Maitane Iturrate Garcia
Lindenweg 50, 3003, Bern-Wabern
maitane.iturrategarcia@students.unibe.ch

Data Science Project

Air Pollution – Are Low-Cost Gas Sensors the Solution to Spatial Data Gaps in Air Quality Monitoring?

Conceptual Design Report

6 October 2024

Abstract

Air pollution is a global environmental and health issue (WHO, 2021) with a great spatial and temporal heterogeneity. Air pollution assessment is needed in order to estimate the exposure degree of population and ecosystems to gas pollutants. For that purpose, amount of substance fractions (i.e., concentration) of target pollutants (e.g., ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO)) are measured at fixed-location sites within air monitoring networks. These sites are equipped with high quality instruments. However, due to their high costs, the number of sites is limited or inexistent, resulting in spatial gaps.

In this work, performance (e.g., cross-sensitivity with other pollutants, response to environmental conditions such as air temperature or relative humidity) of low-cost gas sensor systems (LCSS) will be evaluated in the laboratory and in atmospheric simulation chambers to assess the suitability of using LCSS to cover the spatial gaps of monitoring stations. Moreover, alternatives to classical laboratory calibrations, such as machine learning approaches to model LCSS responses to air pollution will be explored to optimize resources (i.e., shorten calibration periods and reducing costs) when calibrating LCS.

Table of Contents

1. Project Objectives

The use of LCSS on air applications has exploded in recent years for air pollution applications because of the sensor technology development. However, the limitations that low-cost gas sensors (LCS) present (e.g., poor selectivity, cross-sensitivity, aging, response to environmental variables (mainly temperature and relative humidity)) can overcome their advantages (e.g., low weight, small size, rapid responses, relative low price). One way of minimizing those limitations is to perform frequent calibrations of the LCSS. Classical calibration of "golden units" combined with machine learning (ML) modelling predicting the response of LCSS might contribute to increase the data quality of LCSS and to optimize resources (i.e., shorten calibration periods and reducing costs).

The main goal of this project is to assess the suitability of using low-cost gas sensor systems (LCSS) to cover the existing spatial gaps on air pollution monitoring resulting from the limited number of monitoring stations due to their high costs. To achieve this objective, the performance of LCSS (e.g., cross-sensitivity with other pollutants, response to environmental conditions) will be evaluated using data collected in the laboratory (calibration) and under semi-controlled conditions by running experiments in an atmospheric simulation chamber. Correlations between LCSS response signals and environmental variables will be evaluated as a first step to obtain accurate calibration models. Once the LCSS performance will be assessed, predictive models using supervised machine learning (ML) approaches will be explored by comparing LCSS responses against data from reference instruments. Different model and data quality metrics will be estimated and evaluated to explore the LCSS suitability as part of air pollution networks.

2. Methods

2.1. Data generation

Data used in this project was generated by 10 low-cost gas sensor systems (LCSS) that were developed within the framework of the Innosuisse 36779.1 IP-ENG project: "Novel SI-traceable low-cost sensor systems for air quality monitoring" (Fig. 1). Each LCSS consisted on four low-cost electrochemical gas sensors: carbon monoxide (CO) sensor (Alphasense CO-B4), nitrogen monoxide (NO) sensor (Alphasense NO-B4), nitrogen dioxide (NO_2) sensor (Alphasense NO2-B43F) and ozone-dioxide nitrogen (O_3-NO_2) sensor (Alphasense OX-B431), a humidity sensor, a temperature sensor, a pressure sensor and a particulate matter sensor. All the sensors were

located in a metal box (18.0 cm × 18.4 cm × 12.5 cm; 2.3 kg) provided with a fan to ensure homogeneous mixing of the gas mixture within the box.

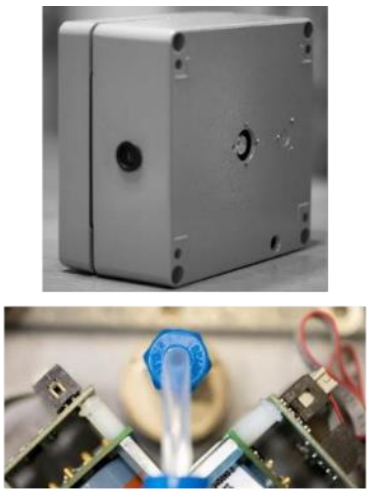

Figure 1: Detail of the low-cost gas sensor systems used in this project (photo by courtesy of D. Calabrese, LNI Swissgass).

Data generated by both, LCSS and reference instruments during the calibrations performed before and after the experiments described in subsection *2.2 LCSS calibration* and the experiments described in *2.2 Experimental design* are needed to answer the research questions of this project.

2.2. LCSS calibration

The ten LCSS were calibrated before and after the experiments at the climatic chamber of the Federal Institute of Metrology METAS (Wabern-Bern, Switzerland). Room temperature (23.0 ± 2.0 °C) was used during the calibration because of a technical issue of the panel controlling the conditions within the climatic chamber. A detail description of the calibration set-up and the reference instruments used can be found in Tancev et al. (2022).

2.3. Experimental design

To collect the data needed for this project, a series of experiments were run between 20.04-05.05.2023 at the European Photoreactor (EUPHORE; Fig. 2)) facility of the pan-European Aerosol, Clouds and Trace Gases Research Infrastructure (ACTRIS).

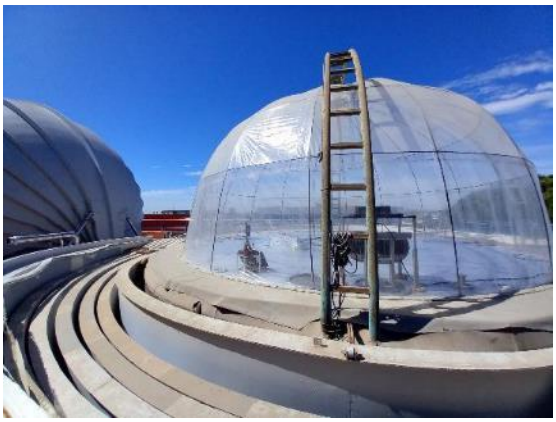

Figure 2: Photo of the European Photoreactor (EUPHORE) facility where the low-cost gas sensor systems (LCSS) were placed during the experiments.

Four groups of experiments were performed (8 experiments in total; Fig. 3):

a) lab conditions (no aerosols, no sunlight, similar temperature and relative humidity than the laboratory conditions during the calibration of the LCSS.
b) gas compounds (lab conditions) and aerosols
c) simulation of rural, suburban and urban atmospheres (with and without sunlight)
d) ambient air

Detailed description of the experimental protocols can be found in Appendix 1.

