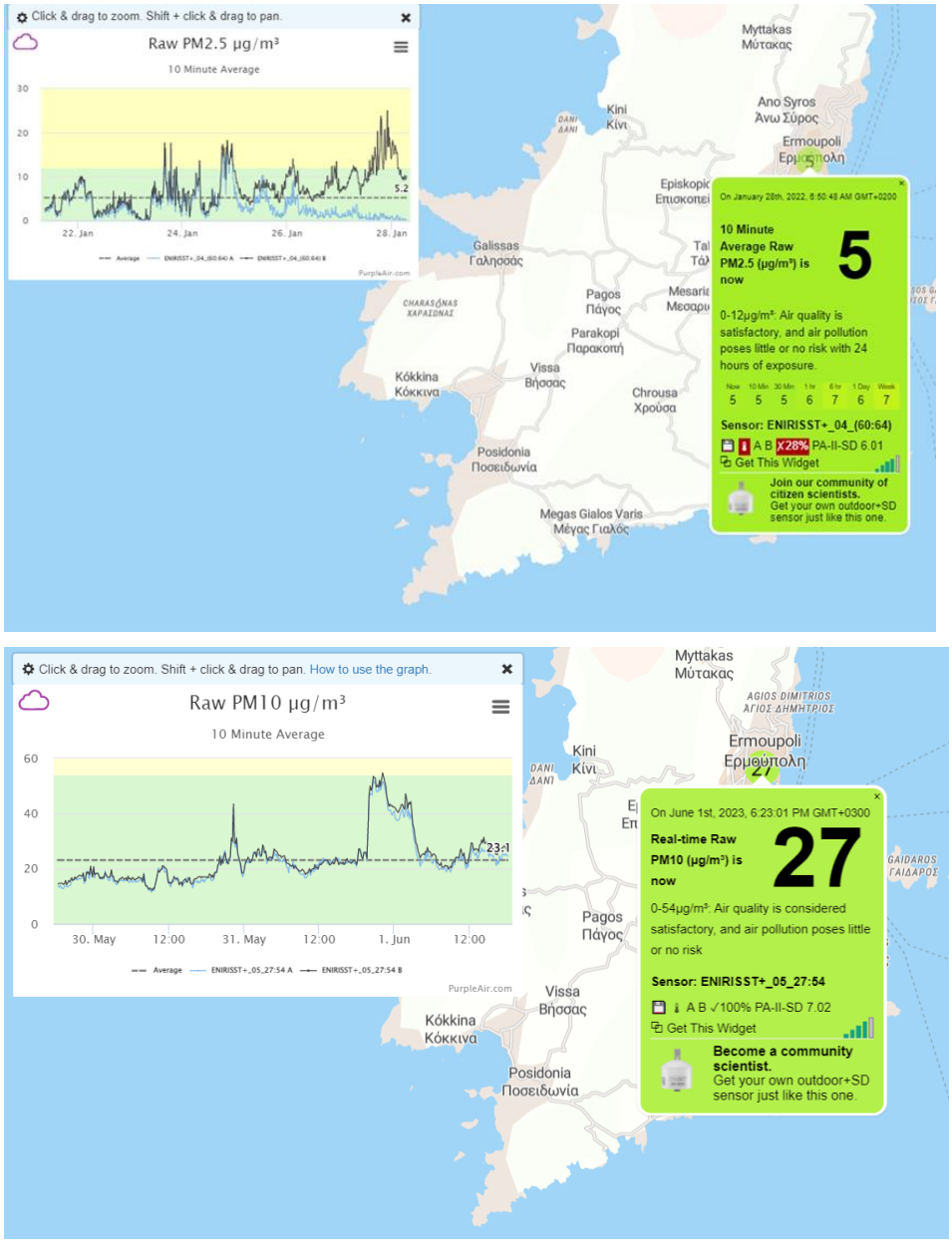


Figure 10. The two channels A and B of a Purpleair outdoor sensor located in Ermoupoli, Syros, and their confidence score (28% for the first case and 100 % for the second)

A limitation of using the percentage of the confidence score as a degradation metric is that it does not account for the possibility that channels A and B might both degrade in a similar manner. Therefore, we rely on a second approach, using collocated reference monitoring measurements, to evaluate this aspect of possible degradation.

