CitiObs - ENHANCING CITIZEN OBSERVATORIES FOR HEALTHY, SUSTAINABLE, RESILIENT, AND INCLUSIVE CITIES

DELIVERABLE 2.3

‘ValAir’ and ‘MapAir’ toolkits

LEAD AUTHORS: VASILEIOS SALAMALIKIS & PHILIPP SCHNEIDER (NILU)

DISCLAIMER

This document contains material, which is the copyright of certain CITIOBS beneficiaries, and may not be reproduced or copied without permission.

The information appearing in this document has been prepared in good faith and represents the views of the authors. Every effort has been made to ensure that all statements and information contained herein are accurate; however, the authors accept no statutory, contractual or other legal liability for any error or omission to the fullest extent that liability can be limited in law.

This document reflects only the view of its authors. Neither the authors nor the Research Executive Agency nor European Commission are responsible for any use that may be made of the information it contains. The use of the content provided is at the sole risk of the user. The reader is encouraged to investigate whether professional advice is necessary in all situations.

No part of this document may be copied, reproduced, disclosed, or distributed by any means whatsoever, including electronic without the express permission of the CITIOBS project partners. The same applies for translation, adaptation or transformation, arrangement or reproduction by any method or procedure whatsoever.

COPYRIGHT MESSAGE

ACKNOWLEDGEMENT

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them

DOCUMENT DESCRIPTION

Delivery date:	08/01/26		
Type*:	OTHER	**Dissemination Level**:**	PU - Public
Contributing WP:	WP2		
Lead Partner Organisation:	NILU		
Lead author(s):	Vasileios Salamalikis (NILU) Philipp Schneider (NILU)		
Contributor(s):	Shobitha Shetty (NILU) Amirhossein Hassani (NILU) Paul D. Hamer (NILU) Kerstin Stebel (NILU) Terje Koren Berntsen (University of Oslo) Nuria Castell (NILU) Peter Bult (RIVM) Sjoerd van Ratingen (RIVM)		
Reviewer(s):	Nuria Julia (CREAF) Sjoerd van Ratingen (RIVM)		

VERSION LOG

Version	Date	Partner	Content and changes
0.1	01.11.2025	NILU	Initiate drafting the deliverable
0.2	10.12.2025	NILU	Submit the deliverable for internal review

0.3	18.12.2025	RIVM	Complete the initial review process
0.4	19.12.2025	NILU	Prepare a revised version incorporating reviewers' comments and concerns
0.5	31.12.2025	CREAF	Complete the second review process
1.0	08.01.2026	NILU	Prepare the final version of the deliverable and submission to ECAS portal

CITIOBS

CitiObs is a four-year project funded under Horizon Europe by the European Commission. CitiObs will consolidate and apply tools and practice-based knowledge for co-creating data, knowledge and local action via Citizen Observatories (COs): these tools will enhance existing and new citizen observatories to engage people from a diverse range of communities, add value to environmental observations in the urban context, increase and validate citizen observations of the urban environment as part of the existing in-situ Earth Observation systems, co-create inclusive local actions for sustainability and ensure that CO data contributes to research and policy development towards the objectives of the European Green Deal. To ensure broad use, the CitiObs tools and approaches will be developed in co-creation with COs in 5 Frontrunner cities, finetuned with 30 Implementer cities and showcased to 50 Fellow cities.

CitiObs will support citizen observatories in distinct cities to create/enhance/or scale up inclusive and diverse citizen observatories, fostering in particular an active role of citizens in the observation of the urban environment using low-cost sensor technologies and wearables, with a particular focus on air quality and related environmental measures. CitiObs will formalise, valorise and legitimise the role of citizen observations.

The CitiObs methodology of using a large-scale demonstration, co-design and coaching approaches with CO stakeholders (citizens, scientists, policy/decision makers) in 5+30+50 cities in Europe explicitly builds on the Responsible Research & Innovation (RRI) dimensions as founding principles. Ethics consideration will be addressed consistently across all Work Packages.

- WP1. Social dimensions of Citizen Observatories for transition governance
- WP2. Tools, Technologies, and Data Services for Citizen Observatories
- WP3. Co-creation of data and actions for healthy, sustainable and resilient cities with Citizen Observatories
- WP4. Impact creation, Communication, Dissemination and Exploitation
- WP5. Project management
- WP6. Ethics

EXECUTIVE SUMMARY

This report summarises activites, methodologies, and results from the ValAir and MapAir toolkits developed as part of the CitiObs project.

The ValAir toolkit offers a comprehensive overview of the algorithms developed through the CitiObs project for the quality control and correction of citizen-generated low-cost sensor (LCS) data. ValAir builds on FILTER (Framework for Improving Low-Cost Technology Effectiveness and Reliability), a suite of algorithms designed to harmonize, quality-check, flag, and perform in-situ corrections on crowd-sourced sensor measurements. By addressing heterogeneity and variable reliability across different LCS networks, FILTER ensures that crowd-sourced data are both validated and consistent.

The initial version of FILTER was developed for static $PM_{2.5}$ sensors. The FILTER methodology has been also expanded to examine data quality challenges arising from wearables, mobile sensors, and noise sensors. All FILTER versions follow a common processing pipeline, generating individual quality flags based on distinct statistical tests that indicate the reliability of each measurement. This modular and hierarchical design allows users to tailor the quality control and correction process to the characteristics of their data and the requirements of specific applications.

The MapAir work presented in this report shows that combining observations from low-cost sensor networks with advanced data fusion, data assimilation, and machine-learning techniques enables the production of high-quality, high-resolution air-quality maps that go beyond what traditional monitoring systems can provide. The MapAir algorithms demonstrate how point-based sensor measurements, when carefully quality-controlled and merged with model and satellite information, can be transformed into spatially complete and physically consistent pollution fields at both urban and regional scales. The S-MESH framework extended with low-cost sensors further illustrates how integrating validated low-cost sensor data can improve the accuracy of daily $PM_{2.5}$ mapping across Europe, particularly in densely populated and highly polluted areas where information about local pollution patterns is important. These approaches show potential of substantial societal benefits: they support more reliable exposure assessments, allow authorities to pinpoint emission hotspots, inform public health responses, and empower communities with transparent and actionable insights into local air quality.

TABLE OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

1 INTRODUCTION

1.1 Purpose of the document

This document provides an overview of the activities, methodologies, and results from the ValAir and MapAir toolkits developed as part of the CitiObs project. It aims to present in detail the algorithms used for quality control and correction of low-cost sensor data, and to illustrate how data-fusion and machine learning methodologies combine diverse air-quality datasets, including low-cost sensor networks, official monitoring stations, atmospheric models, and satellite observations, into spatially continuous, decision-ready air pollution information.

1.2 Scope of the document

This document covers the scope of Task 2.2.3 of the CitiObs project, focusing on the ValAir and MapAir toolkits. It provides an extensive overview of the quality control and correction procedures applied to low-cost sensor data and describes how modern data-fusion and machine learning techniques integrate heterogeneous air quality datasets—including low-cost sensor networks, regulatory monitoring stations, atmospheric models, and satellite observations—into spatially complete, decision-ready air pollution information.

1.3 Structure of the document

The document is organised as follows:

- Section 1 - Introduction: description of the purpose and scope of the document and its structure.
- Section 2 – Data quality considerations: general overview of low-cost sensors data quality and basic introduction to ValAir.
- Section 3 - ValAir: description of the quality control and correction methodologies applied to the low-cost sensor data in the CitiObs project.
- Section 4 - MapAir: description of how modern data-fusion and machine-learning methods can transform heterogeneous air-quality data (e,g, LCS networks, regulatory monitors, atmospheric models, and satellite retrievals) decision-ready air pollution information.
- Section 5 - Summary: description of the key points addressed by the ValAir and MapAir toolkits.

2 DATA QUALITY CONSIDERATIONS

Uptake of citizen generated data in (government) decision making is often hampered by a lack of trust in the data's validity: "the data has insufficient quality" the refrain often goes. Yet, what exactly is meant with *quality* and in what respect the data is *insufficient* is seldom specified. By extension, the steps that may be taken to address this perceived lack of quality are difficult to identify.

In this document we lay out CitiObs' perspective on the current state of data quality determination for citizen generated data and reflect on what steps may be taken toward resolving the debate surrounding data quality. We focus specifically on low-cost particulate matter sensors for air quality monitoring. Nonetheless, some of the best practices we suggest for citizen air quality monitoring data may apply more broadly to other air pollutants and environmental monitoring domains as well.

What is data quality anyway

The first hurdle we observe is that data quality is an abstract concept that can have a very different meaning to different people. A (citizen) scientist may take data quality to mean the measurement uncertainty, while a legal specialist may interpret quality in terms of how well they can trace the provenance of a dataset back to measurement devices. A policy advisor, by contrast, may interpret quality to mean that the data is current and actionable.

The bottom line is that the quality of data is only "good" when it meets the expectation of the user and their specific use case. In other words: data quality is always relative to the application. The concept of data quality is therefore not about achieving the best possible measurement, but rather about collecting measurements of *known* quality against a *known expectation*.

The importance of metadata

To a large degree, questions of data quality come down to providing extra information context about how a data set came to be. This is why metadata plays a central role, as metadata provides information about the context of the measurements. Common metadata attributes of interest are the specific device doing the measuring, along with information about when and where the data was collected. More specific information is always better. To allow metadata to be processed by automated systems, it is also good practice to adhere to a standard exchange format. To that end, CitiObs promotes the use of the SensorThings API as the default exchange format for both measurement data and its associated metadata.

The knowns and unknowns of low-cost sensors

Very often data quality is interpreted in the narrower sense of how well a measurement result corresponds with the true condition being measured. In other words, what is the uncertainty, the bias, the precision, and/or the accuracy of the measurement?

Ideally, aspects like the uncertainty of the measurement data are also known, quantified, and communicated through the metadata. Unfortunately, the low-cost air quality sensors available to citizens are subject to a number of nuisances, for instance: they drift over time, are sensitive to humidity, report both real and spurious spikes in the measured concentrations. What's more, all these influences can vary in magnitude depending on the specific sensor make and model, its production batch, and even the specific orientation of its deployment. Given all these influences, a single low-cost sensor does not (and indeed cannot) capture information about its own uncertainty. This is perhaps the largest challenge to a constructive discussion about the uptake of citizen data.

One pathway to try and address this issue is to combine the measurement of individual low-cost sensors into a network and estimate quality indicators through post-processing of that network as a whole. This is the processing that is promoted by the CitiObs project.

Post-processing and the application data quality flags

By processing a network of many sensors we can attempt to infer how well the sensors are performing. For example, we can look at how sensor measurements vary in space and in time. We can compare them with each other, but also with other sources of data, such as official monitoring networks.

A viable approach to determining quality indicators could be to annotate individual sensor measurements with quality flags. A quality flag captures the logical outcome of a conditional test. For example, we could ask if a sensor measurement is positive. The outcome of this test is either true or false and can be stored as a logical flag in measurement metadata.

Which conditional tests are useful to express as a quality flag is inherently dependent on the intended use case. The combined experience in CitiObs suggests a few broad categories of quality flags that may be useful to consider generally:

- Physical range: consider if the sensor is reporting a measurement that is plainly unphysical, for instance a negative value.
- Flatline: consider if the sensor is spuriously reporting repeated constant values

- Outlier: If one sensor reports a concentration that is several times higher than reported by the rest of the network, we call it an outlier. Such an outlier could indicate a bad measurement but could also indicate a local source of emission. Whether or not you want to filter out such measurements depends greatly on the research question.
- Correlation: Does a given sensor correlate with other sensors nearby? If not, then it could indicate a bad measurement but could also indicate a local source of emission.
- Similarity: How does the sensor compare to other sources of data, such as nearby official measurements or other sensing devices?

By offering such quality flags along with the measurement data, the end user retains the freedom to select the level of quality that suits their use case. For instance, if one is interested in using sensor data to refine annual average modelled concentration maps, then it is important to ensure that the sensor data is representative of the large-scale background. Thus, a similarity check that evaluates the sensors against the official reference measurement should be included. If the use case of the sensors is to look for episodic localized sources of emission, however, we probably do not want to include an outlier filter, as that could accidentally remove the very datapoints we are looking for.

Because these quality flags annotate individual measurements, then can be evaluated independently for the original measurement result and any post-processed calibration outcome. Because the act of calibration tends to correct measurement results toward reference measurements, one would expect that the calibration outcomes score better on similarity. In this sense we see how quality flags express information about the (estimated) quality of the data, while it is the calibration routine that acts to "improve" the data.

In practice

The adoption of quality flags offers a flexible approach to establishing the data quality of low-cost sensors. The specific rules and recipes needed to calculate such flags, however, are highly context dependent. For example, the FILTER method established within the CitiObs project (Hassani et al., 2025) proposes a specific recipes recipe for $PM_{2.5}$, but other air pollutants or other physical properties will necessitate a different ruleset.

The advantage of quality flags is that offers a flexible way to make the discussion around data quality concrete and negotiable. They are aimed mainly at users that want to (re-) analyze the dataset as a whole and embed them into other products. For users that wish to just consume the outcome of a single sensor. It may be helpful to reduce all computed flags to single *plausibility*

metric (https://www.samenmeten.nl/data/plausibiliteit-van-fijnstofmeting) that qualitatively indicates how likely it is that this particular measurement reflects the true conditions.

3 VALAIR

The absence of a harmonized, quality-controlled, and corrected dataset for low-cost sensor (LCS) measurements—integrating data from multiple LCS networks, particularly those operated by citizens—has limited the use of LCS data in air quality research, public health assessments, and environmental policy. To address this gap, the CitiObs project developed **ValAir**, a comprehensive toolkit for quality control (QC) and correction of LCS measurements. Its core component, **FILTER** (**F**ramework for **I**mproving **L**ow-Cost **T**echnology **E**ffectiveness and **R**eliability), provides a set of algorithms to unify, quality check, flag, and perform in-situ correction of outdoor, crowd-sourced sensor observations.

The original version of FILTER (Hassani et al., 2025) was specifically developed for QC and correction of outdoor static $PM_{2.5}$ sensors. It offers algorithms to harmonize, validate, and correct citizen-operated sensor data, addressing the heterogeneity and variable reliability of measurements across different LCS networks. FILTER has been applied to data from the largest networks, namely sensor.community (https://sensor.community/en/) and PurpleAir (https://www2.purpleair.com/), two of the most extensive citizen-driven air quality monitoring networks across Europe.

Building upon this, the initial FILTER methodology has been expanded to address additional quality control challenges related to mobile sensors and wearables measuring $PM_{2.5}$, as well as noise measurements. For mobile and wearable sensing, the original framework has been adapted to account for the dynamics of sensor movement while preserving the core principles of transparency, scalability, and sensor independence; it also introduces dedicated steps to address motion-related artifacts, intermittent time series, and location-specific uncertainties. For noise sensing, the methodology incorporates the fundamental components of FILTER but is adapted to noise measurements obtained from low-cost sensors, analyzing sound levels expressed in dB(A). It operates on a single integrated noise metric representing overall acoustic exposure rather than individual frequency components, with the A-weighting filter applied to raw decibel measurements to account for the frequency sensitivity of human hearing by emphasizing mid-frequency sounds (approximately 2–5 kHz) and attenuating very low and very high frequencies (Murphy and King,

2022). Consequently, A-weighted levels more closely reflect perceived loudness and potential hearing damage than unweighted sound pressure levels.

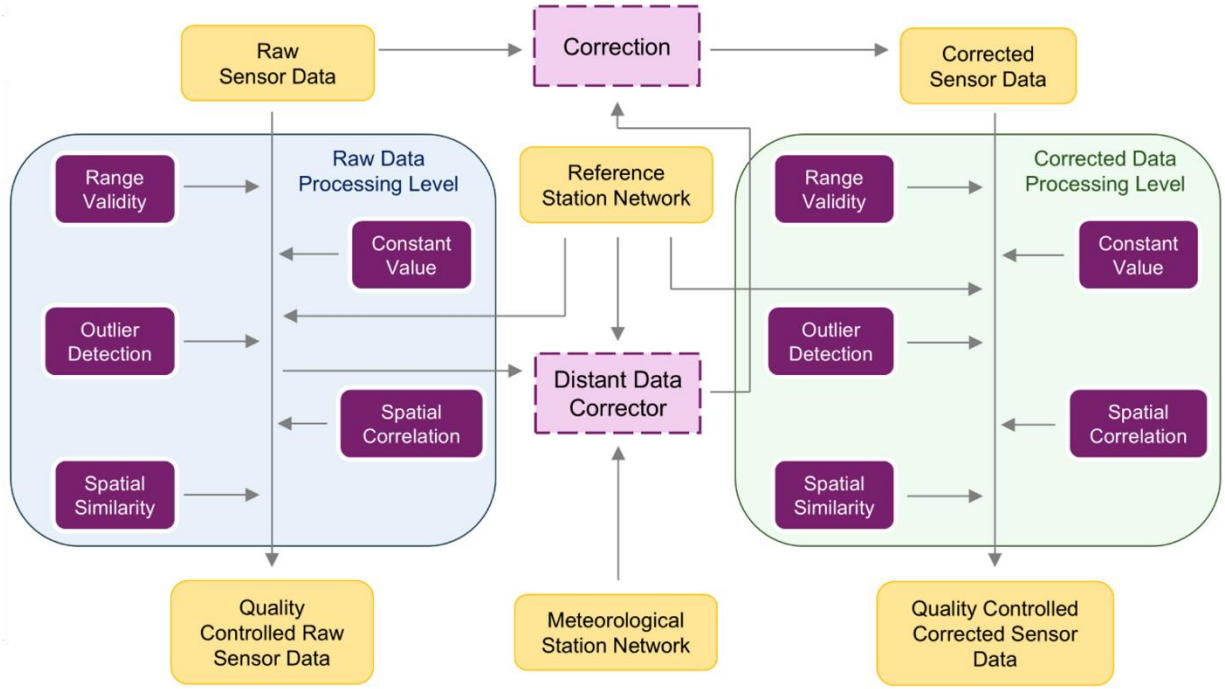

Figure 1: *Graphical representation of the FILTER framework. The names of the statistical checks are presented in a generic form, with the corresponding algorithms to be adapted for each specific FILTER version.*

All variants of the FILTER framework share a common processing pipeline (Figure 1), in which both basic and advanced statistical checks are applied sequentially to the sensor data. The quality control and correction workflow is illustrated in Figure 1. While the statistical steps are similar across all three FILTER variants, the underlying algorithms differ following the specific characteristics of each type of LCSs measurement. Based on the outcomes of these checks, individual quality flags are generated to indicate the reliability of each measurement. This modular and hierarchical design allows users to customize the QC and correction process according to the data characteristics and the needs of each specific application.

Depending on the FILTER version, the framework incorporates two processing levels, namely 'Raw' and 'Corrected', allowing the quality control procedure to be applied at either stage. It is important to clarify the use of term 'Raw'. In the broader field of environmental monitoring, 'Raw' data from LCSs often refers to the direct measurand (e.g., voltage, current, resistance, conductance, or sound pressure) that is subsequent transformed into mass concentration and noise levels through proprietary algorithms. In practice, however, the term 'Raw' here refers to the measurements obtained from the online repositories of the LCS data providers. These values

are assumed to have undergone minimal to no quality assurance, although some basic processing—such as averaging, internal corrections, conversions, or laboratory calibration—may already have been applied, either within the sensor firmware or during ingestion on the hosting servers. The data referred to here as 'Raw' corresponds most closely to Level-1 data, representing intermediate geophysical quantities (Schneider et al., 2019).

3.1 QC procedure

This section describes the quality control procedures implemented in FILTER for both static and mobile sensors. It presents in detail the steps followed to assess and ensure the reliability of sensor measurements, including the mathematical description of each step and the extraction of the corresponding quality flags.

3.1.1 QC 0: Spatiotemporal Attributes

This check focuses on the evaluation of the spatiotemporal attributes associated with each measurement record and is specifically applicable to mobile sensors. Since one of the defining characteristics of mobile LCS data is the explicit geolocation of each observation, ensuring the validity and consistency of both coordinates and timestamps is a prerequisite for any subsequent analysis.