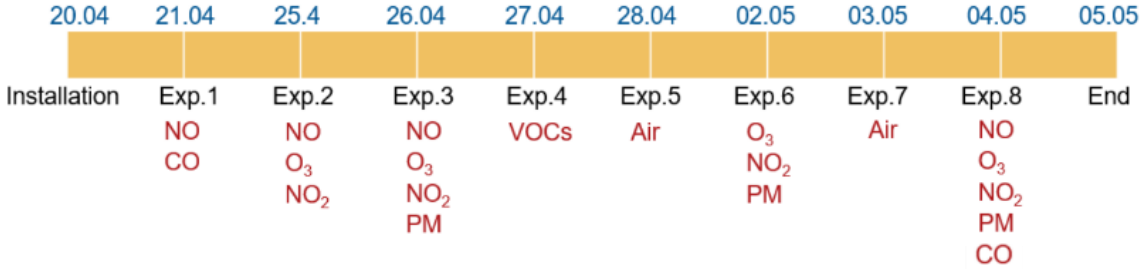

Figure 3: Scheme of the different experiments performed to assess the performance of the low-cost gas sensor systems (LCSS) in the atmospheric simulation chamber (EUPHORE).

The different reference instruments used during the EUPHORE experiments are shown in Table 1.

Table 1: Reference instruments used during the low-cost gas sensor system (LCSS) experiments. Analyzers of carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NO_x), ozone (O_3), particulate matter of 2.5 µm and 10 µm of diameter ($PM_{2.5}$, PM_{10}), volatile organic compounds (VOCs) and hydrogen peroxide (H_2O_2) are listed.

Compound/variable	Method	Model	Manufacturer
CO	Gas filter correlation	TE48C	Thermo Scientific
CO	Fourier-Transform Infrared (FTIR)	Nicolet 6700	Thermo Scientific
NO, NO_2, NO_x	Chemiluminescence	T200U	Teledyne API
NO_2	Cavity Attenuated Phase Shift (CAPS)	T500U	Teledyne API
O_3	Ultraviolet (UV) absorption spectroscopy	Serinus 10	Ecothec
Aerosols ($PM_{2.5}$, PM_{10})	Scanning Mobility Particle Sizer (SMPS)	Model 3938	TSI Incorporated
Aerosols ($PM_{2.5}$, PM_{10})	Tapered Element Oscillating	Model 1400a	Thermo Scientific
VOCs	FTIR	Nicolet 6700	Thermo Scientific
H_2O_2	FTIR	Nicolet 6700	Thermo Scientific
Pressure	Differential pressure sensors	Air-DB-VOC	Sirsa
Temperature	Electrical resistance of a platinum wire	Pt-100 RTD	RS components
Relative humidity	Dewpoint mirror measuring system	TS-2	Walz

2.4. Data Analysis

2.4.1. Infrastructure and tools

All data will be stored on a server-based infrastructure owned by the Federal Institute of Metrology METAS after data collection. Python will be the programming language used for this project. Therefore, additional required infrastructure comprises local installations of Python (using Spyder or Anaconda Integrated Development Environment (IDE)) and Google Colab.

2.4.2. Software libraries

The following Python libraries will be mainly used for this project:

1. NumPy (https://numpy.org/): for data manipulation; it offers a broad number of mathematical tools for manipulating and operating on large, multidimensional arrays and matrices.

2. pandas (https://pandas.pydata.org/): for data analysis and manipulation; it allows manipulating numerical tables and time series.
3. matplotlib (https://matplotlib.org/): for data visualization; matplotlib is a plotting library. The Graphical User Interface (GUI) of this library is provided by matplotlib.pyplot.
4. plotly (https://plotly.com/python/): for data visualization; it is a graphing library which makes interactive graphs at publication-quality level.
5. statsmodels (https://www.statsmodels.org/stable/index.html): for statistical analysis; this library provides functions for the estimation of many statistical models, conducting statistical tests and statistical data exploration. The interface of this library is given by statsmodels.api.
6. SciPy (https://scipy.org/): for data analysis; it contains modules for optimization, linear algebra, integration, interpolation and clustering among others.
7. scikit-learn (https://scikit-learn.org/stable/): for data analysis using machine learning approaches, including support-vector machines, k-means, gradient boosting, random forests and DBSCAN, among others; it provides tools for classification, regression and clustering algorithms.
8. datetime: for manipulating date and time data; this Python module supplies classes for manipulating and formatting dates and times.
9. npTDMS (https://nptdms.readthedocs.io/en/stable/): for manipulating files with TDMS format (format of the files produced by the LCSS which contain metadata and raw data); npTDMS is a cross-platform package built on top of the NumPy library, that allows reading from TDMS files as numpy arrays and writing numpy arrays to TDMS files.

2.4.2. Statistical methods and modelling

Effects of environmental variables (i.e., temperature, relative humidity, pressure and aerosol concentration) on the response signal of LCSS will be explored. For that purpose, linear correlations between LCSS response signals and environmental variables will be evaluated using the Pearson correlation coefficient.

For this project, supervised machine learning (ML) approaches will be used to predict the LCSS response to the amount of substance fractions of different pollutants. For that purpose, the relationship between LCSS response signals – taking into account the influence of environmental variables such as temperature and relative humidity – and the response of the reference instruments used during the calibration and experiments will be modelled. An increasing

complexity sequence of ML approaches (e.g., linear regression (LR) → multiple linear regression (MLR) → random forest (RF)), will be evaluated to find the model that predicts the best the LCSS response signal, using parameters such as the determination coefficient (R^2), root mean square error (RMSE) and mean absolute error (MAE). Data will be split into a training dataset (80 % of the data) and a validation dataset (20 %) for the ML modelling.

3. Data

For this project, three different datasets will be needed:

1. **LCSS dataset**: this dataset comprises the response signal – recorded every 15 seconds – of the ten low-cost gas sensor systems (Tancev et al., 2022) used during the experiments. The dataset encompasses three data subsets corresponding to the data collected during the pre-calibration, post-calibration and experimental measurements. The originated data is saved in TDMS* format by the software controlling the LCSS, which was written using LabVIEW (version Q3 (2023), National Instruments NI-Emmerson). The generated data files have variable size depending on the experiment duration. These files include:
    a. metadata: provides names and properties of all objects in the segment, as well as index information that is used to locate the raw data for this object in the segment.
    b. **LCSS raw data**: one data tdms sheet for each LCSS (Table 2).
    c. **reference instrument raw data**, which comprises date-time stamp, experiment number, stability flag, temperature (mV), pressure (mv), relative humidity (RH), and CO, NO, NO_2 and O_3 response signals for channels 1 and 2 of the low-cost gas sensors (one column per pollutant and channel) (Table 3). Channel 2 corresponds to the signal background, which will be removed from the response signal given by channel 1 (an additional column including the difference between channels 1 and 2 will be added to the dataset during the data pre-processing).