In addition to the confidence score provided by PurpleAir, it is good practice the user itself to perform a quality control of the data by comparing channels A and B readings. A methodology to do this is proposed by Barkjohn et al (2021) by comparing the daily average concentrations from the two channels using the equation:

$$24\ h\ percent\ difference = \frac{(A-B)\times 2}{(A+B)}$$

where A is the measurement from channel A and B is the measurement from channel B. In case the percentage difference is larger than 61%, values of this day are further checked in order the values to be removed.

Secondly, the hourly concentrations from the two channels are checked based on the methodology proposed by Bi et al (2020) for the discard of outliers. The equation used is:

$$APB = \left|\frac{PM2.5B - PM2.5A}{PM2.5A}\right| \times 100\%$$

The measurements with the top 5% APB values are removed.

Is the data reasonable?

The next step is to examine the actual values in the data set and see whether they seem reasonable (they make sense). You can inspect the fluctuation in the PM concentrations by plotting them over time. You can also inspect the range of values running a frequency distribution in case you are using a statistical package, by sorting the cases and inspecting them visually if you are using a spreadsheet package, or with software-specific procedures such as the Filter option in Microsoft Excel.

One reason for unreasonable data is a failure in the sensor itself or in data transmission. For example, in the case of figure 11, the sensor has failed to transmit data between 12-13 and 14-19 of April and these values should be excluded from further analysis.

In case that unrealistically high or low concentrations are recorded for a considerable period, the readings of the sensor should be compared with nearby ones to check if the same pattern appears. If not, then perhaps (in case of high values) a source near the sensor gives high levels of pollution and must be further investigated. To have an idea of what levels of pollution are considered as high or low, you can consult the European Air Quality Index concentrations (as daily averages) that correspond to good, fair, moderate, or poor air quality (figure 12).

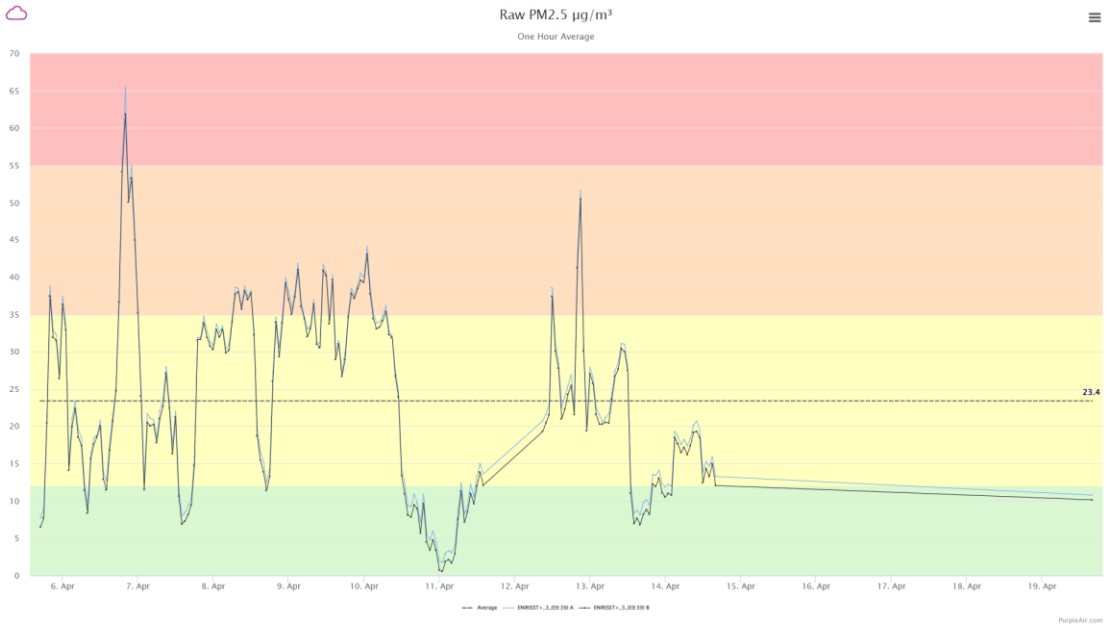


Figure 11. Plotting of sensor readings showing data transmission failure between 12-13 and after 14 of April

Pollutant	Index level (based on pollutant concentrations in µg/m3)					
	Good	Fair	Moderate	Poor	Very poor	Extremely poor
Particles less than 2.5 µm ($PM_{2.5}$)	0-10	10-20	20-25	25-50	50-75	75-800
Particles less than 10 µm (PM_{10})	0-20	20-40	40-50	50-100	100-150	150-1200

Figure 12. The European Air Quality Index based on 24-hour running means for PM_{10} and $PM_{2.5}$ (Source: EEA (n.d.))