A typical issue arises when multiple measurements from the same device are recorded with identical timestamps, which may indicate synchronization errors or faulty data logging. Similarly, errors in geolocation such as missing, duplicated, or unrealistic coordinates can lead to misleading representations of the measured concentrations. Two main checks are suggested:

- Coordinate validation: Each measurement is first evaluated against an expected region of interest (ROI), which could correspond to a neighbourhood, city boundary, or national extent depending on the study context. Measurements located outside the ROI are flagged as low-quality data, as they likely reflect erroneous or corrupted location information.
- Speed plausibility check: The implied travel speed of the sensing platform is calculated by the ratio between the distance between two consecutive measurements calculated through the great-circle algorithm and the corresponding time interval. The speed values are then compared against plausible ranges that depend on the platform carrying the sensor. For example:
    - pedestrians: 0–15 km h^{-1},
    - cyclists: 0–45 km h^{-1},

- cars (urban driving conditions): 0–140 km h^{-1}

Measurements with calculated speeds outside of these ranges are flagged as unreliable, as they may indicate GPS errors, temporal mismatches, or other technical artifacts. This QC step follows a binary flagging system: '0' = data point fails the QC test (considered invalid) and '1' = data point passes the QC test (considered valid).

3.1.2 QC 1: Range Test

The *Range Test* evaluates whether the recorded LCS measurements fall within a physically plausible domain. An observation outside the predefined limits is flagged as invalid with a value of '0', while observations within the acceptable range are flagged as valid (Flag = 1). If X_t denotes the measurement at time t. The flag for the *Range Test* is defined as:

$$f_t^{\text{range}} = \begin{cases} 0 & \text{Out-of-range: } X_t < th_l \text{ and } X_t > th_u, \\ 1 & \text{Within range: } th_l \leq X_t \leq th_u \end{cases}$$

where th_l is the lower threshold and th_u the upper threshold. The *Range Test* is implemented across all FILTER versions, with the following default thresholds:

Static $PM_{2.5}$: 0 – 1,000 µg m^{-3}

Mobile $PM_{2.5}$: 0 – 1,000 µg m^{-3}

Noise Level: 0 – 120 dB

For $PM_{2.5}$ measurements, a value is considered valid if it lies between 0 and 1,000 µg m^{-3}, a range that reflects both the operational limits of low-cost optical sensors and the expected variability of urban ambient air pollution. Concentrations above the upper limit are highly unlikely in typical urban environments, even under exceptional conditions such as wildfire smoke or dust storm intrusions, while negative values have no physical interpretation and typically indicate measurement or retrieval errors.

For noise level measurements, values are considered valid if they fall between 0 and 120 dB, consistent with the performance range of low-cost acoustic sensors and the typical variability of urban sound levels (Murphy and King, 2022). In situations where sensors are deployed in extremely noisy environments (e.g., near airports or industrial zones), these limits may be adjusted accordingly to reflect local conditions.

3.1.3 QC 2: Quality control for Constant or Flatlined Sensor Measurements

The *Constant Value* step addresses the detection of constant value reporting, which can indicate sensor malfunction, signal saturation, or data transmission errors. This step flags any sensor that continuously records almost identical values over a rolling time window. It uses an n-step moving window also considering the number of available (non-missing) measurements within each temporal window and it is updated successively at each time step. A quality flag is assigned to indicate the validity of each measurement.

If X_t denotes the measurements at a given location within a temporal window $\mathcal{T}$, where $t \in \mathcal{T}$, then the window is defined as:

$$\mathcal{T} = [t_i - n\Delta t + 1, t_i]$$

Here, $n\Delta t$ represents the length (duration) of the temporal window $\mathcal{T}$, and Δt is the temporal resolution of X_t time series. Within each temporal window, the statistical range (Maximum – Minimum) is computed to assess the variability or the stability of the measurements,

$$Range_{\mathcal{T}^*} = \max_{t \in \mathcal{T}^*}(X_t) - \min_{t \in \mathcal{T}^*}(X_t)$$

where $\mathcal{T}^* \subseteq \mathcal{T}$ represents the available (non-missing) values within the temporal window $\mathcal{T}$. The corresponding quality flag, f_t^{constant}, for the constant Value test is then defined as:

$$f_t^{\text{constant}} = \begin{cases} 0 & \text{Insuffient data: } n_{\mathcal{T}^*} < n_{min}, \\ 1 & \text{Constant: } Range_{\mathcal{T}^*} \leq Range_{th} \text{ and } n_{\mathcal{T}^*} \geq n_{min}, \\ 2 & \text{Non-constant: } Range_{\mathcal{T}^*} > Range_{th} \text{ and } n_{\mathcal{T}^*} \geq n_{min}. \end{cases}$$

where:

- $n_{\mathcal{T}^*}$: number of available (non-missing) values within $\mathcal{T}$.
- n_{min}: minimum number of non-missing measurements required to evaluate the test, typically given as $p \times n\Delta t$, where p is the minimum percentage of data completeness.
- $Range_{th}$: threshold to assess the flatlined behavior.

For static LCSs, and in the case of hourly $PM_{2.5}$, the default duration of the time window is 8 h, with a minimum data availability requirement of 6 h to perform the statistical comparison. In contrast, for mobile LCSs, the *Constant Value* test requires specific adaptations, as the variability of $PM_{2.5}$ concentrations depends on both the measurements' temporal resolution and the deployment platform (e.g., pedestrian, bicycle, or vehicle). For example, for a dataset recorded at a temporal resolution of 10 sec, the test can be applied using a rolling window corresponding to

3 min of data, i.e., 18 consecutive measurements. If fewer than 12 valid measurements are available within the window, the data point under evaluation is flagged as '0', indicating insufficient data coverage, as the test cannot be reliably evaluated statistically. For both static and mobile $PM_{2.5}$ sensors, the threshold used to compare the calculated range is set 0.1 μg m^{-3}. This value is selected because, under normal atmospheric conditions, such a small variation in PM concentrations over the given time intervals is highly unlikely.

In addition, the relevant threshold for noise level measurements is set to 1 dB, while flatline behavior in noise time series is relatively rare. This is because the microphones in noise LCSs are highly sensitive and can detect even small fluctuations in sound levels caused by ambient noise, wind, or electronic variations. Therefore, the presence of simultaneous constant values indicates sensor malfunction, data logging failure or communication errors rather than environmental noise stability. Depending on the temporal resolution of the noise data, the duration and data completeness should be adjusted accordingly. For example, for noise time series with a 1 min temporal resolution, the default duration of the rolling windows is 10 min, with a minimum data availability requirement of 8 min to evaluate the test.

3.1.4 QC 3: Spatiotemporal outlier detection

The detection of abnormal data in the LCSs time series is conducted out across both the temporal and the spatial domains. The outlier detection algorithm integrates a temporal consistency check with a spatial comparison step to distinguish real pollution events and sensor-related anomalies. The test operates over an n-step rolling temporal window, which is updated successively at each time step, and creates a quality flag that determines the validity of each individual measurement.

A similar detection process is applied both for static $PM_{2.5}$ and noise sensors, although different statistical indicators are used to detect outliers (Table 1). For mobile sensors, this check is excluded from the quality control process. The temporal window for the measurements X_t at a given sensor location is $\mathcal{T} = [t_i - n\Delta t + 1, t_i], t \in \mathcal{T}$ where $n\Delta t$ represents the duration (length) of $\mathcal{T}$, and Δt is the temporal resolution of the X_t time series. Within each $\mathcal{T}$, a statistical indicator is computed and compared against a predefined threshold to detect potential outliers. Since some measurements may be missing or flagged as invalid from the previous QC checks, the computation is performed over $\mathcal{T}^* \subseteq \mathcal{T}$ which includes only the available (non-missing) data.

Table 1: *Overview of the statistical indicators used in the outlier detection test.*

Sensors	Statistical Indicators
Static $PM_{2.5}$	$Z_t = \frac{\lvert X_t - m_{\mathcal{T}^*} \rvert}{MAD_{\mathcal{T}^*}}$ $m_{\mathcal{T}^*} = median_{t \in \mathcal{T}^*}(X_t)$ $MAD_{\mathcal{T}^*} = median_{t \in \mathcal{T}^*}(X_t - m_{\mathcal{T}^*})$ $m_{\mathcal{T}^*}$: the median, $MAD_{\mathcal{T}^*}$: the median absolute deviation within $\mathcal{T}^*$
Noise Level	$Z_t = \frac{\left\lvert X_t - L_{eq_{\mathcal{T}^*}} \right\rvert}{SD_{\mathcal{T}^*}}$ $L_{eq_{\mathcal{T}^*}} = 10 \log_{10}\left(\frac{1}{n_{\mathcal{T}^*}} \sum_{t \in \mathcal{T}^*} 10^{X_t/10}\right)$ $SD_{\mathcal{T}^*} = \sqrt{\frac{1}{n_{\mathcal{T}^*} - 1} \sum_{t \in \mathcal{T}^*} \left(X_t - L_{eq_{\mathcal{T}^*}}\right)^2}$ $L_{eq_{\mathcal{T}^*}}$: the energetic mean, $SD_{\mathcal{T}^*}$: the standard deviation within $\mathcal{T}^*$

The quality flag for the outlier detection test, f_t^{outlier}, is defined as:

$$f_t^{\text{outlier}} = \begin{cases} 0 & \text{Insufficient data: } n_{\mathcal{T}^*} < n_{min}, \\ 1 & \text{Outlier: } Z_t \geq q \text{ and } n_{\mathcal{T}^*} \geq n_{min}, \\ 2 & \text{Non-Outlier: } Z_t < q \text{ and } n_{\mathcal{T}^*} \geq n_{min}. \end{cases}$$

where:

- $n_{\mathcal{T}^*}$: number of available (non-missing) values within $\mathcal{T}$.
- n_{min}: minimum number of non-missing measurements required to evaluate the test, typically given as $p \times n\Delta t$, where p is the minimum percentage of data completeness.
- q: threshold for detection the abnormal values

Measurements for which Z_t surpasses a pre-defined threshold are initially classified as outliers (Flag = '1'), whereas those below the threshold are considered valid (Flag = '2'). When the data

coverage within the window is insufficient to ensure a reliable statistical assessment, the corresponding measurements are flagged with '0', indicating that the test could not be evaluated.

The specific values of the threshold q and the temporal window length are adapted according to the characteristics of the noise and $PM_{2.5}$ datasets. More specifically, for static $PM_{2.5}$ sensors with 1 h resolution, the default configuration employs a 360 h rolling window, a minimum data coverage of 90 h, and a detection threshold of $q = 10$. A detailed description of the derivation and validation of these parameters is presented in Hassani et al., 2025.

In the case of noise sensor data, the parameters are defined analogously but adjusted to reflect the typically higher variability and finer temporal structure of acoustic measurements. On the other hand, the noise level follows in general a normal distribution and therefore a z-score statistic is computed. A typical threshold of $q = 5$ is recommended. For noise time series with a 1 min temporal resolution, the default temporal window length is set 120 min, with a minimum data coverage of 90 min to evaluate the test.

Once temporal outliers are identified in the sensors' time series, a spatial consistency check is performed to determine whether these anomalies are also observed at nearby monitoring locations. If neighboring sensors within a defined spatial radius exhibit similar anomalies at the same timestamps, the corresponding measurements are reclassified as non-outliers (Flag updated from '1' to '2'), as they may reflect a specific pollution event. Conversely, if no comparable anomalies are detected among neighboring sensors, the measurements remain flagged as outliers, indicating potential sensor-specific anomalies.

To derive meaningful spatial comparisons, the analysis is limited to neighboring sensors located within maximum geographical distances. more specifically:

- Static $PM_{2.5}$ sensors:

  If one or more neighboring sensors within a 3 km radius, or at least two neighboring sensors within a 30 km radius, are also flagged as outliers, then the measurement from the sensor of interest is not considered an outlier.

- Noise sensors:

  If the measurements from one or more neighboring sensors within a short proximity are also flagged as outliers, then the measurement from the sensor of interest is not considered an outlier.

The treatment of specific measurements as outliers depends on the underlying research question. For example, in analyses targeting hyperlocal phenomena such as residential wood burning, observations that may be considered outliers in other contexts can represent relevant local signals, highlighting the context-dependence of the outlier detection process.

3.1.5 QC 4: Spatial Correlation

In this QC step, the spatial coherence of LCS measurements is assessed in relative terms through the Pearson's R correlation coefficient between a target sensor and neighboring sensors or reference stations. The calculation of spatial correlation is again context-dependent and should be adapted to the spatial scale of each use case. For example, the parameterization of the spatial correlation test needs to be adjusted accordingly when the spatial resolution of interest is at the continental, city, neighborhood, or street level.

The *Spatial Correlation* test for the static and mobile $PM_{2.5}$ sensors is largely similar, with the main difference being the pre-processing of sensor measurements prior to calculating the correlation coefficient. Since mobile sensors often record data at sub-hourly intervals, their measurements are first aggregated to an hourly resolution using the median operator to reduce high-frequency variability. For static sensors, this step is not necessary, as the entire FILTER methodology has been designed for hourly $PM_{2.5}$ data, which is the typical temporal resolution provided by sensor suppliers. However, if static sensors record $PM_{2.5}$ data at sub-hourly intervals, the measurements are averaged to an hourly resolution during the initial stage before applying the QC methodology.

Correlation is calculated over a 30-day rolling window. To reduce computational cost, the test is performed once per day at 23:00 UTC, and the resulting quality flag is applied to all measurement hours (0–23) for that day. However, if sufficient computational resources are available, the temporal windows can also be initiated every hour. Based on this, the rolling window is $\mathcal{T} = \{t \mid t \in [t_n - (T-1) \cdot \Delta t,\ t_n]\}$, where t_n is reference time at the end (i.e., 23:00) of day n, T is the duration of the temporal window (30 days x 24 h/day), and Δt is the temporal resolution of the sensor time series.

For the correlation coefficient calculations, only valid measurements are used, meaning that data flagged as invalid by previous QC tests are replaced with NAs. The same QC procedure is also applied to neighboring sensors before their inclusion in the calculation. The selection of neighboring sensors for each temporal window differs between static and mobile sensors. For static LCSs, the spatial network is fixed, and only neighboring sensors within a 30 km radius are considered. The network of mobile sensors is not fixed, and neighbor distances vary depending

on the region of interest; therefore, a region-specific threshold is recommended instead of a fixed-radius neighbor selection. For example, in city-scale analyses, neighbors may be defined within a 10 km radius around the city center.

During each temporal window $\mathcal{T}$, the Pearson correlation coefficient between the measurements at locations i and j is calculated as:

$$R_{ij}(\mathcal{T}) = \frac{\sum_{t\in\mathcal{T}^*}(X_{it}-\bar{X}_i)\left(X_{jt}-\bar{X}_j\right)}{\sqrt{\sum_{t\in\mathcal{T}^*}(X_{it}-\bar{X}_i)^2}\sqrt{\sum_{t\in\mathcal{T}^*}\left(X_{jt}-\bar{X}_j\right)^2}}$$

where $\mathcal{T}^*$ represents the timestamps with available (non-missing) paired measurements, and $\bar{X}_i$ and $\bar{X}_j$ are the corresponding averages within the window.

For every pair of locations, the separation distance d_{ij} is also computed. All $R_{i,j}$ are then compared against an expected correlation threshold, $R_{th}(d_{ij}, s)$, which depends on both the separation distance and the season. The expected thresholds are obtained from a pre-computed look-up table derived from reference stations located in 27 EU member states, as well as UK, Norway and Switzerland. For each calendar season, DJF (December–January–February), MAM (March–April–May), JJA (June–July–August), and SON (September–October–November), as described in (Hassani et al., 2025), a third-degree polynomial function is extracted relating the median R to the separation distance between reference stations (Figure 2). These functions provide the baseline correlation expected for a given separation and seasonal context. The correlation threshold (R_{th}) is then estimated based on the scale of the region, with a conservative relaxation factor of 10–20% to avoid overly strict rejection of valid data.

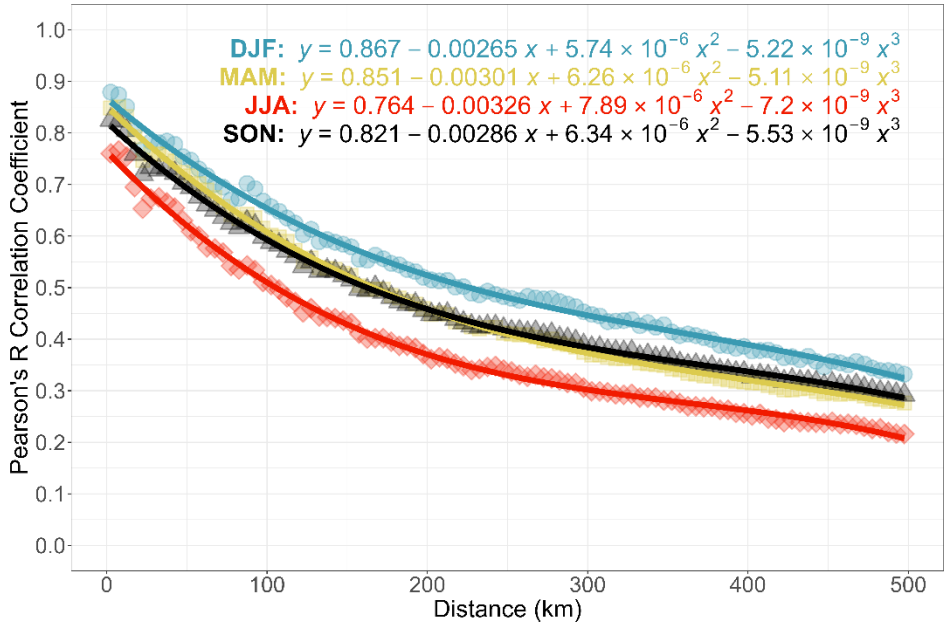

Figure 2: *Pearson Correlation Coefficient as a function of separation distance between reference stations' $PM_{2.5}$ data. The correlation calculated based on monthly paired reference observations from 2005 to 2023, including stations located in the 27 European Union Member States, as well as the UK, Norway, and Switzerland. DJF: December, January, February, MAM: March, April, May, JJA: June, July, August, SON: September, October, November.*

For each temporal window, neighboring sensors are identified based on two main conditions:

- Data completeness: A minimum number of valid paired measurements, n_{min}, is required within the window to ensure statistical reliability. For both the static and mobile sensors, data coverage of $n_{min} = 90\ h$ is suggested for a 30-day temporal window. In the case of shorter windows, the required data coverage may be adjusted accordingly.
- Spatial proximity: Only sensors located within defined maximum distances (radius ρ) are considered as potential neighbors.
    - Static sensors: Local-specific maximum distance of 3 km (inner radius ρ_{local}) and region-specific maximum distance of 30 km (outer radius $\rho_{regional}$),
    - Mobile sensors: Region-specific maximum distance ρ. For city-scale applications, ρ can be set a 10 km radius around the city center.

As in the previous QC checks, the *Spatial Correlation* test assign quality flags on the sensor measurements. The expressions for corresponding quality flag are shown in Table 2.

Table 2: *Quality flags assigned in the Spatial Correlation test for static and mobile $PM_{2.5}$ sensors.*

Sensors	Statistical Indicators

Static $PM_{2.5}$	$f_i^{\text{correlation}}(n) = \begin{cases} 0 & \text{Insufficient data: } n_{\mathcal{T}^*} < n_{min}, \\ 1 & i \text{ is isolated: } N_j(\rho_{\text{local}}) < 1 \text{ or } N_j(\rho_{\text{regional}}) < n_{\text{min, neigh}}, \\ 2 & \text{Non-correlated: } N_{R_{ij} > R_{\exp}(d_{ij},s)} < 1 \quad d_{ij} < \rho_{\text{local}} \\ & \quad or \\ & N_{R_{ij} > R_{\exp}(d_{ij},s)} < n_{\text{min, neigh}} \quad d_{ij} < \rho_{\text{regional}} \\ 3 & \text{Correlation: } N_{R_{ij} > R_{\exp}(d_{ij},s)} \geq 1 \quad d_{ij} < \rho_{\text{local}} \\ & \quad or \\ & N_{R_{ij} > R_{\exp}(d_{ij},s)} \geq n_{\text{min, neigh}} \quad d_{ij} < \rho_{\text{regional}} \end{cases}$, $n_{\text{min, neigh}}$: the minimum number of neighbors
Mobile $PM_{2.5}$	$f_i^{\text{correlation}}(n) = \begin{cases} 0 & \text{Insufficient data: } n_{\mathcal{T}^*} < n_{\text{min}}, \\ 1 & i \text{ is isolated: } N_j(\rho) < 1, \\ 2 & \text{Non-correlated: } N_{R_{ij} > R_{\exp}(d_{ij},s)} < 1 \\ 3 & \text{Correlation: } N_{R_{ij} > R_{\exp}(d_{ij},s)} \geq 1 \end{cases}$

The *Spatial Correlation* test assigns quality flags to sensor measurements based on data availability, isolation, and correlation with neighboring sensors. Measurements are flagged as '0' when the number of valid measurements within a window does not meet the data completeness criteria, making the calculation of correlation coefficient with neighboring sensors unreliable. A flag of '1' is assigned if the sensor is isolated, meaning that no valid neighbors are available within the defined geographical range. Measurements are flagged as '2' (non-correlated) or '3' (correlated) depending on the number of cases in which the calculated correlation coefficient exceeds the expected correlation threshold. The radius of influence—region-specific for mobile sensors, and both local- and region-specific for static sensors—also plays a role in assigning the appropriate quality flag (see Table 2).