* With the TDMS format, data is organized in a three-level hierarchy of object. File-specific information is included at the top level, which is comprised of a single object. Each file can contain an unlimited number of groups and an unlimited number of channels per group. Every TDMS object is uniquely identified by a path: i.e., a string including the name of the object and the name of its owner in the TDMS hierarchy. Each name is enclosed by quotation marks and separated by a forward slash.

Table 2: Structure of the low-cost gas sensor system (LCSS) data subset

Column number	Column name	Column units	Column description
01	Epoch	No units	Date and time of the data point collection (dd/mm/yyyy hh:mm:ss.000 AM/PM)
02	ExperimentNum	No units	Number of the experiment that generated the data observation
03	Stable	No units	Flag that indicates whether measurement conditions (i.e., temperature, relative humidity) during the data point collection were stable
04	Temperature (mv)	mV	Response signal (voltage) of the temperature sensor included in the LCSS
05	Pressure (mV)	mV	Response signal (voltage) of the pressure temperature sensor included in the LCSS
06	Rel_humidity (mV)	mV	Response signal (voltage) of the relative humidity sensor included in the LCSS
07	CO_OP1 (mV)	mV	Response signal (voltage) given by the first channel of the carbon monoxide (CO) sensor included in the LCSS
08	CO_OP2 (mV)	mV	Response signal (voltage) given by the second channel of the carbon monoxide (CO) sensor included in the LCSS
09	NO_OP1 (mV)	mV	Response signal (voltage) given by the first channel of the nitrogen monoxide (NO) sensor included in the LCSS
10	NO_OP2 (mV)	mV	Response signal (voltage) given by the second channel of the nitrogen monoxide (NO) sensor included in the LCSS
11	NO2_OP1 (mV)	mV	Response signal (voltage) given by the first channel of the nitrogen dioxide ($NO_{2)}$ sensor included in the LCSS
12	NO2_OP2 (mV)	mV	Response signal (voltage) given by the second channel of the nitrogen dioxide ($NO_{2)}$) sensor included in the LCSS
13	O3_OP1 (mV)	mv	Response signal (voltage) given by the first channel of the ozone (O_3) sensor included in the LCSS
14	O3_OP2 (mV)	mv	Response signal (voltage) given by the second channel of the ozone (O_3) sensor included in the LCSS

Table 3: Structure of the calibration reference instrument data subset

Column number	Column name	Column units	Column description
01	Epoch	No units	Date and time of the data point collection (dd/mm/yyyy hh:mm:ss.000 AM/PM)
02	ExperimentNum	No units	Number of the experiment that generated the data observation
03	Stable	No units	Flag that indicates whether measurement conditions (i.e., temperature, relative humidity) during the data point collection were stable
04	Temperature (°C)	°C	Response signal (voltage) of the temperature sensor included in the LCSS
05	Pressure (mbar)	mbar	Response signal (voltage) of the pressure temperature sensor included in the LCSS
06	Humidity (%)	%	Response signal (voltage) of the relative humidity sensor included in the LCSS
07	Air_Flow (ml/min)	mL/min	Response signal (voltage) given by the first channel of the carbon monoxide (CO) sensor included in the LCSS
08	CO_Flow (ml/min)	mL/min	Flow rate of the mass flow controller (MFC) controlling the CO flow added to the calibration gas mixture
09	NO_Flow (ml/min)	mL/min	Flow rate of the MFC controlling the NO flow added to the calibration gas mixture
10	NO2_Flow (ml/min)	mL/min	Flow rate of the NFC controlling the NO2 flow added to the calibration gas mixture
11	CO (ppm)	µmol/mol	CO amount of substance fraction measured by the Picarro analyzer (CO and H_2O reference instrument)
12	CO2 (ppm)	µmol/mol	CO_2 amount of substance fraction measured by the Picarro analyzer (CO and H_2O reference instrument)
13	CO2_dry (ppm)	µmol/mol	CO_2 amount of substance fraction measured under dry conditions by the Picarro analyzer (CO and H_2O reference instrument)
14	CH4 (ppm)	µmol/mol	CH_4 amount of substance fraction measured by the Picarro analyzer (CO and H_2O reference instrument)
15	CH4_dry (ppm)	µmol/mol	CH_4 amount of substance fraction measured under dry conditions by the Picarro analyzer (CO and H_2O reference instrument)

Column number	Column name	Column units	Column description
16	RH (%)	%	Relative humidity measured by the Picarro analyzer (CO and H_2O reference instrument)
17	NO (ppb)	nmol/mol	NO amount of substance fraction measured by the Thermo42i analyzer (NO and NO_2 reference instrument)
18	NO2 (ppb)	nmol/mol	NO_2 amount of substance fraction measured by the Thermo42i analyzer (NO and NO_2 reference instrument)
19	NOx (ppb)	nmol/mol	NO_x amount of substance fraction measured by the Thermo42i analyzer (NO and NO_2 reference instrument)
20	O3 (ppb)	nmol/mol	O_3 amount of substance fraction measured by the Thermo Fisher 49C analyzer (O_3 reference instrument)

2. **Reference instrument dataset**: response of all the reference instruments installed in the atmospheric simulation chamber (EUPHORE). This dataset comprises five data subsets: FTIR data, monitor data, PM2.5 data, SMPS data and TEOM data. Data files formats are .xlsx.

3. **Air quality dataset**: air quality data collected by the monitoring station of Paterna-CEAM, which belongs to the air quality surveillance and control network of Valencia (Spain). Average daily values were downloaded directly from the network website (https://mediambient.gva.es/es/web/calidad-ambiental/datos-on-line). Hourly average values were directly provided by the CEAM, which is responsible for the quality assessment and quality control (QA/QC) of data generated by the monitoring station. Data files have .csv format.

Figure 4 shows an example of response signal during an experiment for the NO low-cost gas sensors installed in all the LCSS used in this project.

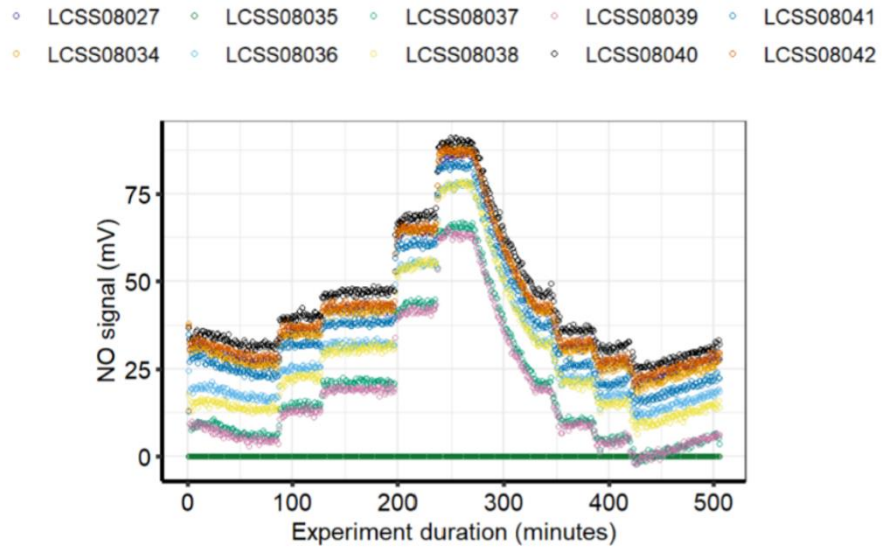

Figure 4: Response signal (voltage, in mV) measured during one of the experiments ran at the simulation chamber EUPHORE by the NO sensors installed on the low-cost gas sensor systems (LCSS). The responses show the high unit-to-unit variability.