Missing data

The final step before beginning data analysis is to examine the amount of missing data and its patterns. Nearly every data set ever collected has included missing data. Despite the ubiquity of missing data, however, it is not a simple problem to deal with. Your first goal is to find out how much missing data there is. The second is to examine the patterns of missing data across multiple variables. For instance, is data frequently missing on particular sets of variables? Are there cases with lots of missing data, while others are entirely or primarily complete? What are the different reasons why data is missing?

Data can be missing for many reasons, and it is useful if the reasons are recorded within the data set. The method to accomplish this varies with different systems but is nearly always

possible, often by using specific data codes to differentiate among them, using values (such as –8 and –9) that cannot appear as true values for the variable in question. For the Purpleair sensors, common reasons are equipment failure or problems in data transmission via WiFi. Please look at the Troubleshooting section for more information about missing values.

Missing data poses two problems: it reduces the number of cases available for analysis, thereby reducing statistical power (your ability to find true differences in the data) and may also introduce bias into the data. The first point is because, all things being equal, statistical power is increased as the number of cases increase, so any loss of cases may result in a loss of power.

Missing data is often divided into three types:

- Missing completely at random (MCAR),
- Missing at random (MAR), and
- Nonignorable.

MCAR means that the fact of a piece of data being missing is not related to either its own value or the value of other variables in the data set. This is the easiest type of missing data to deal with, since the complete cases may be considered to be a random sample of the entire data set. This is the case of missing data with the Puprleair sensors, since common reasons for missing data are equipment failure or data transmission failure.

MAR data means that the fact that the data is missing is not related to its own value but is related to the values of other variables in the analysis. For instance, electrochemical sensors that measure gases such as NOx, O_3 or CO give less accurate data in extreme weather conditions (approx. >30 Celsius, <30% RH) and such data are not reported by certain quality assurance procedures (e.g. please check https://www.aqmesh.com/techsupport/aqmesh-technical-support/faqs/extreme-environment-flagging/) resulting in missing values of the MAR type.

Nonignorable refers to data whose missingness is related to its own value. Nonignorable missing data is the type most likely to introduce bias into a statistical analysis. This could be the case (for example) if a sensor stops transmitting data in case of very low or very high pollutant concentrations. To our experience, Purpleair sensors do not appear to have this problem in low or medium polluted areas.

Because the most common methods of statistical analysis assume that you have complete, unbiased data, if a data set has a large quantity of missing data, you will have to decide how to deal with it. The most preferable solution is to make an extra effort to collect the missing data by following up with the source, which solves the problem by making the missing data no longer missing. In case of data transmission problems, the SD edition of the Purplair sensors includes an onboard SD (Secure Digital) logger to record and store data without a WiFi connection. The SD can store data of about 2 years.

Another solution is to include a dummy (0, 1) variable in your analysis that indicates that data was missing, along with an imputed value replacing the missing data. There are different methods of data imputation and the most sophisticated may require calling in a statistical consultant or using software designed specifically for dealing with missing data (e.g., SPSS MVA module). A simple imputation method is to substitute the missing values with a value such as the hourly mean of a selected period (a season or the whole year). However, this is not recommended, as it nearly always results in an extreme underestimate of variance.

Finally, in cases of large amounts of missing data, the solution may be to drop the respective cases or variables and change the scope of the analysis. For example, if there are frequent data transmission failures during winter due to electricity power outages, then the analysis may focus on the other seasons.

Sensor field (in situ) evaluation of the precision and the accuracy

Optical sensor methods do not directly measure mass concentrations; rather, they measure light scattering of particles having diameters typically > ~0.3 mm. Several assumptions are made to convert light scattering into mass concentrations that can introduce errors in the results (de Souza, 2023). The conversion methodology of each sensor manufacturer is usually proprietary and not open to the user.