3.1.6 QC 5: Similarity Test

The *Similarity Test* evaluates whether the sensor's values are consistent with expected spatial patterns for the region, focusing on absolute values rather than correlation. Similarity is assessed by comparing the LCSs measurements with nearby reference or official station data, aiming to identify observations that deviate substantially from the expected environmental patterns. The algorithms used to assess similarity differ across the static and mobile $PM_{2.5}$ sensors, and they are discussed separately. Most parts of this test employ rolling windows, consistent with the other statistical tests described earlier. For each sensor, the temporal window for the measurements X_t is defined as $\mathcal{T} = [t_i - n\Delta t + 1, t_i]$, $t \in \mathcal{T}$, where $n\Delta t$ represents the duration (length) of the window, and Δt is the temporal resolution of the X_t time series.

3.1.6.1 Static $PM_{2.5}$ sensors

The target statistic for assessing the similarity between LCSs and the neighboring reference sites up to a specified radius is the Euclidean distance, D. For static LCSs, the spatial network is fixed, and only reference stations within a 30 km radius are considered. The Euclidean distance, D_{ij}, is calculated between the measurements (X_{it}) at location i and its neighboring reference stations j (with Y_{jt}), along with the separation distance d_{ij}:

$$D_{\mathcal{T}}(i,j) = \sqrt{\sum_{t \in \mathcal{T}_n^*} \left(X_{it} - Y_{jt}\right)^2}$$

with $\mathcal{T}^*$ the available (non-missing) measurements within the temporal window.

As in the *Spatial Correlation* test, D_{ij} is compared against an expected similarity threshold, $D_{\text{th}}(d_{ij}, s)$, which varies with both the inter-station distance and the calendar season. The seasonal thresholds are obtained through fitting third-degree polynomial functions that relate the median D to the geographical distance between reference stations (Figure 3). The expected similarity threshold is determined according to the spatial scale of the region, and a relaxation factor of 10–20% is applied to prevent overly strict exclusion of valid measurements.

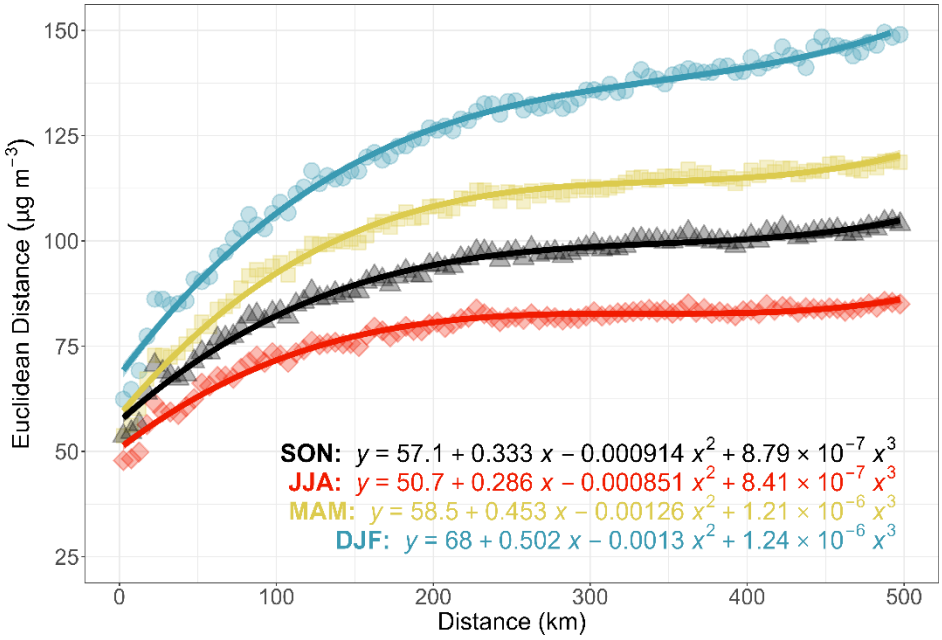

Figure 3: *Euclidean Distance (in µg m^{-3}) as a function of separation distance between reference stations' $PM_{2.5}$ data. The correlation calculated based on weekly paired reference observations from 2005 to 2023, including stations located in the 27 European Union Member States, as well as the UK, Norway, and Switzerland. DJF: December, January, February, MAM: March, April, May, JJA: June, July, August, SON: September, October, November.*

The quality flag for the site i at each timestamp t is then defined as:

$$f_{it}^{\text{similarity}} = \begin{cases} 0 & \text{Insufficient data: } n_{\mathcal{T}^*} < n_{min}, \\ 1 & \text{Isolated } i\text{: } N_j(\rho) < 1, \\ 2 & \text{Non-Similar with the nearby reference stations } j\text{: } N_{D_{\mathcal{T}}(i,j) < D_{\text{th}}(d_{ij},s)} < 1, \\ 3 & \text{Similar with the nearby reference stations } j\text{: } N_{D_{\mathcal{T}}(i,j) < D_{\text{th}}(d_{ij},s)} \geq 1. \end{cases}$$

The key criterion for interpreting the quality flag is the number of pairwise D_{ij} values that exceed the expected threshold. If at least one sensor-reference pair meets this condition, the timestamp is assigned to a flag of '3', indicating spatial similarity. Conversely, if all D_{ij} fall below the threshold, the sensor at that timestamp is considered dissimilar to the nearest references, and the observation is flagged as '2'. As in the other spatial QC tests, flags of '0' or '1' indicate insufficient data coverage within the temporal window and an isolated sensor, respectively. A window length of 168 h and a minimum data requirement of 90 h are recommended for this test.

3.1.6.2 Mobile $PM_{2.5}$ sensors

The *Similarity Test* for mobile sensors can be implemented in two different stages, namely the '*Sensor-Reference Similarity Test*' and '*Chain Test*', depending on the availability of sensor and reference data. In the first stage, the mobile sensor's measurements are compared with those from the nearest reference site using rolling temporal windows. The second stage is applied when higher-accuracy nodes (either reference stations or calibrated sensors) are available; in this case, the test evaluates rendezvous events, e.g., periods where the mobile sensor and a higher-accuracy node provide temporally coincident measurements.

3.1.6.2.1 Sensor – Reference Similarity Test for mobile sensors

For mobile sensors, measurements are first aggregated to hourly values using the median operator, as these sensors typically record data at a finer temporal resolution. The hourly measurements are then temporally aligned with the corresponding official station data, producing paired datasets for direct comparison. The statistical index within each rolling window $\mathcal{T}$ is computed in two steps: first, the absolute difference between the sensor (X_{it}) and reference (Y_{it}) measurements is calculated; second, the median absolute deviation is computed within $\mathcal{T}$.

$$r_{it} = |X_{it} - Y_{it}|$$

$$MAD_{it} = median_{t \in \mathcal{T}^*}(r_t - m_{\mathcal{T}^*}), \quad m_{\mathcal{T}^*} = median_{t \in \mathcal{T}^*}(r_{it})$$

QC flags are assigned according to the following criteria. Observations within windows that do not meet the required data completeness are flagged as insufficient data (Flag = '0'). Sensors

without nearby reference sites in the vicinity are treated as isolated and assigned Flag = '1'. Differences that exceed the predefined similarity threshold are flagged as not similar (Flag = "2"), whereas measurements that fall below the acceptable threshold value are considered valid and assigned Flag = '3'. This framework ensures that only data points showing reasonable agreement with reference measurements are retained as highly reliable.

The quality flag categories described above can be summarized as follows:

$$f_{it}^{\text{similarity}} = \begin{cases} 0 & \text{Insufficient data: } N_{\mathcal{T}_n^*} < n_{\min}, \\ 1 & \text{Isolated } i\text{: } N_j(\rho) < 1, \\ 2 & \text{Non-Similar with the nearby reference station } j\text{: } MAD_{it} > q, \\ 3 & \text{Similar with the nearby reference station } j\text{: } MAD_{it} > q. \end{cases}$$

Here, the window length n, the data completeness threshold $n_{\min}$, and the threshold q to compare sensor-reference differences are treated as tuning parameters and can be adjusted based on the characteristics of the mobile sensor monitoring campaign. However, suggested values are 72 h window length, $n_{\min} = 24$ h, and a threshold of $q = 5$ (in µg m⁻³).

3.1.6.2.2 Chain Test

This QC procedure is recommended only when the sensor network includes at least one node with higher accuracy (e.g., a regulatory reference station or a calibrated mobile device) than the sensor of interest. These high-accuracy nodes act as anchors against which the performance of LCSs can be evaluated.

Interference is defined as the situation in which two or more sensors are in close spatial proximity and temporal coincidence. In this case, interference is detected for a sensor i and a high accuracy node j if:

- their separation distance (d) is lower than a pre-defined distance (d_{max}), and
- their measurements occur within a maximum time difference, $|t_i - t_j| < \Delta t_{max}$.

During such overlaps, the sensors are assumed to measure comparable air masses. Using distance calculations, all pairs of interfering sensors are identified. The evaluation is performed in fixed (non-rolling) temporal windows $\mathcal{L}$. The window length $n_{\mathcal{L}}$ can range from one week to several months, depending on sensors' mobility and interferences. A data completeness threshold $n_{\min}$ is also applied to ensure robust comparisons within each window. All measurements within a window receive the same QC flag based on the evaluation result.

If a sensor directly interferes with a high-accuracy node, its measurements are compared to those of the reference. In cases where a sensor does not directly overlap with a high-accuracy node, it can still be evaluated indirectly through another sensor that has been validated against a reference. For example, if Sensor A overlaps with Sensor B, and Sensor B has been validated against a reference, then Sensor A's measurements can be compared to those of Sensor B using the same approach.

In both direct and indirect evaluations, the normalized root mean square deviation ($nRMSD$) is calculated between the sensor (X_{it}) and the comparison node (reference or calibrated sensor) measurements (Y_{jt}) within each window $\mathcal{L}$ to quantify agreement:

$$RMSD_{ij}(\mathcal{L}) = \sqrt{\frac{1}{n_{\mathcal{L}^*}} \sum_{t \in \mathcal{L}^*} \left(X_{it} - Y_{jt}\right)^2}, \quad \overline{Y_j(\mathcal{L})} = \frac{1}{n_{\mathcal{L}^*}} \sum_{t \in \mathcal{L}^*} Y_{jt}$$

$$nRMSD_{ij}(\mathcal{L}) = \frac{RMSD_{ij}(\mathcal{L})}{\overline{Y_j(\mathcal{L})}}$$

with $n_{\mathcal{L}^*}$ the total number of temporal coincidences within $\mathcal{L}$.

The calculated ratio determines the QC flag, $f_i^{\text{chain}}(\mathcal{L})$, which is representative for all measurements within the temporal window, $f_{it}^{\text{chain}} = f_i^{\text{chain}}(\mathcal{L})$. The quality flag for the *Chain Test*, is then formulated as follows,

$$f_i^{\text{chain}}(\mathcal{L}) == \begin{cases} 0 & \text{Insufficient data: } N_{\mathcal{L}^*} < n_{\min}, \\ 1 & \text{Indirect Evaluation (validated sensor) - low agreement:} \\ & \text{Direct Evaluation (high-accuracy node) - low agreement: } nRMSD_{ij}(\mathcal{L}) \geq \tau, \\ 2 & \text{Indirect Evaluation (validated sensor) - moderate agreement: } nRMSD_{ij}(\mathcal{L}) < \tau, \\ 3 & \text{Direct Evaluation (high-accuracy node) - High agreement: } nRMSD_{ij}(\mathcal{L}) < \tau. \end{cases}$$

with τ the threshold for the test and $n_{\min}$ the minimum required number of paired observations within $\mathcal{L}$.

Several parameters need to be adjusted when implementing this test. The suitability of the *Chain test* parameters depends on both the properties of the sensor data (e.g., temporal resolution) and the characteristics of the measurement campaign (e.g., availability of high-accuracy reference stations, frequency of sensor coincidence events, etc.). For sensors with 1 min data resolution, the following parameters can be used as starting points: d_{max}~$250\,m$, $\Delta t_{max} = 3\,min$, $n_{\mathcal{L}} = 1\,week - several\,months$, $n_{\min} = 30$, and $\tau = 0.68\ (68\%)$.

3.2 Correction procedure

Two processing levels are defined, namely 'Raw' and 'Corrected', and the QC procedure can be evaluated at either level. At the 'Corrected' processing level, LCS observations are assigned quality flags using the same QC framework described in Section 3.1, with the key distinction that the quality control tests are applied to the corrected sensor measurements rather than the raw data.

LCSs are typically factory-calibrated; however, their performance under real-world deployment conditions often deviates from true ambient levels (Kang et al., 2022). For particulate matter, such discrepancies largely arise because particle composition, emission sources, concentration levels, and meteorological conditions in the field differ from the controlled laboratory conditions used for calibration (Kuula et al., 2020). The standard and most reliable practice is to conduct co-location experiments in which sensors are deployed alongside reference-grade instruments. Sensor measurements are then recalibrated to better reflect local conditions (Levy Zamora et al., 2023). More recent studies (Bagkis et al., 2022; Considine et al., 2021; Hofman et al., 2022) have focused on:

a) Real-time remote/distant correction: sensor measurements are adjusted using nearby reference stations, or spatially interpolated correction factors generated by reference-sensor pairs (Wesseling et al., 2024).

b) In-situ calibration: calibration models are developed using one sensor batch and are then assumed to be transferable and applied across various sensors in a network.

As most low-cost sensor (LCS) networks are operated by citizens, several recurring challenges arise in sensor deployment and long-term operation:

- Sensors are often handled by non-specialists, which can lead to deviations from recommended maintenance and operational protocols.
- Repeated recalibration is required throughout the sensor lifetime to maintain measurement accuracy and compensate for sensor drift.
- Significant inter-sensor variability within a sensor network limits the transferability of calibration models, particularly across different manufacturers, sensor types, and environmental settings.
- Environmental conditions (e.g., ambient humidity and air temperature), can strongly affect sensor response.

- Advanced machine learning (ML) and artificial intelligence (AI) based calibration models do not consistently outperform simpler statistical approaches (e.g., simple and multiple linear regression), while introducing additional computational cost and methodological complexity without yielding substantial improvements in accuracy and overall performance.

In FILTER, correction for static $PM_{2.5}$ sensors is performed using nearby official monitoring stations in a dynamic manner, employing rolling temporal windows to reflect the varying conditions of the deployment environments. It is worth noting that the term 'calibration' is avoided here, as the correction procedure does not involve co-location experiments. Furthermore, the correction is sensor-specific, meaning that for each sensor, a tailored statistical model is developed to adjust its measurements. This approach is adopted because sensor networks may include sensor units from different manufacturers, exhibiting varying levels of accuracy and performance when compared with nearby reference instruments.

As previously noted, a dynamic distance-based correction algorithm is proposed for the adjustment of $PM_{2.5}$ measurements. The LCSs data is corrected using a 30-day rolling temporal window. To reduce computational cost, the correction is performed once per day at 23:00 UTC, and the derived correction function is then applied to all hourly measurements (0–23) of that specific day. The rolling temporal window is $\mathcal{T} = \{t \mid t \in [t_n - (T-1) \cdot \Delta t, t_n]\}$, where t_n is reference time at the end (i.e., 23:00) of day n, T is the duration of the temporal window (30 days x 24 h/day), and Δt is the temporal resolution of the sensor time series. A correction efficiency flag is also generated and assigned to all corresponding measurement hours. If sufficient computational resources are available, the temporal window can alternatively be updated on an hourly basis. Moreover, depending on data availability and sensor temporal resolution, shorter rolling windows may also be employed. Correction is sensor-type agnostic, requiring no prior information on the sensor type, and it is evaluated independently for each individual sensor.

The correction includes four (4) steps:

1. Extraction of nearest meteorological and reference sites: The selection of neighboring reference and meteorological sites depends on data availability within the temporal window and the separation distance from each sensor i. For meteorological stations, the maximum radius of influence is set to 30 km, whereas for reference air-quality stations it is defined as a function of both distance and season and is extracted from the third-order polynomial relationships shown in Figure 3. Two observation nodes are considered similar in terms of $PM_{2.5}$ concentration levels if their Euclidean distance D remains within 10% of

the estimated value at 0 km. Specifically, if for a given season D intersects the vertical axis at α μg m^{-3}, the corresponding distance to select reference sites and apply the correction is determined by solving the third-order polynomial for a concentration of 1.1α μg m^{-3} (after applying the relaxation parameter). Using this approach, the resulting radii of influence for the four calendar seasons are approximately 11.5 km for DJF (December–January–February), 12.7 km for MAM (March–April–May), 20 km for JJA (June–July–August), and 17 km for SON (September–October–November).

2. Reference and meteorological data: In the case of multiple reference and meteorological stations fall within the pre-defined distances described above, spatially aggregated time series are derived using the Inverse Distance Weighted (IDW) method. If Y_{jt} denote the measurements of a specific parameter of reference or meteorological stations j at $t \in \mathcal{T}$, then the weighted time series Y_{wt} is calculated as:

$$Y_{wt} = \frac{\sum_j w_{ij} \cdot Y_{jt}}{\sum_j w_{ij}} \quad w_{ij} = \begin{cases} \left(\frac{\rho - d_{ij}}{\rho \cdot d_{ij}}\right)^2 & d_{ij} < \rho, \\ 0 & d_{ij} \geq \rho. \end{cases}$$

where d_{ij} is the separation distance between i and j, ρ is the radius of influence, and w_{ij} is the spatial weight calculated through the modified Shepard method.

3. Extraction of low-variability periods: Only periods exhibiting minimal spatial variability are selected for correction, as sensors and reference stations are not co-located and are therefore expected to show more similar spatial patterns during low-variability conditions. In urban environments, $PM_{2.5}$ typically follows a distinct diurnal cycle, with lower spatial variability at night and higher variability during day rush hours due to varying emission sources. In order to identify low-variability periods, the reference measurements within the temporal window $\mathcal{T}$ are grouped by hour and the corresponding statistical variance is calculated for each hourly group. Then, a subset of hours with the lowest variability (e.g., the lowest 10) is then selected for developing the correction model.

4. Development of the correction algorithm: The statistical correction model is developed by applying Multiple Linear Regression (MLR) based on the low-variability periods described above. The MLR model includes reference measurements (Y_{it}) as the dependent variable and the sensor data (X_{it}) together with simple meteorological parameters ($M_{n,it}$), such as air temperature and relative humidity, as independent predictors. The model is expressed as:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \sum_{n \in M_{n,it}} \beta_n M_{n,it} + \epsilon_t$$

where $\beta_0, \beta_1, \ldots, \beta_n$ are the regression coefficients and ϵ_t represents the random error term.

This model aims to adjust the sensor data by correcting the sensor gain and accounting for local meteorological variability. In this way, the model corrects the general tendency of the sensor measurements without forcing them to be identical to the reference values, thereby preserving the local information provided by the LCSs. To ensure robustness, two additional steps are applied to extract the final MLR correction model.

1. Outlier detection to the model residuals (ϵ_t): A robust outlier detection method based on the median absolute deviation is applied to the model residuals. Observations that deviate considerably from the central tendency are classified as outliers and removed from subsequent analysis. Data points are considered outliers if:

   $$|\epsilon_t - median(\epsilon_t)| \geq 3 \times MAD_t, \quad MAD_t = median(|\epsilon_t - median(\epsilon_t)|)$$

   Only the remaining, reliable timestamps are retained for building the correction model.

2. Model refinement using significant predictors: The regression model is updated by retaining only those predictors that have statistically significant influence on reference measurements. This is achieved by testing the significance of the regression coefficients ($\beta_0, \beta_1, \ldots, \beta_n$) at a X% confidence level (e.g., 95%) using the statistical Student's t test. Since the correction is performed using rolling temporal windows, the included predictors may differ from one window to another, as different combinations of input parameters may be statistically significant.