Some examples of response signal data distribution (i.e., histograms) are shown in Figs. 5-7 for the sensors of temperature (Fig. 5), relative humidity (Fig. 6) and CO (Fig. 7) installed in the LCSS08035.

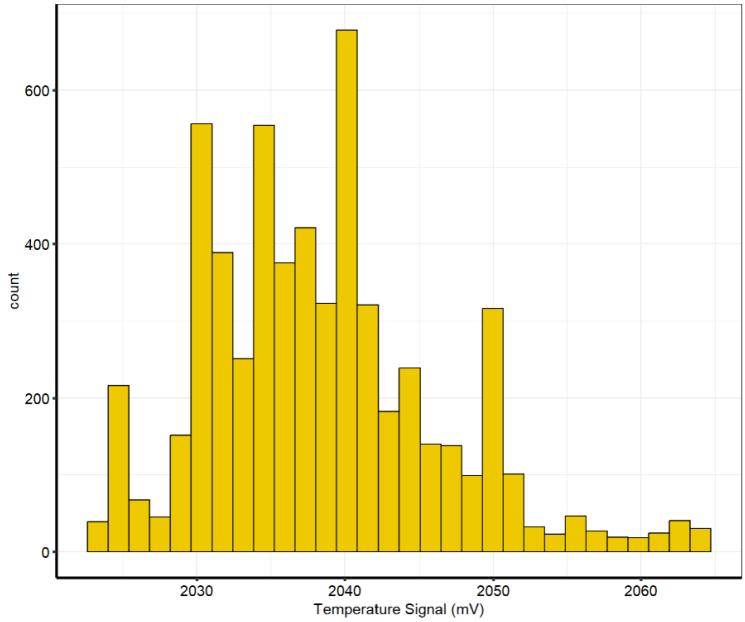

Figure 5: Histogram of the response signal (in voltage, mV) given by the temperature sensor installed on one of the low-cost gas sensor systems (LCSS08035).

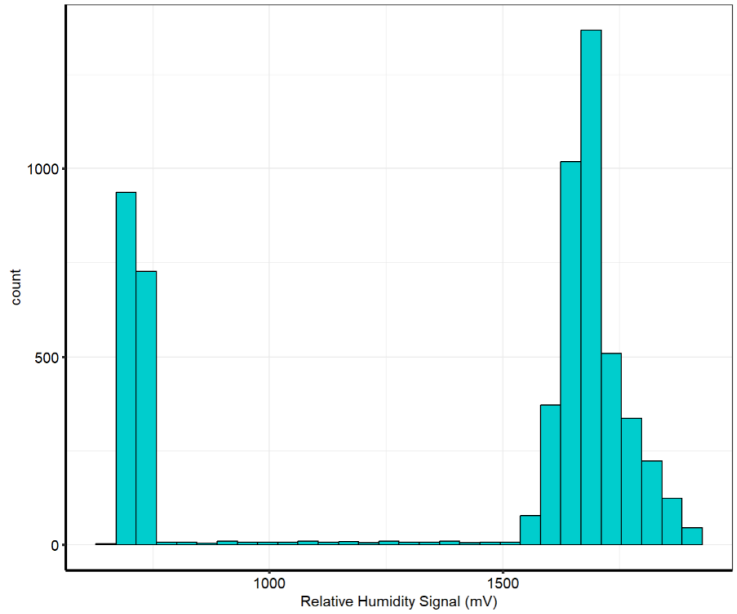

Figure 6: Histogram of the response signal (in voltage, mV) given by the relative humidity sensor installed on one of the low-cost gas sensor systems (LCSS08035). The two groups correspond to the two relative humidity levels generated during the LCSS calibration (20 % and 60 %).

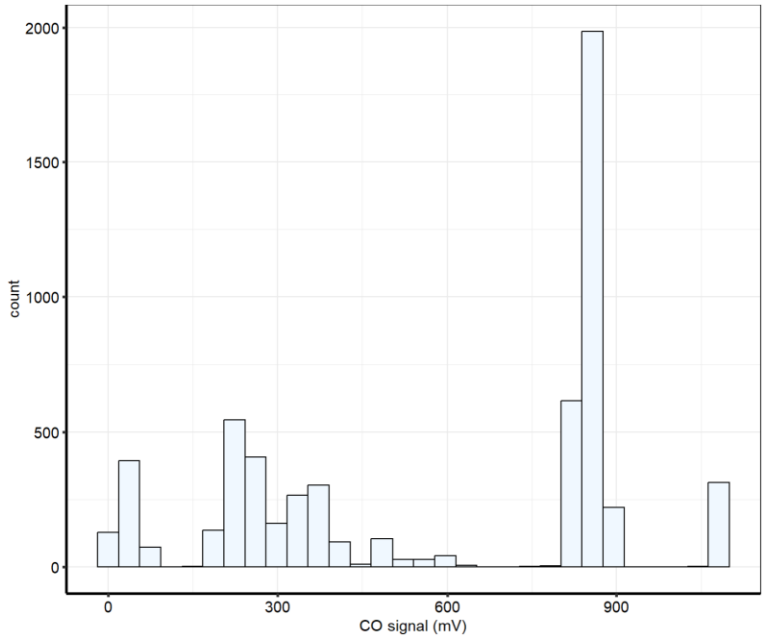

Figure 7: Histogram of the response signal (in voltage, mV) given by the CO sensor installed on one the low-cost gas sensor systems (LCSS08035). The two groups correspond to the two amount of substance fraction levels generated during the LCSS calibration (50 nmol/mol and 850 nmol/mol).

Preliminary linear correlation found between the LCSS response signals for CO, NO, NO2 and O3 and the relative humidity generated during experiments 2 and 3 are shown in Figs. 8 and 9.