In addition, it is known that atmospheric particles absorb water as the atmospheric relative humidity (RH) increases (Kosmpopoulos et al, 2020). This leads to increases in both their size and mass. Light scattering by the particles also increases. The reported PM concentrations should not include this water. High-quality instruments are equipped with dryers or heaters to remove particle water. Low-cost sensors, such as PurpleAir, do not include dryers or heaters. Therefore, PM concentrations reported by LCSs can be highly biased due to hygroscopic growth of particles when ambient RH is high (de Souza, 2023). As a result, PurpleAir sensors constantly overpredict fine particle concentrations in most locations and under higher humidity compared to the regulatory-grade monitors that are operated at the same location (EPA, n.d.).

The accuracy and precision of measurements are defined as follows (EPA, 2001):

- Accuracy - the degree of correctness with which a measurement reflects the true value of the parameter being assessed.
- Precision- the degree of variation in repeated measurements of the same quantity of a parameter.

For example, if ten measurements for a given parameter are taken at the same time at the same location by the same method, the accuracy would be indicated by how well the average of the ten measurement results reflects the actual concentration present and the precision would be indicated by the variation in the results of the ten measurements. A classic illustration of precision and accuracy is depicted in figure 13.

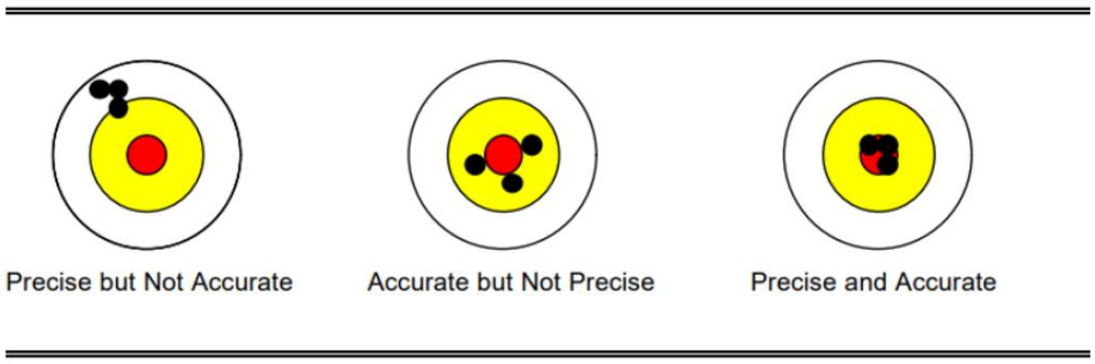


Figure 13. Precision and accuracy

The precision and accuracy of the sensors of the ENIRISST+ network was examined. Hourly averages of PM2.5 concentrations were used in the analysis. At first the precision (repeatability of the readings of the 8 sensors) was checked. The 8 sensors were co-located at a suburban site at Chios for a period of 17 days (27.10.2021 – 12.11.2021). In the absence of errors, the concentrations measured by the sensors at pairs should be equal (y=x). A linear regression model was applied to depict the actual relation between sensor readings in pairs. The coefficient of determination R2 was ≥0.9976, the slopes of the linear regression ranged between 0.950-1.04 and the intercept was close to zero. These findings suggest excellent matching between the readings of the pairs of sensors.

Then, one sensor was selected at random, and it was co-located with a reference equivalent beta attenuation monitor at the Thissio Air Monitoring Supersite, operated by the National Observatory of Athens, for a winter period (23/12/2021 - 28/02/2022) to check the accuracy of

the sensors. The concentrations reported by the reference-grade instrument are assumed to be the true values.

We applied linear regression between the sensor readings and the RH (independent variables) and the reference equivalent instrument concentration (dependent variable) and we found that the equation that describes best the relationship between them is:

$y_{ref}=0.423*x_{CF1}-0.117*RH+11.051$

Where: y_{ref} ($\mu g/m^3$) is the concentration measured by the reference-equivalent instrument, x_{CF1} ($\mu g/m^3$) is the concentration measured by the PurpleAir sensor and RH (%) is the relative humidity. The coefficient of determination R^2 that measures the goodness of fit of the model was 0.81.