Finally, corrected values ($\hat{Y}_{it}$) and the associated X% confidence intervals are computed as:

$$\hat{Y}_{it} \pm t_{\alpha/2} \times SE(\hat{Y}_{it})$$

with $SE(\hat{Y}_{it})$ the standard errors of the prediction and $t_{\alpha/2}$ the critical t-value at $\alpha = 1 - X/100$ level of significance.

The effectiveness of the correction model to adjusting the LCSs measurements is evaluated in terms of the root mean squared deviation ($RMSD$) between the initial and the corrected data

relative to the reference observations. Based on the behavior of the $RMSD$, a correction flag, $f_i^{\text{correction}}$, is assigned for each temporal window and timestamp, respectively:

$$RMSD_{i,\mathcal{T}}^{Raw} = \sqrt{\frac{1}{n_{\mathcal{T}^*}} \sum_{t \in T^*} (Y_{it} - X_{it})^2}, \quad RMSD_{i,\mathcal{T}}^{Cor} = \sqrt{\frac{1}{n_{\mathcal{T}^*}} \sum_{t \in T^*} \left(Y_{it} - \hat{Y}_{it}\right)^2}$$

$$f_{it}^{\text{correction}} = \begin{cases} 0 & \text{Insufficient data: } N_{\mathcal{T}_n^*} < n_{\min}, \\ 1 & \text{Isolated } i\text{: } N_j(\rho) < 1, \\ 2 & \text{Correction does not improve the initial data: } RMSD_{i,\mathcal{T}}^{Raw} \leq RMSD_{i,\mathcal{T}}^{Cor}, \\ 3 & \text{Correction improves the initial data: } RMSD_{i,\mathcal{T}}^{Raw} > RMSD_{i,\mathcal{T}}^{Cor}. \end{cases}$$

where $\mathcal{T}^* \subseteq \mathcal{T}$ represents the available (non-missing) data within $\mathcal{T}$.

3.3 Overall Quality and Further Recommendations

This section provides an overall discussion of the data quality outcomes from the QC framework and offers recommendations for how quality-controlled datasets should be used in different application contexts. It summarizes the key findings and presents practical guidance for interpreting the LCSs measurements.

Two main observations emerge from the application of the FILTER framework to **the static $PM_{2.5}$ sensor data**.

1. A considerable data loss is expected when progression from the *Spatial Similarity* step.

   This reduction is most pronounced in the raw data, as larger biases have not yet been corrected. The primary reason for this data loss is the limited availability of nearby official monitoring stations. The extent of reduction varies across study areas: regions with dense networks of sensors and reference stations retain far more usable observations, whereas sparsely monitored areas experience substantial decreases in spatial coverage.

2. Corrected data that has passed the *Spatial Correlation* step generally remains consistent and reliable.

   In many cases, this level of processing is sufficient even without applying the final *Spatial Similarity* check.

This creates a trade-off. While more stringent QC steps improve confidence in the data, an essential consideration for scientific analyses or policy-relevant applications, they also reduce data availability and may limit the usefulness of low-cost sensor measurements for certain real-world uses.

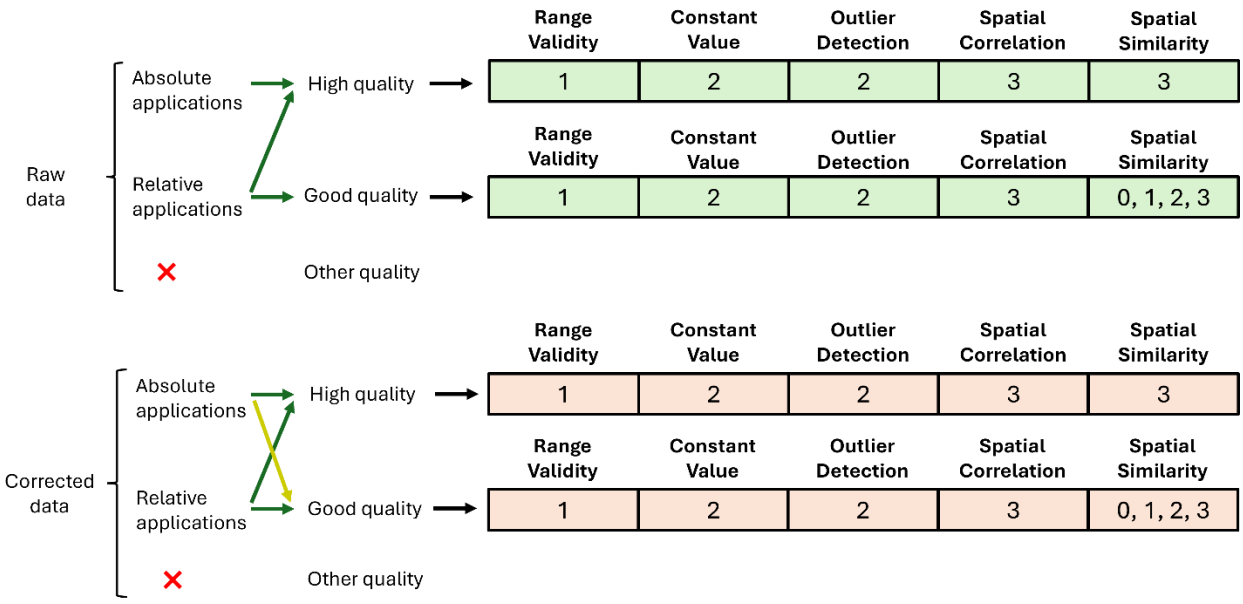

***Figure** 4: Recommended data quality levels and their intended applications for static $PM_{2.5}$ sensors.*

Therefore, two major categories are distinguished (Figure 4):

- <u>Relative Applications</u>

  These focus on variations and trends in $PM_{2.5}$ rather than exact concentrations. Typical examples include tracking short-term fluctuations, evaluating the effectiveness of pollution control measures, analyzing diurnal cycles, or supporting public engagement and awareness initiatives.

- <u>Absolute Applications</u>

  Applications such as regulatory compliance, health risk assessments, AQI reporting, and emission or exposure modelling, require accurate concentration levels. In these contexts, any substantial deviation from true $PM_{2.5}$ levels can lead to incorrect conclusions or policy decisions.

In the case of broad geographic areas (e.g., citywide, regional, or national), the most robust option is to rely on corrected datasets that have passed the full QC chain, including the *Spatial Similarity* check. These datasets provide the highest level of confidence and are suitable for both relative pattern analyses and absolute applications.

In contrast, for smaller-scale applications where the network of nearby sensors is sparse, it may be more practical to use corrected data processed only up to the *Spatial Correlation* step. This

approach preserves a larger number of observations and is generally sufficient for studies focused on relative differences or situations where absolute requirements are less strict.

The same principles are held when working with sensor data at the 'Raw' processing level. Large-area studies and any application that depends on accurate concentration levels should restrict their analyses to raw data that has passed the *Spatial Similarity* step. For applications focused mainly on identifying relative variations, raw data confirming the *Spatial Correlation* requirements can still provide reliable qualitative information.

Based on these distinctions, the individual QC levels can be aggregated into a single plausibility level as.

1. **High Quality** – Raw or Corrected data that passed the *Spatial Similarity* step.
2. **Good Quality** – Data that passed the *Spatial Correlation* step.
3. **Other Quality** – Timestamps with insufficient information to apply the QC steps result in uncertain data quality and should generally be excluded from applications that depend on absolute concentrations. Although the QC framework aims to reduce data loss by lowering the minimum data requirements for each step, gaps in sensor operation still cause many timestamps to fall into this category.

Wearables and mobile sensors operate under dynamic environmental and operational conditions, which often result in higher variability and intermittent data coverage compared to static sensors. Similar conclusions as in static $PM_{2.5}$ sensors could not be extracted at this stage and extensive validation is required. However, the application of the proposed QC pipeline is promoted, although QC thresholds may need to be tailored accordingly to reflect the use-case studies requirements and special characteristics, and also to account for the inherent variability of mobile measurements.

The first three QC steps namely QC 0: Spatiotemporal Attributes, QC 1: Range Test, and QC 2: Quality control for Constant or Flatlined Sensor Measurements, are mandatory and needs to be applied sequentially, since the provide a common baseline of data quality assurance that all datasets should undergo. By contrast, the *Spatial Correlation* and *Spatial Similarity* tests are optional. For applications focused on relative patterns, such as identifying pollution hotspots along travel routes or evaluating short-term exposure trends, mobile sensor data can still provide meaningful insights even if absolute concentration values are less precise. In this case, depending on the availability of mobile sensor data, QC could possibly undergo up to the *Spatial Correlation* step. However, when absolute $PM_{2.5}$ concentrations are required. For example, in exposure

modeling, personal exposure assessments, or urban-scale health risk studies, only measurements that pass the highest QC tier should be considered. Using lower-quality data in these contexts can introduce bias, potentially underestimating or overestimating pollutant exposure and affecting the accuracy of downstream analyses and decision-making.

Finally, **noise sensors** are influenced by environmental variability, localized noise sources and operational conditions, which can result in inconsistent readings, particularly for low-cost devices. For applications, such as identifying noise hotspots (e.g., mapping noisy streets, tracking noise changes over time), tracking diurnal or weekly noise variations, evaluating the impact of interventions, or supporting public awareness campaigns, data that pass the mandatory QC steps, namely QC 1: Range Test, QC 2: Quality control for Constant or Flatlined Sensor Measurements, and QC 3: Spatiotemporal outlier detection, are highly recommended.

4 MAPAIR

"MapAir" constitutes a set of algorithms developed to generate spatially continuous, high-resolution air quality maps from point-based observations provided by air quality sensors. These algorithms are designed to go from point measurements of air quality made by low-cost sensor networks to value-added continuous maps. They do so by integrating multiple sources of information in addition to sensor measurements, first and foremost model data from physical models, e.g. data from local scale models or from the Copernicus Atmosphere Monitoring Service (CAMS), satellite retrievals, and air quality monitoring stations. As such they are able to overcome the inherent limitations of point measurements and produce coherent gridded representations of air pollutant concentrations. By fusing these diverse datasets, the algorithms make it possible to move from scattered observations to spatially complete maps that reflect actual variability in air quality at multiple scales.

A key characteristic of the MapAir algorithms is their adaptability to different spatial resolutions depending on data availability and application needs. At the local scale they are mostly dependent on the spatial resolution of the input model information, but are typically run at ca. 100 m by 100 m spatial resolution. At the regional scale, they produce maps at a ca. 1 km resolution. This level of detail provides a compromise between wide spatial coverage, realistic representation of pollution patterns, as well as computational efficiency.

When sufficient numbers of sensor systems are available in a given area, the algorithms can exploit this information to provide finer spatial detail and more accurate spatial gradients in the resulting maps. In such cases, the algorithms combine the dense local sensor network with auxiliary information to generate targeted, high-detail maps. Within CitiObs the algorithms are demonstrated by producing maps selectively and only where data density and quality allow meaningful refinement.

The outputs produced by the MapAir algorithms are delivered as geospatial datasets in standard gridded formats such as GeoTIFF and NetCDF. They provide best-estimate $PM_{2.5}$ concentration fields that can be directly used for further analysis, including exposure assessment, air quality evaluation, or event-specific diagnostics. The algorithms are applied to generate large-scale sensor-based $PM_{2.5}$ maps across Europe at 1 km resolution for daily averages (and monthly or annual means), as well as ad hoc high-resolution maps for selected urban areas or special events where fine spatial detail is crucial.

The MapAir algorithms used within CitiObs are currently not used for operational deployment. Their purpose is to demonstrate how a structured set of algorithms can integrate multiple data streams to produce spatially consistent and decision-ready air quality information. These algorithms provide a possibility for taking point-based observations from sensor networks and exploiting their information for continuous spatial mapping of air quality.

4.1 Background on data fusion and data assimilation

Integrating low-cost sensor (LCS) data into air quality models using data fusion (Schneider et al., 2017) and assimilation methods (Lahoz and Schneider, 2014) has the potential to significantly improve the accuracy of air quality assessments, particularly at urban scales (Hassani et al., 2023; Schneider et al., 2023). Although LCS devices generally have lower accuracy compared to traditional regulatory-grade monitors (Vogt et al., 2021), their affordability and flexibility facilitate dense network deployments, providing extensive real-time data coverage. This denser coverage is useful in locally correcting the output from model-based estimates.

To maximize their utility, LCS data ideally should undergo careful calibration, robust quality control, and precise uncertainty estimation. This is for example now available through the FILTER framework (Hassani et al., 2025) developed as part of CitiObs. Integrating LCS data with model output through data fusion or assimilation methods allows making use of the complementary strengths of observational and modelled data. Such integration not only adds value to sensor data by enabling robust spatial interpolation but also enhances model accuracy by directly constraining simulations with uncertainty-weighted observational evidence.

The integration of low-cost sensor (LCS) observations with air quality models and other auxiliary datasets is useful to improve our understanding of urban air quality at the local and regional scale. LCS networks provide dense spatial and temporal coverage but suffer from measurement uncertainties, calibration drift, and variable performance depending on environmental conditions. On the other hand, air quality models, including chemical transport and dispersion models, offer physically consistent, spatiotemporally continuous fields but are prone to systematic errors and insufficient resolution to fully capture local variability. Data fusion and data assimilation methods are therefore powerful approaches to combine these complementary strengths while mitigating their respective weaknesses.

Data fusion methods generally operate in a statistical post-processing mode, combining observational data and model outputs after model simulations have been completed. The simplest fusion techniques rely on spatial interpolation methods such as inverse distance weighting or

nearest neighbor algorithms to generate concentration fields between sensor locations. They are typically computationally inexpensive, however these methods can usually not fully account for complex transport patterns or nonstationary variability. There are more advanced statistical fusion approaches that use, for example, geostatistical techniques such as kriging, co-kriging, or kriging with external drift and these have been used to some extent for mapping air quality with low-cost sensing devices (Chiles and Delfiner, 2009; Gressent et al., 2020; Schneider et al., 2017). These methods exploit the spatial autocorrelation within the data and allow model outputs or other auxiliary datasets to serve as covariates. As such they typically generate more physically plausible interpolations and in addition they can provide estimates of uncertainty. In addition to geostatistics, there is land use regression (LUR), which is another widely applied strategy for data fusion. LUR methods build statistical models to relate pollutant concentrations to predictors such as land cover/land-use, emission sources, traffic, and meteorology. LUR models can be trained on LCS data and then used to predict concentrations at unmonitored locations and this has been demonstrated on various oocasions (Adams et al., 2020; Coker et al., 2021; Jain et al., 2021; Lim et al., 2019; Weissert et al., 2020). LUR methods are interpretable and relatively easy to implement, however they are limited in their ability to capture dynamic atmospheric processes. Machine learning techniques offer an even more flexible alternative. Relevant techniques include random forests, gradient boosting, or artificial neural networks (Shetty et al., 2024, 2025). These techniques can easily take care of nonlinear relationships as well as a wide variety of input sources, although they often require a substantial amount of input data. These sources could include sensor data, model outputs, land-use information, and satellite observations (Coker et al., 2021; Guo et al., 2022; Jain et al., 2021; Liang et al., 2023; Lim et al., 2019). ML-based models have in the past been shown to produce high-resolution air quality fields in complex environments. However, the models created by these techniques can be difficult to interpret (although interpretable machine learning methods such as SHAP can help) and require a careful and systematic approach to avoid overfitting. In practice, hybrid approaches combining interpolation, regression, and machine learning are increasingly common, as they enable more robust data fusion products that exploit complementary methodological strengths.

In contrast to fusion, **data assimilation methods** (Bouttier and Courtier, 2002; Fletcher, 2017; Kalnay, 2013; Lahoz and Schneider, 2014) are designed to integrate observations directly into the state of a running model, allowing real-time corrections to the model. Most data assimilation methods stem from decades of research in numerical weather prediction and oceanography. The simplest such approach can be considered optimal interpolation (OI) (Hassani et al., 2023; Mijling, 2020; Schneider et al., 2023). In OI, a weighted combination of the model forecast and the

observations is computed based on their error statistics (i.e their respective uncertainties). Although OI is relatively straightforward to implement, this method requires reliable estimates of both observation and model error covariances, which can be challenging to derive for urban-scale applications. Other data assimilation approaches are based on Kalman filtering. In particular, the ensemble Kalman filter (EnKF) (Evensen, 2003) is very useful when it is difficult to manually derive background error covariance fields as it represents uncertainties by maintaining an ensemble of model forecasts, and thus can dynamically estimate model error covariances. It then updates the model state when new observations become available. This approach is well suited to handle nonstationary error structures and to propagate observational information forward in time, though it can be computationally expensive. One example of using EnKF with LCS data can be found in Lopez-Restrepo et al. (2021).

Variational methods such as three- and four-dimensional variational assimilation (3D-Var and 4D-Var) (Fletcher, 2017) formulate assimilation as the minimization of a cost function that penalizes deviations from both the background model and the observations, weighted by their uncertainties. 4D-Var methods assimilate data over a time window, thereby adjusting the full model trajectory and improving temporal consistency. Although theoretically elegant, variational methods require adjoint or tangent-linear models, which can be complex to implement for high-resolution air quality models. Hybrid methods combining ensemble and variational formulations offer additional flexibility by using ensemble-derived covariances while retaining the optimization framework of variational schemes. Although primarily designed for surface observations or satellites, these systems can be adapted for LCS data, despite limited applications to date. Lopez-Ferber et al. (2022) used 3D-Var with simulated sensor observations.

These fusion and assimilation techniques are particularly valuable for local-scale applications because they enable the combination of spatially dense but uncertain LCS measurements with physically consistent model simulations. As such, they improve the spatial and temporal representativeness of air quality estimates, correct systematic biases in models, and compensate for sensor drift or calibration errors. These methods have high potential as a versatile tool for integrating various data sources, e.g. from low-cost sensors, regulatory monitors, outputs from dispersion models or chemical transport models, and satellite remote sensing observations, into coherent and high-resolution air quality fields. This integrated approach is more and more recognized as useful for supporting exposure assessments, regulatory applications, and local-scale policy interventions.

In CitiObs we have chosen to use Optimal Interpolation for combing sensor data with model information at the local/urban scale and an XGBoost-based machine learning approach for regional scale mapping of low cost sensor data.

As mentioned above, Optimal Interpolation is a comparatively simple assimilation technique, which statistically assigns observation weights based on spatial proximity and error covariance structures, and minimizes deviations between model predictions and measurements. Its flexibility permits application in both online (real-time updating) and offline (post-processing) modes without explicitly modifying internal model states. In CitiObs we use it in post-processing mode. In practice, OI has been effectively utilized for integrating LCS data with local-scale dispersion models (Hassani et al., 2023; e.g. Mijling, 2020; Schneider et al., 2023), and for this reason it has been chosen as the method of choice for local-scale applications in CitiObs.

For regional-scale applications we have adopted a data fusion framework based on machine learning. The S-MESH approach (Shetty et al., 2024, 2025) has been extended as part of the CitiObs project to be also able to integrate quality-controlled observations from large low-cost sensor networks such as those from sensor.community[1] and PurpleAir[2].

4.2 Local-scale air quality mapping with sensor networks using data fusion or data assimilation

4.2.1 Introduction and Background

In the following we provide the description of a basic algorithm for integrating observations from low-cost sensor networks with models for local-scale air quality mapping. It is based on the technique of Optimal Interpolation (Fletcher, 2017; Kalnay, 2013; Lahoz and Schneider, 2014; Mijling, 2020; Schneider et al., 2023).

4.2.2 OI fundamentals

Optimal Interpolation (OI) and geostatistical methods such as universal kriging or kriging with external drift are fundamentally related, both in mathematical formulation and application. These techniques aim to optimally combine model data and observations to produce improved estimates of spatial fields. Here we adopt the OI formulation as described by Kalnay (2013).

[1] https://sensor.community/en/

[2] https://www2.purpleair.com/

The objective of OI is to compute the so-called analysis field, denoted by the vector $\boldsymbol{x}_{\mathrm{a}}$, which represents the best linear unbiased estimate of the true state. The analysis is calculated by adjusting a prior (or "background") estimate with observational information, as expressed in the following equation:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{W}[\mathbf{y}_{\mathrm{o}} - H(\mathbf{x}_b)]$$

Here, $\mathbf{x}_b$ is the background state vector, typically derived from a numerical model; $\boldsymbol{y}_{\mathrm{o}}$ is the vector of observations obtained from a low-cost sensor (LCS) network; H is the observation operator, which maps the background field into the observational space. In our case, where both the background and observations share the same physical units, this operator is implemented via bilinear interpolation. The term $\boldsymbol{y}_{\mathrm{o}} - H(\mathbf{x}_b)$ is commonly referred to as the innovation vector, representing the discrepancy between observed and modeled values.