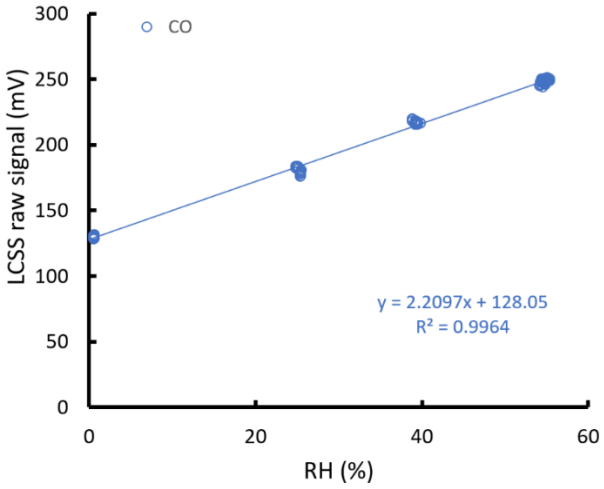

Figure 8: Linear correlation between the CO response signal of the ten low-cost gas sensor systems (LCSS) and the relative humidity (RH) generated during experiment 2.

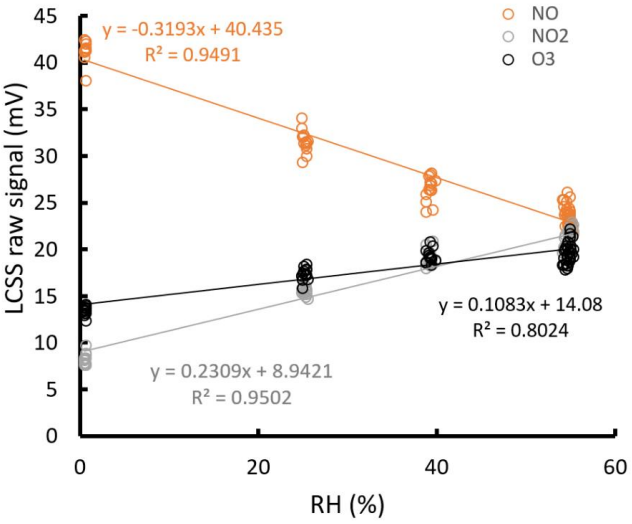

Figure 9: Linear correlation between the NO, NO_2 and O_3 response signals of the ten low-cost gas sensor systems (LCSS) and the relative humidity (RH) generated during experiment 3.

4. Metadata

Detailed metadata will be created and distributed together with the dataset to guarantee the findability/discoverability of the dataset and the reproducibility of the data collection and data analysis. For that purpose, at least the following fields, grouped by aims, will be included in the metadata:

1. **Dataset findability/discoverability**: this section will include all the information needed to make the dataset findable, such as keywords, Digital Object Identifier (DOI) of associated publication(s), information about the dataset (title, DOI, authors…).
2. **Data and data file description**: information on the data and data files will be provided in this section, such as: file format, number of files, headers (if datasets content header rows and the number of header rows), column names, number of columns, column units, data separator, date and time format, decimal punctuation, missing data code, etc.
3. **Data provenance**: this section will include information needed to reproduce the experiments that originated the data. For this project, the main metadata required for that will include:
    a. Chemical compounds (compound name and formula, compound Chemical Abstract Service (CAS))
    b. Analysis technique (method, type, ontology source, capillary column, cold trap material, oven temperature method, carrier gas, flows, etc.)
    c. Material: tubing material, length and diameter, pressure regulator type and manufacturer, calibration standard (composition, amount of substance fraction, purity, expanded uncertainty, manufacturer…).

4. **Data processing and analysis**: this metadata field will include information about the data processing and analysis performed, such as outlier identification and removal, data transformation (whether data were transformed or not, the type of transformation), removal (or not) of rows with missing data, statistical tests performed, machine learning approach used and programming language and libraries used (name, version) among others.
5. Metadata on **licenses** and project **funding**
6. Metadata on **additional resources**: for example, scripts used to open, read and analyze the dataset.

Additional metadata not listed above but required by the selected open repository, where data and metadata will be deposited, and/or metadata that is mandatory and recommended by Metadata Schema v4.3 (DataCite), will be added.

The metadata will be provided in JSON format to ensure human- and machine-readability. Commonly used ontologies will be privileged. If specific ontologies are required, mapping to common ontologies will be provided.

Metadata, together with the data, will be made publicly available upon publication in an open-source journal by uploading them to the select trust repository (e.g. Zenodo). Metadata and data will be published under a Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Before publication, metadata will be stored at the internal server of the Federal Institute of Metrology METAS. During this period, metadata and data requests to the data author will be analyzed on a case-by-case basis. Sharing will be granted whenever the Intellectual Property Rights (IPR) of the data owners are not threatened and/or violated.

5. Data Quality

Even if the existing definitions of data quality (e.g., ISO 8000, 2022, D'Aniello et al., 2018) differ among them, they relate data quality with the degree to which data fulfils requirements or are fit-for-purpose. Therefore, data quality depends on the application of the data. For this work, the data quality dimensions that will be considered are completeness, correctness and accuracy. Regarding **completeness**, at least 95 % of data points should be present in the time series of each measurement for each LCSS. In this work, response signals have to be positive and above a defined threshold value corresponding to the zero air (reference clean air without pollutants). Data **correctness** will be fulfilled when these conditions are met for at least 99 % of the data of each measurement and LCSS. **Accuracy** is another crucial data quality metric in this project, which will be done by the uncertainty of the reference gas mixtures used for the calibration of LCSS (expanded uncertainty $U \leq 2$ % (coverage factor $k = 2$)), the uncertainty of the measurements performed with the reference instruments used in the calibration ($U \leq 5$ %), the uncertainty of the gas mixtures introduced in the simulation chamber ($U \leq 3$ %) and the uncertainty of the LCSS measurements ($U \leq 20$ %).

The experimental design and measurement protocol used for the data collection was defined to comply with the data quality objectives just mentioned. For example, data quality was improved by reducing the uncertainty of the measurement. For that purpose, the following aspects were considered during the data collection:

1. Laboratory LCSS calibration:
    a. Stability: measurements were performed during periods of time long enough to minimize the surface effects of the reactive pollutants in contact with tubings and analyzers and to ensure stability of the readings.
    b. Reference gas mixtures: to calibrate the laboratory reference instruments and to generate the gas mixtures needed to perform the laboratory calibration and performance assessment of the LCSS, reference gas mixtures (NO, NO_2, O_3, CO) of relative low uncertainty were used.
    c. Replicates: calibration measurements under a specific combination of variables (RH, temperature, pollutants) were repeated at least twice.

2. Experimental reference instruments: similar principles than for the laboratory LCSS calibration were used: use of standard gas mixtures of low uncertainty for the calibration of the reference instruments and time duration of each measurement long enough to ensure stability of the measurement.

Moreover, to minimize data gaps, data from 10 LCSS were collected. In case of failure of some of the LCSS, there should be enough data to perform valid statistical analysis – failure of all LCSS at the same time is not expected. During the experiments, periodic visualization of the data collected was done to check for collection and recording issues and, if needed, to repeat the experiment.