By applying the correction equation, in the case that the sensor reports a concentration of 30 $\mu g/m^3$, at a RH of 50%, the corrected concentration is 18 $\mu g/m^3$. In the case that the sensor reports a concentration of 30 $\mu g/m^3$, at a RH of 80%, the corrected concentration is 14 $\mu g/m^3$.

The sensor's raw and corrected data, the reference data and the RH (relative humidity) for 2 days are presented in Table 2. The improvement in accuracy of the corrected values is illustrated in figure 14.

Table 2. The raw and corrected data of the sensor, the reference values and the humidity for a period of 2 days

DateTime (UTC+2)	$PM_{2.5}$ **sensor** **($\mu g/m^3$)**	$PM_{2.5}$ **sensor corrected** **($\mu g/m^3$)**	$PM_{2.5}$ **reference** **($\mu g/m^3$)**	**Humidity (%)**
23/12/2021 2:00	19.1	13.5	15.9	48
23/12/2021 3:00	15.1	12.0	9.2	47
23/12/2021 4:00	15.4	12.3	10.5	45
23/12/2021 5:00	17.7	12.9	10.5	48
23/12/2021 6:00	16.2	12.2	11.3	48
23/12/2021 7:00	14.5	11.6	12.2	48
23/12/2021 8:00	20.0	14.0	7.3	47
23/12/2021 9:00	22.0	14.9	17.8	46
23/12/2021 10:00	19.9	14.5	14.2	42

23/12/2021 11:00	9.1	10.7	8.4	36
23/12/2021 12:00	9.5	11.3	5.4	32
23/12/2021 13:00	10.7	12.0	6.9	31
23/12/2021 14:00	13.9	13.0	6.9	34
23/12/2021 15:00	17.0	14.1	10.9	35
23/12/2021 16:00	19.9	14.9	10.2	39
23/12/2021 17:00	23.0	14.9	13.8	51
23/12/2021 18:00	22.1	14.1	16.0	54
23/12/2021 19:00	36.2	19.8	15.3	56
23/12/2021 20:00	64.2	31.6	24.8	57
23/12/2021 21:00	117.8	54.1	43.9	58
23/12/2021 22:00	158.5	70.9	69.0	61
23/12/2021 23:00	171.3	76.2	81.0	63
24/12/2021 0:00	179.2	79.5	83.4	63
24/12/2021 1:00	175.9	78.1	88.1	63
24/12/2021 2:00	169.0	75.2	87.5	63
24/12/2021 3:00	105.5	48.7	70.3	60
24/12/2021 4:00	102.6	47.5	51.6	59
24/12/2021 5:00	64.6	31.4	47.2	60

24/12/2021 6:00	63.2	30.6	26.2	62
24/12/2021 7:00	39.2	20.3	24.5	62
24/12/2021 8:00	80.5	37.7	31.5	64
24/12/2021 9:00	44.5	23.2	24.1	57
24/12/2021 10:00	32.5	19.2	16.5	48
24/12/2021 11:00	31.9	19.4	18.7	44
24/12/2021 12:00	25.0	16.9	9.5	40
24/12/2021 13:00	16.0	13.5	10.9	37
24/12/2021 14:00	9.1	10.5	5.0	38
24/12/2021 15:00	9.2	10.0	6.1	42
24/12/2021 16:00	14.6	11.8	2.2	46
24/12/2021 17:00	21.2	14.4	9.0	48
24/12/2021 18:00	40.4	22.3	17.6	50
24/12/2021 19:00	56.1	28.9	26.8	50
24/12/2021 20:00	57.1	28.9	29.8	54
24/12/2021 21:00	31.1	17.2	10.1	60
24/12/2021 22:00	47.2	23.9	10.4	61
24/12/2021 23:00	48.4	24.3	16.7	61
25/12/2021 0:00	31.6	17.2	21.3	61