The matrix $\mathbf{W}$, often referred to as the gain matrix, determines the influence of each observation on the analysis and is defined as:

$$\mathbf{W} = \mathbf{B}\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1}$$

In this expression, $\mathbf{B}$ denotes the background error covariance matrix, which characterizes the spatial correlation structure and magnitude of uncertainties in the background field. The matrix $\boldsymbol{H}$ is the linearized form (or tangent linear approximation) of the observation operator H, and $\mathbf{R}$ represents the observation error covariance matrix. For simplicity, $\mathbf{R}$ is typically assumed to be diagonal, implying uncorrelated errors across different observation locations.

The uncertainty associated with the resulting analysis field is captured by the analysis error covariance matrix $\mathbf{P}_a$, calculated as:

$$\mathbf{P}_a = (\mathbf{I} - \mathbf{W}\mathbf{H})\mathbf{B}$$

where $\mathbf{I}$ is the identity matrix, ensuring that the structure of $\mathbf{B}$ is correctly modified to reflect the assimilation of observations.

4.3 Background error covariance

One of the most technically demanding aspects of implementing OI, particularly for high-resolution, urban-scale applications such as ours, lies in specifying the background error covariance matrix $\mathbf{B}$. In our approach, we construct $\mathbf{B}$ based on spatial autocorrelation characteristics of the model field, modulated by a distance-decay function. This formulation allows each observation to influence its surrounding area according to its representativity, effectively

defining a spatial representativity footprint for each sensor location. This footprint governs how the signal from an individual observation spreads across space during the assimilation process.

The background field can be any high-resolution dataset of air quality available for the given location. We have successfully used operational air quality forecasts generated by the EPISODE (Hamer et al., 2020) and uEMEP (Denby et al., 2020; Mu et al., 2022) models. Both models offer high-resolution estimates of urban air pollution levels and serve as a robust and physically grounded background field for assimilation. By integrating this model output with spatially distributed LCS observations through the OI framework, we can produce enhanced air quality maps that are both spatially resolved and observationally constrained.

A key technical and conceptual challenge in the practical implementation of Optimal Interpolation (OI), especially in high-resolution urban-scale air quality applications such as the one presented in this study, is the specification of the background error covariance matrix $\mathbf{B}$. This matrix plays a central role in determining how observational information is propagated spatially within the assimilation system and, consequently, how it modifies the background model fields. The definition of $\mathbf{B}$ must reflect the spatial correlation structure of errors in the model background, which are inherently influenced by the underlying physical processes, emission patterns, and urban morphology.

To address this challenge, we have developed and implemented a method for constructing $\mathbf{B}$ that captures the spatial structure of uncertainties within each hourly model field. This is achieved by quantifying the spatial autocorrelation of pollutant concentrations derived from the model and modulating it with a distance decay function. The decay function serves to gradually reduce the influence of an observation with increasing spatial separation, effectively modelling the loss of correlation over distance. The result is a set of location-specific covariance fields that reflect how observational information from each monitoring site is expected to influence the surrounding model grid. Put differently, the method weights observations not just by geometric distance, but by the difference in modeled background concentration. This ensures that a measurement in a "clean" park doesn't erroneously update a nearby "polluted" highway pixel, preserving the high-resolution gradients of local-scale models.

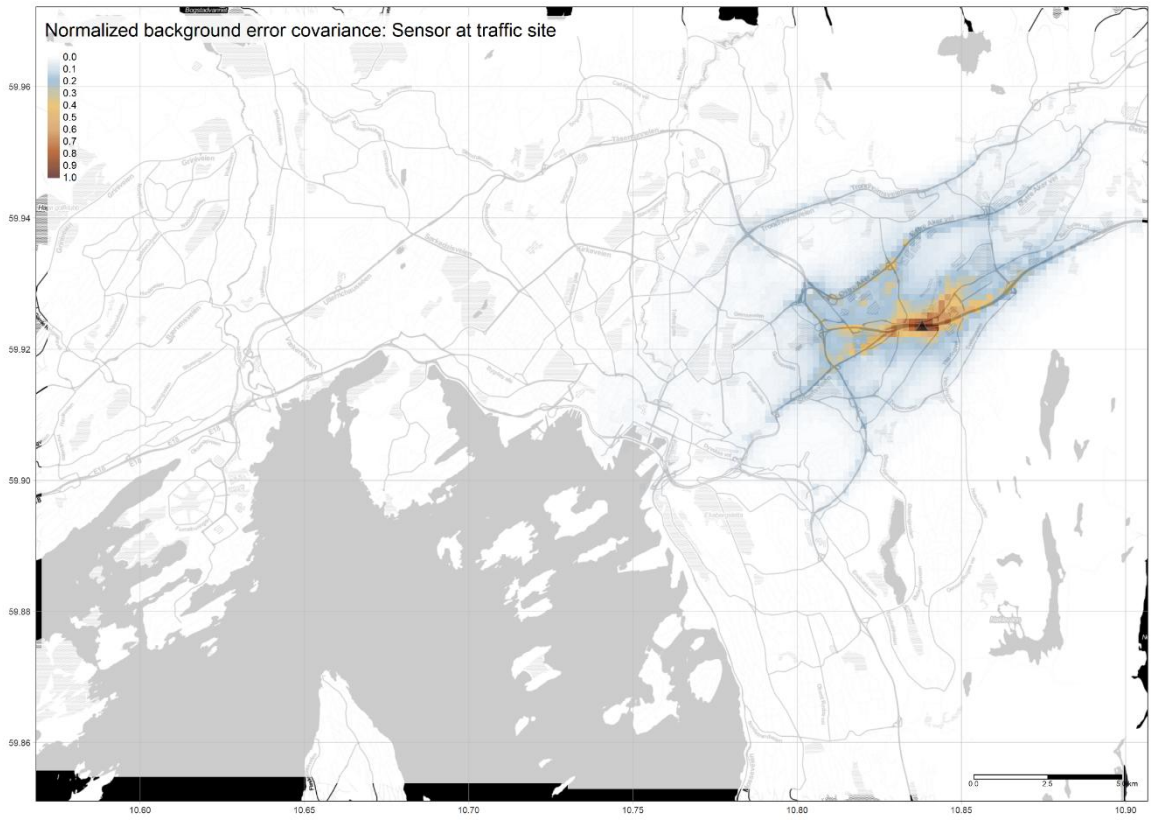

Figure 5: *Normalized background/model error covariance for a point observation at a traffic site in Oslo. Grid cells with high values are strongly affected by the observation (marked with a black triangle), whereas grid cells with low values are only weakly or not at all modified according to this particular observation.*

Examples of the resulting background error covariance fields are illustrated in Figure 5 and Figure 6. Each individual grid cell within the modelling domain is associated with a unique spatial "footprint", defined by the structure of the corresponding row (or column) of the matrix $\mathbf{B}$. These footprints determine the spatial extent and intensity of the impact that a given observation has on nearby grid cells. Conceptually, they describe how the signal from an observation "spreads" in space and influences the a priori model values at surrounding locations.

For example, Figure 5 presents the footprint associated with a traffic monitoring site located adjacent to a major roadway. The background error covariance in this case is highly anisotropic, exhibiting stronger correlations along the road network. This reflects the highly localized nature of traffic emissions and their directional influence, which is often shaped by street orientation and the prevailing wind regime. As a result, the adjustment applied by the OI algorithm is concentrated along the road corridors, aligning with the expected spatial pattern of transport-related pollution. However, as this covariance is re-calculated for each hour and the spatial patterns are directly derived from the model prediction for this hour, the wind field is implicitly also considered,

particularly for high wind speeds. As such, this method could also be considered to be generating "flow-dependent" background error covariances.

In contrast, Figure 6 shows the footprint corresponding to a background observation site situated in a forested area outside the urban core but still close to residential areas. Here, the covariance field is more isotropic, with a closer to circular spread around the observation point. This reflects the more homogeneous and less directional nature of pollution levels in such locations, where sources are sparse, and dispersion is not constrained by complex urban structures.

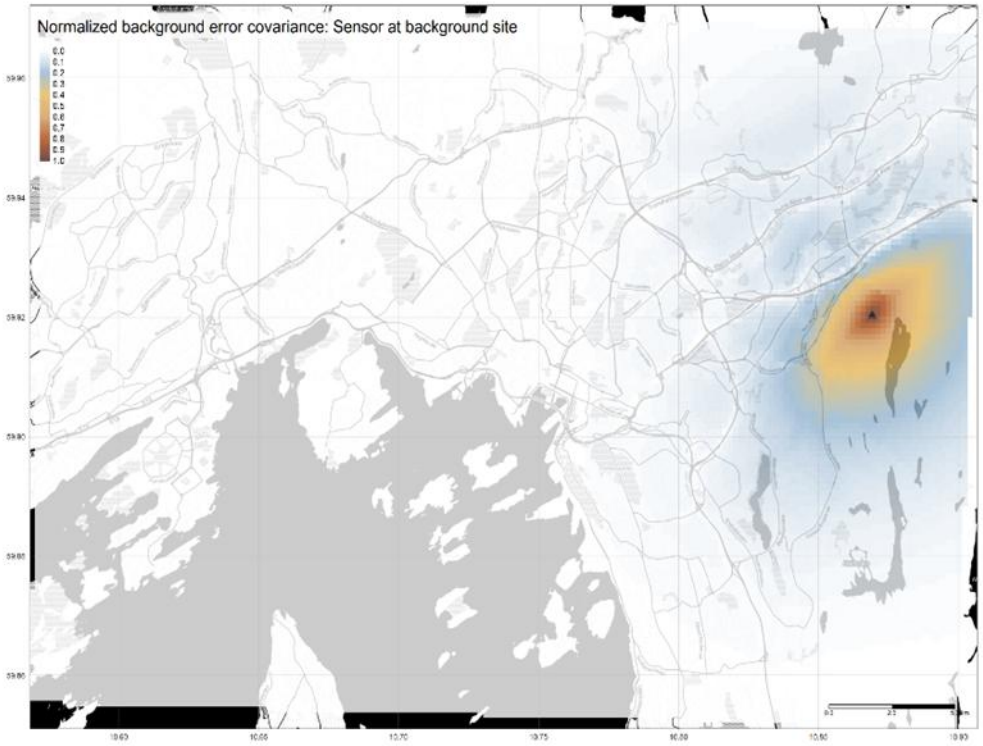

Figure 6: *Normalized background/model error covariance for a point observation at a background site in Oslo. Grid cells with high values are strongly affected by the observation (marked with a black triangle), whereas grid cells with low values are only weakly or not at all modified according to this particular observation.*

These contrasting examples highlight the flexibility and realism introduced by our approach to constructing the background error covariance matrix. By allowing spatially variable and context-dependent covariance structures, the method enhances the ability of the OI framework to assimilate observations in a physically consistent and spatially meaningful manner. This is particularly important in urban environments, where pollution patterns are heterogeneous and influenced by fine-scale processes that standard, stationary covariance assumptions often fail to capture.

4.3.1 Example results

The OI data assimilation approach has been used for assimilating sensor observations in previous work (Hassani et al., 2023; Schneider et al., 2023). In this section we show some examples for various urban areas in Norway.

4.3.1.1 Oslo

An example of the data assimilation procedure for Oslo, Norway, is shown in Figure 7 for a day with significant $PM_{2.5}$ pollution levels (6th January 2024 at 18:00 UTC). Both sensor observations and data from air quality monitoring stations were assimilated in this case.

Here we use modelling information from the uEMEP model (Denby et al., 2020; Mu et al., 2022). The uEMEP background field in the upper-left panel of represents the model's first-guess $PM_{2.5}$ distribution derived from emissions, meteorology, and boundary conditions alone. Concentrations broadly decrease from south-west to north-east, with maximum values along the Oslofjord shoreline and the E18 motorway. Such large-scale gradients are consistent with the location of major traffic corridors, industrial areas near the harbour, and a shallower atmospheric boundary layer over the fjord in winter. Nevertheless, the model field in general is spatially relatively smooth.

The overlaid observations reveal systematic departures from the model. Along the waterfront several reference stations and low-cost sensors report $PM_{2.5}$ concentrations that are lower than the background by 20–40 μg m^{-3}, whereas stations in urban and suburban areas in the centre and north-east of the city show positive differences of comparable magnitude. These mismatches indicate potential errors in local emissions, meteorology, or other modelling issues. For example, an over-prediction near the fjord could arise if the model overestimates residential wood combustion in that sector or if the meteorological driver maintains an inversion that is too strong. Conversely, an under-prediction inland may reflect unaccounted residential wood stoves or traffic congestion after the evening commute.

The data assimilation via OI combines each observation with the background according to prescribed error statistics. In the resulting analysis (upper-right panel), the broad concentration gradient remains but an explicit lower-concentration zone appears along the fjord (particularly in the western part of the domain), and a higher-concentration pocket develops in the northeastern part of the city around Storo/Grefsen/Kjelsås. These spatial features provided in the analysis align with the sign and magnitude of the sensor and station observations while at the same time remaining internally consistent with the model's physical constraints.

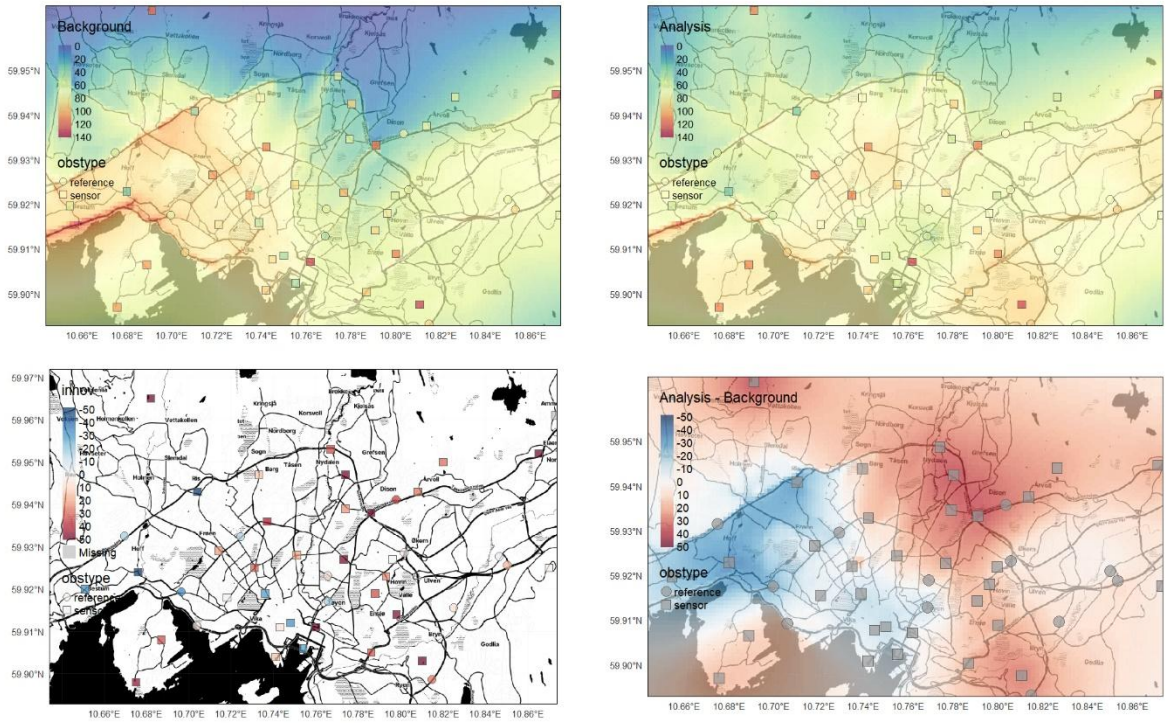

Figure 7: *Combining observations of low-cost sensor systems with model information through data assimilation, here shown for PM2.5 in Oslo for the period of 2024-01-06 at 18:00 UTC. Top left panel: a priori data set i.e. the uEMEP model output (background), and sensor/station observations (symbols); top right panel: the concentration field resulting from the data assimilation (the "Analysis") and the original sensor observations; bottom left panel: the innovation, i.e., the difference between model prediction and sensor observation, at the sensor deployment sites; bottom right panel: difference between analysis and uEMEP model prediction, indicating the spatial patterns of the corrections that were carried out as part of the assimilation. Base map copyright OpenStreetMap contributors and map tiles by Stamen Design, under CC BY 3.0.*

The innovation field in the lower-left panel quantifies the raw model-observation differences. Negative innovations (blue), i.e. locations where the model overestimated the $PM_{2.5}$ concentration compared to the observations, dominate the waterfront, whereas positive innovations (red), i.e. areas in which the model underestimated the true measured concentrations, are clustered in the center and north-east of the domain. Individual innovations exceed ±50 µg m^{-3}, which is large relative to typical wintertime Norwegian urban averages near 15 µg m^{-3}, however the day analysed here had overall very high $PM_{2.5}$ pollution levels. The spatial coherence of the innovations suggests that errors are not random sensor noise but reflect systematic model bias linked to local source patterns or meteorological representation.

The analysis increment (analysis minus background, lower-right panel) is the spatial expression of the corrections applied to the uEMEP a priori field as a result of the measurements from both sensors and air quality monitoring stations. A coherent negative increment extends along the fjord and western Ring 3 corridor, mirroring the aggregated negative innovations. Positive increments of up to 40 µg m^{-3} appear in the north- and south-eastern residential areas. The increment field

is smoother than the innovations because spatial covariances in the background-error model spread each station's information downwind and across adjacent grid cells while damping isolated outliers. As a result, neighbouring districts that lack sensors inherit credible adjustments derived from upwind observations.

Both observation classes (sensors as well as monitoring stations) contribute to these corrections but in different ways. High-accuracy reference stations, though few, carry substantial weight and anchor the analysis at regulatory monitoring sites. The denser low-cost network constrains finer-scale gradients that the reference network alone cannot resolve. Their combined use reduces expected analysis error variance relative to either data source used separately. The assimilation reduces systematic bias, sharpens spatial gradients, and provides a dynamically consistent $PM_{2.5}$ field suitable for exposure assessment, epidemiological linkage, and near-real-time air-quality management. The approach also supplies a quantitative basis for evaluating emission inventories and the meteorological drivers.

To visualize how the pattern can vary over time we also show the same figure for an hour just a day later (2024-01-07 at 20:00 UTC) in Figure 8. Between 6 January 2024 18:00 UTC and 7 January 2024 20:00 UTC the spatial distribution of $PM_{2.5}$ in Oslo, as well as the character of the errors corrected by data assimilation, changed substantially.

In the 6 January background field, the highest concentrations were confined to a narrow south-west coastal corridor that follows the E18 motorway and the inner Oslofjord, while the north-eastern suburbs and forested uplands were comparatively clean. On 7 January the model's maximum shifted inland and stretched along a diagonal band that runs from the south-west along Ring 3 towards Storo/Ensjø and Grorudalen in the east. Consequently, the gradient that had earlier separated the polluted fjord zone from cleaner air was replaced by a broader plume intersecting the urban core. Relatively low pollution levels were simulated in the Frogner area as well as Lillomarka and Østmarka.

The observation-model differences are similar. On 6 January most waterfront sensors and two reference stations reported concentrations 20–40 μg m^{-3} lower than the background, whereas several instruments in the north-east showed the opposite sign; the innovation field therefore displayed a clear two-peak pattern with negative values to the south-west and positive values inland. Twenty-six hours later the sign pattern was nearly reversed: negative innovations clustered across central Oslo, indicating a widespread model over-prediction there, while positive innovations emerged at the eastern periphery and, to a smaller extent, south of the fjord. Individual innovations again reached ±50 μg m^{-3}, but their location shifted by several kilometres.