Finally, a thorough data pre-processing will be performed before proceeding to the data analysis to increase the data quality, which will include data cleaning (e.g., missing data), outlier detection and time synchronization of the different datasets.

6. Data Flow

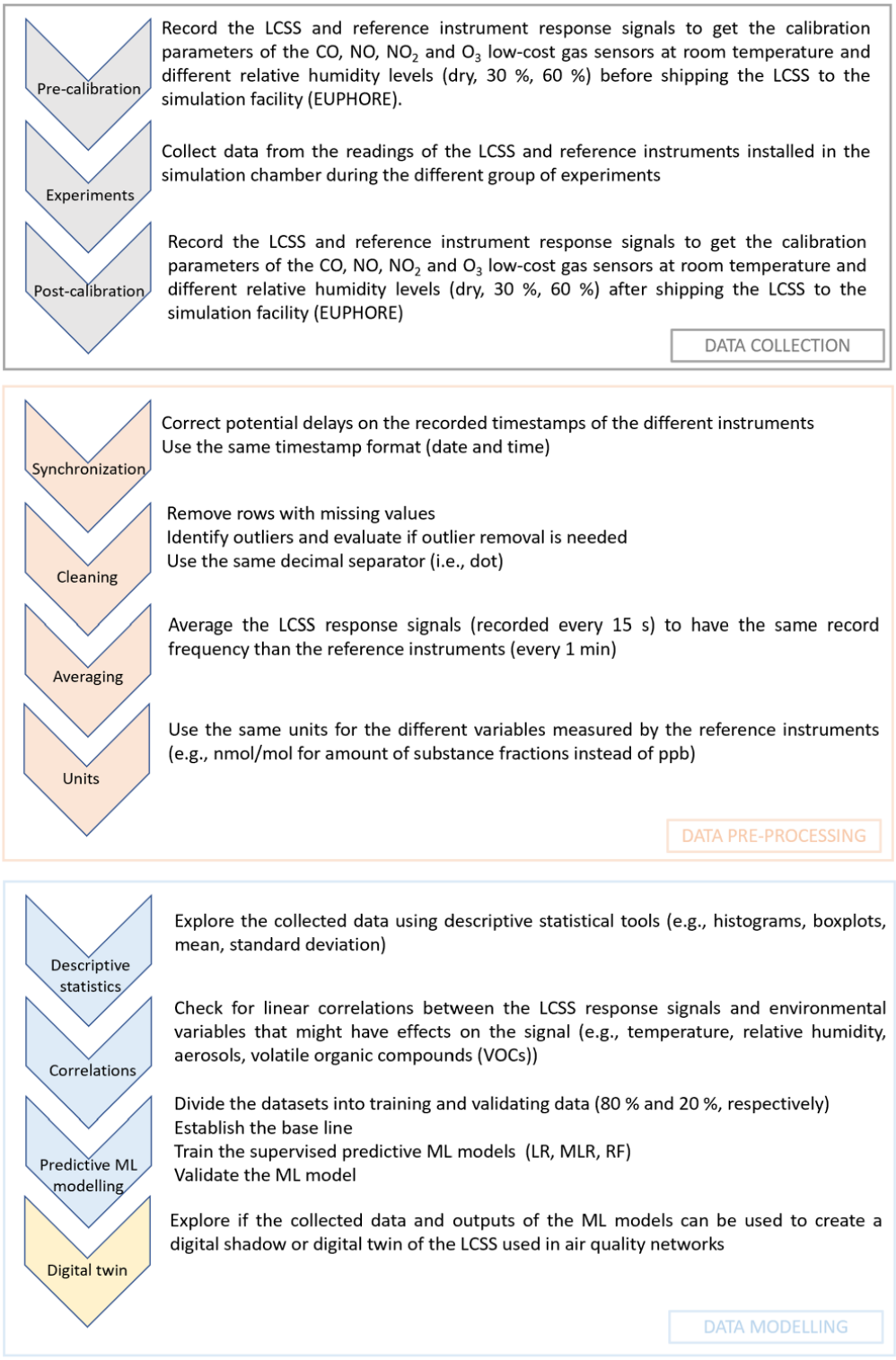

Figure 10: Data flow diagram of the project.

7. Data Model

7.1. Conceptual

By assessing the performance of LCSS in terms of response to environmental variables, such as temperature and relative humidity (i.e., correlation between LCSS response signals and the selected environmental variable), cross-sensitivity issues will be identified and corrected. Once these effects are taking into account, predictive models of LCSS response signals will be defined by comparing these signals against the reference instrument signals using different machine learning (ML) approaches. Defining ML models that accurately predict LCSS responses will not only contribute to get a better understanding on LCSS performance in air quality application, but also to optimize the calibration of LCSS by minimizing costs (e.g., time, consumables).

7.2. Logical

The columns/features needed for the analysis and modeling will be the following (grouped by dataset):

1. LCSS dataset: Epoch/Timestamp (DateTime), experiment number (integer), temperature (float), pressure (float), relative humidity (float), response signals (difference between channel 1 and 2 of each sensor) of CO (float), NO (float), NO2 (float) and O3 (float).

2. Calibration reference instrument dataset: Epoch/Timestamp (DateTime), experiment number (integer), temperature (float), pressure (float), relative humidity (float) and amount of substance fraction of CO (float), NO (float), NO_2 (float) and O_3 (float).

3. Experiment reference instrument dataset: Timestamp (DateTime), experiment number (integer), temperature (float), pressure (float), relative humidity (float), amount of substance fraction of CO (float), NO (float), NO_2 (float), O_3 (float), VOCs, concentration of $PM_{2.5}$, PM_{10} and black carbon.

4. Air quality dataset: Timestamp (DateTime), amount of substance fraction (after unit conversion) of CO (float), NO (float), NO_2 (float), O_3 (float) and PM_{10} (float).

8. Documentation

Because no simultaneous collaboration among multiple users is foreseen at this stage, the project will be documented using Jupyter Notebook 7.2. If in later stages, multi-user collaboration is needed, Google Colab will be privileged for the documentation in case some of the collaborators do not have Jupyter Notebook installed in their computers.

The documentation will include text describing the different sections, the code created to analyze the data conveniently commented, figures and graphs of the main results and other relevant information needed to understand the findings by third parties. Document version controlling and life cycle tracking – with comprehensive description of the changes – will be properly recorded.

9. Risks

This project will use data that were already collected and, thus, risks are minimal. During the data collection, potential risks were mitigated by using 10 LCSS, so in case of failure of some of the systems, the rest would have provided enough data to perform the data analysis. Replicates, real-time data checking and measuring periods allowing for stability were used to contribute minimizing the data collection risks.