25/12/2021 1:00	21.4	12.9	14.8	62

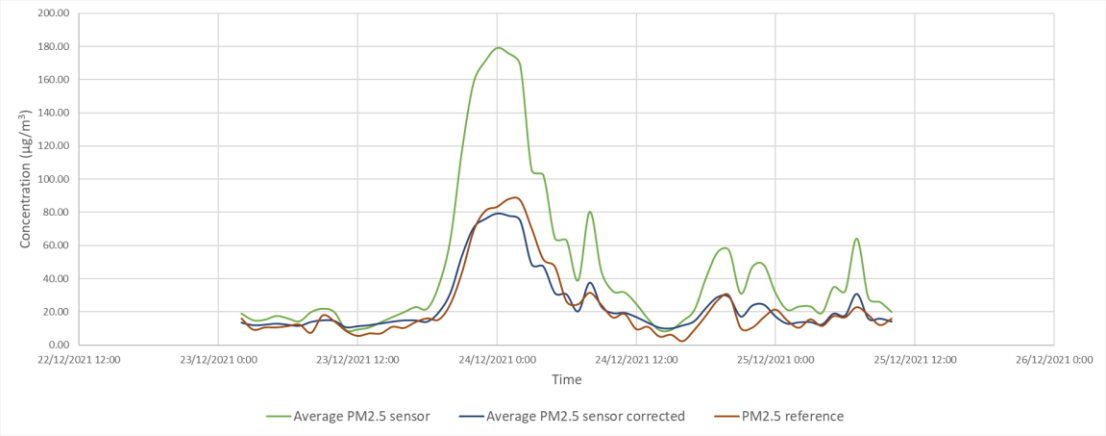


Figure 14. The reference instrument values compared to the raw readings of the sensor and the corrected values

DESCRIPTIVE STATISTICS AND GRAPHICS

Descriptive statistics is the use of statistical and graphic techniques to present information about the data set being studied. Computing descriptive statistics and examining graphic displays of data is an advisable preliminary step in data analysis. Examining the actual distribution of the data collected (as opposed to the distribution you expected) is always time well spent. Descriptive statistics and graphic displays are also the final product in some contexts.

Measures of central tendency (the mean and the median)

Measures of central tendency (the arithmetic mean and median), also known as measures of location, are typically among the first statistics computed for the continuous variables in a new data set.

The mean: For continuous data, the mean is simply calculated by adding up all the values and dividing by the number of values. The mean is not a good indicator of central tendency in cases that there are outliers.

The median is the middle value when a data set is ordered from lowest to greatest. If there are n values, the median is formally defined as the $(n+1)/2$th value. If $n = 7$, the middle value is the $(7+1)/2$th or the fourth value. If there is an even number of values, the median is the average of the two middle values. This is formally defined as the average of the $(n/2)$th and $((n/2)+1)$th value. If there are six values, the median is the average of the $(6/2)$th and $((6/2)+1)$th value, or the third and fourth values. Both techniques are demonstrated below:

Odd number of values: 1, 2, 3, 4, 5, 6, 7 median = 4

Even number of values: 1, 2, 3, 4, 5, 6 median = (3+4)/2 = 3.5

The median is a better measure of central tendency than the mean for data that is asymmetrical or contains outliers. This is because the median is based on the ranks of data points rather than their actual values: 50 percent of the data values in a distribution lie below the median, and 50 percent above the median, without regard to the actual values in question. Therefore, it does not matter if the data set contains some extremely large or small values, because they will not

affect the median more than less extreme values. For instance, the median of all three distributions below is 4:

Distribution A: 1, 1, 3, 4, 5, 6, 7

Distribution B: 0.01, 3, 3, 4, 5, 5, 5

Distribution C: 1, 1, 2, 4, 5, 100, 2000

In a symmetrical distribution (such as the normal distribution), the mean and median coincide. In an asymmetrical or skewed distribution, they differ, and the amount by which they differ is one way to evaluate the skewness of a distribution.

Measures of dispersion

Dispersion refers to how variable or "spread out" data values are: for this reason, measures of dispersion are sometimes called "measures of variability" or "measures of spread." Knowing the dispersion of data can be as important as knowing its central tendency: for instance, two sets of air pollution concentrations of $PM_{2.5}$ may both have the same mean of 20 μg/m^3 but one could have a range of 0 to 40 μg/m^3 (good to poor air quality) while the other has a range of 0 to 70 μg/m^3 (good to very poor air quality, refer to figure 12).