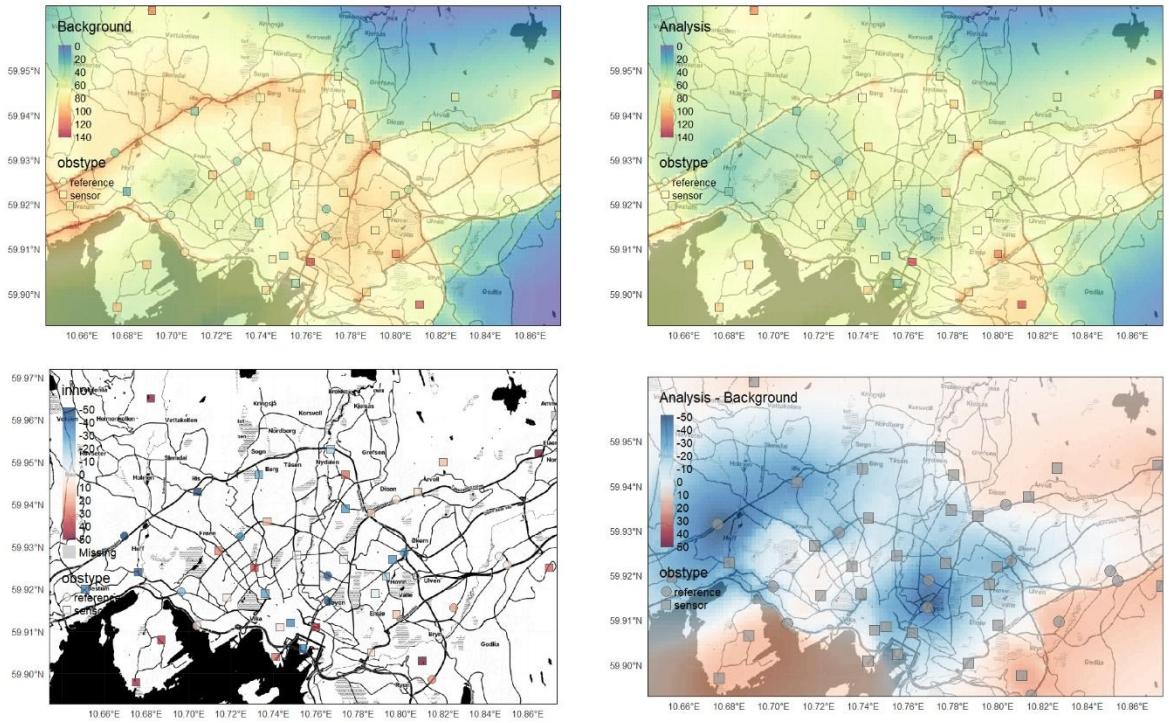

Figure 8: *Combining observations of low-cost sensor systems with model information through data assimilation, here shown for PM2.5 in Oslo for the period of 2024-01-07 at 20:00 UTC. Top left panel: a priori data set i.e. the uEMEP model output (background), and sensor/station observations (symbols); top right panel: the concentration field resulting from the data assimilation (the "Analysis") and the original sensor observations; bottom left panel: the innovation, i.e., the difference between model prediction and sensor observation, at the sensor deployment sites; bottom right panel: difference between analysis and uEMEP model prediction, indicating the spatial patterns of the corrections that were carried out as part of the assimilation. Base map copyright OpenStreetMap contributors and map tiles by Stamen Design, under CC BY 3.0.*

These opposite innovation patterns produced contrasting analysis increments. On 6 January the assimilation lowered the fjord-side concentrations and raised them in the north-east. On 7 January the increment field became predominantly negative across the city centre and the western half of the domain, with areas of positive correction confined to the forested areas and the very southwest of the domain. The smoother and more extensive area of negative increment on the later date shows that the algorithm applied a larger downward correction over a broader region, whereas the previous day's adjustment had been more localised.

The analysis panels corroborate the effect of these corrections. After assimilation on 6 January, the fjord-side bias was largely removed but relatively high concentrations persisted within the city, yielding a west-to-east gradient that was steeper than in the background. By 7 January the analysis field itself exhibits lower concentrations over the city core than 24 h earlier, even though the background predicted higher values there; in contrast, the residential districts in Groruddalen and Bryn in the eastern part of the domain became the relative hotspot. The net result is a

reduction in the population-weighted $PM_{2.5}$ exposure in central Oslo between the two evenings, accompanied by an increase in exposure for residents east of the city centre.

Taken together, the comparison highlights three points. First, winter-time $PM_{2.5}$ fields over Oslo can reorganise within a single diurnal cycle, most likely in response to changes in wind direction, boundary-layer depth and the spatial pattern of domestic wood-burning. Second, the uEMEP model captured the broad synoptic evolution but exhibited alternating sign biases that depended on location and hour, underscoring the need for continual observational correction. Third, data assimilation adapted effectively to these changing errors: it removed a coastal over-prediction on 6 January and, one day later, applied a larger but differently placed downward adjustment across the urban core while selectively increasing concentrations on the eastern fringe. The sequence therefore illustrates the dynamic interaction between model physics, emission inventories, and heterogeneous observations in producing realistic high-resolution air-quality analyses.

4.3.1.2 Kristiansand

For Kristiansand, Norway, the data assimilation has been successfully used with the local LCS network as highlighted in detail in Hassani et al. (2023). Comparisons between averaged $PM_{2.5}$ concentrations measured by static LCSs and urban EMEP (uEMEP) model simulations revealed significant discrepancies, particularly in the Grim area in the west of the domain, associated with residential wood combustion (RWC). Model outputs indicated higher $PM_{2.5}$ concentrations for the winter of 2021–2022, whereas sensor observations recorded higher levels in the winter of 2020–2021.

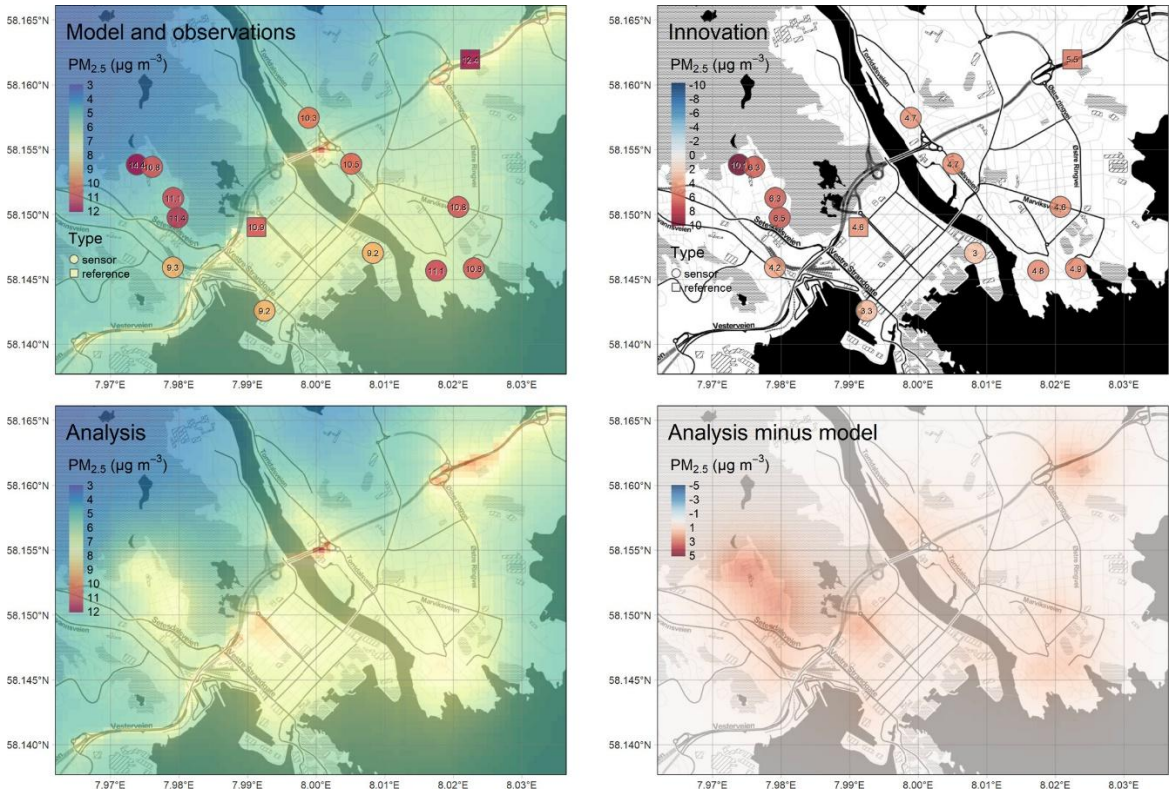

Figure 9: *Combining observations of low-cost sensor systems with model information through data assimilation, here shown for $PM_{2.5}$ in Kristiansand for the period of 2020-12-01 through 2021-02-28. Top left panel: original uEMEP model, a priori data set (background), and sensor observations (symbols); top right panel: the innovation, i.e., the difference between model prediction and sensor observation, at the sensor deployment sites; bottom left panel: the concentration field resulting from the data assimilation (the "Analysis") and the original sensor observations; bottom right panel: difference between analysis and uEMEP model, indicating the spatial patterns of the corrections that were carried out as part of the assimilation. Base map copyright OpenStreetMap contributors and map tiles by Stamen Design, under CC BY 3.0.From (Hassani et al., 2023)*

To address these differences, we used data assimilation based on Optimal Interpolation (OI), integrating Airly $PM_{2.5}$ sensor network observations with uEMEP model outputs. This assimilation method substantially improved the model estimates, especially increasing $PM_{2.5}$ concentrations in the Grim area in the western part of the domain. Moderate adjustments were also applied in the Kvadraturen and Lund regions, resulting in a better match between modeled values and observed data.

A leave-one-out cross-validation (LOOCV) further validated the benefits of data assimilation. At reference monitoring stations, assimilation modestly improved predictions toward the observed values in both winters. However, across the entire sensor network, the improvements were notably larger, resulting in reductions in mean bias, root mean square error, and mean absolute error ranging from approximately 40–56%. These findings underline the effectiveness of using OI-based data assimilation with LCS networks to enhance high-resolution urban air quality mapping. For more details please see Hassani et al. (2023).

In addition to the previous analysis which showed the situation for a temporal average over 3 months, we also show an example of the data assimilation process in Kristiansand for a single hour (Figure 10). At 18:00 UTC on 7 January 2024 the uEMEP background field for the greater Kristiansand region (upper-left panel) displays a very prominent $PM_{2.5}$ hotspot over southwestern part of the domain related to the industrial activity in the Fiskåtangen area with elevation $PM_{2.5}$ of 60-80 µg m^{-3} and a secondary peak in the Kvadraturen area of the city (ca. 40-50 µg m^{-3}).

The observations, however, deviate systematically from that picture. Both reference stations (circles) and low-cost sensors (squares) in the Kvadraturen area report values 10–40 µg m^{-3} lower than the background; only a single instrument north-west of the city centre in the Grim area shows a modest positive departure. These differences appear in the innovation panel (lower-left) as contiguous light blue symbols southwest of the Otra river and a lone red symbol to the north-west in Grim. The spatial coherence of the negative innovations indicates that the model over-estimates the true concentration levels in these areas rather than the observations suffering from random sensor uncertainty. Northeast of the Otra river we see some positive deviations among the sensors but a negative innovation for the reference station.

Optimal interpolation adjusts the field accordingly. In the analysis (upper-right) the Fiskåtangen hotspot remains more or less unchanged due to lack of observations in this area. However, concentrations in the Kvadraturen area are somewhat reduced to 20–40 µg m^{-3}, bringing the field into line with the majority of observations. The analysis increment (lower-right) quantifies this correction: negative values of −10 to −40 µg m^{-3} spread over the eastern neighbourhoods, whereas a small positive lobe (<20 µg m^{-3}) develops north-west of the city where the background was too low. It should be noted that the latter is likely driven primarily by the single sensor in the Grim area, for which the model did not simulate strong spatial gradients and as such this observations has a close to isotropic error covariance structure, which is likely not quite correct in this case (the influence of this sensor should ideally be limited to the Grim area itself and not adjust background values further north).

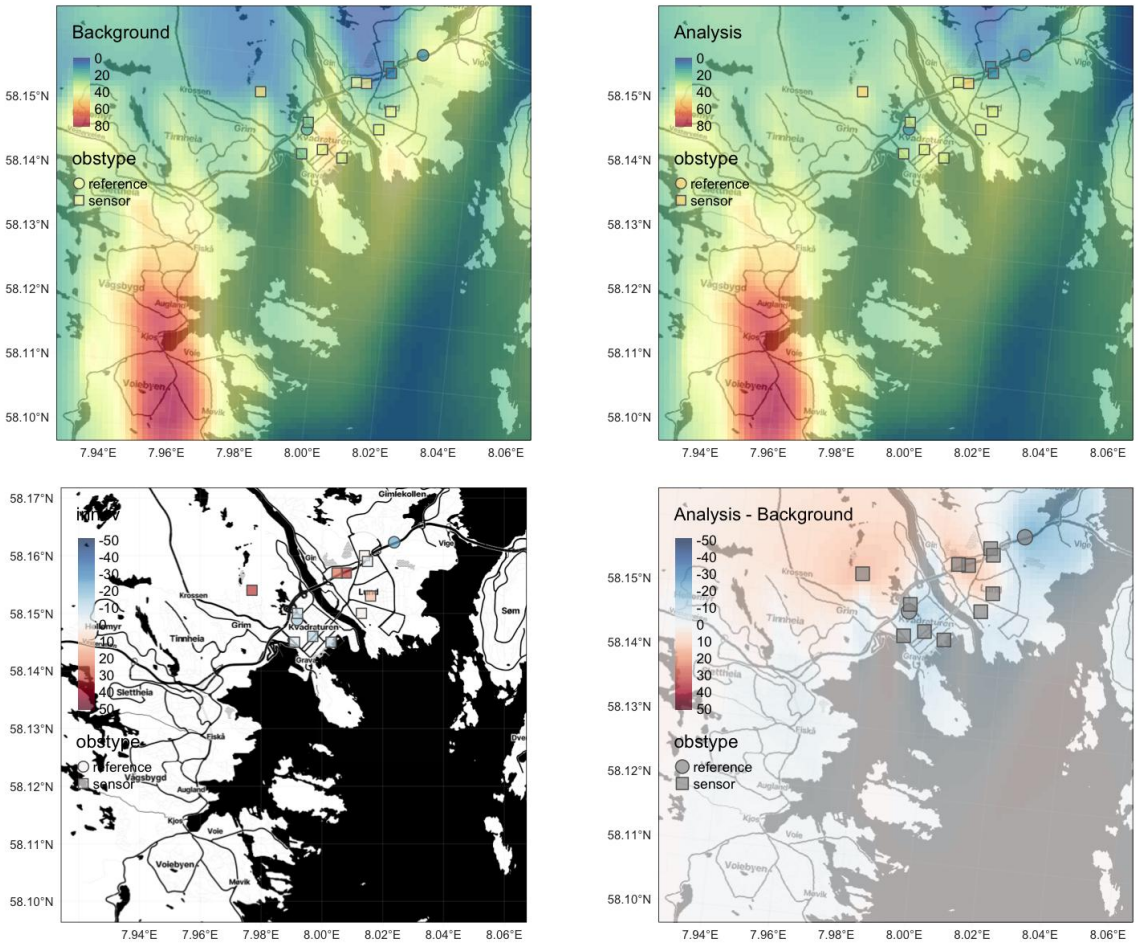

Figure 10: *Combining observations of low-cost sensor systems with model information through data assimilation, here shown for $PM_{2.5}$ in Kristiansand for the hour of 2024-01-07 at 18:00 UTC. Top left panel: original uEMEP model, a priori data set (background), and sensor observations (symbols); top right panel: the innovation, i.e., the difference between model prediction and sensor observation, at the sensor deployment sites; bottom left panel: the concentration field resulting from the data assimilation (the "Analysis") and the original sensor observations; bottom right panel: difference between analysis and uEMEP model, indicating the spatial patterns of the corrections that were carried out as part of the assimilation. Base map copyright OpenStreetMap contributors and map tiles by Stamen Design, under CC BY 3.0.*

Because the assimilation spreads each observation through a flow-dependent error covariance, districts lacking sensors inherit either the concentration from the uEMEP a priori (for regions very far away from observations) or concentrations proportional to corrected upwind grid cells. In operational terms, the assimilation step cuts the mean-absolute model error at the available stations from roughly 25 µg m^{-3} to below 10 µg m^{-3} and corrects model biases, allowing for more accurate exposure assessment. At least for concentration mapping purposes, this also highlights the importance of placing sensors in a mostly regular pattern, covering as many areas of a domain as possible. Without sensor observations in reasonable proximity, the analysis in these regions defaults to the a priori simulations from the model and the approach can thus not add value there.

4.3.1.3 Bergen

We provide an example figure illustrating the impact of the data assimilation routine for Bergen, Norway (Figure 11). At 18:00 UTC on 7 January 2024 the uEMEP first-guess (upper-left panel) depicts a shallow $PM_{2.5}$ trough roughly aligned with the northwest–southeast valley that cuts through Bergen. Concentrations in the city centre and Årstad fall in the 40–60 µg m^{-3} range, tailing off to <20 µg m^{-3} over the surrounding fjords and the hills east of 5.36 °E.

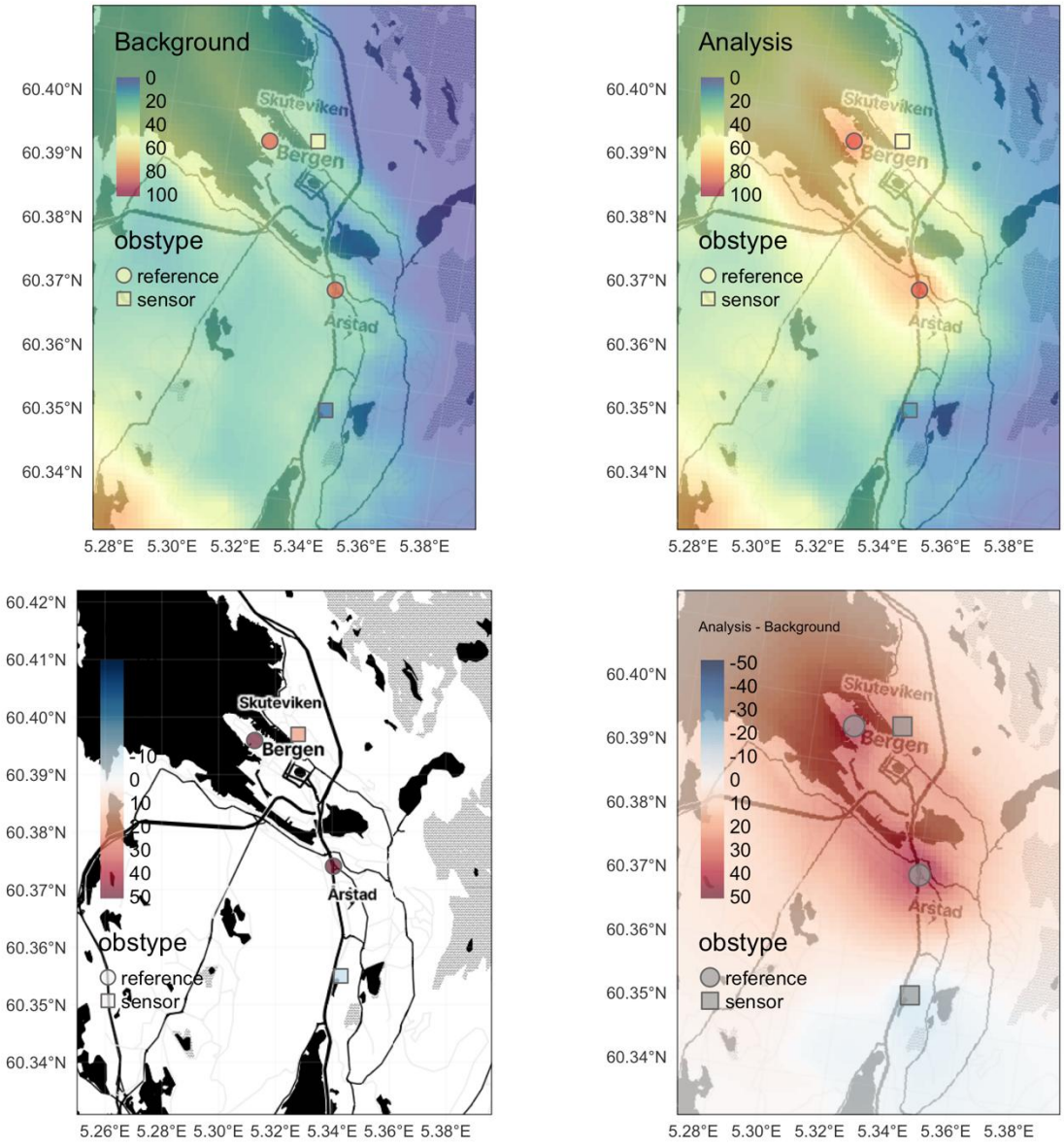

Figure 11: *Combining observations of low-cost sensor systems with model information through data assimilation, here shown for $PM_{2.5}$ in Bergen for the hour of 2024-01-07 at 18:00 UTC. Top left panel: original uEMEP model, a priori data set (background), and sensor observations (symbols); top right panel: the innovation, i.e., the difference between model prediction and sensor observation, at the sensor deployment sites; bottom left panel: the concentration field resulting from the data assimilation (the "Analysis") and the original sensor observations; bottom*

right panel: difference between analysis and uEMEP model, indicating the spatial patterns of the corrections that were carried out as part of the assimilation. Base map copyright OpenStreetMap contributors and map tiles by Stamen Design, under CC BY 3.0.