Additional potential risks may appear during the data analysis phase, such as data loss, non-conclusive results and data quality objectives (DQO) that are not met, among others. As consequence of these risks, the project time schedule (February 2025) might be delayed by 2-3 months and the project costs increased by 5 %. Uncertainty might be affected also by these risks, resulting on expanded uncertainties of the LCSS measurements greater than the target ones. Mitigation measures (i.e., what will be done to decrease the likelihood of the risk occurring) and contingency measures (i.e., what will be done if despite the mitigation the risk still occurs) will be applied to minimize the impact of the risks (Table 4).

Table 4: Potential risks that can take place during the project and the mitigation and contingency measures that will be applied to minimize their impact.

Risk description	Impact and severity of occurrence	Mitigation measures	Likelihood after mitigation	Contingency measures
Data loss	Depending on the amount of data loss and the parameters affected by the loss, the impact will range from a low impact (e.g., analysis can be performed and conclusions extracted) to a very high-impact (e.g. not enough data to perform any analysis or performance) Low severity of occurrence	Frequent back-up copies saved in data storage devices (e.g. hard disk, UBS flash drives) and in METAS/CEAM internal servers. Data versioning control using standard approaches and software for that purpose (e.g. GitHub). Detail metadata and documentation regarding the data collection, pre-processing and analysis	Low	Calibration and experiments will be repeated in order to collect enough data to be able to perform the LCSS data analysis
Non-conclusive/ negative results	Impact will be low to some extent; this project is conceived as an exploratory work. So, even if results and non-conclusive or negative, these findings will be used to provide guidelines to LCS end-users for air quality applications. Medium severity of occurrence	A preliminary study was run before performing the data collection with promising outputs. A thorough data pre-processing will be performed to ensure that missing data and/or outliers are not masking significant results of the statistical analysis/modelling.	Medium	Experiments will be run using LCS types and technologies with lower cross-sensitivity (i.e., more selective)
Expanded uncertainty greater than the target value	The impact will be similar to the impact described in the previous risk (non-conclusive/negative results) Medium-high severity of occurrence.	Making sure that the set data quality objectives are met, data cleaning (e.g. identification and removal of outliers) and increasing the number of	Medium-high	Reference gas mixtures of lower uncertainty will be used for the calibration of LCSS during the repetition of the experiments. Coating of the lines

Risk description	Impact and severity of occurrence	Mitigation measures	Likelihood after mitigation	Contingency measures
		observations will be some of the measures to apply. Alternatively, because the known low quality (e.g., great uncertainty) of LCSS currently available in the market, the targe uncertainty value will be increased to realistic values on a pollutant specific approach.		and longer stability periods will be used to reduce the uncertainty.

10. Conclusions

The aim of this work is to explore the suitability of low-cost gas sensor systems (LCSS) to cover the spatial gaps of monitoring stations when assessing the air pollution of an area (e.g., neighborhood, city, region). The data required for this evaluation was generated during the experimental campaign described in section *2. Methods*. Data exploration preliminary results suggest that the analysis of the available data will be enough to answer the main research question of the project. Furthermore, detailed LCSS performance report and best practice guidelines on LCSS calibration will be elaborated based on the data collected. By following this Conceptual Design Report (CDR), project risks and delays will be minimized.

Acknowledgements

Special thanks to the EUPHORE team: M. Ródenas, T. Gómez, E. Borrás, T. Vera, R. Soler and A. Muñoz for their support during the performance of the experiments in the simulation chamber. This work is part of a project that is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRAIA-2020-1, through the ATMO-ACCESS Integrating Activity under Grant Agreement number: 101008004, as well as part of an internal project of the Gas Analysis Laboratory of the Federal Institute of Metrology METAS.

Statement

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 05.10.2024 Signature: 

References and Bibliography

D'Aniello, G., Gaeta, M., Hong, T., 2018. Effective quality-aware sensor data management. *IEEE Transaction on Emerging Topics in Computational Intelligence*, 2, 65-77. https://doi.org/10.1109/TETCI.2017.2782800

ISO 8000, 2022. International Standard Organization 8000-2 Data Quality – Part 2-Vocabulary. ISO-Technical Committee 184.

Tancev, G., Ackermann, A., Schaller, G., Pascale, C., 2022. Efficient and automated generation of orthogonal atmospheres for the characterization of low-cost gas sensor systems in air quality monitoring. IEEE Transactions on instrumentation and measurement, 71, 1006410, https://doi.org/10.1109/TIM.2022.3198747

WHO, 2021. WHO global air quality guidelines. Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: World Health Organization, 2021. Licence: CC BY-NC-SA 3.0 IGO.

Appendix 1

A1.1. Material of the experimental design

For the experiments in the atmospheric simulation chamber (i.e., European Photoreactor EUPHORE), the different pollutants were added to the gas mixtures generated for the experiments using different methods:

1) NO was added from gas cylinder (estimation of the NO amount fraction introduced in the chamber through the time during which the gas cylinder was opened)
2) CO was added by injection of a syringe filled previously from the CO gas cylinder using a septum for that. The amount fraction was estimated considering the volume of the syringe filled with CO.
3) Water was added to the chamber in liquid form, by applying the Venturi principle, to modify the relative humidity of the gas mixtures.
4) O_3 was added using an ozone generator (SONIMIX, LNI Swissgas). The O_3 amount fraction was estimated considering the generation rate and time.
5) Aerosols were added using the so-called "seeds", which consisted in $(NH_4)_2SO_3$ (5 M) and ClNa (1M), or by burning biomass (e.g., orange tree wood). For the latter, the aerosol concentration was estimated considering the time period during which smoke was introduce in the simulation chamber.
6) 1,3,5-trimethylbenzene (TMB) was added into the simulation chamber by injection. To estimate the amount of substance fraction, the syringe volume filled with TMB was considered.
7) NO_2 was obtained in the chamber by reaction of the compounds NO and O_3 (one-to-one reaction). The amount of substance fraction was estimated by stoichiometry.

A1.2. Experimental protocols

Protocol of Experiment 1

The protocol of Experiment 1, run on 21.04.2023, is described in Table A1.1.

Table A1.1: Description of the protocol for Experiment 1.

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
01	background				30
02	addition	NO	38 s	15	10
03	chamber flushing		2.4 m^3/h for 50 min		50
04	addition	NO	25 s		40
05	addition	NO	18 s	15	30
06	addition	SF_6	6 mL		3
07	addition	CO	40 mL	200	30
08	addition	CO	30 mL		3
09	addition	NO	62 s		36
10	addition	CO	30 mL		3
11	addition	NO	62 s		
12	chamber flushing		3.8 m^3/h for 65 min		15
13	addition	H_2O	11 min; up to 30 %		15
14	addition	H_2O	8 min; up to 40 %		34
15	addition	H_2O	8 min; up to 60 %		42

Protocol of Experiment 2

The protocol of Experiment 2, run on 25.04.2023, is described in Table A1.2.