The range

The simplest measure of dispersion is the range, which is simply the difference between the highest and lowest values. Often, the minimum (smallest) and maximum (largest) values are reported as well as the range (table 3).

Variance, standard deviation and coefficient of variation

The most common measures of dispersion for continuous data are the variance and standard deviation. Both describe how much the individual values in a data set vary from the mean or average value. The variance is the average of the squared deviations from the mean, and the standard deviation is the square root of the variance.

The formula for the variance s^2 of a data set with n observations is:

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Where x_i is the value i from the data set and $\bar{x}$ is the mean of the data set.

In calculating the variance, we have changed from our original units to squared units, which may not be convenient to interpret. For instance, if we were measuring concentrations in μg/m^3, we would probably want measures of central tendency and dispersion expressed in the same units, rather than having the mean expressed in μg/m^3 and variance in squared μg/m^3. To get back to the original units, we take the square root of the variance: this is called the standard deviation. The formula for a sample standard deviation is:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

In general, for two variables measured with the same units, the group with the larger variance and standard deviation has more variability. However, the unit of measure affects the size of the variance: $PM_{2.5}$ concentrations, expressed mg/m^3 rather than $\mu g/m^3$, would have a smaller variance and standard deviation. The coefficient of variation (CV), a measure of relative variability, gets around this difficulty and makes it possible to compare variability across variables measured in different units. The CV is calculated by dividing the standard deviation by the mean, then multiplying by 100:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

The descriptive statistics of sensor data from table 2 are presented in table 3.

Table 3. Descriptive statistics of the corrected data of the PurpleAir sensor $PM_{2.5}$ ($\mu g/m^3$)

Mean	25.8
Median	17.1
Minimum	10.0
Maximum	79.5
Range	70
Variance	403.7
Standard deviation	20.1
Coefficient of variation (CV)	78%

Graphs

Graphics are a tool used in the service of communicating information about data. There are innumerable graphic methods to present data, from the basic techniques included with spreadsheet software such as Microsoft Excel to the extremely specific and complex methods developed in the computer language R. The simplest presentation is often the best.

Line graphs are often used to display the relationship between two variables, often between time on the x-axis and some other variable on the y-axis.

In figure 15, a line graph of the PurpleAir $PM_{2.5}$ corrected concentrations (data from table 2) in relation with time is presented for a 2-day period (Christmas eve and Christmas day 2021). On the first day, the $PM_{2.5}$ concentrations start to rise in the late afternoon (from about 6:00 pm) and reach their peak (80 $\mu g/m^3$) at 12:00 midnight. Then start to decline until 7:00 am. The next day, a smaller peak appears in the morning, at 8:00 am (29 $\mu g/m^3$). The afternoon and night of the second day two small peaks are detected at 7:00 pm (30 $\mu g/m^3$) and 10.30 pm (24 $\mu g/m^3$). This different pattern of air pollution during the two days is probably due to the amount of the

emissions but also to the meteorological conditions. The residential heating system is a considerable source of $PM_{2.5}$ in the winter months, especially when biomass is used in stoves and fireplaces. Peaks usually appear in the afternoon and evening, when people are at their homes and use residential heating (peak of the first day). Road traffic could also give peaks, especially during the day (it could be the morning peak of the second day). The meteorological conditions play an important role since high wind speeds favour the dispersion of pollution and rain washes out the PM from the atmosphere. Perhaps this is the case on the second day of measurements. These are indicative explanations of atmospheric pollution patterns that must be supported by further research.