Only four observations are available: two high-quality reference stations in the city centre of Bergen and in Årstad (circles) and two low-cost sensors, one at Bryggen and one south of the city (squares). Three of the four instruments report $PM_{2.5}$ levels that exceed the background by 15–45 µg m^{-3}, a pattern visible in the innovation panel (lower-left) as uniformly positive symbols. One of the two sensors indicates a slightly lower value than the a priori model field. With exception of the southern area of the domain, the model therefore generally under-represents the $PM_{2.5}$ concentrations during this particular date and time, possibly because it lacks local wood-smoke contributions or because the complex meteorology of the areas was not represented accurately enough.

Optimal interpolation elevates the field for the northern half of the domain, while it corrects the uEMEP concentration field slightly downwards in the south. In the analysis (upper-right) a band of 60–90 µg m^{-3} now links the city centre and Årstad, while concentrations elsewhere remain close to the background. The analysis increment (lower-right) shows positive corrections of +20 to +50 µg m^{-3} around three stations, fading rapidly with distance. Weak negative increments exist around the sensor observation in the south but not have a major impact on the analysis.

Because the observing network is extremely sparse, two reference sites and two sensors over an area with steep orography and complex land–sea flows, the assimilation system can do little more than correct the model field in the immediate vicinity of those points. Large sectors of the fjord, the eastern and western suburbs and the mountains remain largely unconstrained and may still carry the original model bias to some extent. The example nevertheless demonstrates the core functionality of the system: when observations are available the algorithm adjusts the background towards them, weighting each update by the assumed errors of both model and instrument. Expanding the low-cost sensor deployment in Bergen would allow those corrections to extend beyond the narrow corridor sampled here and would give greater confidence in the analysed exposure for neighbourhoods that currently lack measurements.

4.4 Regional-scale air quality mapping with sensor networks using machine learning

4.4.1 Introduction

NILU has developed a machine-learning approach that is able to integrate air quality data of various types and use it in combination to provide improved maps of air quality at the regional

scale. The approach is called S-MESH[3] (Shetty et al., 2024, 2025, 2026) and addresses two persistent needs in operational air-quality mapping: city-scale spatial detail and timely availability. For $PM_{2.5}$, the aim is to deliver daily 1 km hindcasts that are available much sooner than the CAMS regional reanalysis yet comparable in utility, by downscaling the CAMS 24-h forecast into a higher-resolution, bias-corrected daily product. For NO_2, S-MESH is motivated by the fact that satellite instruments such as Sentinel-5P/TROPOMI measure tropospheric columns rather than surface concentrations and pass each location only once per day, so surface fields must be learned by fusing satellite signals with other drivers and station data for being able to provide metrics that are more societally relevant.

Methodologically, S-MESH is a machine-learning fusion framework that is trained on reference-grade monitoring stations and predicts daily surface concentrations on a 1 km grid. For NO_2, it uses an XGBoost regressor with twelve spatiotemporal inputs and a design that avoids spatial leakage: stations are split into non-overlapping train/test sets, the training set is stratified by season, and evaluation uses standard error and correlation metrics. Station observations are temporally aligned to the Sentinel-5P overpass by a weighted two-hour average so the target matches satellite sampling, and selected predictors are transformed to improve normality before model fitting. Model behaviour and driver importance are examined with SHAP to validate that physically meaningful features dominate. The datasets included in the approach mirror these objectives.

Within CitiObs we have extended the existing S-MESH framework for $PM_{2.5}$ (Shetty et al., 2025) for its potential to also integrate quality-controlled observations from low-cost sensor networks. In the following we present the methodology and results of this study. The content of this section is largely adapted from Shetty et al. (2026).

4.4.2 Methodology

The methodology builds on an extended version of the S-MESH framework to evaluate how validated low-cost sensor (LCS) data can improve daily $PM_{2.5}$ estimation over Central Europe at 1 km spatial resolution. S-MESH employs a stacked ensemble of XGBoost models integrating a wide range of environmental data sources, including satellite aerosol optical depth (AOD), meteorological reanalyses, chemical transport model (CTM) forecasts, and ground-based observations. In this study, three variants of S-MESH were implemented. The Baseline model

[3] https://models.nilu.no/models/s-mesh/

reproduces the original S-MESH configuration and is trained solely on regulatory air quality monitoring data. The LCST model uses the same predictors as the Baseline but replaces regulatory $PM_{2.5}$ with corrected LCS measurements as the training target. The LCSI model retains reference stations as targets but incorporates spatially processed LCS data as two additional input features: an inverse-distance-weighted convolution layer of LCS-derived $PM_{2.5}$ and the distance to the nearest LCS. All three models are trained and evaluated over a Central European domain for 2021–2022. The domain in Central Europe was chosen because the density of sensor systems was by far the highest in all of Europe, suggesting an ideal area for evaluation of LCS data within the S-MESH system. The overall concept of the approach can be found in Figure 12.

A comprehensive set of predictor datasets was assembled and harmonized on a common 1 km × 1 km grid. Surface $PM_{2.5}$ observations from regulatory monitoring stations were obtained from the EEA, restricting the dataset to verified hourly values and deriving daily averages from days with at least 75% data completeness. This resulted in 832 stations distributed across urban, suburban, rural, traffic, and industrial environments. LCS measurements were taken from the Sensor.Community[4] and PurpleAir[5] networks in their FILTER-corrected and quality-controlled form, ensuring consistency across sensor types and deployments. Only the highest-quality data, as defined by the FILTER protocol (Hassani et al., 2025), were retained. Original sub-hourly measurements were aggregated to hourly and then daily values. The final dataset contained more than 11,000 unique sensor locations, with strong representation in urban areas. For use in the LCSI model, an inverse-distance-weighted spatial convolution of LCS $PM_{2.5}$ was computed daily, using a distance exponent of two and requiring a minimum of two neighboring sensors within 20 km. This step ensured that the convolution layer reflects meaningful spatial structure rather than isolated, potentially biased measurements. An additional variable quantifying the distance to the nearest LCS was derived to inform the model about local data density and expected reliability of the convolution field.

[4] https://sensor.community/en/

[5] https://www2.purpleair.com/

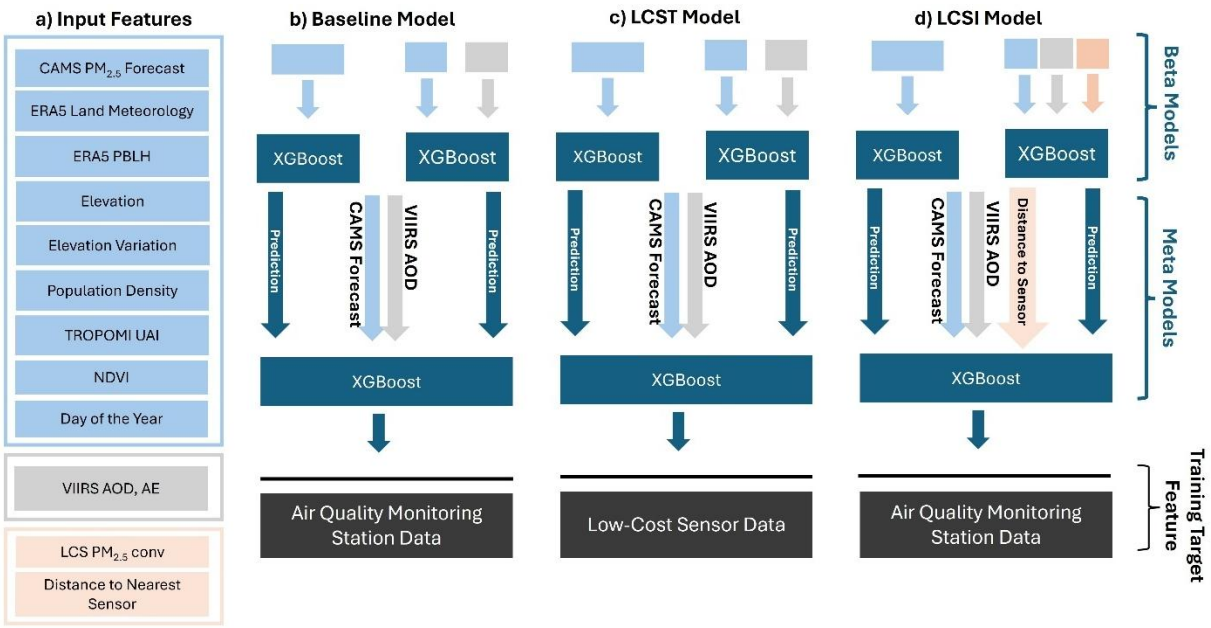

Figure 12: *The S-MESH conceptual framework with stacked XGBoost ML for three model variants (a) summarizes all input features used across the models. These features are grouped and color-coded to simplify their representation in the model diagrams for (b) the Baseline Model, (c) the LCST Model, and (d) the LCSI Model. The black boxes along the bottom row of each model indicate the specific target variables used for training. From Shetty et al. (2026).*

To complement ground-based observations, the study integrates the CAMS regional ensemble forecast product, which provides daily $PM_{2.5}$ estimates at ~10 km resolution based on an ensemble of regional CTMs. These forecasts serve as a key input feature, providing both a physical foundation grounding the prediction and at the same time helping to compensate for missing AOD data under cloudy conditions. The CAMS interim regional reanalysis was used as an external spatial benchmark for evaluating the realism of modeled $PM_{2.5}$ distributions. Satellite-derived AOD and Ångström Exponent from the VIIRS Deep Blue product were included after filtering out low-quality retrievals, while vegetation and land surface characteristics were represented through VIIRS NDVI and MERIT digital elevation model derivatives. Meteorological predictors, including temperature, dew point temperature, surface pressure, precipitation, wind components, and boundary layer height, were extracted from ERA5 and ERA5-Land reanalyses and aggregated to daily means. Additional spatial predictors, namely population density from GPWv4 and the UV Aerosol Index from TROPOMI, were added to capture anthropogenic activity and aerosol type information.

All spatial datasets were resampled to the 1 km modeling grid using bilinear interpolation for coarse-resolution inputs and spatial averaging for finer-scale products. Predictor variables with skewed distributions were transformed as needed for numerical stability. Point-based values for both reference stations and LCS locations were extracted from the harmonized grids. The station

dataset was partitioned into training and testing subsets, with 80% of stations used for training and 20% held out for independent evaluation. Within the training subset, stations were further split to separately train the two first-stage "beta" models and the second-stage meta model of the stacked XGBoost architecture. Hyperparameters were optimized using a combination of Bayesian and random search approaches. One of the beta models was deliberately structured to rely strongly on CAMS forecast $PM_{2.5}$, ensuring spatial continuity in days or regions with limited satellite AOD coverage.

Model performance was assessed using a consistent evaluation dataset covering 2021–2022. Accuracy metrics included mean absolute error, root-mean-square error, mean bias, Spearman correlation, coefficient of determination, and relative absolute error. Spatial representativity was also evaluated by comparing model predictions to CAMS reanalysis patterns. To interpret model behavior and quantify the contribution of different predictors, SHAP (Shapley Additive Explanations) values were computed. Mean absolute SHAP values provided global feature importance, while spatial and temporal SHAP aggregations were used to explore intra-urban, inter-urban, and event-specific dynamics in model feature reliance.

Together, these methodological components enabled a systematic and controlled comparison of three strategies for incorporating LCS data into a regional, satellite-assisted machine learning framework for $PM_{2.5}$ estimation.

4.4.3 Results

The results demonstrate clear differences in performance between the three S-MESH model variants, namely Baseline, LCST, and LCSI, with the LCSI model consistently outperforming the others. Overall accuracy evaluated against an independent set of reference-grade stations shows that LCSI achieves the lowest mean absolute error (2.68 μg m^{-3}) and an almost unbiased mean prediction, improving on both the Baseline model (3.32 μg m^{-3}) and LCST (3.66 μg m^{-3}). The LCST model exhibits the largest negative bias and reduced performance, particularly at higher pollution levels, where relative errors increase substantially. In contrast, LCSI shows lower relative absolute errors across the entire $PM_{2.5}$ concentration range. Correlation-based metrics confirm these trends: while LCST attains an R^2 of only 0.5, the Baseline model increases this to 0.65, and LCSI improves further to 0.74. This highlights the added predictive value of the spatially processed LCS information.

Feature attribution using SHAP values provides insights into the drivers of these performance differences. Across all models, CAMS $PM_{2.5}$ forecasts are among the strongest predictors, along

with meteorological variables such as dew point temperature, boundary layer height, and surface temperature. The LCSI model stands out due to the dominant influence of the LCS convolution layer and, to a lesser degree, the distance to the nearest sensor, both of which capture local-scale pollution variability. In the LCST model, satellite AOD becomes disproportionately important, suggesting that in the absence of spatially explicit LCS inputs, the model relies more heavily on columnar aerosol information to compensate for the noisier LCS-derived training targets.

The density and spatial distribution of LCS strongly influence LCSI performance. Relative errors decrease substantially when multiple LCS sensors are located within 20 km of a test station, and errors remain below 25% up to distances of approximately 10 km from the nearest LCS. Beyond about 100 km, the accuracy of LCSI converges with that of the Baseline model, indicating that the benefits of the LCS convolution layer diminish in sensor-sparse regions. This reflects a broader spatial overlap in the deployments of reference stations and LCS networks, particularly in urban areas, which helps explain similar error patterns in both models.

Spatial analyses during a major pollution episode on 25 March 2022 show that all models capture the broad-scale $PM_{2.5}$ patterns observed in the CAMS reanalysis, including extensive pollution across Poland, northern Germany, the Netherlands, Belgium, and northern Italy (Figure 13). However, LCSI produces sharper, more localized gradients than either the Baseline or LCST models, reflecting its sensitivity to dense urban LCS networks. While this improves local agreement with reference stations, evidenced by the highest occurrence of low-error sites in LCSI, its spatial continuity is weaker than in the Baseline model and CAMS reanalysis, particularly in areas with few or no LCS. SHAP-based spatial analyses within a subdomain around Berlin confirm that LCSI predictions are strongly shaped by the spatial structure of the LCS layer, showing high SHAP values concentrated in urban areas. CAMS forecasts contribute more uniformly across space, and their influence decreases progressively from the Baseline model to LCST and further to LCSI. The LCST model shows additional influence from AOD, which becomes particularly visible in spatial SHAP maps.

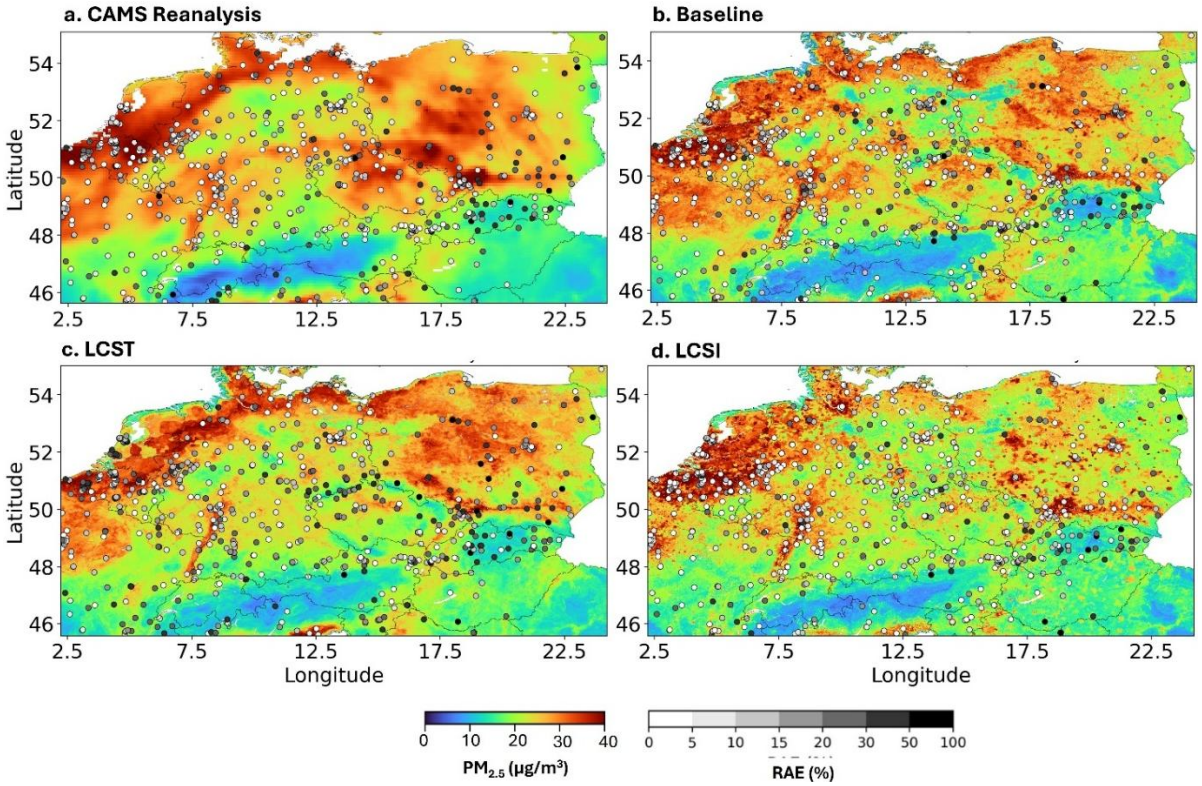

Figure 13: *$PM_{2.5}$ estimates for March 25th, 2022 are shown for a) CAMS regional reanalysis, b) the Baseline model, c) the LCST model (using LCS as the target), and d) the LCSI model (using LCS as an input). These maps illustrate $PM_{2.5}$ levels during one of the episodic pollution days over Europe. Overlaid symbols indicate the relative absolute errors (in %), showing the model deviations from station measurements for that date. From Shetty et al. (2026).*

Monthly mean $PM_{2.5}$ fields for March 2021 in southern Poland reinforce the performance differences. The CAMS reanalysis and LCST model both systematically underestimate concentrations in urban hotspots such as Katowice, whereas the Baseline and LCSI models capture observed spatial patterns more accurately. LCSI shows the strongest spatial agreement with both reference station and LCS-derived monthly means, reflecting finer spatial gradients that stem from the LCS convolution input. Temporal SHAP analyses at an urban station in Katowice for March 2021 highlight the differing model mechanisms. In the LCSI model, the LCS convolution layer becomes the dominant predictor during high-pollution days, whereas the Baseline and LCST models rely more heavily on CAMS forecasts throughout the month. The LCST model's reliance on AOD also becomes evident during these episodes, contributing to its systematic underestimation of $PM_{2.5}$ in urban hotspots.