Table A1.2: Description of the protocol for Experiment 2.

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
01	background				30
02	addition	SF_6	6 mL		24
03	addition	O_3	26 nmol/mol/min for 5 min	up to 130	22
04	addition	O_3	26 nmol/mol/min for 45 s	up to 151	22
05	addition	NO	0.8 nmol/mol/s for 187 s	150	17
06	addition	NO	7 s		30
07	chamber flushing		for 48 min		20
08	addition	O_3	2 min		36

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
09	chamber flushing		for 24 min		28
10	addition	NO	112 s	60	28
11	addition	H_2O	22 min; up to 40 %		30
12	addition	H_2O	22 min; up to 65 %		41
13	chamber flushing		for 34 min		40
14	addition	O_3	3 min	up to 60	7
15	addition	O_3	15 s	up to 60	29
16	addition	H_2O	7 min; up to 60 %		28

Protocol of Experiment 3

The protocol of Experiment 3, run on 26.04.2023, is described in Table A1.3.

Table A1.3: Description of the protocol for Experiment 3.

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
01	background				21
02	addition	SF_6	6 mL		2
03	addition	O_3	4 min 30 s	120	0
04	addition	H_20	for 13 min	up to 45 %	5
05	addition	NO	75 s	up to 60	38
06	addition	$(NH_4)_2SO_4$	1 M for 10 min		
07	addition	H_2O	after 3 min seeds addition started	up to 45 %	36
08	addition	NaCl	0.1 M for 5 min		45
09	addition	O_3	1 min 45 s	up to 60 (O_3 + NO_2)	3
10	addition	H_2O		up to 45 %	1
11	addition	NO	37 s	up to 60 (O_3 and NO_2)	6
12	addition	$(NH_4)_2SO_4$	1 M for 10 min		30
13	addition	NaCl	1 M for 15 min		

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
14	addition	H_2O	after 3 min seeds addition started	up to 45 %	44
15	addition	H_2O	for 5 min	up to 60 %	30
16	chamber flushing		98 min		16
17	addition	NO	70 s + 10 s	up to 50	25
18	addition	CO	40 mL	up to 300	29
19	addition	H_2O	for 18 min	up to 45 %	24
20	addition	$(NH_4)_2SO_4$	1 M for 15min	up to 30 µg/m^3	
21	addition	H_2O	after 2 min seeds addition started	up to 60%	38
22	addition	NaCl	1 M for 20 min	up to 80 µg/m^3	22
23	addition	H_2O	for 2 min	up to 60 %	

Protocol of Experiment 4

The protocol of Experiment 4, run on 27.04.2023, is described in Table A1.4.

Table A1.4: Description of the protocol for Experiment 4. TMB is 1,3,5-trymethylbenzene.

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
01	background				22
02	addition	SF6	6 mL		5
03	addition	TMB	230 uL for 8 min	201	1
04	addition	H_2O_2	5 mL for 1 min		0
05	addition	NO	0.8 nmol/mol/s for 63 s		32
06	addition	H_2O	for 13 min		34
07	addition	$(NH_4)_2SO_4$	0.5 M for 35 min		0
08	addition	H_2O	for 7 min		31
09	addition	H_2O	for 2 min		1
10	chamber opening				22
11	addition	H_2O	for 12 min		60
12	addition	H_2O			150

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
13	chamber closing				25
14	chamber flushing		for 82 min		0
15	chamber filling	outside air			

Protocol of Experiment 5

The protocol of Experiment 5, run on 28.04.2023, is described in Table A1.5.

Table A1.5: Description of the protocol for Experiment 5.

Step	Action	Compound	Addition value	Estimated amount fraction (%)	Measuring time (min)
01	chamber filling stopped	outside air			22
02	background				70
03	analysers' sampling connected outside chamber				4
04	analysers' sampling reconnected to chamber				26
05	chamber opening				77
	addition	SF_6			90
06	addition	NO	7 s		100
07	addition	H2O		up to 60 %	62
08	chamber closing				34
09	addition	$(NH_4)_2SO_4$	1 M for 20 min		

Protocol of Experiment 6

The protocol of Experiment 6, run on 02.05.2023, is described in Table A1.6.

Table A1.6: Description of the protocol for Experiment 6.

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
01	background				22

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
02	addition	O_3	5 min 40 s	150	4
03	addition	O_3	15 s		28
04	addition	NO	3 min 20 s	up to 150	18
05	addition	SF_6	6 mL		27
06	chamber flushing		for 47 min		
	addition	O_3		up to 60	21
07	addition	H_2O		up to 55 %	33
08	addition	O_3	1 min 15 s	up to 60	4
09	addition	NO	20 s	up to 60	37
10	addition	$(NH_4)_2SO_4$	1 M for 20 min		82
11	addition	O_3	2 min 20 s	up to 60	1
12	addition	H_2O		up to 55 %	1
13	addition	NO	1 min	up to 75	7
14	addition	NaCl	1 M for 15 min	up to 55 µg/m^3	40
15	addition	O_3	1 min 40 s	up to 60	7
16	addition	NO	40 s	up to 60	29
17	addition	$(NH_4)_2SO_4$	1 M for 15 min		63
18	addition	H_2O		up to 65 %	30
19	chamber filling	outside air			

Protocol of Experiment 7

The protocol of Experiment 7, run on 03.05.2023, is described in Table A1.7.

Table A1.7: Description of the protocol for Experiment 7.

Step	Action	Compound	Addition value	Estimated amount fraction (%)	Measuring time (min)
01	background				9
02	NOx analyser	zero air			5
03	addition	SF_6	6 mL		6
04	NOx analyser reconnected to chamber				18

Step	Action	Compound	Addition value	Estimated amount fraction (%)	Measuring time (min)
05	chamber opening				172
06	addition	H_2O	for 55 min	up to 55 %	54
07	addition	NO	7 s		88
08	chamber partial closing (strong wind)				2
09	chamber closing				97
10	NOx analyser	zero air			20
11	addition	NO	50 s		

Protocol of Experiment 8

The protocol of Experiment 8, run on 04.05.2023, is described in Table A1.8.

Table A1.8: Description of the protocol for Experiment 8.

Step	Action	Compound	Addition value	Estimated amount fraction (nmol/mol)	Measuring time (min)
01	background				26
02	addition	NO	100 s	up to 80	0
03	addition	SF_6	6 mL		36
04	addition	CO	140 mL	up to 700	26
05	addition	O_3	5 min	up to 85	32
06	addition	O_3	2 min 20 s	up to 60	35
07	addition	H_2O	for 18 min	up to 55 %	68
08	biomass smouldering burning	particles	10 s		21
09	biomass smouldering burning	particles	15 s		90
10	addition	CO	40 mL	up to 200	42
11	addition	NO + O_3			48
12	addition	NO		up to 80	