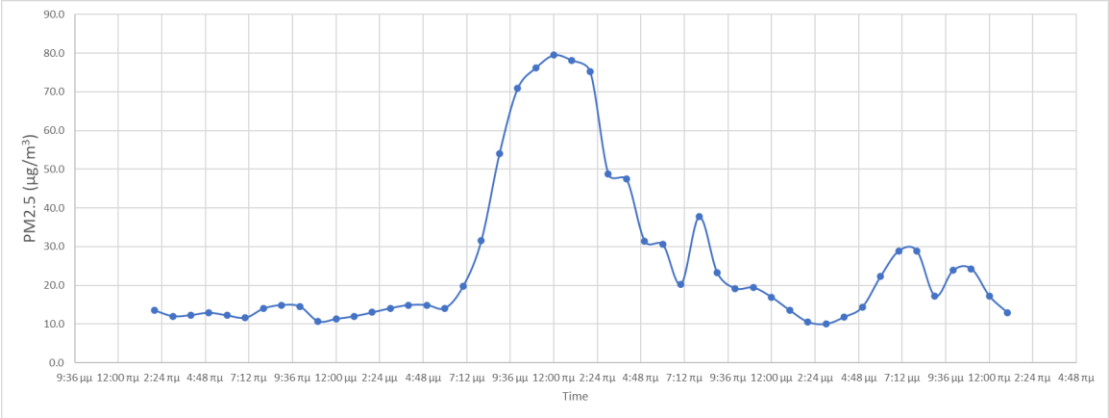


Figure 15. Line graph of the fluctuations of $PM_{2.5}$ concentrations in relation to time

Funding: This research is financed by the Research Infrastructure EN.I.R.I.S.S.T.+ (MIS 5047041), implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

2016 product data manual of PLANTOWER, Digital universal particle concentration sensor, available at https://www.aqmd.gov/docs/default-source/aq-spec/resources-page/plantower-pms5003-manual_v2-3.pdf

Ardon-Dryer, K., Dryer, Y., Williams, J. N., & Moghimi, N. (2020). Measurements of PM 2.5 with PurpleAir under atmospheric conditions. Atmospheric Measurement Techniques, 13(10), 5441-5458.

Barkjohn, K. K., Gantt, B., & Clements, A. L. (2021). Development and application of a United States-wide correction for $PM_{2.5}$ data collected with the PurpleAir sensor. Atmospheric Measurement Techniques, 14(6), 4617–4637. https://doi.org/10.5194/amt-14-4617-2021

Bi, J.;Wildani, A.; Chang, H.H.; Liu, Y. Incorporating low-cost sensor measurements into high-resolution $PM_{2.5}$ modeling at a large spatial scale. Environ. Sci. Technol. 2020, 54, 2152–2162. https://doi.org/10.1021/acs.est.9b06046

Boslaugh, S. (2012). Statistics in a nutshell: A desktop quick reference. O'Reilly Media, Inc.

deSouza, P., Barkjohn, K., Clements, A., Lee, J., Kahn, R., Crawford, B., & Kinney, P. (2023). An analysis of degradation in low-cost particulate matter sensors. Environmental Science: Atmospheres, 3(3), 521-536.

EEA (n.d.). European Air Quality Index, European Environment Agency, available at https://airindex.eea.europa.eu/Map/AQI/Viewer/

EPA (2001). The Precision and Accuracy of Environmental Measurements for the Building Assessment Survey and Evaluation Program Previously, U.S. Environmental Protection Agency, available at https://www.epa.gov/sites/default/files/2014-08/documents/precisionandaccuracy.pdf

EPA (2022). A Guide to Siting and Installing Air Sensors, available at https://www.epa.gov/air-sensor-toolbox/guide-siting-and-installing-air-sensors

EPA (n.d.). How to Evaluate Low-Cost Sensors by Collocation with Federal Reference Method Monitors, National Exposure Research Laboratory Office of Research and Development, available at https://www.epa.gov/sites/default/files/2018-01/documents/collocation_instruction_guide.pdf

Giordano, M. R., Malings, C., Pandis, S. N., Presto, A. A., McNeill, V. F., Westervelt, D. M., ... & Subramanian, R. (2021). From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. Journal of Aerosol Science, 158, 105833.

Kosmopoulos, G., Salamalikis, V., Pandis, S. N., Yannopoulos, P., Bloutsos, A. A., & Kazantzidis, A. (2020). Low-cost sensors for measuring airborne particulate matter: Field evaluation and calibration at a South-Eastern European site. Science of The Total Environment, 748, 141396.

PurpleAir (n.d.). What are channel A and channel B? available at https://community.purpleair.com/t/what-are-channel-a-and-channel-b/3643

Yatkin, S., Gerboles, M, Borowiak, A and Signorini, M. (2022). Guidance on low-cost sensors deployment for air quality monitoring experts based on the AirSensEUR experience, Publications Office of the European Union, Luxembourg, doi:10.2760/14893, JRC130050.