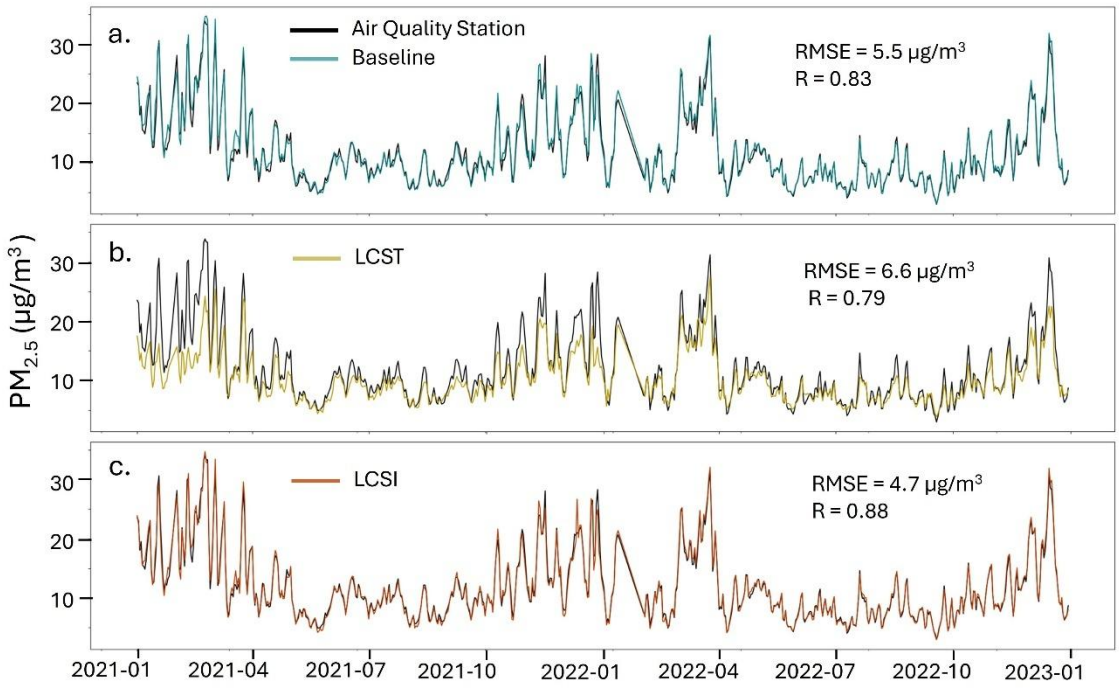

Figure 14: *Time series of median $PM_{2.5}$ levels from the three S-MESH models compared with median station measurements across all test sites for 2021–2022. The three horizontal panels display $PM_{2.5}$ variations estimated by a) the Baseline Model (teal line), b) the LCST Model (golden line), and c) the LCSI Model (orange line), shown alongside station measurements depicted by black lines. From Shetty et al. (2026).*

A two-year time series comparison of the median over all test stations further shows that LCSI best captures both seasonal and day-to-day variability in $PM_{2.5}$, yielding the lowest RMSE (4.7 µg/m³) and highest correlation (0.88) with reference observations (Figure 14). The Baseline model performs moderately well but with higher errors, while LCST systematically underestimates $PM_{2.5}$ and shows the weakest temporal agreement. All models reproduce the pronounced winter peaks and lower summer concentrations, including the markedly elevated levels in March 2022 compared to 2021. Annual mean maps for 2021 confirm that all S-MESH variants improve upon CAMS reanalysis in terms of both spatial detail and agreement with station observations. LCSI provides the highest spatial resolution and best captures urban hotspots and local gradients, especially in Poland and Northern Italy, where CAMS is known to underestimate $PM_{2.5}$. Quantitatively, LCSI achieves an annual MAE of 1.73 µg m^{-3}, compared with 3.2 µg m^{-3} for CAMS. However, LCSI also exhibits sharper spatial gradients and underestimates inter-urban pollution transport in areas lacking LCS coverage, highlighting a limitation associated with uneven sensor distribution. The LCST and Baseline models produce smoother patterns more comparable to CAMS, but only the Baseline matches LCSI in reproducing station-level concentrations in high-pollution areas.

Taken together, the results show that integrating LCS data as a model input (LCSI) substantially enhances the capability of S-MESH to reproduce fine-scale spatial and temporal variability in $PM_{2.5}$ particularly in urban regions where LCS density is highest. Using LCS as a training target (LCST) is less effective, primarily due to the uneven distribution and remaining uncertainties in LCS measurements. Despite some limitations in sensor-sparse regions, LCSI represents the most accurate and spatially detailed approach, demonstrating the added value of large-scale LCS networks for improving high-resolution air quality modelling across Europe.

More information about the CitiObs activities on integrating LCS observations into S-MESH can be found in Shetty et al. (2026).

5 SUMMARY

Low-cost sensors (LCSs) operated by citizens have significantly expanded air quality monitoring by enabling dense, real-time, and spatially rich measurements. However, the absence of harmonized, quality-controlled datasets—especially those integrating heterogeneous data streams from multiple citizen-operated LCS networks—has limited their use in scientific research, public health assessments, and regulatory contexts. To address this challenge, this report discusses **ValAir**, a comprehensive toolkit developed within the CitiObs project for quality control (QC), validation, and correction of LCS measurements. At the core of ValAir lies **FILTER** (Framework for Improving Low-Cost Technology Effectiveness and Reliability), a geospatially scalable system designed to unify, evaluate, and correct outdoor, crowd-sourced sensor data.

The original version of FILTER, introduced by Hassani et al. (2025), was developed specifically for static $PM_{2.5}$ sensors and has been evaluated and validated using sensor.community and PurpleAir datasets across Europe. Building on this, FILTER has been expanded into complementary variants tailored to different sensing modalities like mobile and wearable PM2.5 sensors and noise sensors. While these variants share a common processing chain, each employs algorithms adapted to the specific characteristics of the LCS modality considered.

Across all FILTER versions, the QC procedure follows a structured sequence of statistical tests designed to detect different types of measurement characteristics while preserving genuine environmental signals. The tests include: (i) **Range Check** – identifies physically impossible or instrument-limited values, (ii) **Constant Value (Flatline) Check** – detects sensors or time periods exhibiting flatline behaviour, often associated with hardware or firmware failure, (iii) **Spatio-Temporal Outlier Detection** – identifies abnormal measurement patters using robust rolling statistics and cross-checks with neighbouring sensors to distinguish isolated anomalies from genuine area-wide events; (iv) **Spatial Correlation Test** – evaluates whether a sensor's temporal behavior is consistent with patterns observed at nearby sensors or reference stations; and (v) **Spatial Similarity Test** – compares the absolute levels measured by each sensor against the nearby official monitoring sites. Finally, each test assigns an individual quality flag to every measurement, enabling transparent and traceable QC classification.

The FILTER variants operate at two processing levels namely, 'Raw' and 'Corrected'. 'Raw' refers data provided by the sensor suppliers and may have been pre-processed, including internal averaging or basic firmware adjustments. 'Corrected' values incorporate in-situ adjustments using reference monitoring stations. In this process, correction is sensor-specific and sensor-agnostic,

as each sensor is evaluated and corrected individually, independently of sensor type or differences in sensors units across various suppliers. The correction involves a remote procedure, relying on nearby reference and meteorological stations, and is evaluated dynamically in time to adapt to changing environmental conditions. Importantly, this approach allows consistent correction across heterogeneous sensor networks without relying to co-location with regulatory instruments. This dual-level approach enables flexible integration of FILTER into diverse monitoring strategies and modelling workflows.

The optimal configuration of FILTER variants depends on user needs, the characteristics of the measurement campaigns, the spatio-temporal of the sensor network, and the environmental context of each case study. Default parameters—such as outlier thresholds, spatial radius, rolling window sizes, correlation limits, etc.—should be adjusted to reflect local conditions and application-specific requirements. By adapting these settings, the proposed methodologies can be effectively optimized for a wide range of applications, extending from citizen science to regulatory support and environmental modelling.

The work on **MapAir** presented in this document demonstrates how modern data fusion and machine-learning methods can transform air quality information derived from heterogeneous sources, ranging from LCS networks to regulatory monitors, atmospheric models, and satellite retrievals, into spatially complete, decision-ready pollution maps. The MapAir algorithms, forming the backbone of this effort, show that it is now possible to move far beyond conventional point measurements and to produce continuous, high-resolution fields of $PM_{2.5}$ concentrations for both local urban environments and the wider European domain. These algorithms flexibly integrate multiple data streams and can operate at resolutions from roughly 100 m in cities to 1 km at the continental scale, depending on data density and application needs. The project highlights that when sufficiently many sensors of acceptable quality are present, the resulting air quality maps capture spatial gradients with a fidelity that traditional modelling systems alone cannot reach.

A central contribution of this work is the demonstration that low-cost sensor systems, when properly calibrated, quality-controlled, and combined with physically based model information, can significantly strengthen air quality assessments. LCS networks provide dense geographic coverage, but they require careful uncertainty handling. Through methods such as Optimal Interpolation (OI), the project illustrates how LCS observations can be fused with local-scale air quality models to produce accurate and physically consistent concentration fields. The presented assimilation case studies show that these methods have the potential to correct systematic modelling errors, refine spatial patterns during pollution episodes, and align the modelled fields

more closely with observed conditions. When combined with high-resolution models like EPISODE or uEMEP, data fusion and assimilation approaches such as the ones presented here can enhance both spatial and temporal representativeness, thus allowing city authorities and researchers to rely on more accurate, evidence-based air quality information.

At the regional scale, the extended S-MESH machine-learning framework demonstrates how satellite products, meteorological reanalyses, chemical transport model forecasts, regulatory data, and LCS observations can jointly inform daily 1 km $PM_{2.5}$ fields across Europe. Incorporating LCS information into S-MESH as an input feature leads to substantial accuracy improvements, especially in urban areas where sensor density is highest. The results show lower errors, reduced biases, better representation of high-pollution episodes, and more realistic spatial gradients. This confirms that blending structured physical model information with the fine-scale spatial detail from LCS networks yields a system that is more responsive to local pollution events while still preserving large-scale coherence. The work also identifies current limitations, especially in regions with sparse LCS deployment, emphasising the need for more geographically balanced sensor coverage.

The developments presented here represent an advance toward scalable, flexible, and transparent air quality mapping approaches that can support a wide range of societal needs. By integrating various available data sources, the methods have potential for allowing more accurate exposure assessments for health research, enhance the capability of cities to evaluate local emission sources, and provide timely high-resolution information that is suitable for public communication and community engagement. We demonstrate the feasibility of integrating dense citizen-driven sensor networks with air quality modelling systems. Methods such as those presented here can contribute to a more participatory and data-rich approach to urban environmental management. These innovations also lay the groundwork for future operational services and open-data products that can strengthen air quality governance across Europe.

6 REFERENCES

Adams, M. D., Massey, F., Chastko, K., and Cupini, C.: Spatial modelling of particulate matter air pollution sensor measurements collected by community scientists while cycling, land use regression with spatial cross-validation, and applications of machine learning for data correction, Atmos. Environ., 230, 117479, https://doi.org/10.1016/j.atmosenv.2020.117479, 2020.

Bagkis, E., Kassandros, T., and Karatzas, K.: Learning Calibration Functions on the Fly: Hybrid Batch Online Stacking Ensembles for the Calibration of Low-Cost Air Quality Sensor Networks in the Presence of Concept Drift, Atmosphere, 13, 416, https://doi.org/10.3390/atmos13030416, 2022.

Bouttier, F. and Courtier, P.: Data assimilation concepts and methods, ECMWF, 2002.

Chiles, J.-P. and Delfiner, P.: Geostatistics: modeling spatial uncertainty, John Wiley & Sons, 2009.

Coker, E. S., Amegah, A. K., Mwebaze, E., Ssematimba, J., and Bainomugisha, E.: A land use regression model using machine learning and locally developed low cost particulate matter sensors in Uganda, Environ. Res., 199, 111352, https://doi.org/10.1016/j.envres.2021.111352, 2021.

Considine, E. M., Reid, C. E., Ogletree, M. R., and Dye, T.: Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network, Environ. Pollut., 268, 115833, https://doi.org/10.1016/j.envpol.2020.115833, 2021.

Denby, B. R., Gauss, M., Wind, P., Mu, Q., Grøtting Wærsted, E., Fagerli, H., Valdebenito, A., and Klein, H.: Description of the uEMEP_v5 downscaling approach for the EMEP MSC-W chemistry transport model, Geosci. Model Dev., 13, 6303–6323, https://doi.org/10.5194/gmd-13-6303-2020, 2020.

Evensen, G.: The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dyn., 53, 343–367, https://doi.org/10.1007/s10236-003-0036-9, 2003.

Fletcher, S. J.: Data Assimilation for the Geosciences: From Theory to Application, 1st edition., Elsevier, Amsterdam, Netherlands, 976 pp., 2017.

Gressent, A., Malherbe, L., Colette, A., Rollin, H., and Scimia, R.: Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value, Environ. Int., 143, 105965, https://doi.org/10.1016/j.envint.2020.105965, 2020.

Guo, R., Qi, Y., Zhao, B., Pei, Z., Wen, F., Wu, S., and Zhang, Q.: High-Resolution Urban Air Quality Mapping for Multiple Pollutants Based on Dense Monitoring Data and Machine Learning, Int. J. Environ. Res. Public. Health, 19, 8005, https://doi.org/10.3390/ijerph19138005, 2022.

Hamer, P. D., Walker, S.-E., Sousa-Santos, G., Vogt, M., Vo-Thanh, D., Lopez-Aparicio, S., Schneider, P., Ramacher, M. O. P., and Karl, M.: The urban dispersion model EPISODE v10.0 – Part 1: An Eulerian and sub-grid-scale air quality model and its application in Nordic winter

conditions, Geosci. Model Dev., 13, 4323–4353, https://doi.org/10.5194/gmd-13-4323-2020, 2020.

Hassani, A., Schneider, P., Vogt, M., and Castell, N.: Low-Cost Particulate Matter Sensors for Monitoring Residential Wood Burning, Environ. Sci. Technol., 57, 15162–15172, 2023.

Hassani, A., Salamalikis, V., Schneider, P., Stebel, K., and Castell, N.: A scalable framework for harmonizing, standardization, and correcting crowd-sourced low-cost sensor PM2.5 data across Europe, J. Environ. Manage., 380, 125100, https://doi.org/10.1016/j.jenvman.2025.125100, 2025.

Hofman, J., Nikolaou, M., Shantharam, S. P., Stroobants, C., Weijs, S., and La Manna, V. P.: Distant calibration of low-cost PM and NO2 sensors; evidence from multiple sensor testbeds, Atmospheric Pollut. Res., 13, 101246, https://doi.org/10.1016/j.apr.2021.101246, 2022.

Jain, S., Presto, A. A., and Zimmerman, N.: Spatial Modeling of Daily PM2.5, NO2, and CO Concentrations Measured by a Low-Cost Sensor Network: Comparison of Linear, Machine Learning, and Hybrid Land Use Models, Environ. Sci. Technol., 55, 8631–8641, https://doi.org/10.1021/acs.est.1c02653, 2021.

Kalnay, E.: Atmospheric modeling, data assimilation, and predictability, Cambridge University Press, Cambridge, 2013.

Kang, Y., Aye, L., Ngo, T. D., and Zhou, J.: Performance evaluation of low-cost air quality sensors: A review, Sci. Total Environ., 818, 151769, https://doi.org/10.1016/j.scitotenv.2021.151769, 2022.

Kuula, J., Mäkelä, T., Aurela, M., Teinilä, K., Varjonen, S., González, Ó., and Timonen, H.: Laboratory evaluation of particle-size selectivity of optical low-cost particulate matter sensors, Atmospheric Meas. Tech., 13, 2413–2423, https://doi.org/10.5194/amt-13-2413-2020, 2020.

Lahoz, W. A. and Schneider, P.: Data assimilation: making sense of Earth Observation, Front. Environ. Sci., 2, 16, 2014.

Levy Zamora, M., Buehler, C., Datta, A., Gentner, D. R., and Koehler, K.: Identifying optimal co-location calibration periods for low-cost sensors, Atmospheric Meas. Tech., 16, 169–179, https://doi.org/10.5194/amt-16-169-2023, 2023.

Liang, L., Daniels, J., Bailey, C., Hu, L., Phillips, R., and South, J.: Integrating low-cost sensor monitoring, satellite mapping, and geospatial artificial intelligence for intra-urban air pollution predictions, Environ. Pollut., 331, 121832, https://doi.org/10.1016/j.envpol.2023.121832, 2023.

Lim, C. C., Kim, H., Vilcassim, M. J. R., Thurston, G. D., Gordon, T., Chen, L.-C., Lee, K., Heimbinder, M., and Kim, S.-Y.: Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea, Environ. Int., 131, 105022, https://doi.org/10.1016/j.envint.2019.105022, 2019.

Lopez-Ferber, R., Leirens, S., and Georges, D.: Source Estimation: Variational Method Versus Machine Learning Applied to Urban Air Pollution*, IFAC-Pap., 55, 78–83, https://doi.org/10.1016/j.ifacol.2022.08.052, 2022.

Lopez-Restrepo, S., Yarce, A., Pinel, N., Quintero, O. L., Segers, A., and Heemink, A. W.: Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley, Colombia, Atmosphere, 12, 91, https://doi.org/10.3390/atmos12010091, 2021.

Mijling, B.: High-resolution mapping of urban air quality with heterogeneous observations: a new methodology and its application to Amsterdam, Atmospheric Meas. Tech., 13, 4601–4617, https://doi.org/10.5194/amt-13-4601-2020, 2020.

Mu, Q., Denby, B. R., Wærsted, E. G., and Fagerli, H.: Downscaling of air pollutants in Europe using uEMEP_v6, Geosci. Model Dev., 15, 449–465, https://doi.org/10.5194/gmd-15-449-2022, 2022.

Murphy, E. and King, E. A.: Environmental noise pollution: Noise mapping, public health, and policy, Elsevier, 2022.

Schneider, P., Castell, N., Vogt, M., Dauge, F. R., Lahoz, W. A., and Bartonova, A.: Mapping urban air quality in near real-time using observations from low-cost sensors and model information, Environ. Int., 106, 234–247, https://doi.org/10.1016/j.envint.2017.05.005, 2017.

Schneider, P., Bartonova, A., Castell, N., Dauge, F. R., Gerboles, M., Hagler, G. S. W., Hüglin, C., Jones, R. L., Khan, S., Lewis, A. C., Mijling, B., Müller, M., Penza, M., Spinelle, L., Stacey, B., Vogt, M., Wesseling, J., and Williams, R. W.: Toward a Unified Terminology of Processing Levels for Low-Cost Air-Quality Sensors, Environ. Sci. Technol., 53, 8485–8487, https://doi.org/10.1021/acs.est.9b03950, 2019.

Schneider, P., Vogt, M., Haugen, R., Hassani, A., Castell, N., Dauge, F. R., and Bartonova, A.: Deployment and Evaluation of a Network of Open Low-Cost Air Quality Sensor Systems, Atmosphere, 14, 540, https://doi.org/10.3390/atmos14030540, 2023.

Shetty, S., Schneider, P., Stebel, K., David Hamer, P., Kylling, A., and Koren Berntsen, T.: Estimating surface NO2 concentrations over Europe using Sentinel-5P TROPOMI observations and Machine Learning, Remote Sens. Environ., 312, 114321, https://doi.org/10.1016/j.rse.2024.114321, 2024.

Shetty, S., Hamer, P. D., Stebel, K., Kylling, A., Hassani, A., Berntsen, T. K., and Schneider, P.: Daily high-resolution surface PM2.5 estimation over Europe by ML-based downscaling of the CAMS regional forecast, Environ. Res., 264, 120363, https://doi.org/10.1016/j.envres.2024.120363, 2025.

Shetty, S., Hassani, A., Hamer, P. D., Stebel, K., Salamalikis, V., Berntsen, T. K., Castell, N., and Schneider, P.: Evaluating the role of low-cost sensors in machine learning based European PM2.5 monitoring, Environ. Res., 291, 123558, https://doi.org/10.1016/j.envres.2025.123558, 2026.

Vogt, M., Schneider, P., Castell, N., and Hamer, P.: Assessment of Low-Cost Particulate Matter Sensor Systems against Optical and Gravimetric Methods in a Field Co-Location in Norway, Atmosphere, 12, 961, https://doi.org/10.3390/atmos12080961, 2021.

Weissert, L., Alberti, K., Miles, E., Miskell, G., Feenstra, B., Henshaw, G. S., Papapostolou, V., Patel, H., Polidori, A., Salmond, J. A., and Williams, D. E.: Low-cost sensor networks and land-use regression: Interpolating nitrogen dioxide concentration at high temporal and spatial

resolution in Southern California, Atmos. Environ., 223, 117287, https://doi.org/10.1016/j.atmosenv.2020.117287, 2020.

Wesseling, J., Drukker, D., Gressent, A., Janssen, S., Joassin, P., Lenartz, F., Van Ratingen, S., Rodrigues, V., Sousa, J., and Thunis, P.: Using synthetic data to benchmark correction methods for low-cost air quality sensor networks, Air Qual. Atmosphere Health, 17, 979–996, https://doi.org/10.1007/s11869-023-01493-z, 2024.